2013-04-04

Nuclear Organization in Breast Cancer: A Dissertation

Jason R. Dobson

University of Massachusetts Medical School

NUCLEAR ORGANIZATION IN BREAST CANCER

A Dissertation Presented

By

Jason Russell Dobson

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

April 4th, 2013

Cell Biology

NUCLEAR ORGANIZATION IN BREAST CANCER

A Dissertation Presented
By
Jason Russell Dobson

The signatures of the Dissertation Defense Committee signify
completion and approval as to style and content of the Dissertation

Jane B. Lian, PhD, Thesis Co-Advisor

Janet L. Stein, PhD, Thesis Co-Advisor

Jeanne B. Lawrence, PhD, Member of Committee

Jeffrey A. Nickerson, PhD, Member of Committee

Sharon B. Cantor, PhD, Member of Committee

Louis C. Gerstenfeld, PhD, Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets
the requirements of the Dissertation Committee

Anthony N Imbalzano, PhD, Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies
that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

Cell Biology Program
April 4th, 2013

Dedication

This work, and all the years of work before and after are dedicated to my daughter Alice.

“Unless a man undertakes more than he possibly can do, he will never do all that he can.” – Henry Drummond

“Nothing ever comes to one that is worth having except as a result of hard work” – Booker T. Washington

"The man in black fled across the desert, and the gunslinger followed." – *The Gunslinger* by Stephen King.

Acknowledgements

I would first like to thank my co-advisors Jane and Janet for their patience and hard work over the years trying to keep me focused. I also am deeply grateful for the collaborative environment created by all four of my PIs: Gary, Janet, Jane and Andre. It is no small feat to run a lab with more than one PI (let alone four), and during my many years in their research group they have shown me how much can be accomplished when everyone's talents are combined.

I want to thank my good friends Kaleem, Ricardo and Jitesh for their many years of lunches and coffee/tea breaks. Our discussions were some of my most cherished moments in the laboratory, and they continuously challenged me to become a better scientist. I am also thankful for Yang's great sense of humor, and constant smile; never a dull moment in the Stein lab. I would also like to thank Dana and Shirwin for their friendship over the years. They have both made me feel welcome at all times, and were incredibly supportive as my wife and I started our family together.

I am especially grateful for the opportunity to have collaborated with Jeff and Tony through the NCI Program Project. The discussions I have had with them over the years have been truly inspirational as both a scientist and a person, and I am lucky to have had the opportunity to work so closely with them.

My grandmother, Elaine, has always been able to simultaneously support me and push me to be better. I would not have gotten through graduate school if it weren’t for her. She has taught me so much about who I am, and who I can be. Although she is not here today, I know she is proud.

Most importantly, I have to thank my wife, Lauren, who has been incredibly patient with me and supportive of me. While raising our daughter, Alice, and enrolled in graduate school in different cities, Lauren has always done whatever it took to make things work. Regardless of whatever challenge we faced Lauren’s unwavering support has been truly humbling.

Abstract

The nuclear matrix (NM) is a fibrogranular network of ribonucleoproteins upon which transcriptional complexes and regulatory genomic sequences are organized. A hallmark of cancer is the disorganization of nuclear architecture; however, the extent to which the NM is involved in malignancy is not well studied.

The RUNX1 and RUNX2 proteins form complexes within the NM to promote hematopoiesis and osteoblastogenesis, respectively at the transcriptional level. RUNX1 and RUNX2 are both expressed in breast cancer cells (BrCCs); however, their genome-wide BrCC functions are unknown. RUNX1 and RUNX2 activate many tumor suppressor pathways in blood and bone lineages, respectively, including attenuation of protein synthesis and cell growth via suppression of ribosomal RNA (rRNA) transcription, which appears contrary to Runx-expression in highly proliferative BrCCs. To define roles for RUNX1 and RUNX2 in BrCC phenotype, we examined the involvement of RUNX1 and RUNX2 in rRNA transcription and generated a genome-wide model for RUNX1 and RUNX2-binding and transcriptional regulation. To validate gene expression patterns identified in our screen, we developed a Real-Time qPCR primer design program, which allows rapid, high-throughput design of primer pairs (FoxPrimer). In BrCCs, RUNX1 and RUNX2 regulate genes that promote invasiveness and do

not affect rRNA transcription, protein synthesis, or cell growth. We have characterized *in vitro* functions of Runx proteins in BrCCs; however, the relationships between Runx expression and diagnostic/prognostic markers of breast cancer (BrCa) in patients are not well studied. Immunohistochemical detection of RUNX1 and RUNX2 in BrCa tissue microarrays reveals RUNX1 expression is associated with early, smaller tumors that are ER+ (estrogen receptor), HER2+, p53-, and correlated with androgen receptor (AR) expression; RUNX2 expression is associated with late-stage, larger tumors that are HER2+. These results show that the functions and expression patterns of NM-associated RUNX1 and RUNX2 are context-sensitive, which suggests potential disease-specific roles.

Two functionally disparate genomic sequence types bind to the NM: matrix associated regions (MARs) are functionally associated with transcriptional repression and scaffold associated regions (SARs) are functionally associated with actively expressed genes. It is unknown whether malignant nuclear disorganization affects the functions of MARs/SARs in BrCC. We have refined a method to isolate nuclear matrix associated DNA (NM-DNA) from a structurally preserved NM and applied this protocol to normal mammary epithelial cells and BrCCs. To define transcriptional functions for NM-DNA, we developed a computational algorithm (PeaksToGenes), which statistically tests the

associations of experimentally-defined NM-DNA regions and ChIP-seq-defined positional enrichment of several histone marks with transcriptome-wide gene expression data. In normal mammary epithelial cells, NM-DNA is enriched in both MARs and SARs, and the positional enrichment patterns of MARs and SARs are strongly associated with gene expression patterns, suggesting functional roles. In contrast, the BrCCs are significantly enriched in the silencing mark H3K27me3, and the NM-DNA is enriched in MARs and depleted of SARs. The MARs/SARs in the BrCCs are only weakly associated with gene expression patterns, suggesting that loss of normal DNA-matrix associations accompanies the disease state. Our results show that structural preservation of the *in situ* NM allows isolation of both MARs and SARs, and further demonstrate that in a disorganized, cancerous nucleus, normal transcriptional functions of NM-DNA are disrupted.

Our studies on nuclear organization in BrCC, show that the disorganized phenotype of the cancer cell nucleus is accompanied by deregulated transcriptional functions of two constituents of the NM. These results reinforce the role of the NM as an important structure-function component of gene expression regulation.

Table of Contents

List of Figures and Tables

List of Symbols, Abbreviations or Nomenclature

° :	Degrees
3' :	Three primer end
5' :	Five prime end
ACTB :	Beta-actin
ADAM22 :	A disintegrin and metalloprotease domain family 22
AluI :	Restriction enzyme, which cuts AGCT
AMIGO2 :	Adhesion molecule with immunoglobulin-like domain 2
ANOVA :	Analysis of variance
API :	Application Programming Interface
AR :	Androgen receptor
ASF1B :	Anti-silencing function 1 homolog B
AT :	Nucleotide bases: adenine and thymidine
ATF1 :	Activating transcription factor 1
B23 :	Numatrin
BAM :	Binary sequence Alignment Map format file

BamHI :	Restriction enzyme, which cuts GGATCC
BASH :	GNU project Bourne shell
BCL :	B-cell lymphoma 2
BED :	UCSC file format for genomic intervals
BioPerl :	Perl modules for bioinformatics
BMP :	Bone morphogenetic protein
bp :	Base pairs
BrCa :	Breast cancer
BRCA1 :	Breast cancer associated 1
BrCCs :	Breast cancer cells
BSP :	Bone-sialoprotein
C :	Centigrade
c-terminus :	Carboxyl terminal end of a protein
CD82 :	CD82 antigen
cDNA :	Complimentary DNA, reverse transcribed from RNA
CELSR3 :	Cadherin, EGF LAG seven-pass G-type receptor 3

CHERP :	Calcium homeostasis endoplasmic reticulum protein
ChIP :	Chromatin immunoprecipitation
ChIP-seq :	Chromatin immunoprecipitation and deep-sequencing
CI :	Confidence interval
Ciz1 :	Cyclin-dependent kinase inhibitor 1A (p21) interacting zinc finger protein 1
Coilin :	Cajal body protein
CpG :	Genomic regions in which a cytosine and a guanine are side-by-side
CPU :	Central Processing Unit
CSK :	Cytoskeletal
Ct :	Threshold cycle
CTCF :	CCCTC-binding factor (zinc finger protein)
Cys :	Cysteine
CysPh :	Cytosarcoma phyllodes
DAB :	3,3'-Diaminobenzidine
DamID :	DNA adenine methyltransferase identification
DAVID :	Database for Annotation, Visualization, and Integrated Discovery

DCIS :Ductal carcinoma in situ

DMEM/F12 : Dubelco's modified eagle medium: nutrient mixture 12

DNA : Deoxyribonucleic acid

DNAse I : Deoxyribonuclease I

DYNLT3 : Dynein, light chain, Txtex-type 3

EC2 : Amazon elastic compute cloud

ECM : Extracellular matrix

EDTA : Ethylenediaminetetraacetic acid

EGF : Epidermal growth factor

EGTA : Ethylene glycol tetraacetic acid

EMSA : Electro-Mobility Shift Assay

ENCODE : Encyclopedia of DNA Elements

ER : Estrogen receptor

ERCC4 : Excision repair cross-complementing rodent repair deficiency complementation group 4

ETV1 : Ets variant 1

FFPE : Formalin-fixed, paraffin-embedded

FHOD3 : Formin homology domain 2

FibroAd : Fibroadenoma

FoxPrimer : Real-Time qPCR primer design program and database

FPR1 : Formyl peptide receptor 1

FTP : File Transfer Protocol

FZD7 : Frizzled homolog 7

g : Gravity

GC : Nucleotide bases: guanine and cytosine

GCNT1 : Glucosaminyl (N-acetly) transferase 1

GenBank : Sequence database provided by NCBI

GeneRIF : Gene Reference into Function, provided by NCBI

GI : NCBI GenInfo identifier

GNG2 : Guanine nucleotide binding protein (G protein) gamma 2

GO : Gene ontology

GRPR : Gastrin-release peptide receptor

H2A : Histone 2 A

H2B :	Histone 2 B
H3 :	Histone 3
H3K27me3 :	Histone 3, lysine 37 trimethylation
H3K4me3 :	Histone 3, lysine 4 trimethylation
H4 :	Histone 4
HaeIII :	Restriction enzyme, which cuts GGCC
HAS2 :	Hyaluronan sythase 2
HBEGF :	Heparin-binding epidermal growth factor like growth factor
HEPES :	4-(2-hydroxyethyl)-1-piperaznieethanesulfonic acid
HER2 :	Human epidermal growth factor receptor 2
hg19 :	Human genome annotation reference 19
HMGB3 :	High mobility group box 3
hPhox :	FoxGP
HPRT :	Hypoxanthine-guanine phosphoribosyltransferase
HRP :	Horseradish peroxidase
hsa-mir-30c :	*Homo sapiens* microRNA 30e

hsa-mir-30c-3p :	*Homo sapiens* microRNA 30c, three prime mature end
hsa-mir-30c-5p :	*Homo sapiens* microRNA 30c, five prime mature end
HTML :	HyperText Markup Language
IDC :	Invasive ductal carcinoma
IgG :	Immunoglobulin G
IHH :	Indian hedgehog
IL11 :	Interleukin 11
IP :	Immunoprecipitate
JASPAR :	Transcription factor binding profile database
KAL1 :	Kallmann syndrome 1
Kb :	Kilo-base pairs
kDa :	Kilo Dalton
kg :	Kilogram
ki67 :	ki67 protein
KIAA1199 :	Unnamed protein
LADs :	Lamin-associated domains

LiS : Lithium 3,5-diiodosalicylate

LMBRD2 : Limb region 1 domain containing 2

LPCAT2 : Lysophosphatidylcholine acyltransferase 2

LPXN : Leupaxin

M : Molar

MACS : Model-based analysis of ChIP-seq

MARCH3 : Membrane-associated ring finger 3

MARCH5 : Membrane-associated ring finger 5

MARs : Matrix associated regions

Mb : Mega-base pairs

MBLAC2 : Metallo-beta-lactamase domain containing 2

MCF-7 : Malignant, metastatic breast cancer cells isolated by pleural effusion

MCF10a : Normal mammary epithelial cells, spontaneously immortalized

MDA-MB-231 : Malignant, metastatic breast cancer cell line isolated via pleural effusion

Met : Breast tissue metastasized to lymph mg : Milligram

MG132 : Carbobenzoxy-L-leucyl-L-leucyl-L-leucinal, Z-LLL-CHO, proteasome inhibitor

miRNA : microRNA

mL : Milliliter

mm : Millimeter

mM : Millimolar

MMP : Matrix metalloprotease

mmu-mir-30c : *Mus musculus* microRNA 30c

mRNA : Mature, splice messenger RNA

MspI : Restriction enzyme, which cuts CCGG

MVC : Model, view, controller program architecture

MWCO : Molecular weight cutoff

n-terminus : Amino-terminal end of a protein

NAT : Normal adjacent tissue

NCBI : National center for biotechnology information

NEDD4 :	Neural precursor cell expressed developmentally down-regulated protein 4
NFIC :	Nuclear factor I/C (CCAAT-binding transcription factor
ng :	Nanogram
NIH3T3 :	Mouse embryo fibroblast cell line
NM :	Nuclear matrix
nM :	Nanomolar
NM-DNA :	Nuclear matrix associated deoxyribonucleic acid
NME7 :	Non-metastatic cells 7
NMP-22 :	Nuclear matrix protein 22
NMTS :	Nuclear Matrix Targeting Sequence
NOV :	Nephroblastoma overexpressed
OP :	Osteopontin
p53 :	Tumor protein 53
PAGE :	Polyacrylamide gel electrophoresis
PBS :	Phosphate buffered saline
PCR :	Polymerase chain reaction

PeakSeq : A program for identifying and ranking peak regions in ChIP-seq experiments

PeaksToGenes : Average gene plots and statistical testing program

pH : Hydrogen ion concentration

PLXNA1 : Plexin A1

PML : Promyelocytic leukemia

PNPLA3 : Patatin-like phospholipase domain containing 3

Pol II : RNA polymerase II

PR : Progesterone receptor

pre-mRNA : Unspliced, newly-synthesized messenger RNA

pre-rRNA : Unspliced, newly synthesized ribosomal RNA

PRKRIR : Protein-kinase, interferon inducible double-stranded RNA dependent inhibitor, repressor of p58

PTHRP : Parathryoid receptor protein

PU.1 : ETS domain transcription factor

PvuII : Restriction enzyme, which cuts CAGCTG

qPCR : Quantitative polymerase chain reaction

R398A/Y428A :	Amino acids arginine and tyrosine mutated to alanine
RDBMS :	Relational Database Management System
rDNA :	Repetitive gene unit, which is transcribed into rRNA
RefSeq :	NCBI reference sequence database
RMA :	Robust means average
RNA :	Ribonucleic acid
RNP :	Ribonucleic acid and protein complex
rRNA :	Ribosomal ribonucleic acid
RsaI :	Restriction enzyme, which cuts GTAC
RUNX1 :	*Homo sapiens* Runt-related transcription factor 1
Runx1:	*Mus musculus* Runt-related transcription factor 1
Runx1p1 :	RUNX1 promoter 1 – ChIP primer location
RUNX2 :	*Homo sapiens* Runt-related transcription factor 2
Runx2 :	*Mus musculus* Runt-related transcription factor 2
RUNX3 :	*Homo sapiens* Runt-related transcription factor 3
s :	Seconds

SAM : Sequence Alignment Map format file

SaOS-2 : Osteosarcoma cells

SARs : Scaffold associated regions

SC-35 : Serine/arginine-rich splicing factor 2

SCML2 : Sex comb on midleg-like 2

SDS : Sodium dodecyl sulfate

SEM : Standard error of the mean

SES : Sequence enrichment scaling

siRNA : Small interfering RNA

SMAD : Homologs of *Drosophila* mothers against decapentaplegic and the *C. elegans* small body size

SPP : An R package to identify peak regions of enrichment from ChIP-seq

SQL : Structured Query Language

SRGN : Serglycin

TAE : Tris-acetate-EDTA

TBC1D5 : TBC1 domain family member 5

TBLX1 : Transducin-beta-like, X-linked 1

TCGA : The Cancer Genome Atlas

TE : Tris-EDTA

TGF : Transforming growth factor

TGFBR1 : Transforming growth factor beta receptor 1

TNM : Tumor size, node status, distal metastasis status

TRANSFAC :Database for eukaryotic transcription factors

TRIB2 : Tribbles homolog 2

TSS : Transcriptional start site

TTS : Transcriptional terminal site

TWF1 : Twinfillin

U6 : Non-coding small nuclear RNA

UBF : Upstream binding factor, RNA polymerase I

UBR7 : Ubiquiting protein ligase E3 component n-recognin 7

UCSC : University of California at Santa Cruz

UMASS : University of Massachusetts

UMMS : University of Massachusetts Medical School

UTR :	Untranslated region
UV :	Ultraviolet
v/v :	Volume per volume
VEGF :	Vascular endothelial growth factor
w/v :	Weight per volume
WNT :	Wingless-homolog proteins and signaling pathway
ZCCHC5 :	Zinc finger, CCHC-containing group 5
µL :	Microliter
µM :	Micromolar

List of Multimedia Objects or Files

To view a limited version of FoxPrimer, please visit: www.foxprimer.org

The source code for FoxPrimer is available at: www.github.com/foxprimer

The source code for PeaksToGenes is available at:

www.github.com/peakstogenes

Preface

Portions of this dissertation will appear in the following manuscripts:

Chapter 2:

Jason R. Dobson, Deli Hong, Hai Wu, Jane B. Lian , Janet L. Stein, Jeffery A. Nickerson, Andre J. van Wijnen, Gary S. Stein*. Isolation and characterization of transcriptional associations of nuclear matrix-associated DNA in breast cancer cells. *Manuscript in preparation*.

Chapter 3:

Jason R. Dobson, Gillian Browne, Deli Hong, Maria Libera de la Porta, Andre J. van Wijnen, Janet L. Stein, Gary S. Stein, Jane B. Lian*. Genome-wide analysis of binding and gene expression regulation of Runx1 and Runx2 in MDA-MB-231 cells and association of Runx1 and Runx2 expression with hormone receptors in breast cancer patient samples. *Manuscript in preparation*.

Chapters 3 and 4:

Jason R. Dobson*, Ricardo Medina, Andre J. van Wijnen, Janet L. Stein, Jane B. Lian, Gary S. Stein. FoxPrimer: A web-interface Real-Time qPCR primer design program and database. *Manuscript in preparation*.

Chapters 2, 3, and 5:

Jason R. Dobson*, Andre J. van Wijnen, Jane B. Lian, Gary S. Stein, Janet L. Stein. PeaksToGenes: average gene plot generation and statistical testing. *Manuscript in preparation*.

Appendix A1:

Jason R. Dobson, Hanna Taipaleenmäki, Yu-Jie Hu, Deli Hong, Andre J. van Wijnen, Janet L. Stein, Gary S. Stein, Jane B. Lian, Jitesh Pratap*. hsa-mir-30c targets NOV/CCNB3 to promote the invasion of MDA-MB-231 cells. *Manuscript in preparation*.

* – indicates corresponding author(s)

Research Chapters

CHAPTER 1 INTRODUCTION

General background

The nucleus is the eukaryotic organelle in which genomic DNA is stored, replicated, and transcribed (Alberts et al. 2002). Many factors collectively regulate the transcription of specific genes, the precise temporal expression of which is required for cellular functions such as cell cycle progression, cellular homeostasis, and lineage commitment (Lanctôt et al. 2007, Hager et al. 2009). A critical mechanism for gene expression regulation is the facultative spatial organization of regulatory genomic sequences and transcriptional complexes (Zaidi et al. 2007). In diseases such as breast cancer, many of the regulatory mechanisms controlling gene expression are impaired, which results from and potentiates the disease state (Meaburn et al. 2009, Misteli 2010). In cancer, malformed or irregular nuclei are used as diagnostic and prognostic parameters of disease. Cancer cell nuclei are often characterized as having “folds” or “indentations”. The distribution of heterochromatin is often markedly disrupted in cancer cells as compared to normal cells (Zink et al. 2004b). We hypothesize that these grossly observed nuclear changes correspond to changes in the spatial organization of transcriptional regulatory complexes and genomic sequences. This study is focused on two components that contribute to nuclear organization, the nuclear matrix and RUNX proteins, and their functional associations with gene expression in breast cancer cells.

Nuclear domains associated with transcription

The diameter of the average human nucleus is approximately 6μm (Alberts et al. 2002). However, the human genome is approximately 2 meters in length (Lander et al. 2001). The length of the genome in relation to the size of the nucleus creates a non-trivial space restriction issue. In order to attain sufficient compaction to fit into the nucleus, DNA is wrapped around histone octamers and further folded into higher-order structures (Li et al. 2007). Genes are subsequences of DNA that can be transcribed into RNA and comprise roughly 1% of the human genome (ENCODE Project Consortium 2011). The small percentage of genomic sequence that encodes genes, combined with nuclear volume limitations necessitates the organization of genomic DNA such that genes are accessible to transcriptional regulators. Further compounding the spatial problem, while all cells within an organism have approximately the same genomic sequence, not all genes within the genome are expressed in every cell. Specific genes, herein referred to as phenotypic genes, are only expressed within certain developmental time frames and in defined cellular lineages (Zaidi et al. 2005). Therefore, of the 1% of the human genome encoding genes, only a subset of genes are transcriptionally active in a given cell type. One mechanism for specification of phenotypic gene expression is coordinated spatial organization of genomic sequences and regulatory complexes (Stein et al. 2003, Misteli 2007, Hager et al. 2009). Disruption of key components of nuclear organization and gene expression regulation often results in developmental abnormalities, indicating the importance of nuclear organization as a mechanism for cellular phenotype (Alvarez et al. 2000, Melillo et al. 2001, De Sandre-Giovannoli et al. 2002, Alsheimer et al. 2004, Roshon and Ruley 2005, Frock et al. 2006).

At a gross scale, expression of genes can be bisected into two groups based on the chromatin environment within and surrounding the genes. Genes organized within more compact chromatin structures, referred to as heterochromatin, are unlikely to be expressed as these regions are mostly inaccessible to transcriptional complexes. Genic regions organized within less compact structures, or euchromatin, are permissive for interaction with transcriptional complexes and therefore have the potential to be transcribed (Li et al. 2007). Heterochromatin and euchromatin occupy spatially distinct regions of the nucleus. Light microscopy imaging of cells reveals the nucleus as being organized into three regions: 1) the peripheral heterochromatin, 2) the nucleolus (sometimes more than one, depending on cell type), and 3) the euchromatic space between the nucleolus and peripheral heterochromatin (Berezney et al. 1995, Nickerson et al. 1995). Even at this broad scale, patterns in gene expression form based on whether a particular gene is localized to the euchromatic space, the peripheral heterochromatin, or the nucleolus (Nickerson 2001). While it is convenient to discuss these nuclear areas as if they are separated by a physical barrier, similar to cytoplasmic organelles, it is important to make the critical distinction that nuclear organelles, or compartments, are not separated by a membrane. Nuclear organelles are defined by the concentration of similar structural or functional bodies within the nucleus, not by physical separation via a membrane (Dundr and Misteli 2010).

The nucleolus is a specialized nuclear compartment within which the short arms of human chromosomes 13, 14, 15, 21, and 22 are localized (Sullivan et al. 2001). These chromosomal regions contain tandem-arrayed repeats of the gene encoding ribosomal RNA (rRNA), which, once processed, forms the major catalytic subunit of the ribosome. Specialized proteins, such as RNA Polymerase I (Pol I), are concentrated within the nucleolus to drive the transcription of ribosomal DNA (rDNA) and processing of pre-rRNA into mature ribosomes (McStay and Grummt 2008). The dense concentration of these regulatory factors and ribosomal RNAs causes the visually distinct appearance of the nucleolus, which can be observed via light microscopy or staining techniques that distinguish between RNA and DNA, such as Papanicolaou stain (Papanicolaou and Traut 1997). Ribosomal RNA transcription is the rate limiting step for cellular growth and accounts for the majority of a cell's transcriptional activity (Grummt and Voit 2010). Organization and concentration of regulatory factors within the nucleolus is essential for the precise regulation of rRNA transcription (Grummt 2010).

The nucleolus is not the only place where compartmentalization of transcriptional functions is observed in the nucleus. Typically, RNA Polymerase II (Pol II) transcription of phenotypic genes occurs between the nucleolus and the peripheral heterochromatin in the euchromatic space (Herman et al. 1978, Jackson and Cook 1985, Jackson et al. 1993, Wansink et al. 1993, Niedojadlo et al. 2011). Peripheral heterochromatin is generally transcriptionally silent, with the exception of genes within the peripheral heterochromatin proximal to nuclear pores that are sometimes observed to be highly expressed in specific instances (Mateos-Langerak et al. 2007, Papantonis and Cook 2010). The transcription of most phenotypic genes is executed by Pol II in coordination with a multitude of regulatory factors, which can be cell cycle, cell type, or developmentally specific (Stein et al. 2011). Pol II transcription occurs at a defined number of sites, primarily within the euchromatic space, which are referred to as transcription factories (Jackson et al. 1998, Cook 1999). These transcription factories are relatively static in position, and contain a high concentration of transcription factors and RNA processing factors (Ghamari et al. 2013). The inclusion of RNA processing factors is likely due to the observation that RNA splicing occurs co-transcriptionally (Xu and Cook 2008).

Similar to the transcription factories, the chromosomes themselves are not free-floating entities within the nucleus. Chromosomes occupy distinct territories relative to one another, which are preserved throughout multiple cell divisions (Lanctôt et al. 2007, Geyer et al. 2011). Transcriptionally silent chromosomal regions are largely localized to the peripheral regions of the nucleus, while transcriptionally active regions of chromosomes are found within the euchromatic space (Andrulis et al. 1998). Live cell imaging of genes has demonstrated that silent, or non-expressed genes, are quite static within the chromosome territory; in contrast, expressed genes are dynamic in their range of movement within the chromosome territory (Kosak et al. 2002, Zink et al. 2004a, Hewitt et al. 2004). This freedom of motion observed for expressed genes is thought to arise from the transient nature of transcription. Genes are typically not constitutively expressed; rather they are transcribed in bursts. It has been hypothesized that these transient events occur concurrently with the movement of a gene into a transcription factory (Deng et al. 2012).

It is thought that transcription factors regulate genes by binding to *cis*-elements such as promoters and enhancers (Ong and Corces 2011, Spitz and Furlong 2012). Yet, this model may not fully explain how genes are recruited to transcription factories. The average gene length in the human genome is approximately 3,000 base pairs, which is similar to the average length of a gene in a simpler organism like *Drosophila melanogaster* (modENCODE Consortium et al. 2010, ENCODE Project Consortium 2011). A major difference between less-complex organisms and humans is the variance in gene length. Human genes (and many other mammals) can be tens to hundreds of kilobase pairs in length. Transcription factors typically bind a sequence of DNA that is roughly 10 bases in length (Wang et al. 2012). It seems unlikely that a single binding event would be sufficient to drive a stable interaction between a gene that may be several kilobase pairs in length and a transcription factory. It is therefore more likely that a combination of binding events, driven by multiple DNA-binding factors, promotes the recruitment of a gene to the transcription factory. This idea is consistent with the observations that gene promoters can be occupied by multiple DNA-binding proteins and their cognate co-factors to regulate gene transcription (Lian et al. 2006). One recent study has demonstrated that, in the case of longer genes, multiple interactions can be observed between the gene body and the transcription factory. These binding events result in multiple sub-loop structures forming while the gene is being actively transcribed (Larkin et al. 2012). How these DNA-binding proteins are concentrated within transcription factories and the binding events required for recruitment of a genic sequence to transcription factories are active areas of research that are likely to provide key insight into the regulation phenotypic gene expression.

Observation and isolation of the nuclear matrix

Early light microscopic images of mammalian cells defined the nucleus as dense chromatin and nucleoli suspended in transparent “nuclear sap” or karyolymph (Fawcett 1966). Use of the term karylomph implied that everything within that area was randomly diffusing throughout the nucleus. Thin sectioning of cells combined with electron microscopy allowed observation of two distinct nuclear structures containing genomic DNA: the previously observed densely-stained compacted heterochromatin and the less compact euchromatin (Nickerson et al. 1995). Electron microscopic imaging of regressive EDTA-stained embedded thin sections revealed a fibrogranular network of RNPs between the densely-stained heterochromatin (Monneron and Bernhard 1969, Bernhard 1969, Fakan and Bernhard 1971). This fibrogranular network of RNPs is highly associated with active transcription, as it is largely comprised of newly-synthesized pre-mRNAs (Bachellerie et al. 1975, Fakan and Nobis 1978, Fakan and Hughes 1989). This observed network of perichromatin fibrils has been defined as the “in situ nuclear matrix” (Berezney 1984).

Berezney and Coffey isolated and characterized the nuclear matrix from rat liver tissue (Berezney and Coffey 1974, 1977). The previously observed *in situ* nuclear matrix can be isolated from nuclei through nuclease digestion followed by high salt and detergent extraction. The residual fraction maintains the ultrastructural and biochemical properties of the intact perichromatin fibrils. When histones are extracted from chromosomes, genomic DNA extends outward from the nuclear matrix in long (up to 100Kbp) loops (Cook and Brazell 1975, 1980, Benyajati and Worcel 1976, Paulson and Laemmli 1977, Marsden and Laemmli 1979, Vogelstein et al. 1980). These observations of genomic DNA being tethered to the nuclear matrix at the bases of loops were very exciting and suggested that the nuclear matrix may play a major role in the architectural organization of nuclear bodies. Experimental approaches centered on the nuclear matrix found that many critical nuclear activities and the factors executing these functions are associated with the nuclear matrix. These functions include but are not limited to replication (Dijkwel et al. 1986, Velden and Wanka 1987, Tubo and Berezney 1987), DNA looping attachment sites (Gasser and Laemmli

1987), and transcription (Feldman and Nevins 1983, Lewis and Lebkowski 1984, Zehnbauer and Vogelstein 1985, van Wijnen et al. 1993). In the case of transcription, the number of nuclear matrix fibrils correlates with transcriptional activity, which supports the idea that the nuclear matrix plays a major role in transcriptional regulation (Petrov and Sekeris 1971).

DNA associated with the nuclear matrix

To understand the mechanisms by which the nuclear matrix organizes DNA into loops and how these binding events are associated with nuclear functions, many experimental approaches were designed to isolate and characterize DNA sequences associated with the nuclear matrix. The two most commonly used techniques for isolation of matrix-associated DNA rely upon similar approaches in which histones are extracted and non-matrix-associated DNA is removed via nuclease digestion (Nickerson et al. 1995). Histones are solubilized and removed via high concentrations of sodium chloride or the chaotropic agent and detergent lithium 3,5-diiodosalicylate (LiS); the residual structures after extraction are referred to as the nuclear matrix and the nuclear scaffold, respectively (Berezney and Coffey 1977, Mirkovitch et al. 1984). DNA regions isolated via sodium chloride or LiS are therefore referred to as matrix-associated regions (MARs) and scaffold-associated regions (SARs), respectively. However, because the experimental approach and residual structure are similar, these approaches were considered interchangeable and the matrix-associated sequences were referred to as S/MARs (Boulikas 1995). These genomic regions attached to the nuclear matrix have been observed in the introns of genes, as well as within both proximal and distal regions flanking gene bodies (Mirkovitch et al. 1987, Georgiev et al. 1991, Boulikas 1995). Both cell-type and developmental time point differences in matrix-associated DNA near genes have been observed (Dworetzky et al. 1992, Bidwell et al. 1993, 1994, van Wijnen et al. 1993), so it is therefore critical to understand how these matrix-attachment events are regulated and how these genomic regions contribute to phenotypic gene expression regulation when associated with the nuclear matrix.

S/MARs are proposed to have common sequence-based elements. There is not a consensus sequence motif for S/MARs in the traditional sense of a transcription factor motif; rather S/MARs share a set of biophysical features. Most commonly, S/MARs are characterized by sequences that allow for bending or curving of DNA, inverted repeat regions, and stretches of AT bases (Boulikas 1995). The AT-rich S/MARs are well-characterized in terms of how frequently these sequences are experimentally observed to be associated with the nuclear matrix and in terms of the identification of proteins that preferentially bind to AT-rich sequences. These AT-rich regions also have a unique physical property in their ability to unwind under torsional stress, and have been referred to as base unpairing regions (BURs) (Bode et al. 1992). Proteins such as: heterogeneous nuclear ribonucleoprotein U (HNRNPU) (also known as scaffold attachment factor A (SAF-A)), special AT-rich sequence binding protein 1 (SATB1), and high mobility group AT-hook 1 (HMGIY) are localized within the nuclear matrix and bind to AT-rich S/MARs (Bode et al. 1992, Dickinson et al. 1992, Fackelmayer et al. 1994, Belle et al. 1998, Liu et al. 1999). These proteins bind S/MARs and organize genomic DNA to promote the expression of genes. SATB1 is particularly interesting in that it normally functions to organize and regulate the expression of cytokine genes in thymocytes (Cai et al. 2006). However, when ectopically expressed in breast cancer cells, SATB1 promotes tumor progression, invasion and metastasis (Han et al. 2008). The observation that the transcriptional role of a S/MAR-binding protein, such as SATB1, can be cell-type dependent suggests that unknown factors may contribute to which S/MARs are available for binding.

RUNX proteins in development and breast cancer

Runt-related transcription factors (RUNX1, RUNX2, and RUNX3) are a family of proteins that share a DNA-binding domain (Runt), which is highly conserved in *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, and

Drosophila melanogaster (Kagoshima et al. 1993, van Wijnen et al. 2004). Runx1 is required for definitive hematopoiesis (Wang et al. 1996) and Runx2 is required for osseous development (Komori et al. 1997). Genetic ablation of either *Runx1* or *Runx2* results in embryonic or post-natal death, respectively (Wang et al. 1996, Komori et al. 1997). Runx3 appears to play roles in gut and neuronal development. However, due to conflicting phenotypes for independently developed *Runx3*-null mouse models of development, the predominant developmental role of Runx3 is unclear (Levanon et al. 2003). This dissertation is focused on the transcriptional roles of RUNX1 and RUNX2 in breast cancer, and will not focus on the functions of RUNX3.

While the DNA-binding domain of RUNX proteins is found near the N-terminus, RUNX proteins also share a unique C-terminal domain called the nuclear matrix targeting sequence (NMTS), which is conserved in mammals (Zeng et al. 1997, 1998, Tang et al. 1999). Runx2 (previously known as NMP-2) was identified as a nuclear matrix-specific protein that binds to and regulates the osteocalcin (*Bglap2*) promoter during osteoblast differentiation, which suggested that nuclear matrix targeting of RUNX proteins may be important for function (Bidwell et al. 1993, Merriman et al. 1995). Genetic insertion of a stop codon in *Runx1* or *Runx2*, before the C-terminal portion containing the NMTS results in developmental phenocopies of the corresponding null models (Choi et al. 2001, Dowdy et al. 2010). Cell-based studies have further demonstrated the importance of RUNX protein targeting to the nuclear matrix in the transcriptional regulation of phenotypic genes (Zaidi et al. 2001, 2006, Vradii et al. 2005). The functional and developmental abnormalities observed when RUNX proteins lack the NMTS domain indicate that targeting to the nuclear matrix is essential for the fidelity of RUNX protein function.

RUNX proteins have been described as both oncogenes and tumor suppressors, depending on the cellular context (Cameron and Neil 2004, Pratap et al. 2010, Chimge and Frenkel 2012). Traditionally, RUNX1 has been classified as a tumor suppressor due to the frequency of mutations and translocations of *RUNX1* in lymphomas (Song et al. 1999, Li et al. 1999, Miething et al. 2007, De Braekeleer et al. 2009, Mangan and Speck 2011). There is also evidence to suggest that ectopic or increased RUNX1 protein expression promotes the development of certain cancers such as epithelial tumors, endometrial cancer, and acute lymphoid leukemia (Niini et al. 2000, Harewood et al. 2003, Planagumà et al. 2004, 2011, Abal et al. 2006, Planaguma et al. 2006, Doll et al. 2009, Hoi et al. 2010). RUNX1 is expressed in the epithelial cells of normal mammary glands (Blyth et al. 2010, Wang et al. 2011a, Janes 2011). In breast cancer, RUNX1: 1) appears to function primarily as a tumor suppressor protein as mutations in *RUNX1* are frequently observed in patients (The Cancer Genome Atlas Network

2012), 2) *RUNX1* transcript levels are significantly decreased in patients with metastatic breast cancer (Ramaswamy et al. 2003), and 3) in cell-based models of breast cancer development, the *RUNX1* gene locus is deleted during oncogenic transformation (Kadota et al. 2010). RUNX2 suppresses the growth of proliferating osteoblasts, and it has been suggested that in this context, RUNX2 can function as a tumor suppressor (Pratap et al. 2003, Galindo et al. 2005, Young et al. 2007b). Similarly, RUNX1 regulates the growth of hematopoietic cells (Bakshi et al. 2008); RUNX proteins bind to rDNA repeats and regulate the transcription of rRNA, which is the rate-limiting step in protein synthesis and cell growth (Grummt and Voit 2010). Although RUNX2 appears to inhibit osteoblastic cell growth, in multiple cancer types such as lymphoma, osteosarcoma, colon, pancreatic and prostate, increased RUNX2 levels are associated with the disease state, indicating that RUNX2 is likely functioning in an oncogenic manner within these cellular lineages (Stewart et al. 1997, Vaillant et al. 1999, Blyth et al. 2006, Kayed et al. 2007, Kuo et al. 2009, Akech et al. 2010, Sase et al. 2012). In normal mammary epithelial cells, low levels of RUNX2 can be detected, and a role for regulating milk production has been proposed (Inman and Shore 2003, Shore 2005). In breast cancer patients, detection of RUNX2 is associated with poor prognosis and clinical outcome, and in cell-based experiments, RUNX2 promotes the invasive and osteolytic properties of bone metastatic breast cancer cells (Pratap et al. 2008, Das et al. 2009, Onodera et al. 2010). These observations, taken together, suggest that RUNX1 may function as a breast cancer tumor suppressor, while RUNX2 may function as a breast cancer oncogene; however, the roles of RUNX1 and RUNX2 in breast cancer development have not been directly compared.

In breast cancer cells, RUNX2 is associated with the nuclear matrix. Runx2 targeting to the nuclear matrix is required for the Runx2-dependent transcription of genes responsible for the osteomimetic and bone resorptive phenotype of the MDA-MB-231 breast cancer cells (Barnes et al. 2004, Javed et al. 2005). A commonly observed translocation event in acute myeloid leukemia between chromosome 8 and 21, t(8;21), results in a fusion protein with the N-terminus of RUNX1 (including the Runt DNA-binding domain) and ETO. The AML-ETO fusion protein interacts with the nuclear matrix, however, the subnuclear distribution of AML-ETO is distinct from the distribution of normal RUNX1 expressed from the wild-type *RUNX1* allele (Barseguian et al. 2002).Although the predominant roles for RUNX1 and RUNX2 in tumorigenesis appear to differ, these observations demonstrate the importance of proper subnuclear targeting to the nuclear matrix for the execution of these RUNX-dependent functions.

Expression of a nuclear matrix-associated protein in cancer cells is often clinically relevant, and is not unique to RUNX proteins. For example, B23 is a nucleolar protein associated with the nuclear matrix that is highly expressed in prostate cancer (Subong et al. 1999). A nuclear matrix-specific variant of Ciz1, which normally interacts with and regulates the localization of p21, promotes cell growth *in vitro*, and is sufficient to identify early stage lung cancer (Higgins et al. 2012). NMP-22 is expressed in bladder cancer and is used to detect bladder cancer in patients (Soloway et al. 1996, Shariat et al. 2004, Zink et al. 2004b). Understanding the functions of nuclear matrix-associated proteins that are expressed in cancer cells, such as RUNX proteins, may lead to novel means of diagnosis and/or treatment.

There have been no studies to date analyzing the functions of ectopically-expressed RUNX proteins in breast cancer cells on a genome-wide scale. This is a critical point, as most of the previous studies that have focused on the role of RUNX2 in human breast cancer cells have relied on overexpression of wild-type or mutant *Mus musculus* Runx2 (Barnes et al. 2004, Javed et al. 2005, Pratap et al. 2008, Chimge et al. 2011). We are primarily interested in defining the native relationships between matrix-associated factors like RUNX proteins and the regulation of gene expression. Therefore, an approach in which RUNX proteins are depleted via RNAi was deemed preferable to overexpression for studying the functional contributions of RUNX proteins to the invasive phenotype of breast cancer cells.

In studies of Asian breast cancer patients, expression of RUNX2 was found to be associated with poor clinical outcome and estrogen receptor (ER) negative hormone status (Das et al. 2009, Onodera et al. 2010). To date, RUNX1 expression in breast cancer patients has not been examined histologically on a broad scale. Detection of RUNX1 has been characterized by The Human Protein Atlas in limited samples where no clinical inferences can be drawn; in these studies, RUNX1 is observed to be strongly expressed in both normal and tumor samples (Uhlén et al. 2005, Pontèn et al. 2008, Uhlen et al. 2010). It is therefore of great interest to understand the functions of ectopic RUNX proteins in breast cancer cells and the extent to which RUNX expression is associated with breast cancer patient tissue samples.

Controversial transcriptional associations of matrix-associated DNA

Many studies investigating the transcriptional role of DNA associated with the insoluble matrix/scaffold have experimentally prepared “nuclear halos” before digestion of looped (or non-matrix-associated) DNA (Adachi et al. 1989, Gerdes et al. 1994, Maya-Mendoza and Aranda-Anzaldo 2003, Ostermeier et al. 2003, Heng et al. 2004, Keaton et al. 2011). While SARs and MARs have previously been described interchangeably, some recent studies have begun to make functional distinctions between these two classes of matrix-associated DNA and their associations with transcription. Technical improvements in the ability to quantitatively measure specific sequences of nucleic acids in a high-throughput manner appear to be driving a reinvestment of effort into understanding the transcriptional associations of matrix-associated DNA. Isolation of the matrix-associated DNA using LiS buffer reveals a positive correlation between gene expression and association of upstream genomic elements with the nuclear matrix (Linnemann et al. 2007, Keaton et al. 2011). These observations are similar to what was previously observed in smaller-scale or non-gene-specific experiments (Herman et al. 1978, Ciejek et al. 1983, Jost and Seldran 1984, Jackson and Cook 1985, Cook 1989, Ogata 1990, Jackson et al. 1993, Wansink et al. 1993). In contrast, when high concentrations of sodium chloride are used to extract chromatin prior to nuclease digestion, a strong correlation between proximity to the nuclear matrix and gene silencing can be observed (Maya-Mendoza and Aranda-Anzaldo 2003, Rivera-Mulia and Aranda-Anzaldo 2010, Trevilla-García and Aranda-Anzaldo 2011). In these studies, proximity to the nuclear matrix is measured by site-specific PCR of DNA associated with the nuclear matrix following timed digestion with DNAse I. These observations that proximity to the nuclear matrix correlates with gene silencing are in stark contrast with many previous studies demonstrating that the nuclear matrix is the site of active transcription. In an attempt to reconcile these observed transcriptional associations for matrix-associated DNA isolated via LiS or sodium chloride, Linnemann and colleagues performed both isolation methods in parallel, followed by hybridization of matrix-associated sequences to a chromosome-wide oligonucleotide array (Linnemann et al. 2008). In this study, to differentiate between the two extraction methods, sequences isolated via sodium chloride or LiS were referred to by the original descriptions of the structures isolated, MARs and SARs, respectively. Genome-wide transcript levels were measured by an Affymetrix cDNA array, and compared to the chromosome-wide profile of matrix-associated DNA isolated by sodium chloride or LiS. Here, it was observed that enrichment of sodium chloride-isolated MARs within genes is negatively correlated with gene expression, while enrichment of LiS-isolated SARs in upstream genomic elements is positively correlated with gene expression. The positions of MARs and SARs were also different; SARs were typically observed 5’ of genes and in gene-rich regions, while MARs were typically in gene-poor regions or within a silenced gene body. The Linnemann study is the only large-scale study comparing quantitative positional measurements of matrix-associated DNA and gene expression between the two isolation methods. It is, however, not the first study to compare the isolation methods. Belgrader and colleagues performed a rigorous analysis of the structural and biochemical differences between nuclear matrices isolated via sodium chloride and LiS (Belgrader et al. 1991). This comparative study also included the more gentle isolation of the nuclear matrix using low concentrations of ammonium sulfate. They observed gross differences in the ultrastructural preservation of the residual nuclear matrix following extraction via these three methods. Matrices isolated via sodium chloride had non-apparent nucleoli and a high degree of aggregate-like structures bound to the fibrogranular nuclear matrix. When the nuclear matrix was isolated using LiS, the nucleolus was visible, but diffuse in appearance; LiS-isolated nuclear matrices had fewer aggregates than sodium chloride-isolated nuclear matrices, but these aggregates were still frequently observed. Nuclear matrices isolated via ammonium sulfate, have extremely well-preserved nucleoli, and very few, if any, aggregates form on the filamentous matrix structure. It has been proposed the studying structural components of the nucleus requires the preservation of structure to the highest degree possible (Jackson and Cook 1995, Nickerson 2001). The generation of data via approaches that best recapitulate the physiological states of intact biological systems is critical for our understanding of changes in phenotypic gene expression regulation. It is possible that the matrix-associated DNA may contain artifacts or may be missing information when isolated via techniques that do not preserve the structural integrity of the nuclear matrix. We hypothesize that isolation of matrix-associated DNA in a manner that preserves the architectural integrity of the nuclear matrix will provide a high level of insight into the native interactions between chromosomes and the nuclear matrix. Applying this approach at a genome-wide scale to compare gene expression profiles and matrix-associated DNA profiles from normal and cancer cells will therefore allow for quantitative investigation into the role of the nuclear matrix in transcriptional regulation of phenotypic gene expression.

General question to be addressed

Cancer cell nuclei are generally disorganized. Are components of the nuclear matrix and associated DNA segments similarly disorganized? If so how do these observations relate to transcriptional functions?

Specific questions to be experimentally addressed

Is genomic DNA associated with the nuclear matrix enriched in particular sequences or regions? Does the pattern of matrix-associated DNA correlate with the enrichment of euchromatin or heterochromatin? How does enrichment of matrix-associated DNA compare to gene expression patterns? Are there differences in matrix-associated DNA between a normal mammary epithelial cell and a malignant metastatic breast cancer cell?

In blood and bone cells, RUNX proteins control cell growth and proliferation through the attenuation of protein synthesis via transcriptional regulation of ribosomal RNA (Young et al. 2007a, Bakshi et al. 2008, Ali et al. 2010, 2012). Breast cancer cells are highly proliferative in the presence of RUNX1 and

RUNX2 proteins. Are the mechanisms of Runx-mediated growth regulation operative in breast cancer cells? What genes are regulated by RUNX proteins in a breast cancer cell? Where, in the genome, are RUNX proteins binding in breast cancer cells? To what extent are these binding events associated with gene expression regulation?

RUNX1 and RUNX2 are expressed in breast cancer cell lines, what is their expression pattern in tumor tissue from breast cancer patients? Is the expression level of RUNX1 or RUNX2 in human breast cancer tissue associated with any particular type, grade, or stage of breast cancer? Does the presence of growth factor or hormone receptors such as HER2, estrogen receptor (ER), progesterone receptor (PR), or androgen receptor (AR) affect the expression intensity of RUNX1 or RUNX2 in tissue from human breast cancer?

Experimental design

To understand functional relationships between parameters of nuclear organization and gene expression regulation, we examined DNA sequences associated with the nuclear matrix in normal breast epithelial cells compared to metastatic breast cancer cells. To address this question we developed and describe a method for the isolation of nuclear matrix-associated DNA (NM-Seq) that preserves nuclear integrity. NM-seq utilizes next-generation sequencing to facilitate a genome-wide unbiased comparison between matrix-associated DNA in normal mammary epithelial cells (MCF10a) and malignant metastatic breast cancer cells (MDA-MB-231). Structure-function relationships between gene expression patterns and nuclear organization are defined by comparing the profiles of matrix-associated DNA in MCF10a and MDA-MB-231 cells with the binding profiles of several histone modifications (H3K4me3, H3K27me3, and H3K9me2) and with genome-wide transcriptome data.

In this study, we also examine the phenotypic roles of the nuclear matrix-associated proteins RUNX1 and RUNX2 in MDA-MB-231 malignant metastatic breast cancer cells. We examine the extent to which RUNX proteins regulate cell growth through the transcriptional regulation of ribosomal RNA. On a genome-wide scale, we define the genes responsive to RUNX protein levels by knocking down endogenous RUNX proteins. We further characterize RUNX proteins in breast cancer cells by comparing RUNX binding on a genome-wide scale with gene expression patterns to establish positional relationships for functional binding of RUNX proteins. To understand potential clinical impacts of these *in vitro* findings, we examine the expression of RUNX1 and RUNX2 in tumor tissue from more than 125 North American breast cancer patients using immunohistochemistry and compare the detected levels of RUNX proteins with several histopathological markers of breast cancer as well as the growth/hormone receptors HER2, ER, PR, and AR.

We describe two open-source software packages that were specifically developed to assist in the understanding of general parameters of nuclear organization. FoxPrimer is a Real-Time qPCR primer design tool and database with a web interface. FoxPrimer is designed for high-throughput creation, storage, and retrieval of high quality primers for the validation of omics-type datasets. FoxPrimer was used to validate gene expression data for Runx-regulated genes in breast cancer cells. PeaksToGenes is a command-line interface program designed to create average gene plots and run statistical tests to define associative patterns in binding of nuclear features near genes. PeaksToGenes was used to model regulatory binding events for both matrix-associated DNA and Runx proteins.

CHAPTER 2 ISOLATION AND CHARACTERIZATION OF NUCLEAR MATRIX ASSOCIATED DNA IN BREAST CANCER CELL LINES

Authors and contributions

Jason R. Dobson, Deli Hong, Hai Wu, Jane B. Lian , Janet L. Stein, Jeffery A. Nickerson, Andre J. van Wijnen, Gary S. Stein.

NM-seq designed by JRD, JBL, JLS, JAN, AJVW, and GSS.

NM-seq optimized and executed by JRD.

Immunofluorescence imaging performed by DH.

Chromatin immunoprecipitation performed by HW.

ChIP-seq libraries prepared by JRD.

Computational analysis performed by JRD.

Introduction

Our understanding of how genomic DNA is organized within the nucleus of a eukaryotic cell has increased dramatically in recent years with the advent of deep sequencing technologies and research consortia such as ENCODE (ENCODE Project Consortium 2011). One of the most fundamental questions under investigation is addressing how higher-order structure of genomic DNA participates in gene regulation.

The concept of nuclear organization has evolved through many observations of non-random localization of functional regulatory factors in the nucleus (Lanctôt et al. 2007, Misteli 2007, Zaidi et al. 2007, Stein et al. 2011). Many critical functional activities such as splicing, transcription, replication, DNA repair and RNA processing are associated with a nuclear structural scaffold comprised of a fibrogranular network of RNPs, the ‘nuclear matrix’ (Berezney et al. 1995, Nickerson 2001). While the nuclear matrix is associated with many nuclear functions, this study is focused on transcription and the spatial relationships of genes and transcription with the nuclear matrix.

The nuclear matrix is a filamentous structure of RNPs (Fey et al. 1986). Using regressive EDTA staining, this network of RNPs can be observed in intact cells by electron microscopy (Monneron and Bernhard 1969, Bernhard 1969). The nuclear matrix, associated proteins and DNA can be biochemically isolated using a variety of methods such as high salt extraction and nuclease digestion (Berezney and Coffey 1977), low salt 3,5-diiodosalicylic acid, lithium salt (LiS) buffer extraction followed by nuclease digestion (Mirkovitch et al. 1984), electroelution of intact nuclei (Jackson and Cook 1988), or a more “gentle” nuclease digestion and salt extraction procedure (Fey et al. 1986, He et al. 1990). Common to all approaches is the digestion of DNA and extraction of soluble chromatin resulting in the isolation of a residual nuclear matrix.

Transcription occurs at distinct foci or transcription factories within the nucleus, which are associated with the nuclear matrix (Jackson and Cook 1985, Jackson et al. 1993, 1998). Transcription factors such as RNA Pol II, co-factors, or lineage specific transcriptional regulators are similarly organized into transcription factories within which actively transcribed genes are localized (Stein et al. 2011). DNA-bound proteins such as topoisomerase or transcription factors such as hormone receptors or SATB1 recruit DNA to the nuclear matrix in a sequence-specific manner (Van Steensel et al. 1991, Dworetzky et al. 1992, Dickinson et al. 1992, van Wijnen et al. 1993). Identifying genomic sequencing enriched in these transcription factories may provide insight into novel matrix-associated proteins functioning to recruit genes to transcription factories. It was hypothesized that computational methods may be able to predict DNA associated with the nuclear matrix based on a consensus sequence (Boulikas

1995, Singh et al. 1997), however, recent experimental approaches have demonstrated that the sequences of DNA associated with the nuclear matrix do not appear to have a conserved sequence motif (Wilson and Coverley 2013). These experimental observations are concordant with the growing database of specific sequences that are recognized by transcription factors (Heinemeyer et al. 1998, Bryne et al. 2008). In the absence of a predictive model for matrix-associated DNA, these sequences must therefore be experimentally defined for a given cell population or physiological state.

To study DNA associated with the nuclear matrix, two common approaches have been taken to isolate and identify matrix-associated DNA. Isolation is achieved through a step-wise process. First, chromatin is extracted via high salt or LiS buffer. This extraction step results in an intermediate “nucleoid” or “nuclear halo” structure in which DNA is observed to form loops emanating from a residual structure (Keaton et al. 2011, Trevilla-García and Aranda-Anzaldo 2011). It is thought that DNA at the base of these loops is matrix-associated DNA (Gasser and Laemmli 1987), therefore loop DNA is separated from matrix DNA by cutting DNA near the base of the loops with either DNAse I or restriction endonucleases. Regions of DNA enriched in association with the nuclear matrix can be measured by fluorescent *in situ* hybridization (FISH) (Gerdes et al. 1994), polymerase chain reaction (PCR) (Maya-Mendoza and Aranda-Anzaldo 2003) or more recently by hybridization to an oligonucleotide tiling array (Heng et al. 2004).

A confusing nomenclature was developed to describe the types of DNA regions associated with the nuclear matrix based on the experimental approach used to extract chromatin. Isolation via high salt or LiS buffer was described as isolating the nuclear matrix or the nuclear scaffold, respectively. Therefore, DNA sequences isolated from the nuclear matrix or the nuclear scaffold were referred to as matrix associated regions (MARs) and scaffold associated regions (SARs), respectively. Initial studies on transcriptional roles for matrix-associated DNA were typically focused on a single gene or cluster of genes, and DNA isolated via either of these approaches appeared to similarly enriched for transcriptionally active genes (Ciejek et al. 1983, Jost and Seldran 1984, Ogata 1990, Gerdes et al. 1994). Therefore, the MARs and SARs were considered reasonably interchangeable and led to the convention of referring to matrix-associated DNA sequences as S/MARs (Nickerson et al. 1995).

Recent PCR-based quantification of matrix-associated DNA of broad gene loci such as the albumin locus led Aranda-Anzaldo and colleagues to propose that matrix-associated DNA is a transcriptionally repressive element rather than a transcriptionally activating element (Maya-Mendoza and Aranda-Anzaldo 2003, Rivera-Mulia and Aranda-Anzaldo 2010, Trevilla-García and Aranda-Anzaldo 2011). These proposed repressive functions for matrix-associated DNA conflict with many previous studies indicating that the nuclear matrix-associated DNA is enriched in actively transcribed genes (Herman et al. 1978, Ciejek et al. 1983, Jost and Seldran 1984, Jackson and Cook 1985, Cook 1989, Ogata 1990, Jackson et al. 1993, Wansink et al. 1993, Gerdes et al. 1994). The advent of genomic tiling microarrays allowed for identification of S/MARs on a broad chromosome-wide scale (Linnemann et al. 2007, Keaton et al. 2011) and later allowed Krawetz and colleagues to make a direct comparison between the positions of MARs (isolated via high salt) and SARs (isolated via LiS) and gene expression patterns on a broad scale (Linnemann et al. 2008). We are inclined to emphasize the results of this approach as this is the only study in which the chromosome-wide quantification of matrix-associated DNA isolated via high salt or LiS buffer has been compared. Linnemann et al. demonstrate that matrix-associated DNA sequences isolated via high salt (MARs) are located within the gene bodies of poorly or non-expressed genes. In contrast, matrix-associated sequences isolated via LiS (SARs) are located 5' of actively transcribed genes. This study suggests that the type of DNA (in relation to transcription) is dependent on the isolation method employed. Belgrader et al. examining the fibrogranular structure of the nuclear matrix via electron microscopy follow multiple extraction methods (Belgrader et al. 1991). High salt extraction resulted in a non-apparent nucleolus and a high degree of aggregate structures on the matrix fibrils; LiS extraction causes a diffuse nucleolar structure to form and a moderate level of aggregate structures. Using a gentler extraction of chromatin via ammonium sulfate, Belgrader et al. observed a high degree of nucleolar preservation and few, if any, aggregates on the matrix fibrils. It has been argued that methods used to isolate the nuclear matrix should be evaluated by the protocol's ability to preserve the integrity of structures, such as the RNP fibrils, which can be observed in the intact nucleus (Nickerson 2001), and we hypothesize that characterization of DNA sequences isolated from a structurally intact nuclear matrix will allow for a high degree of insight into the functional properties of matrix-associated DNA.

To study the association of genomic DNA with the nuclear matrix under conditions in which structural integrity is preserved, we have optimized an experimental approach for biochemical isolation of the nuclear matrix based on the Penman (He et al. 1990, Nickerson et al. 1997, Wan et al. 1999) approach for nuclear matrix isolation. The Penman method, as demonstrated by Belgrader et al., better preserves the structural organization of the nuclear matrix as compared to the high salt and LiS extraction methods. To improve upon previous approaches, which quantified matrix-associated DNA via hybridization-based arrays, we employed recent technological advances in deep sequencing to study the genome-wide patterns of nuclear matrix-associated DNA utilizing long (100bp) paired-end sequencing. This method of sequencing gives us a high volume of reads as well as a high degree of confidence in the positional mapping of reads to the genome. Thus, we can more accurately identify specific sequences enriched in association with the nuclear matrix. We have named this method nuclear matrix sequencing (NM-seq).

In cancers, proteins that are specific to the disease state have been identified to associate with the nuclear matrix, however, it is not known if genomic sequences associated with the nuclear matrix are similarly altered (Getzenberg et al. 1991, 1996, Partin et al. 1993, Samuel et al. 1997). To understand how matrix-associated DNA might be involved with a disease such as breast cancer and to demonstrate a potential application of NM-seq, we applied our technique of isolating nuclear matrix-associated DNA to two breast cell lines: MCF10a immortalized normal mammary epithelial cells and MDA-MB-231 mesenchymal-like metastatic breast cancer cells.

Considering first the non-tumorigenic MCF10a cells, our results demonstrate that NM-seq can identify matrix-associated DNA in both gene-rich/actively-transcribed regions and gene-poor/weakly-transcribed regions, which suggests we have preserved both repressive (MAR) and active (SAR) interactions with the nuclear matrix. When comparing the normal and cancer cells, we observed disparate patterns of nuclear matrix-associated DNA between the MCF10a and MDA-MB-231 cells with respect to gene-rich versus gene-poor association and GC content of matrix-associated DNA. Integrating gene expression data with matrix-associated DNA enrichment patterns, we found that matrix-associated DNA is enriched in regions flanking transcribed genes in MCF10a cells, while in MDA-MB-231 cells poorly or non-expressed gene regions are more associated with the nuclear matrix. These results suggest that the malignant metastatic breast cancer cell line MDA-MB-231 has major differences in the functional association of genomic DNA with the nuclear matrix and in gene expression regulation when compared to the normal mammary epithelial cell line MCF10a.

Results

Preparation of nuclear matrix fraction while preserving nuclear structure

Isolation of the nuclear matrix fraction is typically achieved through step-wise nuclease digestion and salt extraction (Nickerson 2001). Recent approaches used to study matrix-associated DNA (NM-DNA) have biochemically extracted the matrix by salt extracting nuclei, making so-called "halo" preparations, followed by nuclease digestion (Maya-Mendoza and Aranda-Anzaldo 2003, Heng et al. 2004, Linnemann et al. 2007, Rivera-Mulia and Aranda-Anzaldo 2010, Keaton et al. 2011, Trevilla-García and Aranda-Anzaldo 2011). Biochemically, these types of preparations do isolate the nuclear matrix fraction, but they also disrupt nuclear integrity (Belgrader et al. 1991). Structural integrity of the nucleus is critical for maintenance of proper DNA-matrix interactions. DNA sequences associated with the nuclear matrix are highly dynamic and subject to dramatic changes when nuclei are salt extracted (Craig et al. 1997). Using a stabilization reagent, such as copper ion, prevents these salt-induced shifts in matrix-associated DNA (Belgrader et al. 1991).

Given the highly sensitive nature of DNA-matrix interactions, in an attempt to preserve native DNA-matrix interactions we designed a protocol based on previous approaches demonstrating a high degree of structural preservation of the nuclear matrix. The nuclear matrix can be purified while preserving nuclear structure using formaldehyde stabilization prior to nuclease digestion and salt extraction (Nickerson et al. 1997). We used this experimental approach as the starting point in our experimental optimization to isolate NM-DNA. The procedure we developed for isolating the nuclear matrix uses formaldehyde-stabilized nuclei, restriction enzyme digestion and high salt extraction **(Figure 2.1)**, and preserves both nuclear structure and functional domain localization **(Figure 2.2)**.

After nuclear matrix isolation, formaldehyde stabilization is reversed via incubation at 55°C for 16 hours in elution buffer (1% SDS, 100mM Sodium bicarbonate), DNA is isolated and randomly sheared via sonication, and finally, libraries are prepared for deep sequencing **(Figure 2.1 A)**. To control for biases in chromosome copy number, restriction enzyme cut sites, sample preparation (size selection and PCR) and sequencing, DNA isolated from digested nuclei prior to salt extraction was used as “Input” control to define enrichment regions for NM-DNA. Applying this novel approach to our model cell lines resulted in 166,861,270 mapped reads and 255,049,970 mapped reads for the MCF10a and the MDA-MB-231 NM-DNA samples, respectively. Each mapped read pair represents a high-confidence region as it is mapped from long (100bp) paired-end reads.

During the nuclear matrix isolation procedure, we see a significant extraction (>90%) of genomic DNA as measured by DAPI incorporation **(Figure 2.2 A)**. Further validating the effectiveness of this protocol to preserve nuclear integrity, we also observe that the spatial distribution and staining intensity of several matrix-associated proteins (Coilin, NPAT, PML, SC-35, UBF, and Pol II) are maintained throughout nuclear matrix isolation **(Figure 2.2 B)**.

Figure 2.1 Diagram of nuclear matrix-associated DNA isolation

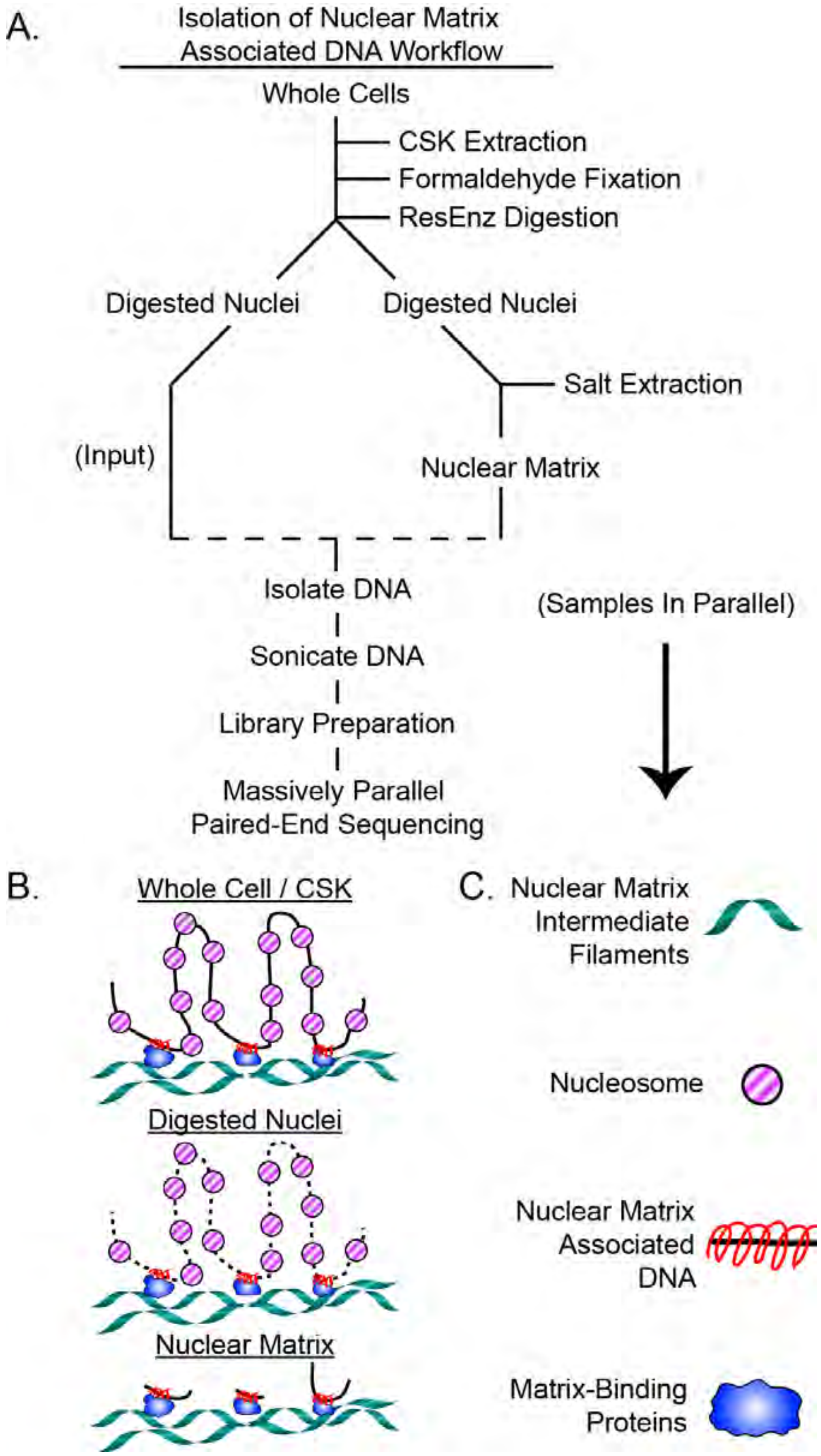


Figure 2.1 Diagram of nuclear matrix-associated DNA isolation

(A) Diagram depicting the workflow of nuclear matrix isolation. **(B)** Representation of matrix-DNA interactions in whole cell/CSK extracted nuclei, digested nuclei, and after salt extraction. **(C)** Legend for symbols/drawings used in **(B)** to represent key components of nuclear organization during isolation of the nuclear matrix fraction.

Figure 2.2 Nuclear integrity and functional domains remain intact during isolation of the nuclear matrix fraction.

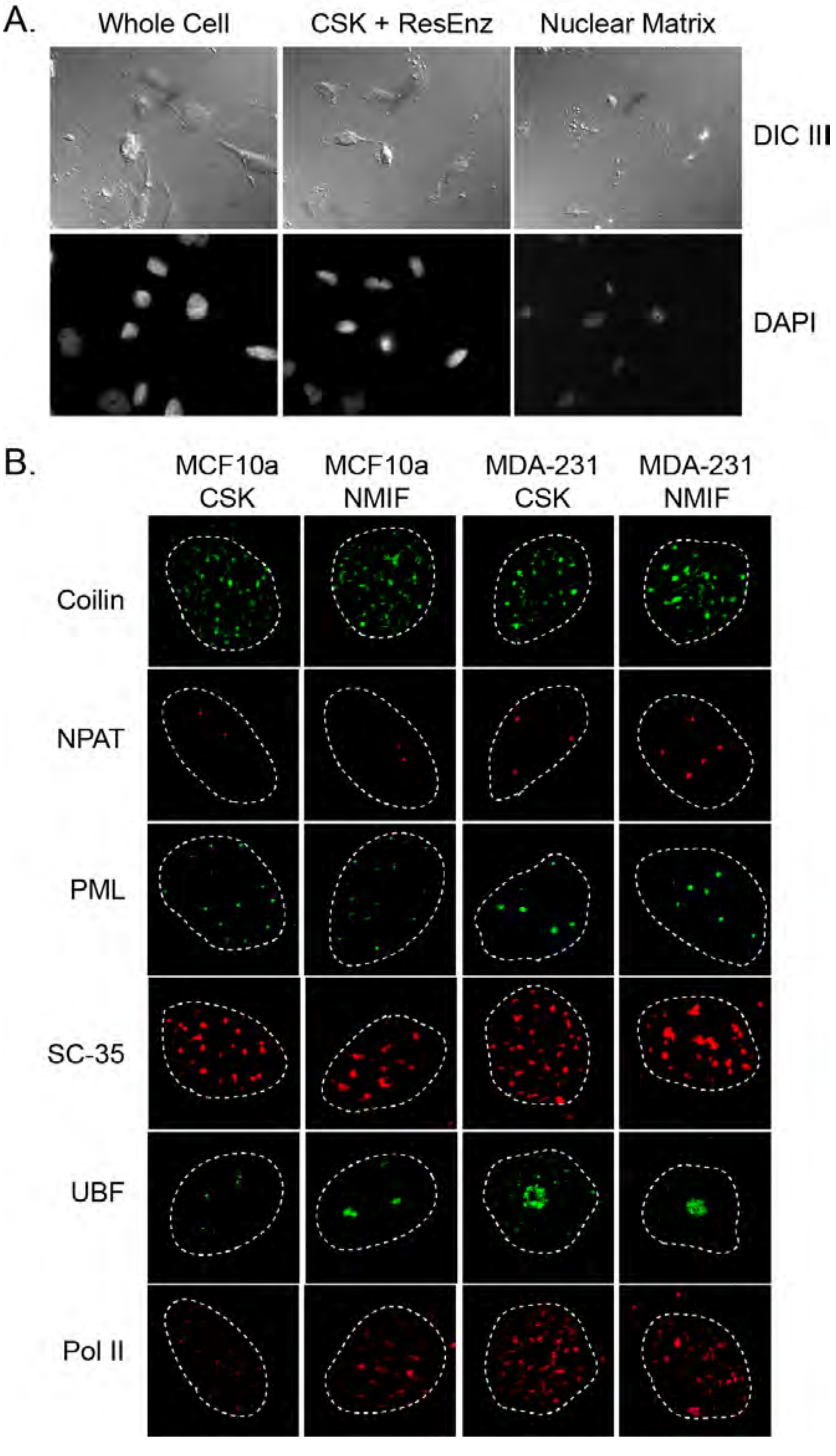


Figure 2.2 Nuclear integrity and functional domains remain intact during isolation of the nuclear matrix fraction.

(A) DIC III phase contrast images and DAPI staining of MDA-MB-231 cells at three stages of isolation of nuclear matrix: whole cell, CSK-extracted nuclei digested with restriction enzymes, and nuclear matrix. **(B)** Confocal immunofluorescence micrographs of critical nuclear proteins in MCF10a and MDA-MB-231 cells after CSK extraction or following full nuclear matrix isolation (NMIF). Proteins examined include: Coilin (Cajal bodies / modification of small nuclear/nucleolar RNAs), NPAT (nuclear protein, ataxia-telangiectasia locus - histone gene transcription), PML (promyelocytic leukemia - PML bodies), SC-35 (serine/arginine-rich splicing factor 2 - pre-mRNA splicing), UBF (upstream binding transcription factor, RNA polymerase I - rRNA transcription), and Pol II (RNA polymerase II – mRNA synthesis).

Defining enriched regions for nuclear matrix-associated DNA

When performing deep-sequencing for chromatin immunoprecipitation (ChIP-seq), it is common protocol to define enriched regions for a given pulldown using some form of “peak calling” algorithm (Landt et al. 2012). Many of the most popular “peak calling” algorithms (MACS (Zhang et al. 2008), SPP (Kharchenko et al. 2008), PeakSeq (Rozowsky et al. 2009)) operate on a major assumption: enrichment of a single protein will produce a defined genomic interval width from which to build a peak-shift model for defining the peak regions. This assumption cannot be made when trying to define enriched regions from matrix-associated DNA, as isolation of the nuclear matrix enriches for an unknown number of proteins associated with DNA in widths of unknown length (Davie 1995, Samuel et al. 1997). To address this issue we developed a more direct method of defining genomic regions enriched (or depleted) in association with the nuclear matrix. At the most basic level, we are examining the ratio of NM-DNA over the DNA isolated from the digested nuclei fraction (DigNuc) in specific genomic regions defined below. This method of measuring NM-DNA enrichment as the ratio of matrix-associated DNA over non-extract DNA is similar to methods applied when using genomic-tiling arrays (Linnemann et al. 2007, 2008, Keaton et al. 2011). Due to the advantageous genome-wide nature of our approach, the DigNuc fraction provides information about regions where genomic sequences are over- and under-represented; using this information we can therefore determine how to best scale the ratio of NM-DNA over DigNuc when defining enriched regions. We used the sequence enrichment scaling (SES) algorithm, as it empirically determines an appropriate scaling factor to be applied to the input (DigNuc) channel based on the sorting statistic (Diaz et al. 2012). Briefly, SES compares the input (DigNuc) and IP (NM-DNA) reads across the entire genome to identify systematic biases caused by the processing samples and deep sequencing. These genome-wide biases are weighted and converted into a scaling factor, which is then applied when calculating the enrichment of the IP (NM-DNA) sample over the input (DigNuc) in a specific region of the genome.

Sequence complexity of nuclear matrix-associated DNA is different between MCF10a and MDA-MB-231 cells

Structural matrix-associated DNA elements such as matrix-attachment regions (MARs) and scaffold-attachment regions (SARs) are procedurally isolated using salt conditions quite different than those we have used in our nuclear matrix isolation protocol (Berezney and Coffey 1977, Mirkovitch et al. 1984). Based on a comparative study by Linnemann et al. these sequence elements have distinct associations with nuclear functions (Linnemann et al. 2008). MARs appear to be highly enriched in gene-poor or transcriptionally silent regions, while SARs are highly associated with gene-rich, actively-transcribed genomic regions. It is thought that each of these element types can be defined by sequence motifs;

however, computational efforts have proven not to predict sequences isolated from cells (Wilson and Coverley 2013). While the order of operations (salt extraction, nuclease digestion) and detergents/salts used affects which type of DNA element will be enriched in the matrix fraction, the procedures used to isolate the aforementioned sequence elements significantly disrupt nuclear structure and organization.

NM-seq preserves the integrity of nuclear organization; we therefore feel it is critical to report the sequence complexity of enriched sequences isolated from our model cell lines. By partitioning the genome into 200bp non-overlapping windows, we calculated the ratio of NM-DNA to DigNuc input (NM-DNA enrichment) and the corresponding percent of G plus C bases (GC content) in each window. Plotting the pairs of NM-DNA enrichment versus GC content in a scatterplot and calculating the Pearson correlation coefficient (r), we observe that in MDA-MB-231 cells the regions most associated with the nuclear matrix are AT-rich and that GC-content and NM-enrichment are slightly negatively correlated (r = -0.24). In the MCF10a cells, we observe that AT/GC content is relatively evenly spread across NM-DNA-enriched and NM-DNA-depleted regions, with a very weak positive correlation trend (r = 0.12) **(Figure 2.3)**. These trends can be further visualized by separating the 200bp non-overlapping genomic regions into ten groups based on the rank of the NM-DNA enrichment in the region (1 = most enriched, 10 = least enriched) and plotting the distribution of GC percentages for the regions in each group **(Figure 2.4)**.

In the MDA-MB-231 cells (breast cancer cell line), NM-DNA is enriched in AT-sequences, while for the MCF10a (normal mammary epithelial) cell line, NM-DNA appears to be AT/GC neutral. In the human genome, genic regions are GC-rich or GC-neutral, while gene-poor regions are AT-rich (Lander et al. 2001). The observed associations of NM-DNA sequence complexity combined with the distributions of AT-rich and GC-neutral regions in the human genome suggests that NM-DNA may be enriched in different genomic regions within normal cells and malignant metastatic cells.

Figure 2.3 Nuclear matrix-associated DNA weakly correlates with AT-rich sequence in MDA-MB-231 cells, but not in MCF10a cells

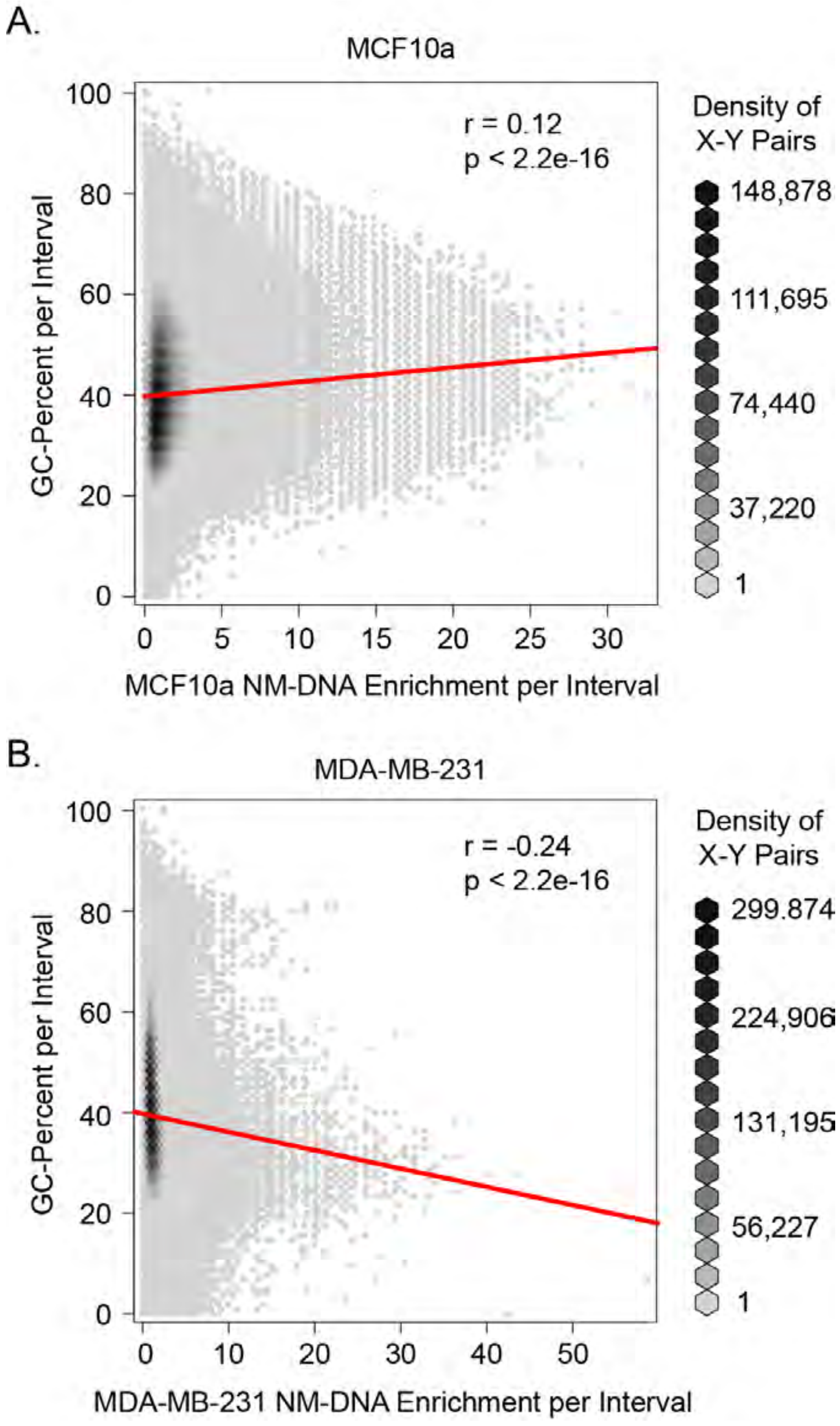


Figure 2.3 Nuclear matrix associated DNA weakly correlates with AT-rich sequence in MDA-MB-231 cells, but not in MCF10a cells

(A and B) Using genome-wide non-overlapping 200bp intervals, nuclear matrix-associated DNA (NM-DNA) enrichment (NM-DNA / Digested Nuclei) was measured in MCF10a cells **(A)** and MDA-MB-231 cells **(B)** and is represented on the x-axis. For each 200bp interval, the corresponding percent of G plus C bases within the interval is plotted as the dependent variable on the y-axis. These x-y pairs are plotted for the entire genome using a hexagonal density scatterplot where an increased number of shared x-y observations are represented as darker hexagons (right panels). Pearson r and p-value were calculated for each cell line with the correlation line plotted in red; the Pearson r and p-value are indicated in the upper right of the scatterplot for each cell line.

Figure 2.4 GC content as a function of nuclear matrix-associated DNA enrichment rank shows a similar trend as the scatter plot; matrix-associated DNA in GC-neutral in MCF10a cells, and AT-rich in MDA-MB-231 cells.

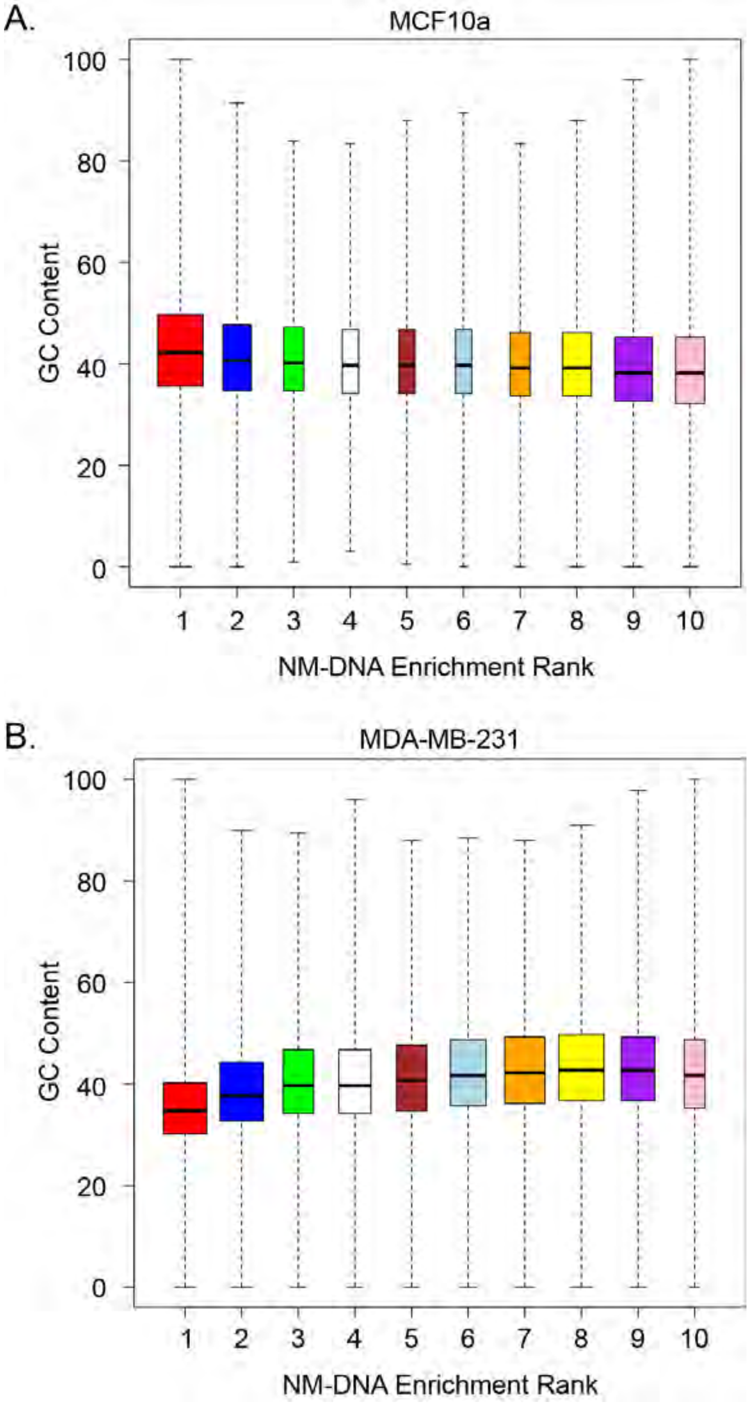


Figure 2.4 GC content as a function of nuclear matrix-associated DNA enrichment rank shows a similar trend as the scatter plot; matrix-associated DNA in GC-neutral in MCF10a cells, and AT-rich in MDA-MB-231 cells.

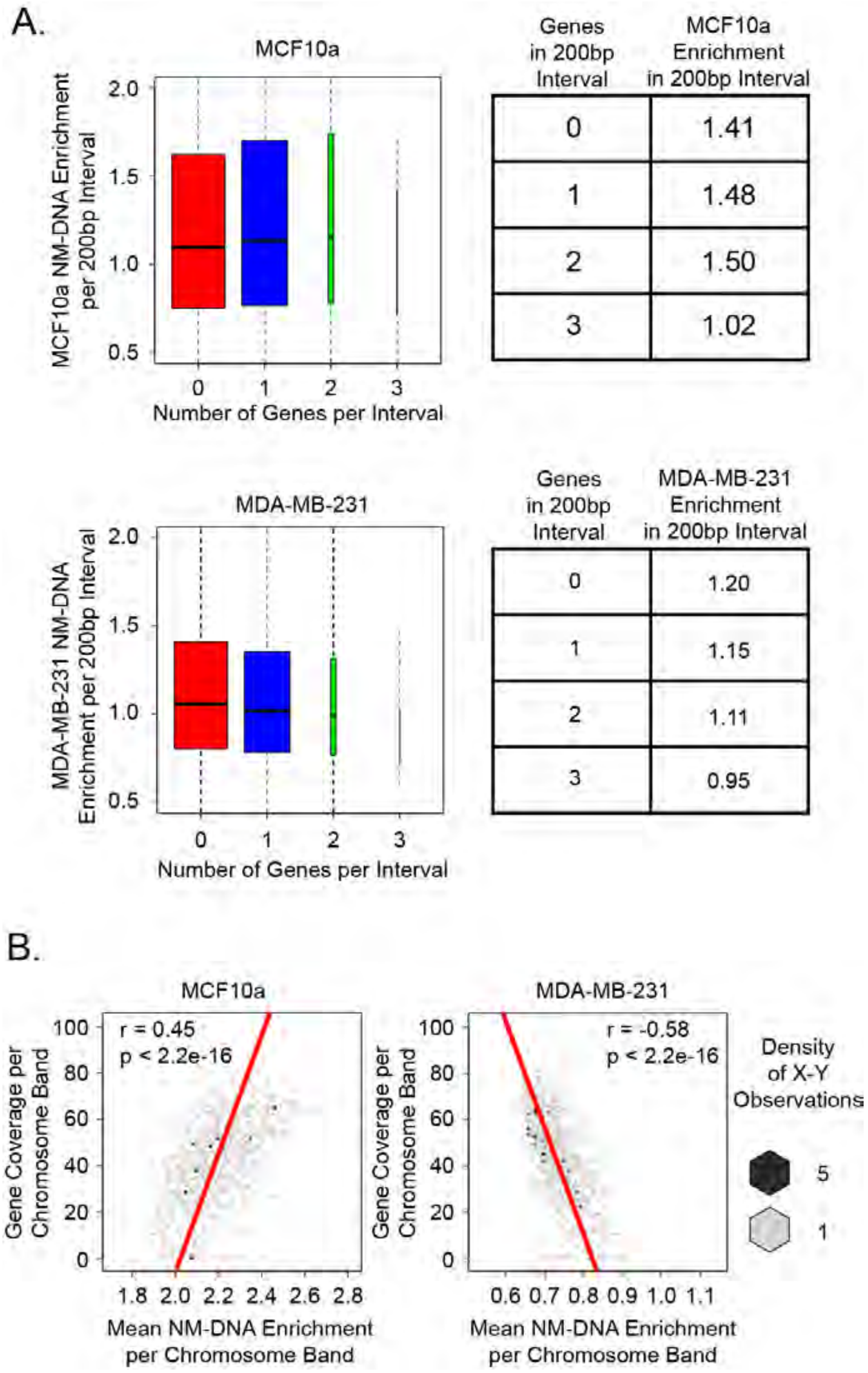
(A and B) For each cell line (MCF10a **(A)** and MDA-MB-231 **(B)**), the NM-DNA enrichment ratios in 200bp genome-wide non-overlapping intervals were measured and subdivided into 10 approximately equal groups based on ranks (rank 1 = top ~10% NM-DNA enrichment ratios, rank 10 = bottom ~10% NM-DNA enrichment ratio). For each of these ranks, the GC content was measured for the intervals in each rank (defined by NM-DNA enrichment) and the resultant GC distribution was plotted (y-axis) in box and whiskers format. Because there were some ties in NM-DNA enrichment ratios, the number of observations of NM-DNA enrichment/GC content was slightly different in each column therefore the width of the box and whiskers plot reflects the number of observations made.

Nuclear matrix-associated DNA is enriched in gene-poor regions in MDA-MB-231 cells and gene-rich regions in MCF10a cells

We next wanted to understand the positions of NM-DNA isolated from NM-seq compare to gene positions. Using the same 200bp windows for measuring GC-content and NM-DNA, we measured the number of RefSeq genes that are found in each 200bp window (Pruitt et al. 2009). This is a greedy definition in which we are defining a gene as being within a given 200bp interval if any portion of the gene body overlaps with the 200bp interval. In these 200bp non-overlapping intervals, we find there are up to three genes within the interval and will refer to the number of genes per interval as a class of 200bp intervals. Instead of plotting this data as a scatterplot, we instead plotted box and whisker plots of the distribution of NM-DNA enrichment values found in each class of 200bp intervals for each cell line **(Figure 2.5 A)**. As there are very few regions in the genome in which there are two or three genes per 200bp interval, the width of the box and whisker plots correlates with the number of 200bp intervals containing 0, 1, 2 or 3 genes. These box and whisker plots are cropped to show the slight differences in the median (horizontal line) NM-DNA enrichment for each class of 200bp interval. The tables to the right of each box and whisker plot show the mean NM-DNA enrichment value per class of 200bp interval **(Figure 2.5 A right)**.

To further investigate where in the genome NM-DNA isolated via NM-seq is enriched relative to genes, we used cytologically-defined chromosomal band coordinates as regions of interest (Meyer et al. 2013). These bands are visualized by Giemsa staining, where the density of genes within each cytological band is inversely correlates with the observed intensity of staining, and have been used for comparison of hybridization of matrix-associated DNA in a previous study (Craig et al. 1997). For each cytological band, we calculated the enrichment ratio of NM-DNA to digested nuclei as well as the percentage of the cytological band that is defined as a gene by RefSeq (Pruitt et al. 2009). By plotting these paired observations per cytological band in a scatter plot and calculating the Pearson r, we observe that in MCF10a cells NM-DNA enrichment is correlative with gene-dense regions, while in MDA-MB-231 cells NM-DNA enrichment is correlated with gene-poor regions **(Figure 2.5 B)**. When the relationships between gene-rich and gene-poor are viewed at the resolution of cytological bands, there are strong relationships between NM-DNA enrichment and gene density, however, when viewed in finer detail (200bp), these trends are not readily apparent. This observation suggests that NM-DNA-enrichment associations with gene-density are broad and may be involved in mediating large-scale organization of genomic DNA. These observations suggest that our hypothesis based on differences in sequence complexity that matrix-associated DNA in each cell line would be enriched in different genomic locations is correct. Observed differences in both sequence complexity and gene-rich versus gene-poor association lead us to hypothesize that NM-DNA is likely to be differentially associated with transcriptional activities between normal and cancer cell lines.

Figure 2.5 Nuclear matrix-associated DNA is broadly associated with gene-rich regions in MCF10a cells and with gene-poor regions in MDA-MB-231 cells. On a small (200bp) scale, these associations are not observed.



Genes in 200bp Interval	MCF10a Enrichment in 200bp Interval
0	1.41
1	1.48
2	1.50
3	1.02

Genes in 200bp Interval	MDA-MB-231 Enrichment in 200bp Interval
0	1.20
1	1.15
2	1.11
3	0.95

Figure 2.5 Nuclear matrix associated DNA is broadly associated with gene-rich regions in MCF10a cells and with gene-poor regions in MDA-MB-231 cells. On a small (200bp) scale, these associations are not observed.

(A) Box and whisker plots of the distributions of nuclear matrix associated DNA (NM-DNA) enrichment (NM-DNA / Digested Nuclei) values (y-axis) plotted based on the number of genes found within the corresponding 200bp intervals (x-axis) for MCF10a cells (upper panel) and MDA-MB-231 cells (lower panel). The width of the box and whisker plots corresponds to the number of times X number of genes are found within 200bp intervals in the human genome. Plots are cropped on the y-axis to show weak trends for the median values of NM-DNA enrichment based on the number of genes in each interval. **(B)** Scatterplots of percent of gene coverage per chromosome band (y-axis) and NM-DNA enrichment per chromosome band (x-axis) in MCF10a cells (left panel) and MDA-MB-231 cells (middle panel). These x-y pairs are plotted for the entire genome using a hexagonal density scatterplot in which an increased number of shared x-y observations are represented as darker hexagons (right panel). Pearson r and p-value was calculated for each cell line with the correlation line plotted in red and the Pearson r and p-value in the upper right of the scatterplot for each cell line.

Genome-wide associations of nuclear matrix-associated DNA and chromatin modifications appears to be different in MDA-MB-231 and MCF10a cells

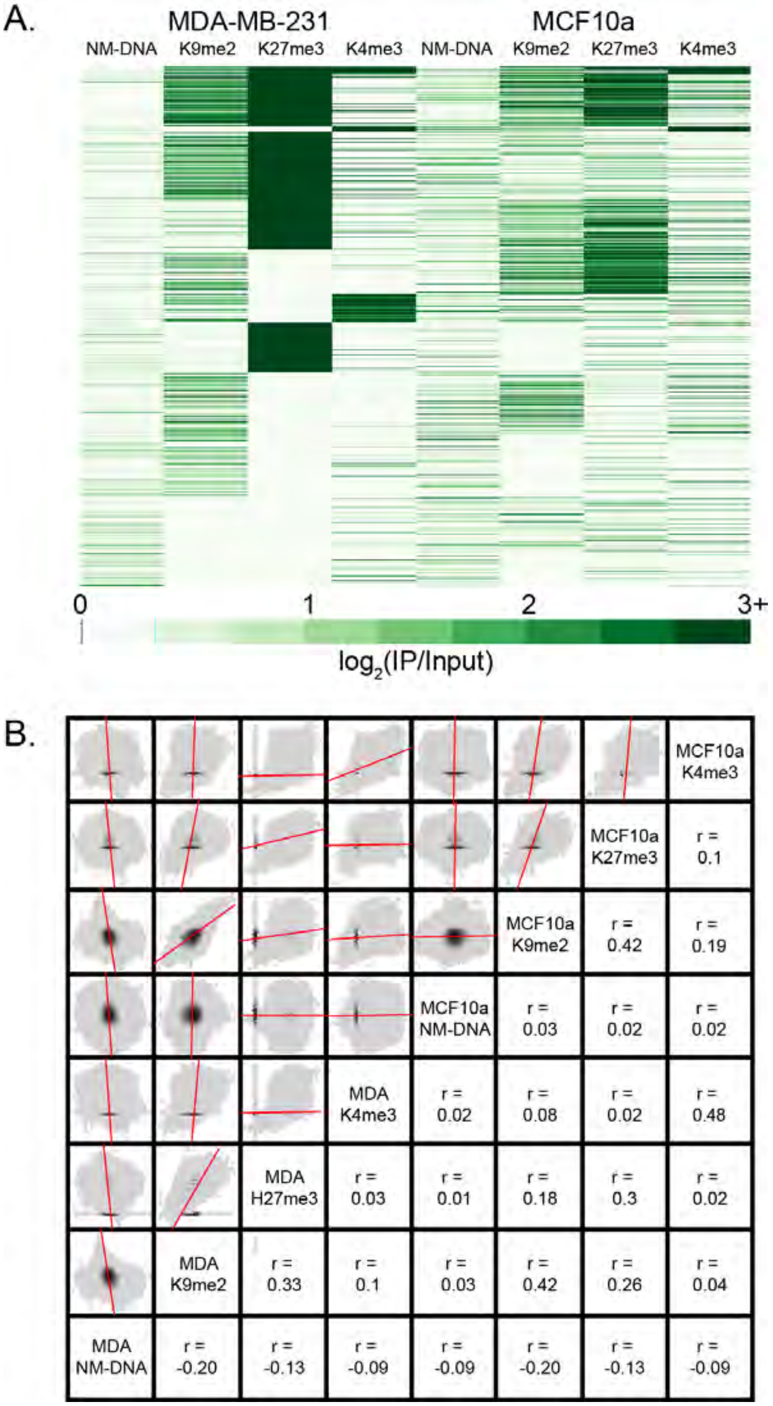
It is not clear whether NM-DNA is associated with transcription repression or transcriptional association (Maya-Mendoza and Aranda-Anzaldo 2003, Linnemann et al. 2007, 2008, Alva-Medina et al. 2011, Keaton et al. 2011, Trevilla-García and Aranda-Anzaldo 2011). Modifications of histone 3 are strongly associated with transcriptional states (Greer and Shi 2012). We therefore sought to investigate how NM-DNA isolated via NM-seq is associated with several modifications of histone 3 on a genome-wide scale. Positional comparison of NM-DNA enrichment and histone enrichment has never been executed at genomic resolution, and NM-seq allows for this type of interrogation of nuclear architecture. In both the MCF10a and MDA-MB-231 cells, we performed ChIP-seq for the histone modifications histone 3 lysine 4 tri-methylation (H3K4me3), histone 3 lysine 27 tri-methylation (H3K27me3) and histone 3 lysine 9 di-methylation (H3K9me2) and compared the genome-wide enrichment profiles of these histone modifications to NM-DNA enrichment patterns. This approach is designed to give us some idea of the extent to which matrix-associated DNA is associated with functionally diverse histone modifications using our novel isolation methods.

We took two unbiased approaches, k-means clustering and Pearson r correlation, to discover potential relationships between genomic enrichment of NM-DNA and of these chromatin modifications **(Figure 2.6)**. The enrichment of NM-DNA, H3K4me3, H4K27me3, and H3K9me2 in 200bp, non-overlapping genome-wide coordinates was measured in both MCF10a and MDA-MB-231 cells. In **(Figure 2.6 A)**, each row is a 200bp non-overlapping interval in which we have measured the enrichment of NM-DNA, H3K9me2, H3K27me3, and H3K4me3 in each cell line. The columns correspond to the histone mark or NM-DNA in each cell line. Once this matrix is constructed, k-means clustering is applied to identify patterns in the enrichment of these marks. While clustering of the enrichment of the histone marks appears to segregate nicely, it does not appear that NM-DNA enrichment patterns are affected by the patterns of enrichment of histone 3 modifications **(Figure 2.6 A)**.

We then performed pairwise comparisons between each histone 3 modification and NM-DNA enrichment, and present these data as a scatter plot correlation matrix **(Figure 2.6 B)**. Along the diagonal from bottom left to top right are the names for cell line and sample. The left half of the matrix shows density scatter plots of enrichment between the pairs of samples, and the right half of the matrix contains the Pearson correlation coefficient (r) between the sample pairs. Here we observed that NM-DNA did not correlate with histone 3 modifications within either cell line nor did NM-DNA in MCF10a cells correlate with NM-DNA in MDA-MB-231 cells.

Combining the observations from both approaches, we find that the enrichment patterns of H3K4me3, H3K27me3, and H3K9me2 are very similar across the two cell lines **(Figure 2.6 A)** and each modification is mildly correlated across cell lines **(Figure 2.6 B)**. Given previous reports that matrix-associated DNA generally lacks nucleosomes (Boulikas 1995), we were surprised to not find negative correlations between NM-DNA enrichment and these chromatin marks. We observe essentially no correlation (positive or negative) between NM-DNA enrichment and these histone marks. This study is the first in which the positional enrichment of histone modifications has been directly compared to the positional enrichment of matrix-associated DNA on a genome-wide scale, so we suggest that, as compared to previous studies, the scale of NM-seq allows for a more accurate representation of the positional relationships between histone 3 modifications and NM-DNA. If matrix-associated DNA were nucleosome-free, any location in which a core histone (histone 3) is found should be depleted of matrix-associated DNA. This analysis uncovers many locations where a histone protein and NM-DNA are enriched, which suggests that NM-DNA is not, as a rule, nucleosome-free. We cannot rule out mutual exclusivity in a shared location. ChIP-seq and NM-seq are measuring frequencies of interaction between a protein and genomic sequence or a genomic sequence and the nuclear matrix, respectively. This analysis demonstrates that there is not a clear relationship between the positional enrichments of nuclear matrix and modifications of histone 3.

Figure 2.6 On a genome-wide scale, nuclear matrix-associated DNA is not negatively-correlated with the histone modifications H3K4me3, H3K27me3 and H3K9me2 in MDA-MB-231 or MCF10a cells.



							MCF10a K4me3
						MCF10a K27me3	r = 0.1
					MCF10a K9me2	r = 0.42	r = 0.19
				MCF10a NM-DNA	r = 0.03	r = 0.02	r = 0.02
			MDA K4me3	r = 0.02	r = 0.08	r = 0.02	r = 0.48
		MDA H27me3	r = 0.03	r = 0.01	r = 0.18	r = 0.3	r = 0.02
	MDA K9me2	r = 0.33	r = 0.1	r = 0.03	r = 0.42	r = 0.26	r = 0.04
MDA NM-DNA	r = -0.20	r = -0.13	r = -0.09	r = -0.09	r = -0.20	r = -0.13	r = -0.09

Figure 2.6 On a genome-wide scale, nuclear matrix-associated DNA is not negatively-correlated with the histone modifications H3K4me3, H3K27me3 and H3K9me2 in MDA-MB-231 or MCF10a cells

(A) Using genome-wide non-overlapping 200bp intervals, the enrichment of NM-DNA (NM-DNA / Digested Nuclei), and the enrichment of the histone modifications H3K4me3, H3K27me3, and H3K9me2 (IP / Input) were measured and plotted as a heat map. Each horizontal row is the same 200bp interval. For each nuclear feature measured (columns), the $\log_2$ of the enrichment values was plotted and then clustered using the k-means expectation-maximization algorithm. The color scale for $\log_2$(enrichment) is found below the heat map. **(B)** Pairwise matrix of scatterplots comparing the enrichment of each nuclear mark in each cell line. Scatterplots (upper left panels) are hexagonal density scatter plots where darker dots indicate an increased number of paired x-y observations. Pearson r values and p-values were calculated for each pairwise comparison. The line of best fit is plotted in red on the scatterplots and the Pearson r and p-values are displayed in the mirrored panels (bottom left panels).

Discriminative motif discovery suggests DNA-binding proteins may participate in the context-dependent functions of nuclear matrix-associated DNA

While NM-DNA enrichment in MCF10a and MDA-MB-231 patterns may be similar in terms of non-correlation with the histone marks examined, the direct comparison of NM-DNA enrichment between MCF10a and MDA-MB-231 cells shows no correlation (r = -0.09) **(Figure 2.6 B)**. Although the general trends of sequence complexity and correlations between gene-rich versus gene-poor enrichment are different between these two cell lines, direct comparison of positional enrichment of NM-DNA between these two cell lines shows that there are some genomic sequences that are associated with the matrix in both cell lines as well as some matrix-associated sequences that are unique to each cell line. This observation, combined with our comparative analyses between NM-DNA enrichment and sequence complexity, as well as associations with gene-rich versus gene-poor regions, leads us to hypothesize that in each cell line there may be different proteins responsible for recruiting DNA to the nuclear matrix. To determine what DNA-binding proteins may be responsible for recruitment of DNA to the nuclear matrix, we looked at sequence motifs enriched in nuclear matrix-associated DNA. We defined four types of sequence regions to perform motif discovery: most associated with the nuclear matrix in MDA-MB-231 cells and least associated with the nuclear matrix in MCF10a cells, most associated with the nuclear matrix in MCF10a cells and least associated with the nuclear matrix in MDA-MB-231 cells, least associated with the nuclear matrix in both cell lines, and most enriched in both cell lines. We then performed discriminative motif discovery on these four sets of genomic intervals to determine which types or families of proteins may be responsible for the disparate enrichment patterns of matrix-associated DNA between these two cell lines **(Figure 2.7)**. Discriminative motif discover takes a set of sequences and compares against a background of sequences to discover sequence motifs that are significantly overrepresented in the set of target sequences (Heinz et al. 2010). For each category of genomic regions, a different set of motifs were found. This suggests that the expression of certain DNA-binding proteins may result in the observed differences in NM-DNA. It is interesting that a motif similar to the RUNX motif is discovered in the regions enriched in NM-DNA in both cell lines, as RUNX proteins are dependent on interaction with the nuclear matrix for many functions and have been reported to be expressed and functional in both cell lines (Zeng et al. 1997, 1998, Barnes et al. 2004, Shore 2005, Kadota et al. 2010). Breast cancer associated 1 (BRCA1), whose motif is enriched in MCF10a NM-DNA, associates with the nuclear matrix and participates in DNA repair, which takes place on the nuclear matrix (Huber and Chodosh 2005). It is not clear what function the v-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian) (MAF) family of proteins, a motif enriched in NM-DNA-enrichment regions in MDA-MB-231 cells, may have.

Figure 2.7 Discriminative motif discovery for sequences most associated with the nuclear matrix in MCF10a and MDA-MB-231 cells suggests that cell-type-specific proteins may participate in differences in matrix-DNA associations.

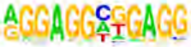
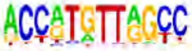
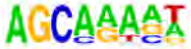
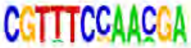
Source	Motif Discovered	Positive Sequences	Background Sequences	p-value	Similar to:
Enriched in both		10,484 / 114,960 (9.12%)	7,039 / 114,635 (6.14%)	1e-340	SP1, ETS:RUNX1
Enriched MCF10a		121,420 / 1,305,589 (9.30%)	93,870 / 1,305,567 (7.19%)	1e-1735	YY1, FOXA1, NFIC, BRCA1
Enriched MDA-MB-231		581,620 / 1,311,136 (44.36%)	329,661 / 794,556 (41.49%)	1e-949	MAF Family
Depleted in both		237 / 473,333 (0.05%)	0 / 520,500 (0.00%)	1e-252	RFX, NFAT, STAT

Figure 2.7 Discriminative motif discovery for sequences most associated with the nuclear matrix in MCF10a and MDA-MB-231 cells suggests that cell-type-specific proteins may participate in differences in matrix-DNA associations.

Four classes of sequences were defined empirically based on NM-DNA enrichment percentiles within MCF10a and MDA-MB-231 cells. Using the paired NM-DNA enrichment values and genomic coordinates for the previously defined genome-wide non-overlapping 200bp intervals, four categories were defined: enriched in both (intervals where NM-DNA enrichment is in the top 10% of each cell line), enriched in MCF10a (intervals where NM-DNA enrichment is in the top 10% in MCF10a cells and bottom 10% in MDA-MB-231 cells), enriched in MDA-MB-231 (intervals where NM-DNA enrichment is in the top 10% in MDA-MB-231 cells and bottom 10% in MCF10a cells), and depleted in both (intervals where NM-DNA enrichment is in the bottom 10% in both cell lines). The sequence logo for each *de novo* motif and the known motifs to which each discovered motif is similar are presented for each of the four classes of NM-DNA enrichment.

NM-DNA in MCF10a cells is associated with higher gene expression, while in MDA-MB-231 cells NM-DNA is associated with lower gene expression

In MDA-MB-231 and MCF10a cells, we observe opposite trends in NM-DNA enrichment with sequence complexity and gene-rich versus gene-poor regions. These observations lead us to hypothesize that NM-DNA may be differentially associated with gene expression patterns between the normal and cancer cell lines. Therefore, we examined the relationships between gene expression and NM-DNA enrichment. RNA isolated from each cell line was subjected to Affymetrix Human Gene 1.0ST arrays to obtain a quantitative signature of gene expression in each cell line. Transcript detection levels were normalized across arrays using the robust means algorithm (RMA) (*for more details, please see methods*) (Irizarry et al. 2003). We began with the cytological band level, and measured the mean RMA-normalized transcript level for all genes within the cytological band and compared these values to the mean NM-DNA enrichment in the cytological band. Again, we observed opposite results for the MCF10a and MDA-MB-231 cells: in MCF10a cells NM-DNA enrichment is somewhat correlated (Pearson r = 0.32) with transcript levels, while in MDA-MB-231 cells NM-DNA enrichment is somewhat negatively correlated (Pearson r = -0.32) with transcript levels **(Figure 2.8)**.

Figure 2.8 Nuclear matrix-associated DNA enrichment correlates with higher levels of gene expression in MCF10a cells and with lower levels of genes expression in MDA-MB-231 cells.

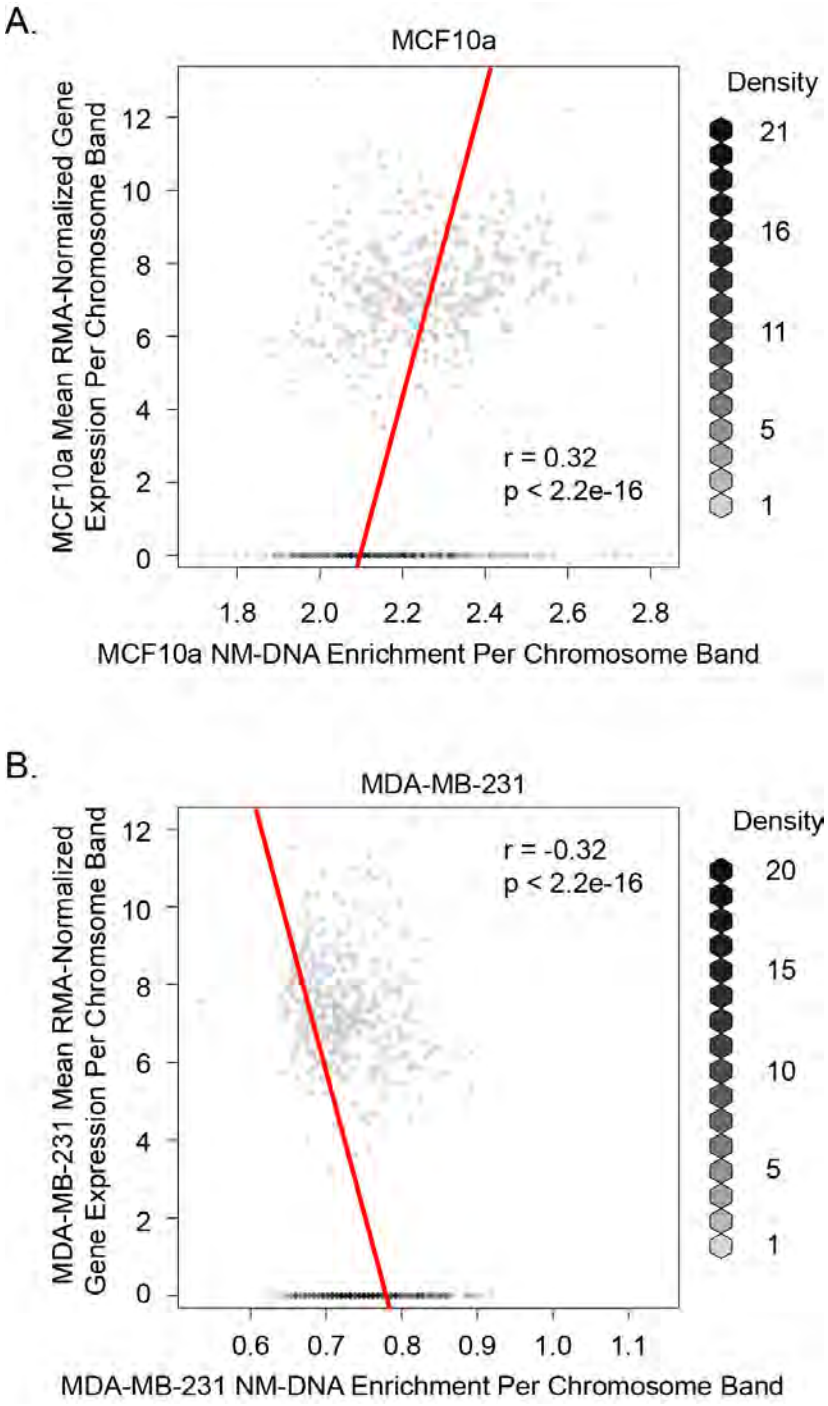


Figure 2.8 Nuclear matrix associated DNA enrichment correlates with higher levels of gene expression in MCF10a cells and with lower levels of genes expression in MDA-MB-231 cells.

(A and B) For each chromosomal band, the mean RMA-normalized transcript detection value (extracted from Affymetrix Human Gene 1.0ST arrays) for all genes found within the given chromosomal band was plotted as the y-coordinate and the corresponding enrichment of NM-DNA was plotted as the x-coordinate. These x-y coordinates were plotted as hexagonal density scatterplots where the darker hexagons represent an increased number of shared x-y observations. The Pearson r and p-value for MCF10a **(A)** and MDA-MB-231 **(B)** were calculated and the line was plotted in red over the scatter plots; the Pearson r and p-value are indicated in the bottom-right corner of the scatterplot in **(A)** and the top-right corner of the scatterplot in **(B)**.

Using a gene-centric approach, NM-DNA is differentially associated with gene expression and chromatin states between MCF10a and MDA-MB-231 cells.

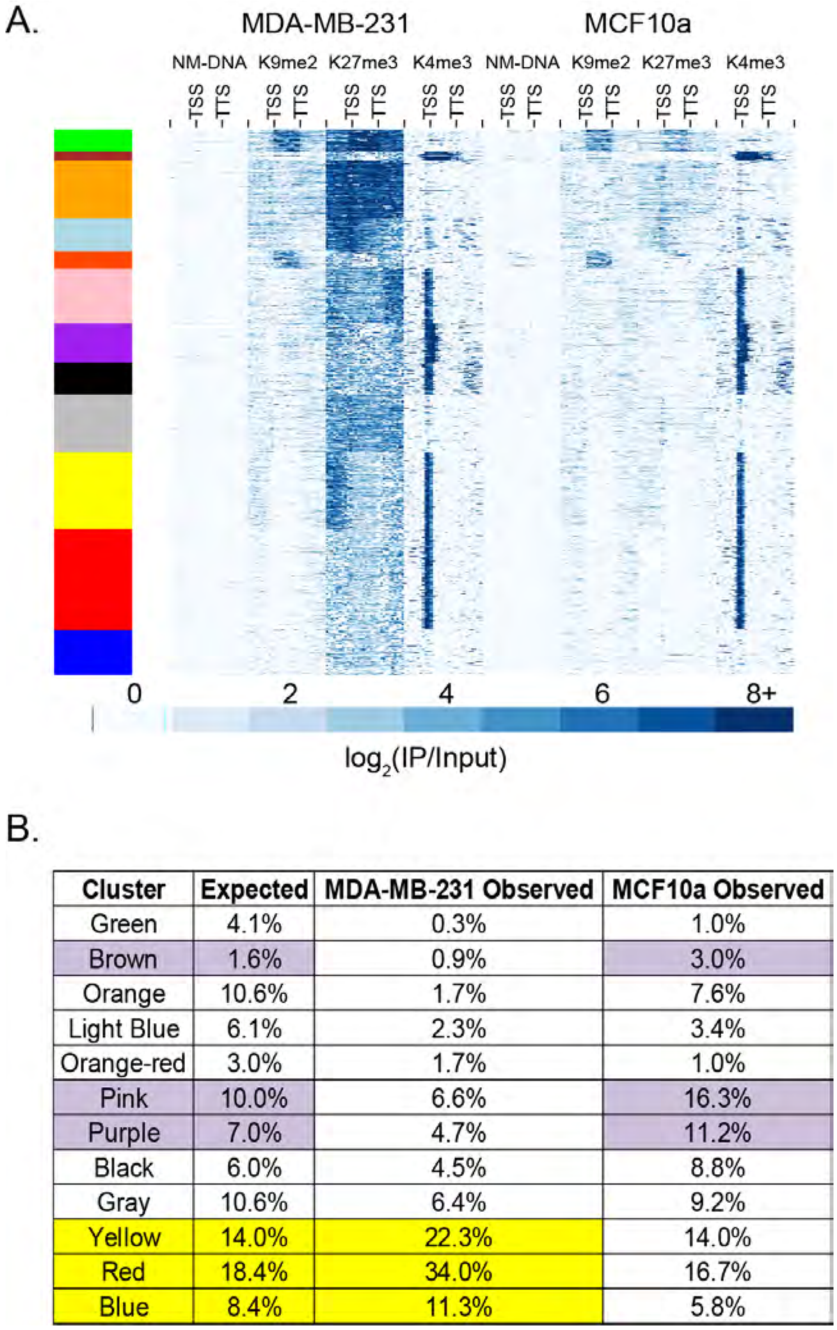
As shown in Figure 2.8, NM-DNA is differentially associated with active gene expression in the MCF10a and MDA-MB-231 cells. In order to understand how NM-DNA enrichment differs at the gene level, we measured the enrichment of NM-DNA within 10Kb of all genes in the hg19 RefSeq genome (Pruitt et al. 2009). We initially clustered these binding patterns to see if genes that are differentially expressed between MCF10a and MDA-MB-231 cells have a particular pattern of NM-DNA-enrichment, however, no meaningful clusters were identified (data not shown). To further define parameters of nuclear organization, we performed ChIP-seq for histone H3K9me2, H3K27me3, and H3K4me3. The enrichment ratios of these chromatin modifications were combined with the NM-DNA enrichment data and k-means clustering was performed **(Figure 2.9 A)**.

Given our previous observations, it is expected that we observe very little NM-DNA enrichment near gene bodies in MDA-MB-231 cells. Similar to the observations made in Figure 4.6, the patterns of the chromatin modifications (H3K4me3, H3K27me3, and H3K9me2) are even more uniform across cell lines at the single gene level. As observed in **(Figure 2.6)**, H3K27me3 is detected at a significantly higher level in MDA-MB-231 cells as compared to MCF10a cells at both the genome-wide and gene-centric scales. An *in vitro* comparison of

H3K27me3 enrichment in normal (human mammary epithelial cells (HMEC)) and breast cancer cells (HCC1954) revealed that H3K27me3 is significantly enriched in the cancer cell line as compared to the normal cell line in both genic and intergenic regions (Hon et al. 2012). These observations are similar to the observations presented here where MDA-MB-231 cells are significantly more enriched for H3K27me3 as compared to MCF10a cells in both genic and intergenic regions. These cell line studies are in contrast with data observed from patient samples where reduction of H3K27me3 is associated with poor patient survival (Greer and Shi 2012). These disparate observations of H3K27me3 enrichment between patients and derived cell lines may be due to either detection methods or cell line derivation, and are not examined in this study.

To investigate whether NM-DNA enrichment near genes may be involved the phenotypic differences between MCF10a and MDA-MB-231 cells, we defined genes that are differentially expressed between MCF10a and MDA-MB-231 cells. The genes that are significantly more or less expressed in each cell line are likely to be related to these phenotypic differences between these two cell lines. To test whether the genomic regions within and surrounding these differentially expressed genes are more or less associated with a pattern of chromatin marks and NM-DNA enrichment we used a chi-squared test. For each cluster identified by k-means clustering (colors on left of **Figure 2.9 A**), this test determines the extent to which these genes are over or underrepresented. This is done by defining an expected value (based on the distribution (represented as a percentage) of all genes within each cluster) and comparing these expected values to the observed distribution of genes defined as more expressed in each cell line. We can see that the genes more expressed in MCF10a cells are primarily found in the purple, pink and brown clusters, while the genes more expressed in MDA-MB-231 cells are primarily found in the yellow, red and blue clusters. In the MCF10a cells, the yellow, red and blue clusters have increased H3K9me2 enrichment compared to the purple, pink and brown clusters, while in the MDA-MB-231 cells, the opposite is observed **(Figure 2.9 B)**. In the case of NM-DNA, the MCF10a cells show more enrichment flanking the gene bodies in the purple, pink and brown clusters compared to the yellow, red and blue clusters. However, in the MDA-MB-231 cells there does not appear to be any difference in the enrichment pattern of NM-DNA between these groups of clusters. In MCF10a cells, the observation that the enrichment pattern of NM-DNA near genes is related to gene expression suggests that NM-DNA may have a functional role in gene expression regulation. In MDA-MB-231 cells, NM-DNA enrichment patterns do not appear to be a strongly associated with gene expression patterns. These observations lead us to hypothesize that the functional role of NM-DNA in gene expression regulation is somehow disrupted in the cancer cell line.

Figure 2.9 Distinct patterns in the enrichment of NM-DNA and H3K27me3 for genes differentially expressed between MDA-MB-231 and MCF10a cells.



Cluster	Expected	MDA-MB-231 Observed	MCF10a Observed
Green	4.1%	0.3%	1.0%
Brown	1.6%	0.9%	3.0%
Orange	10.6%	1.7%	7.6%
Light Blue	6.1%	2.3%	3.4%
Orange-red	3.0%	1.7%	1.0%
Pink	10.0%	6.6%	16.3%
Purple	7.0%	4.7%	11.2%
Black	6.0%	4.5%	8.8%
Gray	10.6%	6.4%	9.2%
Yellow	14.0%	22.3%	14.0%
Red	18.4%	34.0%	16.7%
Blue	8.4%	11.3%	5.8%

Figure 2.9 Distinct patterns in the enrichment of NM-DNA and H3K27me3 for genes differentially expressed between MDA-MB-231 and MCF10a cells.

(A) For each annotated transcript in RefSeq, the enrichment (NM-DNA / Digested Nuclei or IP / Input) of NM-DNA, H3K4me3 (K4me3), H3K27me3 (K27me3) and H3K9me2 (K9me2) was measured in 30 defined relative genomic regions: 10Kb 5'-transcriptional start site (TSS) to the TSS in 1Kb steps, transcription termination site (TTS) to 10Kb 3'-TTS in 1Kb steps, and TSS to TTS. These enrichment ratios for the relative genomic coordinates were concatenated horizontally to combine each nuclear feature from both cell lines so that each horizontal row in the heat map represents one RefSeq transcript. The colors on the left side of the heat map correspond to clusters discovered by k-means expectation maximization. The enrichment ratios were $\log_2$-transformed and the colors corresponding to the $\log_2$-transformed enrichment ratios are found below the heat map. **(B)** Distributions are presented as a percent of total number of genes for each of the three gene lists for each cluster identified by k-mean clustering. Chi-squared table of the distribution of all genes in RefSeq (Expected), genes more expressed in MDA-MB-231 cells (MDA-MB-231 Observed) and genes more expressed in MCF10a cells (MCF10a Observed).

Direct comparison of matrix-associated DNA enrichment with gene expression patterns

Because we observed some differences in how NM-DNA and H3K27me3 are associated with gene expression patterns between MCF10a and MDA-MB-231 cells, we directly tested whether these binding profiles are significantly different near gene bodies using PeaksToGenes (*please see chapter 5 for more complete methods description*).

Using the internal spike-in controls defined by Affymetrix as negative controls, we separated the RefSeq genes in the array into two categories: expressed (greater than the mean of the RMA-normalized expression values of the negative control probes) and non-expressed (less than or equal to the mean of the RMA-normalized expression values of the negative control probes). Then, we performed the Wilcoxon Rank Sum Test for the observed signal enrichment of H3K4me3, H3K27me3, H3K9me2, and NM-DNA in each relative genomic window, comparing the expressed genes versus the non-expressed genes.

The histone modification H3K4me3 is associated with transcriptional initiation (Li et al. 2007), and in both MCF10a and MDA-MB-231 cells the genes defined as expressed are significantly more associated with H3K4me3 in regions proximal to the TSS as compared to genes defined as non-expressed **(Figure 2.10 A)**. The repressive histone modification H3K27me3 (Li et al. 2007) is significantly more associated with genes defined as non-expressed in both MCF10a and MDA-MB-231 cells when compared to expressed genes **(Figure 2.10 B)**.

The widely observed functions of H3K4me3 and H3K27me3 combined with their strong associations near expressed and non-expressed genes in both cell lines suggest that our defining of expressed versus non-expressed genes is valid.

H3K9me2 is thought to be associated with repression (Hawkins et al. 2010), and in both MCF10a and MDA-MB-231 cells, enrichment of this chromatin modification within the gene body is very strongly associated with non-expressed genes **(Figure 2.10 C)**. There is one slight difference, however, in the regions flanking the gene bodies: in MDA-MB-231 cells, non-expressed genes are significantly more associated with H3K9me2 enrichment, while in MCF10a cells there is a much weaker association of H3K9me2 enrichment. This analysis provides evidence to suggest that H3K9me2 may be differentially associated with gene expression between MCF10a and MDA-MB-231 cells.

Figure 2.10 Histone marks linked to transcriptional activation or repression are appropriately associated with expressed or non-expressed genes, respectively.

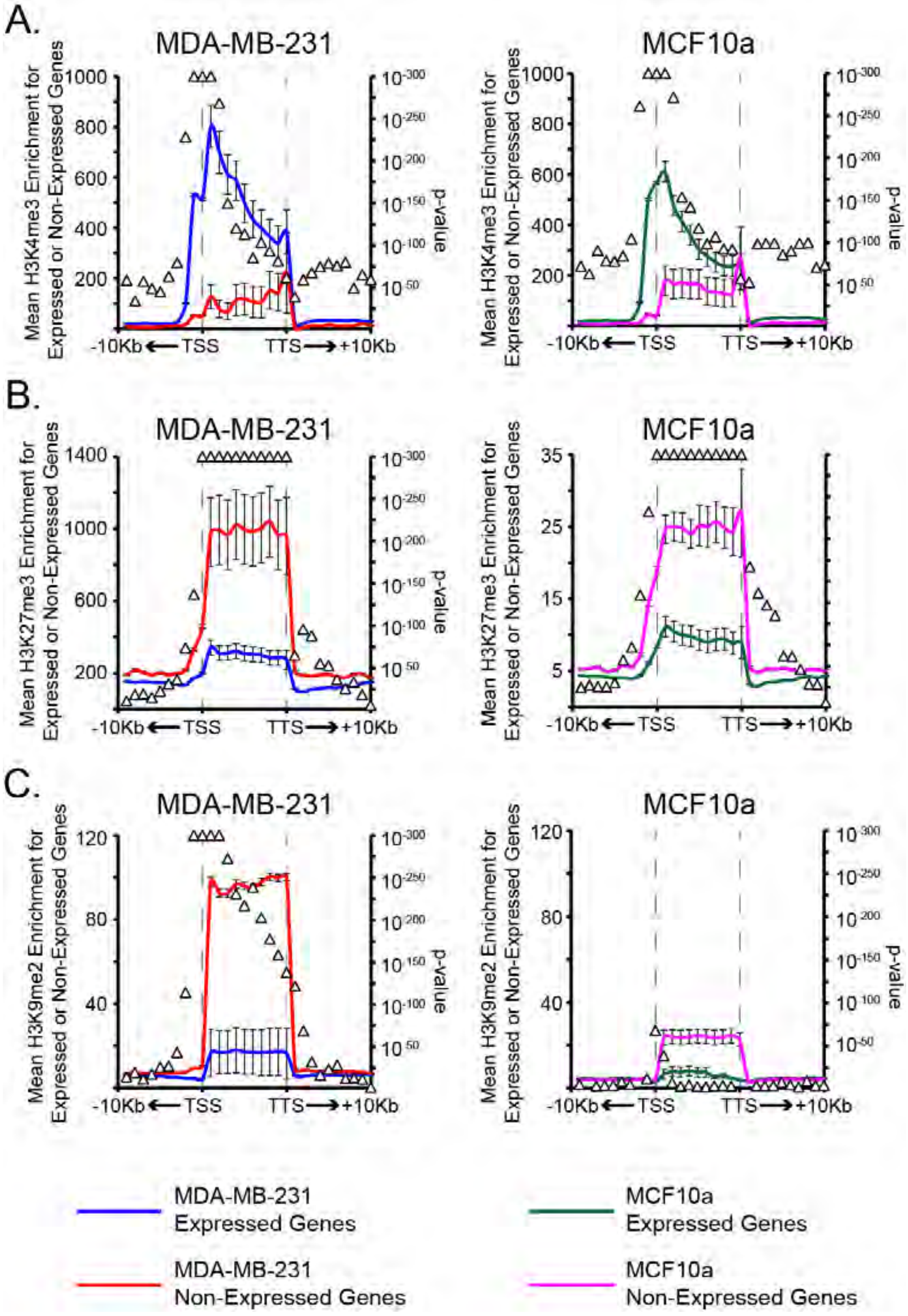


Figure 2.10 Histone marks linked to transcriptional activation or repression are appropriately associated with expressed or non-expressed genes, respectively.

(A, B, and C) For each annotated transcript in RefSeq, the enrichment (IP / Input) of H3K4me3 **(A)**, H3K27me3 **(B)** and H3K9me2 **(C)** was measured in 30 defined relative genomic regions: 10Kb 5'-transcriptional start site (TSS) to the TSS in 1Kb steps, transcription termination site (TTS) to 10Kb 3'-TTS in 1Kb steps, and TSS to TTS in 10 approximately equal intervals (scaled to 1Kb in length). Using a Wilcoxon Rank Sum Test, contrast test types were run to contrast genes expressed in a cell line versus genes not expressed in a cell line. Lines represent the mean enrichment of the nuclear mark being measured (left y-axis) in the relative genomic region (x-coordinate) for the set of genes being contrasted (color-coded and defined below graphs). Triangles correspond to the approximate p-value (95% CI) for the Wilcoxon Rank Sum Test run at each relative interval (right y-axis). All panels: MDA-MB-231 (left) and MCF10a (right). Error bars are SEM.

Comparing NM-DNA enrichment in MCF10a cells near expressed versus non-expressed genes, we observed that NM-DNA is significantly enriched flanking the gene bodies of actively expressed genes (green line) and within the gene bodies of non-expressed genes (pink line) **(Figure 2.11 A – right)**. The magnitudes of the mean differences between NM-DNA enrichment within MCF10a expressed (green) and MCF10a non-expressed (pink) were quite high. In regions flanking gene bodies (especially 5'-TSS), the magnitude of the mean differences of NM-DNA enrichment near MCF10a expressed (green) and MCF10a non-expressed (pink) were not large. However, using p-values approximated from the Wilcoxon Rank Sum Test (triangles), we observed that these small differences in flanking regions were highly reproducible, and we think it is therefore important to understand what these types of matrix-DNA interactions may mean.

In MDA-MB-231 cells, NM-DNA was enriched near the TSS of expressed genes (blue) and near the TTS and 3'-end of non-expressed (red) genes **(Figure 2.11 A – left)**. Compared to the enrichment patterns observed in MCF10a cells **(Figure 2.11 A – right)**, the mean magnitudes of MDA-MB-231 expressed (blue) and MDA-MB-231 non-expressed (red) genes were lower than their respective expression counterparts in the MCF10a cells. This is expected given the gene-poor preferences for NM-DNA enrichment observed in MDA-MB-231 cells. Further comparison of these two plots revealed very different patterns in the mean enrichment positions of NM-DNA relative to expressed and non-expressed genes. While the magnitude of the mean NM-DNA enrichment within MDA-MB-231 non-expressed (red line) genes was greater than expressed genes (blue line), the p-values associated with these differences were much higher when compared to the plots for MCF10a cells. Further inspection revealed differences in NM-DNA enrichment within the 5'-end of gene bodies of expressed and non-expressed genes between MDA-MB-231 and MCF10a cells. In MCF10a cells, there was a clear distinction between NM-DNA enrichment within and outside of the gene bodies of non-expressed genes (pink line), while in MDA-MB-231 cells NM-DNA was only statistically significantly enriched in the 3'-half of gene bodies of non-expressed genes (red line). Another major difference between these two cell lines was the differences in how NM-DNA was enriched in regions flanking expressed and non-expressed genes. In MCF10a cells, some of the most significant differences in the enrichment patterns of NM-DNA are observed in intergenic regions 5' of the TSS, in contrast NM-DNA in MDA-MB-231 cells is similarly enriched near expressed and non-expressed genes. The one consistency between these two cell lines is that NM-DNA enrichment within the gene body is associated with lower expression levels, but even that observation is weakly conserved in the MDA-MB-231 cells.

When considering genes that are differentially expressed between MCF10a and MDA-MB-231 cells for either H3K9me2 **(Figure 2.11 B)** or NM-DNA **(Figure 2.11**

C), we observe that the trends in association of NM-DNA and H3K9me2 with gene expression are even more exaggerated. The genes used for comparison were defined using expression data from Affymetrix and a contrast test was performed to define genes that are differentially expressed between MCF10a and MDA-MB-231 cells. From this test two lists of genes were defined: genes more expressed in MCF10a (orange line) cells and genes more expressed in MDA-MB-231 cells (purple line). When performing the Wilcoxon Rank-Sum Test using PeaksToGenes, the background list of genes is the rest of the RefSeq mRNAs for each respective list in each cell line (brown line – MDA-MB-231, cyan line – MCF10a). In MCF10a, regions flanking the genes expressed more in MCF10a cells as compared to MDA-MB-231 (orange line) are significantly enriched in NM-DNA **(Figure 2.11 B – right)**, and weakly associated with H3K9me2 **(Figure 2.11 C – right)** as compared to the rest of the MCF10a transcriptome (cyan line). Both H3K9me2 and NM-DNA are enriched within the gene bodies of the rest of the MCF10a transcriptome (cyan line) as compared to genes more expressed in MCF10a compared to MDA-MB-231 cells (orange line). Comparing the MCF10a NM-DNA profiles in **(Figure 2.11 C – right)** to **(Figure 2.11 A – right)**, we observed the p-values associated with the differences in NM-DNA enrichment within gene bodies and flanking gene bodies are higher when comparing genes more expressed in MCF10a cells to the rest of the MCF10a transcriptome as compared to MCF10a expressed versus MCF10a non-expressed genes. Of note is the observed magnitude of NM-DNA enrichment in regions flanking genes more expressed in MCF10a cells (orange line) being increased as compared to the rest of the MCF10a transcriptome (cyan line) **(Figure 2.11 C – right)**. This suggests that in normal cells, regions flanking highly expressed genes have increased association with the nuclear matrix.

In MDA-MB-231, the entire region near genes more expressed in MDA-MB-231 cells as compared to MCF10a (purple line) is significantly depleted in association of H3K9me2 as compared to the rest of the MDA-MB-231 transcriptome (brown line) **(Figure 2.11 B – left)**. This is different from what is observed in MCF10a cells **(Figure 2.11 B – right)** where we observed that NM-DNA enrichment in regions flanking gene bodies of MCF10a more expressed genes (orange line) is not significantly different from the enrichment near the rest of the MCF10a transcriptome (cyan line). Interestingly, the pattern of 5' gene body enrichment of NM-DNA for MDA-MB-231 expressed genes (blue line) **(Figure 2.11 A – left)**, is not observed when comparing genes more expressed in MDA-MB-231 cells (purple line) to the rest of the MDA-MB-231 transcriptome (brown line) **(Figure 2.11 C – left)**.

Consistently observed in both cell lines are the strong associations of both H3K9me2 and NM-DNA within the gene bodies of either non-expressed or less-expressed genes. In terms of matrix-associated DNA, these types of sequences would be functionally characterized as MARs (Linnemann et al. 2008). In both cell lines, NM-DNA enrichment near the TSS is associated with expressed genes. These types of sequences would be functionally characterized as SARs (Linnemann et al. 2008). It is interesting that SARs are also found in regions flanking gene bodies of expressed genes, however, this association between NM-DNA and gene expression is only observed in the normal mammary epithelial cell line MCF10a. These analyses demonstrate the ability of our protocol to isolate and enrich for two functional species of DNA associated with the nuclear matrix. While these types of matrix-associated DNA have different roles in gene expression regulation, analyzing enrichment patterns of matrix-associated DNA from our protocol in a gene-centric manner allows for unbiased categorization of both MARs and SARs.

Figure 2.11 Meta-gene analysis reveals significant differences in the associations between NM-DNA enrichment and H3K9me2 enrichment and gene expression patterns in MDA-MB-231 and MCF10a cells.

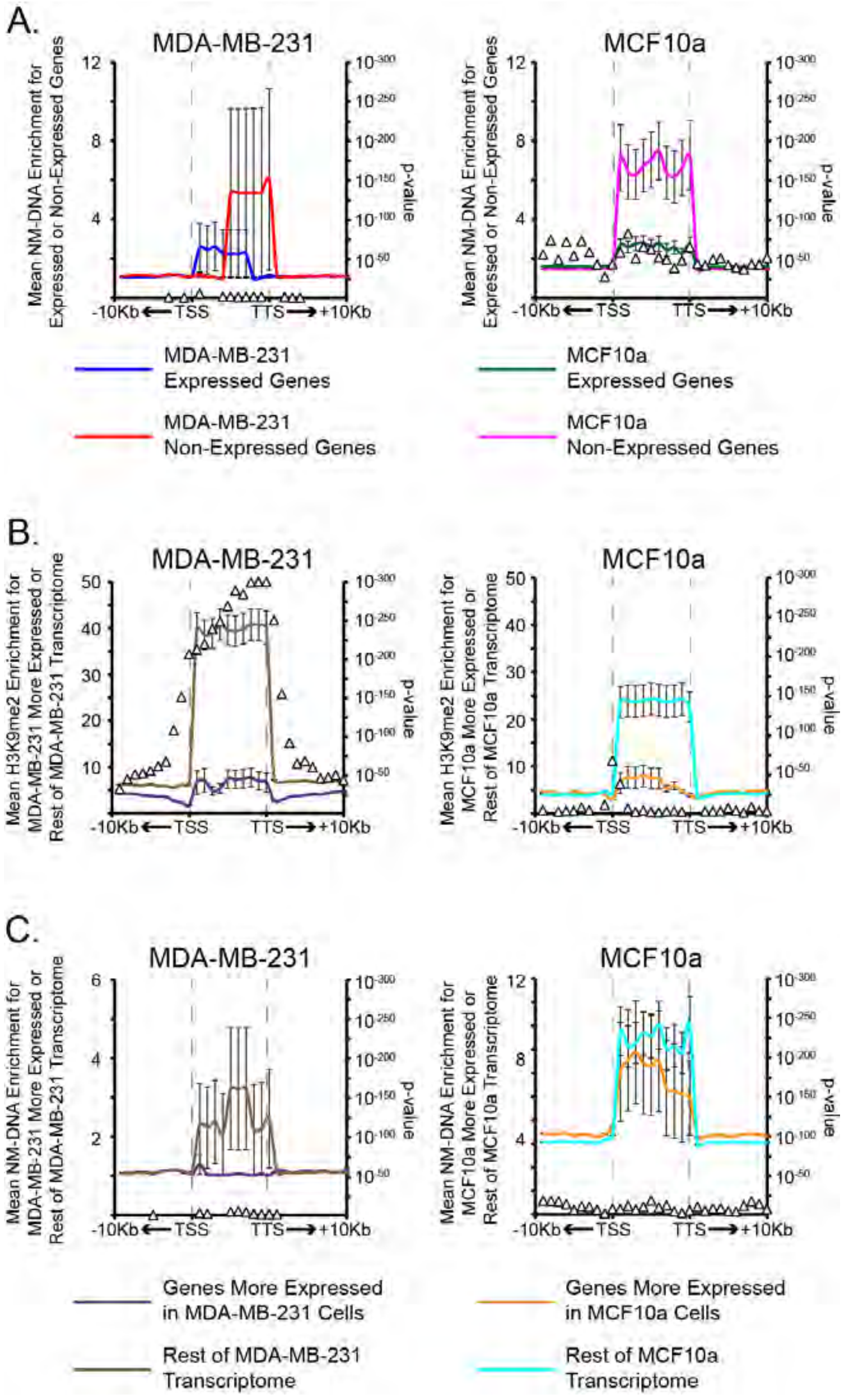


Figure 2.11 Meta-gene analysis reveals significant differences in the associations between NM-DNA enrichment and H3K9me2 enrichment and gene expression patterns in MDA-MB-231 and MCF10a cells.

(A, B and C) For each transcript annotated in RefSeq, the enrichment (NM-DNA / Digested Nuclei or IP / Input) of NM-DNA, or H3K9me2 was measured in 30 defined relative genomic regions: 10Kb 5'-transcriptional start site (TSS) to the TSS in 1Kb steps, transcription termination site (TTS) to 10Kb 3'-TTS in 1Kb steps, and TSS to TTS in 10 approximately equal intervals (scaled to 1Kb in length). Using a Wilcoxon Rank Sum Test, several contrast test types were performed: **(A)** genes expressed in a cell line versus genes not expressed in a cell line, **(B & C)** more expressed in one cell line versus the rest of the transcriptome. Lines represent the mean enrichment of the nuclear mark being measured (left y-axis) in the relative genomic region (x-coordinate) for the set of genes being contrasted (color-coded and defined below graphs). Triangles correspond to the approximate p-value (95% CI) for the Wilcoxon Rank Sum Test run at each relative interval (right y-axis). All panels: MDA-MB-231 (left) and MCF10a (right). Error bars are SEM.

Discussion

Here we describe NM-seq, which is a method for the isolation of genomic DNA associated with the nuclear matrix in a manner that preserves nuclear integrity. When applied to a normal mammary epithelial cell (MCF10a) and a malignant metastatic breast cancer cell (MDA-MB-231), we demonstrated that NM-seq was capable of isolating sequences that can be characterized as either SARs or MARs, based on the positions of these sequences relative to gene expression patterns. When viewing the enrichment pattern of NM-DNA on a whole chromosome for both MCF10a and MDA-MB-231 cells, we observed that there was a somewhat inverse pattern of enrichment **(Figure 2.12 A)**. However, it is important to note that the most enriched regions in MDA-MB-231 cells were similarly enriched in MCF10a cells (although significantly less than other enriched regions in MCF10a cells). This observation, combined with the observed massive increases in H3K27me3 enrichment, suggests that MDA-MB-231 cells are deficient in normal structure-function relationships.

Figure 2.12 Chromosome-wide pattern of NM-DNA enrichment in MCF10a and MDA-MB-231 cells and model for transcriptional associations in each cell line

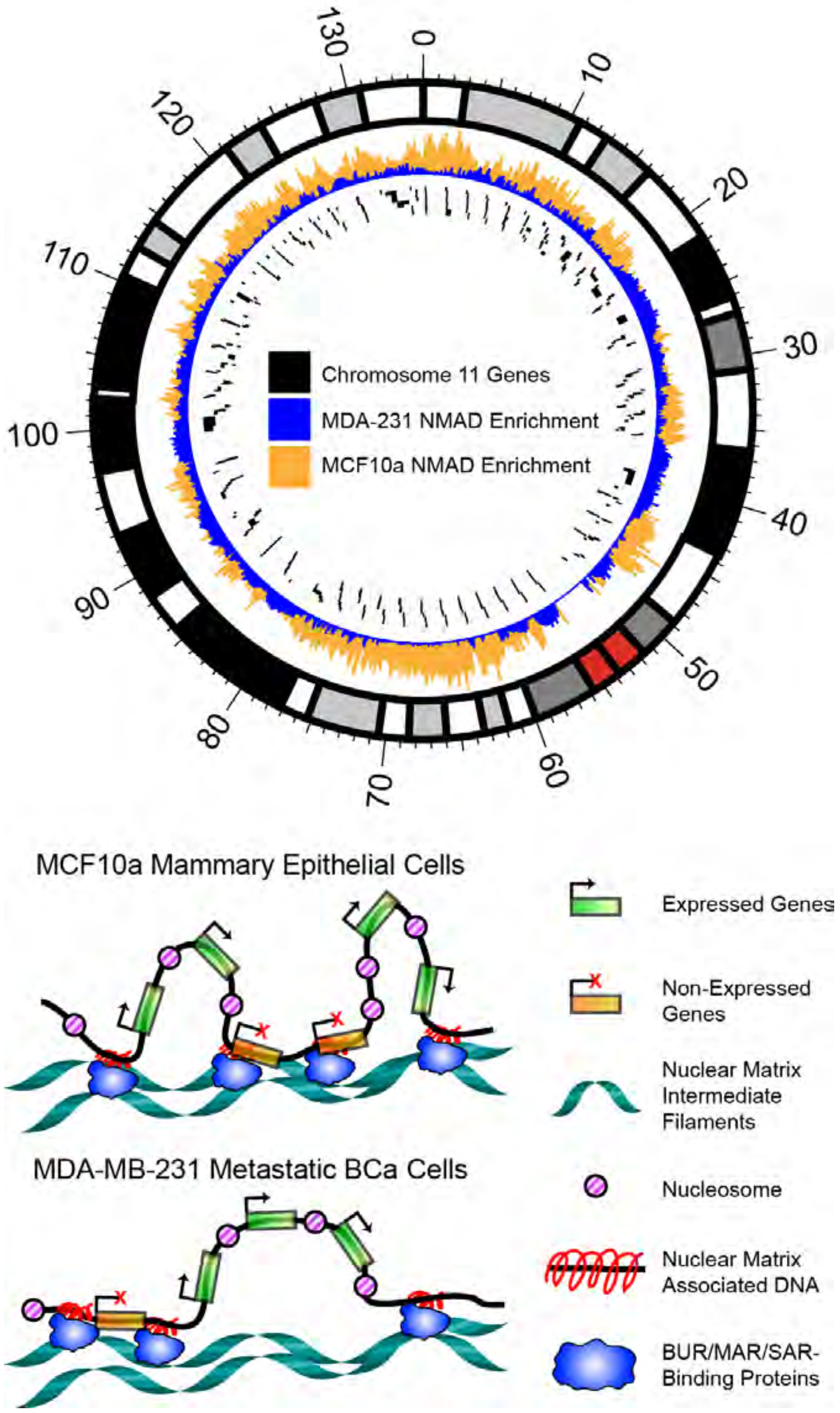

0
10
20
30
40
50
60
70
80
90
100
110
120
130
Chromosome 11 Genes
MDA-231 NMAD Enrichment
MCF10a NMAD Enrichment
MCF10a Mammary Epithelial Cells
MDA-MB-231 Metastatic BCa Cells
Expressed Genes
Non-Expressed Genes
Nuclear Matrix Intermediate Filaments
Nucleosome
Nuclear Matrix Associated DNA
BUR/MAR/SAR-Binding Proteins

Figure 2.12 Chromosome-wide pattern of NM-DNA enrichment in MCF10a and MDA-MB-231 cells and model for transcriptional associations in each cell line

(A) Circos histogram plot of chromosome 11 NM-DNA enrichment in MDA-MB-231 cells (blue) and MCF10a cells (orange). Outer ring depicts ranges of Giemsa stain positivity per chromosomal band; red = centromeres and telomeres, white to black (and intermediate grays) indicate intensity of Giemsa staining (Meyer et al. 2013). Inner bars represent gene bodies. **(B)** Cartoon showing less frequent DNA-matrix interactions in MDA-MB-231 cells leading to alterations in gene expression.

Our results demonstrate that there are major differences in the structure-function relationships between a normal mammary epithelial cell line (MCF10a) and a malignant metastatic breast cancer cell line (MDA-MB-231). Marked reorganization of nuclear heterochromatin is a “hallmark of cancer”, and indicates that the cancer cells are fundamentally different from their normal counterparts in terms of nuclear organization (Hanahan and Weinberg 2000). While this study is limited to two immortalized breast lines, the observed correlations between expression of matrix-associated proteins and cancer progression suggests that changes in DNA associated with the nuclear matrix would follow suit (Samuel et al. 1997, He et al. 2008). We suggest that this study is limited in terms the inference of general roles for the associations between nuclear matrix-associated DNA and transcriptional states because we have only used two cell lines. This study demonstrates a potential application of NM-seq in understanding differences between the transcriptional associations of NM-DNA in normal and disease-state cells. Integrative combinations of NM-seq with ChIP-seq for modifications of histone H3 further demonstrates how NM-seq can be used to investigate the relationships between chromatin states and matrix-associated DNA.

To further study how matrix-associated DNA may be involved with nuclear organization, we propose that the application of NM-seq to several tier 1

ENCODE cell lines (ENCODE Project Consortium 2011) would yield great insight. There are many nuclear functions known to be enriched in association with the nuclear matrix such as replication, transcription, RNA splicing, and DNA repair (Nickerson 2001). Because it is likely infeasible for one research group to perform the required experiments to investigate all these parameters in one cell line, using ENCODE cell lines allows for incorporation of these data without having to execute these experiments in house. This same logic can be applied to non-ENCODE datasets that are publicly available such as lamin-associated domains (LADs) (Guelen et al. 2008, Peric-Hupkes et al. 2010). LADs are defined using Lamin B1 that is modified to methylate adenine bases where Lamin B1 is bound (Guelen et al. 2008), which utilizes an approach called DNA adenine methylation identification (DamID) (Germann and Gaudin 2011). We hypothesize that LADs may have a significant overlap with NM-DNA in the same cell line because Lamin B1 is a matrix-associated protein and is known to bind S/MARs (Ludérus et al. 1992). However, it remains to be seen where these overlaps will occur and how the enrichment of NM-DNA within these regions compares to the global NM-DNA enrichment profile.

Our motif discovery results suggest that multiple transcription factors may be involved in NM-DNA interactions with the nuclear matrix. Runx1 and Runx2 are required for definitive hematopoiesis and osteoblast maturation respectively

(Wang et al. 1996, Komori et al. 1997). Truncation of the C-terminus (required for matrix interaction) of Runx1 or Runx2 results in developmental phenocopies of the null phenotype (Choi et al. 2001, Dowdy et al. 2010). We therefore investigated the extent to which RUNX1 or RUNX2 (from ChIP-seq in Chapter 3) binding is enriched in NM-DNA enriched regions in the MDA-MB-231 cells. Although there is no overlap between the datasets beyond what can be considered random (data not shown), the development of these experimental approaches allows researchers to address these nuclear matrix-centric hypotheses in a rapid and thorough manner consistent with the types of -omics analyses currently being implemented in a variety of developmental and disease-specific contexts.

In this limited two cell line approach, we are considering MCF10a normal mammary epithelial cells as a model for normal associations between NM-DNA and transcriptional activity. In **Figure 2.11 A**, we observed large differences in the magnitude of the mean NM-DNA enrichment within the gene bodies of MCF10a expressed (green line) and MCF10a non-expressed (pink line) genes. Conversely, we observed small differences in the magnitude of the mean NM-DNA enrichment in regions flanking the gene bodies of MCF10a expressed (green line) and MCF10a non-expressed (pink line) genes. While the observed differences in magnitude of NM-DNA enrichment in both gene-centric locations (gene bodies and flanking regions), the p-values from the Wilcoxon Rank Sum Test were very similar across all regions. What these p-values are effectively saying is that the chances that these observed differences are some kind of outlier are very low. A potential mechanism that may explain why we observed highly reproducible differences in NM-DNA enrichment in both expressed and non-expressed genic regions is the dynamic movement of genes to and from transcription factories (Lanctôt et al. 2007, Ronneberger et al. 2008). We must first consider that when we are measuring enrichment of NM-DNA, what we are really measuring is the frequency with which sequences of DNA are interacting with the nuclear matrix. The more often a region interacts with the nuclear matrix, the more DNA will be sequenced from this region, and thus we will calculate this region to be more highly enriched in association with the nuclear matrix. Immunofluorescence studies of the movements of expressed versus non-expressed genes demonstrated that expressed genes are highly dynamic in their range of motion within their chromosome territory, while non-expressed genes are relatively static in position (Andrulis et al. 1998, Kosak et al. 2002, Zink et al. 2004a, Hewitt et al. 2004). Further, expressed genes are not constitutively expressed; rather they are thought to be transcribed in bursts when associated with transcription factories (Dundr and Misteli 2010, Deng et al. 2012). Given that the nuclear matrix is enriched in newly synthesized RNA and is localized within euchromatic regions, we expect actively transcribed genes to have some association with the nuclear matrix. We observed that expressed genes are slightly enriched in association with the nuclear matrix in regions flanking gene bodies **(Figure 2.11 A – right)**, however, the magnitude of these differences was weak. Considering that genes are only transiently associated with transcription factories, we propose that the nuclear matrix serves as a platform to organize actively transcribed sequences through interactions with sequences flanking the transcribed regions. The transient nature of genes with transcription factories explains the highly reproducible (low p-value), yet weak magnitude differences in NM-DNA enrichment observed for regions flanking expressed genes. These matrix-DNA interactions would therefore serve as an organizational platform for the recruitment of transcription and splicing factors to be spatially enriched in order to promote accurate and precise transcription in response to physiological demands. Conversely, the strong enrichment of NM-DNA within the gene bodies of non-expressed genes may be explained by the lack of movement of non-expressed genes. Whereby lack of gene movement results in more stable interactions with the nuclear matrix (or NM-DNA enrichment). These non-expressed genes would typically be observed in the regions of peripheral heterochromatin. While the nuclear matrix is less abundant in the peripheral heterochromatin regions, it is still present. These repressive regions may have more overlap with the previously discussed LADs; however, this is only speculation.

This study is the first to analyze matrix-associated DNA on a genome-wide scale at such a fine resolution (single base). It is also the only study that has made direct positional comparisons between matrix-associated DNA and histone modifications. These histone modifications (H3K4me3, H3K27me3, and H3K9me2) have not previously been analyzed on a genome-wide scale in either cell line. These advances, combined with the integration of transcriptome-wide gene expression analysis, make the impact of this study highly significant. Using a genome-wide unbiased approach, the results of this study reproduce and extend many of the seminal experiments demonstrating the functional associations between matrix-associated DNA and gene expression (Ciejek et al. 1983, Jost and Seldran 1984, Jackson and Cook 1985, Cook 1989, Jackson et al. 1993, Linnemann et al. 2008, Rivera-Mulia and Aranda-Anzaldo 2010, Trevilla-García and Aranda-Anzaldo 2011). Rather than showing that the nuclear matrix is enriched for actively-transcribed loci, without knowing which loci are which, or showing that a single gene or small cohort of genes are associated with the nuclear matrix in a location-specific manner, we have shown that on a genone-wide scale there is are location-specific associations between the nuclear matrix and regions flanking transcribed genes. This is an important breakthrough in our ability to understand how the nuclear matrix is involved in the organization of transcriptional domains within the nucleus. We propose that NM-seq can be utilized as a tool to study the functional contribution of the nuclear matrix to many nuclear functions, not only transcription. Because so many groups are rapidly adopting genome-wide approaches for studying nuclear organization-based questions, we further propose that NM-seq should be considered as a complimentary tool for addressing these problems.

Materials and Methods

Cell Culture

MCF10a cells were cultured in phenol-free DMEM/F12 media supplemented with EGF, insulin, cholera toxin and hydrocortisone and subcultured as described (Debnath et al. 2003).

MDA-MB-231 cells were grown and subcultured as previously described (Pratap et al. 2008).

Immunofluorescence microscopy

Cells were grown on coverslips coated in 0.5% w/v gelatin as previously described prior to extraction (Ali et al. 2010, 2012). Fixed cells/nuclei/matrices were then incubated with primary antibodies (NPAT, Coilin, PML, SC-35, UBF and Pol II) prior to incubation with AlexaFluor-conjugated secondary antibodies as previously described (Ghule et al. 2009, Ali et al. 2010, 2012). Immunofluorescence was executed in parallel to preparation of nuclear matrices for genomic DNA extract and deep-sequencing.

ChIP-seq

ChIP was performed on subconfluent cells as previously described (Lee et al. 2006) with the following modification: after sonication, buffer C was exchanged for FA buffer (50mM HEPES-KOH, pH 7.5, 140mM Sodium Chloride, 1mM EDTA, 1% v/v Triton X-100, 0.1% w/v Sodium Deoxycholate, 0.1% w/v SDS, 1x Roche cOmplete EDTA-free protease inhibitor cocktail, 25nM MG132 in Nuclease-Free Water) prior to antibody pulldown using Amicon columns (10,000kDa MWCO). After genomic DNA was isolated for both specific IPs and Input DNA, libraries were prepared for paired-end multiplexed tag Solexa/Illumina sequencing using the Illumina DNA Sample Prep Kit v2.0 according to manufacturer instructions.

Isolation of nuclear matrix

This section is not presented as methods, but rather in protocol format.

Buffers:

4x CSK Salts

400mM Sodium Chloride, 4mM EGTA, 40mM PIPES pH 6.8 12mM Magnesium Chloride in Nuclease Free Water. This buffer is stable (maintains pH) at 4°C for approximately one month.

4x Digestion Buffer

200mM Sodium Chloride, 4mM EGTA, 40mM PIPES pH 6.8 12mM Magnesium Chloride in Nuclease Free Water. This buffer is stable (maintains pH) at 4°C for approximately one month.

CSK Extraction Buffer

1x CSK Salts, 300mM Sucrose, 0.5% v/v Triton X-100 in Nuclease Free Water.

CSK Crosslinking Buffer

1x CSK Salts, 300mM Sucrose, 1%w/v Formaldehyde in Nuclease Free Water.

CSK Wash Buffer

1x CSK Salts, 1mM EDTA, 1mM EGTA, 300mM Sucrose in Nuclease Free Water

Nuclear Matrix Digestion Buffer

1x Digestion Buffer, 300mM Sucrose, 0.5%v/v Triton X-100, 600U/mL MspI, 600U/mL HaeIII, 600U/mL RsaI, 600U/mL AluI, 600U/mL BamHI, 600U/mL PvuII in Nuclease Free Water.

Digestion Wash Buffer

1x Digestion Buffer, 300mM Sucrose, 10mM EDTA, 10mM EGTA in Nuclease Free Water.

Matrix Extraction Buffer

1x Digestion Buffer, 300mM Sucrose, 10mM EDTA, 10mM EGTA, 250mM Ammonium Sulfate in Nuclease Free Water.

Crosslink Reversal Buffer

100mM Sodium Bicarbonate, 0.5%w/v SDS, 10mM EDTA, 10mM EGTA in Nuclease Free Water.

Sonication Buffer

10mM Tris-Cl, pH 8.0, 100mM Sodium Chloride, 1mM EDTA, 1mM EGTA, 0.1%w/v Sodium Deoxycholate, 0.5%w/v N-Lauroylsarcosine in Nuclease Free Water.

Day 1

All buffers for this section should be prepared several hours prior to removing cells from the incubator. With the exception of the Crosslink Reversal Buffer, all buffers should be kept on ice in nuclease-free tubes/bottles.

It is important that the sucrose in the matrix buffer and the glycerol in the enzyme buffers are properly dissolved in solution before beginning.

Add the formaldehyde to the CSK Crosslinking Buffer immediately before adding to cells.

1. Grow adherent cells to passage (near confluence) density on a 100mM tissue culture plate.

2. Remove plates from incubator and place on ice. Be sure that the plates are partially buried into the ice, and not just resting on top. Be careful not to push the plates in so far that ice will fall into the plates, and try to make the plate surface as level as possible for even buffer distribution.

3. Aspirate media, gently wash cells twice with ice cold PBS to remove all media, serum, and non-adherent cells.

4. Add 5mL CSK Extraction Buffer to each plate. Incubate on ice for 10 minutes. During incubation, gently rotate ice bucket in air to ensure the buffer is evenly dispersed across the cells.

5. Aspirate CSK Extraction Buffer and discard. Wash cells gently twice with CSK Wash Buffer.

6. Add 5mL CSK Crosslinking Buffer. Incubate on ice for 10 minutes. During incubation, gently rotate ice bucket in air to ensure the buffer is evenly dispersed across the cells.

7. Add 250µL freshly-prepared 2.5M Glycine to each plate, gently rotate to mix. Incubate for 10 minutes on ice to stop formaldehyde crosslinking, gently rocking ice bucket as before.

8. Aspirate buffer from cells, and appropriately discard. Gently wash cells twice with CSK Wash Buffer.

9. Add 1mL Nuclear Matrix Digestion Buffer to each plate. Place cells in 30°C incubator for 1 hour. Every 15 minutes, rotate the plates to ensure buffer is evenly distributed.

10. For plates to be used for digested nuclei:

(a) Scrape plates and all buffer into 50mL Amicon column with 10,000MWCO.

(b) Add 10mL Digestion Wash Buffer to tube.

(c) Spin at max speed for column until volume reaches approximately 400µL.

(d) Add 10mL Crosslink Reversal Buffer.

(e) Spin at max speed for column until volume reaches approximately 400µL.

(f) Transfer top solution to 1.75mL Axygen tube.

(g) Add 4µL of 20mg/mL Proteinase K to tube (final concentration of 200µg/mL).

(h) Incubate in a 65°C water bath for 16 hours.

For plates to be used for nuclear matrix preparation:

(a) Place plates back on ice.

(b) Aspirate Nuclear Matrix Digestion Buffer.

(c) Wash gently with Digestion Wash Buffer twice.

(d) Add 5mL Matrix Extraction Buffer. Incubate on ice for 10 minutes. Gently rock ice bucket as described above.

(e) Aspirate Matrix Extraction Buffer. Gently wash cells twice with Digestion Wash Buffer.

(f) Scrape residual matrices on plates in 400µL Crosslink Reversal Buffer into 50mL Amicon column with 10,000mwco.

(g) Add 10mL Crosslink Reversal Buffer to tube. Spin at max speed for column until volume reaches approximately 400µL. Repeat.

(h) Transfer top volume to 1.75mL Axygen tube.

(i) Add 4µL of 20mg/mL Proteinase K to tube (final concentration of 200µg/mL).

(j) Incubate tube in a 65°C water bath for 16 hours.

Days 2: Genomic DNA Isolation

Make a fresh preparation of phenol:chloroform:isoamyl alcohol (25:24:1) and chloroform:isoamyl alcohol (24:1).

1. Pre-spin 2.0mL Phase-Lock Heavy (Qiagen) tubes.

2. Transfer solution from Axygen tube into Phase-Lock Heavy tube.

3. Add 400μL phenol:chloroform:isoamyl alcohol. Vigorously mix for 15s.

4. Spin at 10,000g for 10 minutes.

5. Add 400μL chloroform:isoamyl alcohol. Vortex by tapping the bottom of the tube while holding the top for 30s. Spin again.

6. Transfer supernatant into 1.75mL Axygen tube.

7. Add 8μL 10mg/mL RNAse A to tube (final concentration of 200μg/mL). Incubate tube for 2 hours in 37°C water bath.

8. Add 4μL 20mg/mL Proteinase K to tube (final concentration of 200μg/mL). Incubate tube for 2 hours in 55°C water bath.

9. Pre-spin 2.0mL Phase-Lock Heavy tubes.

10. Transfer solution from Axygen tube into Phase-Lock Heavy tube.

11. Add 400μL phenol:chloroform:isoamyl alcohol. Vigorously mix for 15s.

12. Spin at 10,000g for 10 minutes.

13. Add 400μL chloroform:isoamyl alcohol. Vortex by tapping the bottom of the tube while holding the top for 30s. Spin again.

14. Transfer supernatant into 1.75mL Axygen tube. Add 800μL ice-cold 100% Ethanol. Mix by inverting tube.

15. Place tube at -20°C overnight.

Day 3: Sonication

Freshly prepare Sonication Buffer and store on ice.

Freshly prepare 80% Ethanol and store at -20°C for several hours before washing DNA pellet.

Sonication conditions need to be empirically determined as not all sonicators are the same. Presented here is the methods used in our lab.

1. Remove tubes from -20°C, spin at 4°C at full speed for 20 minutes.

2. Gently remove and discard supernatant.

3. Add 1mL ice cold 80% ethanol. Spin at 4°C at full speed for 10 minutes. Remove supernatant. Repeat.

4. Add 500µL Sonication Buffer to tube, allow DNA to go into solution on ice.

5. Suspend tube in ice water bath in sonication cabinet.

6. Position tube so that the micro-tip is a few millimeters from the bottom of the tube and perfectly centered.

7. Run the following sonication cycle 8 times:

(a) Total sonication time of 20 seconds, oscillating between 1 second of sonication and 2 seconds of rest.

(b) 30 seconds of rest in between each cycle.

8. Add 1mL ice-cold 100% Ethanol. Place tube at -20°C overnight.

Gel Verification

1. Cast a 2% TAE-agarose gel.

2. Remove tubes from -20°C, spin at 4°C at full speed for 20 minutes.

3. Gently remove and discard supernatant.

4. Add 1mL ice cold 80% ethanol. Spin at 4°C at full speed for 10 minutes. Remove supernatant. Repeat.

5. Add 100µL TE to pellet.

6. Quantify DNA using Qubit dsDNA HS Assay (Invitrogen).

7. Run a small (~200ng) amount of DNA (if you can spare it) for several hours at 30V. Use a 100bp DNA ladder for a marker. Gel can be stained with ethidium bromide or other in-gel DNA visualizing reagent. Our lab used ethidium bromide.

8. Visualize DNA using UV; if distribution is centered higher than 400bp, repeat sonication steps. A 400bp band will be selected from the gel during library preparation for sequencing. It is important that the distribution of band sizes is

centered at 400bp so that the slice selected will have a high probability to be representative of the entire population of DNA fragments.

Library Preparation

From this point, all DNA is treated as ChIP DNA for the preparation of genomic DNA libraries for deep-sequencing on the Illumina Hi-Seq 2000. All libraries are prepared according to the manual provided with the Illumina TruSeq DNA Sample Preparation v2 Guide.

Buffers

All buffers should be made fresh the day of the experiment unless otherwise noted, and are made with nuclease-free water in nuclease free tubes/bottles. All reasonable precautions to maintain a nuclease-free environment when preparing/using these buffers should be made.

2.55M Sucrose

Sucrose (g):	1710g
Nuclease-free Water	900mL

Final Molarity	2.55M

The following is from the Lamond lab protocol for sucrose buffer preparation

http://www.lamondlab.com/f7nucleolarprotocol.htm: The stock solution is stable indefinitely at 4°C. This procedure can be carried out at RT. There is no need to heat up the solution to help disolve the sucrose. Heating up an incompletely dissolved sucrose solution can lead to charring of sucrose and affect the quality of the sucrose solution.

Prepare the sucrose buffer as follows:

1. Weigh out 1710 g sucrose. Keep it aside in a clean container.

2. Put exactly 900ml water and a magnetic bar in a 5 liter beaker. Put the beaker on a stirrer and start stirring.

3. Add 1/3 of the sucrose into the beaker. Make sure the magnetic bar is rotating freely. Stir for 1 hour.

4. Add another 1/3 of the sucrose into the solution. Again make sure the rotation of the stir bar is not impaired. Stir for another 1 hour..

5. Add the remaining sucrose. Stir for another 1 hour, or until all the sucrose has gone into solution. The final volume should be exactly 2 liters.

Read Mapping

100bp paired-end reads were mapped to the hg19 build of the human genome using Bowtie2 (Langmead and Salzberg 2012).

Read Conversion

Mapped reads were converted to BED-format reads for analysis using SAMTools (Li et al. 2009) and BedTools (Quinlan and Hall 2010)

Discriminative Motif Discovery

Homer (Heinz et al. 2010) motif discovery software was used to discover enriched motifs. For each set of test intervals, a GC/CpG/length-matched set of intervals was generated by Homer and used as background.

GC Content

GC content was measured using custom Perl scripts reliant upon BioPerl (Stajich et al. 2002) to calculate GC content from FASTA-format sequence files.

Giemsa Chromosome Bands

Coordinates of Giemsa bands and RefSeq gene positions were extracted from the UCSC Genome Browser for hg19 (Meyer et al. 2013).

Scatterplots and Correlation Coefficient Calculation

Hexagonal plots and Pearson R correlation coefficient calculations were done in R using the “hexbin” package and the “cor.test” function.

Heat Maps and Clustering

K-means clustering and heat maps were generated in R. Heat maps were generated using the “heatmap.2” function from the “gplots” library. K-means clustering was performed using the “kmeans” function of the “stats” core package.

Affymetrix Arrays and Analysis

Human Gene 1.0ST arrays from Affymetrix were used to measure gene expression levels in MCF10a and MDA-MB-231 cells. cRNA amplification and hybridization to the array was performed by the UMASS Medical School Genomic Core as previously described (Dowdy et al. 2010).

Analysis was performed in R to execute normalization, quality control, transcript-level reporting, annotation, and contrast tests. The “affy” (Gautier et al. 2004) package was used to read in raw fluorescent values from arrays; values were normalized across all arrays using quantile normalization (Bolstad et al. 2003), robust means average (RMA) background correction and median polish (Irizarry et al. 2003). Quality control plots were generated to ensure the arrays did not have any artifacts and post-processing values were in similar ranges. Transcripts were annotated using the “annotate” package, the “hugene10sttranscriptcluster.db” package. Contrast tests were generated and performed using the “limma” package.

Circos Plot

Circos plot of NM-DNA-enrichment on human chromosome 11 was done using Circos (Krzywinski et al. 2009).

CHAPTER 3 GENOME-WIDE ANALYSIS OF BINDING AND GENE EXPRESSION REGULATION OF RUNX1 AND RUNX2 IN MDA-MB-231 CELLS AND ASSOCIATION OF RUNX1 AND RUNX2 EXPRESSION WITH HORMONE RECEPTORS IN BREAST CANCER PATIENT SAMPLES

Authors and contributions

Jason R. Dobson, Gillian Browne, Deli Hong, Maria Libera de la Porta, Andre J. van Wijnen, Janet L. Stein, Gary S. Stein, Jane B. Lian.

Experimental approach designed by JRD, AJVW, JLS, GSS, and JBL.

Optimization and staining of histological samples performed by JRD.

Histological scoring of tissue microarrays performed by JRD and GB.

Immunofluorescence staining and imaging of RUNX1, RUNX2 and UBF performed by JRD and DH.

Histological imagining of samples from UMASS Medical School performed by MLP.

ChIP, ChIP-seq, siRNA transfection, invasion assays, qPCR, Western blotting, metabolic labeling, growth curves, Affymetrix analysis, ontological enrichment, and ChIP-seq analysis performed by JRD.

Introduction

Transcription factors and their cognate coregulatory partners are critical for cell-type specific expression of phenotypic genes. In cancer, many transcription factors with functions required for normal developmental processes function in tumor cells to promote cancer progression. For example, transcription factors that mediate signaling pathways for growth and tissue formation, including the TGF-beta, BMP, and WNT pathways, are highly active (Zaidi et al. 2007). Here, we address the functional activities of a transcription factor family that is required for developmental processes and whose expression is deregulated in breast cancer cells, though their distinct functional roles in contributing to the progression of tumor growth and later metastatic events are not clear.

Runx1 and Runx2 are developmentally required for the emergence and maturation of the hematopoietic (Wang et al. 1996) and osseous (Komori et al. 1997) lineages, respectively; they are also associated with tumor growth and metastasis (Pratap et al. 2006). In their native cellular context, RUNX proteins participate in multiple critical signaling pathways and regulate the expression of phenotypic genes and ribosomal RNA (Young et al. 2007a, 2007b, Bakshi et al. 2008). Further, the result of RUNX binding to a gene promoter may be either activation or repression, depending on the co-regulatory proteins it recruits (Cameron and Neil 2004). These observations indicate that the presence of a

RUNX protein within a given cell does not necessarily predict the outcome on the transcriptional program.

In breast and prostate cancer cell lines, both of which are predisposed to bone metastasis, RUNX2 plays a key transcriptional role in promoting both the osteomimetic and osteolytic activities of tumor cells (Barnes et al. 2004, Javed et al. 2005, Pratap et al. 2008, Akech et al. 2010). The mechanisms by which RUNX2 promotes metastatic bone disease have been well characterized *in vitro* and include the activation of bone/ECM-adhesion proteins (OP, BSP) and invasion-related factors (MMPs, VEGF), as well as mediating the TGF-beta/SMAD signaling pathway and the vicious cycle of tumor growth and bone resorption (IHH/PTHRP) (Pratap et al. 2010, Chimge and Frenkel 2012).

RUNX1, on the other hand, appears to function as a tumor suppressor in the MCF10a mammary epithelial cells and derived tumorigenic cell lines (Kadota et al. 2010, Wang et al. 2011a, Janes 2011). In MDA-MB-231 cells stably transfected with estrogen receptor alpha (ER-α), RUNX1 has been proposed to function as a tethering factor for ER-α to cooperatively promote estrogen stimulation (Stender et al. 2010). However, the functions of RUNX1 in the context of parental, ER-negative MDA-MB-231 cells are unknown.

In Asian patients, RUNX2 expression is an independent predictor of disease and is primarily associated with estrogen receptor negative (ER-) breast cancers (Das et al. 2009, Onodera et al. 2010). In a small scale study of European breast and breast cancer tissues, the Human Protein Atlas found that RUNX2 was not expressed in either normal or tumor tissues (Uhlén et al. 2005, Pontèn et al. 2008, Uhlen et al. 2010). These observations lead us to hypothesize that due to the small sample size of the Human Protein Atlas study, the observation of RUNX2 not being expressed may not be representative of the expression profile of RUNX2 in all patients. An alternative explanation is that the expression of RUNX2 may be unique to Asian patients as a large scale study of the expression of RUNX2 in American or European patients of breast cancer has yet to occur.

While the *in vitro* functions of RUNX1 in breast cells may not be well described, there is evidence to suggest that in breast cancer patients RUNX1 functions as a tumor suppressor. RUNX1 transcript is significantly lower in metastatic breast cancer cells as compared to primary tumors (Ramaswamy et al. 2003), and *RUNX1* is significantly mutated in patients of breast cancer (The Cancer Genome Atlas Network 2012). In the Human Protein Atlas study mentioned above, RUNX1 was strongly detected in normal breast tissue and in tumor tissue, however, the sample size was small (Uhlén et al. 2005, Pontèn et al. 2008, Uhlen et al. 2010). These *in vivo* observations or RUNX1 expression patterns and mutational frequencies seemed counterintuitive to the observation that RUNX1 was highly expressed in the malignant metastatic MDA-MB-231 cells, so it is therefore necessary to examine both the functions of RUNX1 in MDA-MB-231 cells and the expression pattern of RUNX1 in patient samples of breast cancer.

To investigate the roles of RUNX1 and RUNX2 in breast cancer, we examined the extent to which RUNX expression is associated with breast cancer progression and transition into early metastatic disease. To accomplish this, we used cell line models of breast cancer as well as human breast cancer patient tissue. Using MDA-MB-231 cells as a model of metastatic breast cancer cells to define the functions of endogenously expressed RUNX proteins, we characterized the patterns of genome-wide RUNX1 and RUNX2 binding and genome-wide transcriptome response to RNAi-mediated knockdown of RUNX proteins. Here we report that while RUNX1 and RUNX2 both appear to play roles in transcriptionally regulating genes involved in the invasive phenotype of MDA-MB-231 cells, only RUNX1 appears to be bound near the promoters of these genes. In human patients, while neither RUNX1 nor RUNX2 appear to be independent prognostic markers of disease, we do see RUNX1 expression as primarily associated with early stages of disease, while RUNX2 expression is primarily found in middle-late stage breast cancer. Integrating common markers of breast cancer prognosis (ER, PR, HER2, and AR), we find that RUNX1 expression is correlated with AR expression, and association of RUNX1 with either ER or HER2 is dependent on AR status. These results suggest that the expression of RUNX proteins in breast cancer is highly sensitive to hormone receptor status.

Results

Runx proteins do not regulate ribosomal RNA transcription, protein synthesis, or cell growth in MDA-MB-231 cells

In the hematopoietic and osteoblastic lineages, RUNX1 and RUNX2, respectively, control cell growth by regulating the rate limiting step of protein synthesis by suppressing transcription of ribosomal RNA. RUNX proteins regulate rRNA transcription through binding to the rDNA repeats and are observed to colocalize with upstream binding factor (UBF) at the periphery of nucleoli (Young et al. 2007a, Bakshi et al. 2008, Ali et al. 2010, 2012). Both RUNX1 and RUNX2 have been observed to be expressed in the malignant metastatic MDA-MB-231 cells, which are highly proliferative (Barnes et al. 2004, Javed et al. 2005, Pratap et al. 2008, Stender et al. 2010). It is therefore of interest to investigate whether the growth-suppressive function of RUNX proteins are operative in breast cancer cells.

We began by detecting the levels the levels of RUNX proteins in adjacent normal breast tissue from human patients by immunohistochemistry **(Figure 3.1 A)**. The

Human Protein Atlas has previously observed that RUNX1 can be detected in normal breast tissue, while RUNX2 cannot (Uhlén et al. 2005, Pontèn et al. 2008, Uhlen et al. 2010). While examination of RUNX protein levels in patient tissues does not provide a good indication of the potential for RUNX proteins to regulate rRNA transcription, it does allow for a comparison of the levels of RUNX1 and RUNX2 in normal breast tissue. This approach further indicates the extent to which we can reproduce the observations of others. Immunohistochemical staining of RUNX1 and RUNX2 in adjacent normal mammary epithelial cells showed that RUNX1 was strongly expressed while RUNX2 was not detected at all **(Figure 3.1 A)**.

Immunofluorescence (IF) is a method of immunodetection of an antigen via a fluorescent-conjugated secondary antibody and is generally considered more sensitive than detection with a horseradish peroxidase (HRP)-conjugated secondary antibody. This method allows for examination of the subnuclear localization of RUNX proteins and the extent to which RUNX proteins colocalize with upstream binding factor (UBF), which is a required member of the RNA polymerase I (Pol I) complex and for rRNA transcription (Voit et al. 1995). Therefore, we used IF to detect the level and subnuclear localization of RUNX1, RUNX2, and UBF in: 1) non-tumorigenic breast cells (MCF10a), 2) malignant, poorly-invasive breast cancer cells (MCF-7), and 3) malignant, highly-invasive breast cancer cells (MDA-MB-231) **(Figure 3.1 B)**. By qualitatively comparing of the observed levels of RUNX1 and RUNX2 in normal breast tissue **(Figure 3.1 A)** to the levels detected by immunostaining in MCF10a cells **(Figure 3.1 B – top row)**, we similarly observed that RUNX1 is expressed, while RUNX2 barely detected. Weak detection of RUNX2 in the non-tumorigenic MCF10a cells and no detection of RUNX2 in adjacent normal breast tissue is likely a technical issue as RUNX2 has been observed in normal MCF10a cells by electro-mobility shift assay (EMSA) (Inman and Shore 2003, Shore 2005). These results also give us an indication of the threshold of sensitivity for detection of RUNX proteins breast cancer patient tissue, which will be presented later. Using IF in non-tumorigenic breast and breast cancer cell lines, we observed that RUNX proteins were not co-localized with UBF **(Figure 3.1 B)**. This suggests that RUNX1 and RUNX2 are unlikely to be functioning to regulate rRNA transcription. However, these IF results only describe the potential for RUNX proteins to regulate rRNA transcription based on spatial proximity to UBF, and do not provide direct evidence.

To further investigate the roles of endogenous RUNX1 and RUNX2, we chose to use the MDA-MB-231 cell line, as this cell line has the highest endogenous levels of both RUNX1 and RUNX2 of those examined **(Figure 3.1B – bottom row)**. Utilization of this cell line, in which both RUNX1 and RUNX2 are endogenously expressed at high levels, also allows us to test the degree to which the functions of RUNX1 and RUNX2 are overlapping. RUNX2 regulates the transcription of rRNA by binding to the rDNA repeats in SaOS-2 cells, therefore SaOS-2 cells were used as a control (Young et al. 2007a, Ali et al. 2010, 2012). Although we do not observe overlap in the IF signals of RUNX proteins and UBF in MDA-MB-231 cells, we wanted to directly test the extent to which RUNX proteins are bound to ribosomal genes. Therefore, using chromatin immunoprecipitation followed by sequence-specific Real Time qPCR we measured the binding of RUNX1 and RUNX2 to genomic regions upstream (rDNA A) and within (rDNA B/C) ribosomal genes in MDA-MB-231 cells and SaOS-2 cells **(Figure 3.1 C)**. In this experiment, UBF was used as a positive control for immunoprecipitation of ribosomal DNA, while normal rabbit immunoglobulin G (IgG) was used as a control for immunospecificity. UBF is more strongly bound to rDNA repeats as compared to RUNX proteins (Young et al. 2007a, Bakshi et al. 2008, Ali et al. 2010, 2012), therefore the percent of input recovered by UBF immunoprecipitation (green bars) is plotted on the right y-axis, while normal rabbit IgG, RUNX1, and RUNX2 are plotted on the left y-axis. In SaOS-2 cells, immunoprecipitation with RUNX1 (yellow bars) or RUNX2 (green bars) recovered more input DNA at rDNA B and rDNA C as compared to immunoprecipitation with normal rabbit IgG (pink bars), which indicates that both RUNX1 and RUNX2 are bound to the rDNA repeats in SaOS-2 cells. In MDA-MB-231 cells, we observed that RUNX1 immunoprecipitation recovered more DNA at rDNA B and rDNA C than normal rabbit IgG, indicating that RUNX1 is likely bound to rDNA. We also observed that while the mean amount of rDNA B and rDNA C recovered by RUNX2 immunoprecipitation was more than with rabbit IgG in MDA-MB-231 cells, there was a high degree of variability (error bars for replicates) in the amount of rDNA B/C recovered by RUNX2 immunoprecipitation. This suggests that if RUNX2 is bound to rDNA repeats in the MDA-MB-231 cells the interaction between RUNX2 and rDNA may be weak or only occurring in a subset of MDA-MB-231 cells.

Figure 3.1: RUNX1 and RUNX2 do not colocalize with UBF and are weakly bound to ribosomal DNA in breast cells.

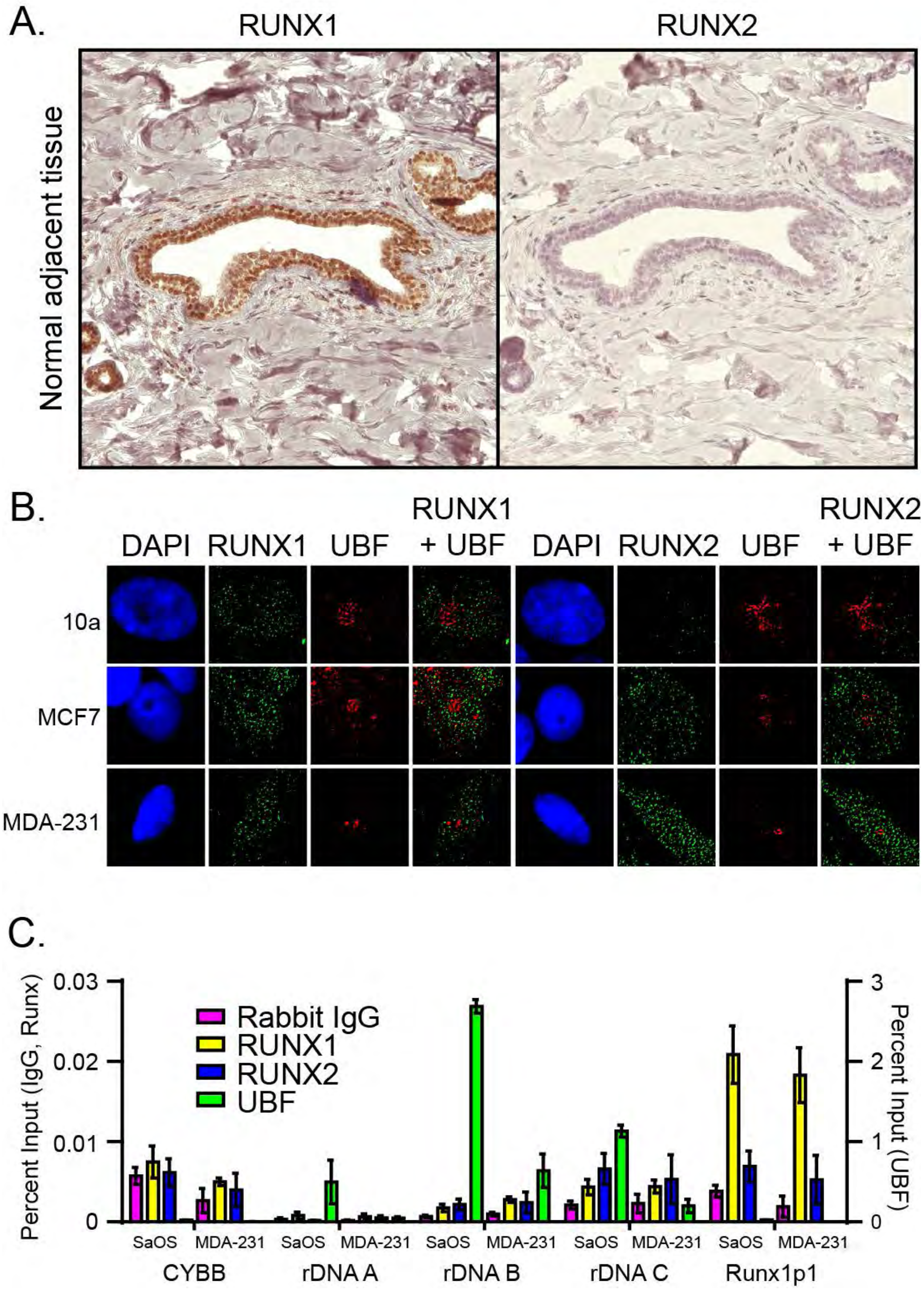


Figure 3.1: RUNX1 and RUNX2 do not colocalize with UBF and are weakly bound to ribosomal DNA in breast cells.

(A) Immunohistochemical staining of RUNX1 and RUNX2 in patient samples of normal adjacent breast tissue. **(B)** Immunofluorescence staining of RUNX1, RUNX2 and UBF in non-tumorigenic (MCF10a), malignant, poorly-invasive (MCF-7) and malignant, highly-invasive (MDA-MB-231) breast cell lines. **(C)** Chromatin immunoprecipitation using normal rabbit IgG, RUNX1, RUNX2, and UBF from SaOS-2 osteosarcoma and MDA-MB-231 bone metastatic breast cancer cells followed by qPCR for regions proximal to the TSS of the human rDNA repeat (Ali et al. 2010, 2012). CYBB – cytochrome b-245, beta polypeptite / GP91-PHOX; rDNA A/B/C ribosomal DNA regions, refer to Ali et al. 2010 Figure 1D; Runx1p1 – RUNX1 promoter 1 (p1) region. Bars represent mean percent of input and error bars equal SEM for two technical replicates and two biological replicates.

Combining our ChIP observations with our IF observations indicates a possible disconnect between the anti-nucleolar localization of RUNX proteins in whole cells observed by IF and the binding of RUNX proteins to rDNA repeats in biochemically extracted chromatin. IF and ChIP provide evidence for the location of a protein not the function of a protein, so the extent to which RUNX proteins regulate protein synthesis in breast cells has not been thoroughly addressed. Therefore, using the MDA-MB-231 cells as a model of breast cancer cells and SaOS-2 cells as a model for normal RUNX function, we investigated the functional contributions of RUNX proteins in the regulation of cell growth, rRNA transcription, and protein synthesis **(Figure 3.2)**. To simultaneously measure the extent to which RUNX proteins regulate cell growth and rRNA transcription, we performed a time course of RNA interference (RNAi) knockdown of the endogenous RUNX proteins in both MDA-MB-231 and SaOS-2 cells. We observed that the short interfering RNAs (siRNAs) targeting RUNX1 and RUNX2 were sufficient to significantly reduce the protein detected in whole cell lysates at 24, 48, and 72 hours in SaOS-2 and MDA-MB-231 cells **(Figure 3.2 A)**. We also saw two interesting phenomena in the levels of RUNX proteins, which were uniquely observed in the MDA-MB-231 cells: 1) as MDA-MB-231 cell density increased over time, the levels of RUNX1 and RUNX2 decreased, and 2) the levels of RUNX2 were significantly lower in MDA-MB-231 cells as compared to SaOS-2 cells. By counting the number of cells in culture after 24, 48, and 72 hours post-transfection of siRNA, we measured the effects of RUNX proteins on growth **(Figure 3.2 B)**. Here, we observed that RUNX proteins do not significantly affect growth in either MDA-MB-231 cells or SaOS-2 cells. RUNX2 has previously been observed to affect the growth of SaOS-2 cells, but only in the absence of histone deacetylase 1 (HDAC1) (Ali et al. 2012), so not seeing a growth affect in a RUNX knockdown still does not rule out RUNX-regulation of rRNA transcription and protein synthesis. We therefore examined the levels of pre-ribsomal RNA (pre-rRNA) after 24, 48, and 72 hours **(Figure 3.2 C)**. pre-rRNA is rapidly processed into mature rRNA, and is used as a proxy for measuring rRNA transcription rates (Grummt and Voit 2010). Knockdown of RUNX1 and RUNX2 in Kasumi-1 and SaOS-2 cells, respectively, has resulted in reduction of pre-rRNA levels (Young et al. 2007a, Bakshi et al. 2008, Ali et al. 2010, 2012). We observed that knockdown of RUNX2 in SaOS-2 caused a significant increase in pre-rRNA levels cells at 24, 48, and 72 hours post-transfection, and knockdown of RUNX1 in SaOS-2 caused an increase only at the 24 hour time point **(Figure 3.2 C – left)**. In contrast, knockdown of RUNX proteins in MDA-MB-231 cells had no significant effects on pre-rRNA levels at any time post-transfection **(Figure 3.2 C – right)**. Pulse-labeling proliferating cells with ^{35}S-tagged amino acids allows for a measurement of the rate of protein synthesis in the cell population. 48 hours post-transfection of RUNX siRNAs, SaOS-2 and MDA-MB-231 cells were pulse labeled with radiolabeled methionine and cysteine to measure the effects of RUNX proteins on the rate of protein synthesis **(Figure 3.2 D)**. Each replicate experiment was quantified by densitometry and combined **(Figure 3.2 D – right)**. We observed that RUNX1 and RUNX2 significantly affected the rate of protein synthesis in SaOS-2 cells, but not at all in MDA-MB-231 cells. These observations are consistent with the results for pre-rRNA levels following knockdown of RUNX proteins **(Figure 3.2 C)**.

Our combined observations in MDA-MB-231 cells show that RUNX proteins do not co-localize with components of RNA Pol I-mediated transcription (UBF) by immunofluorescence **(Figure 3.1 B)**, are not strongly bound to the rDNA by ChIP **(Figure 3.1 C)**, and do not appear to affect pre-rRNA levels **(Figure 3.2 C)** or protein synthesis rates **(Figure 3.2 D)** by RNAi knockdown of Runx proteins. These results clearly demonstrate that Runx proteins do not regulate the transcription of rRNA in MDA-MB-231 cells. Transcription and processing of rRNA are rate limiting steps for cellular growth. The observation that Runx proteins are not affecting these processes is consistent with the result that Runx proteins do not affect the growth of MDA-MB-231 cells **(Figure 3.2 B)**.

Figure 3.2: RUNX1 and RUNX2 in MDA-MB-231 cells does not appear to affect cell growth, ribosomal RNA transcription or protein synthesis.

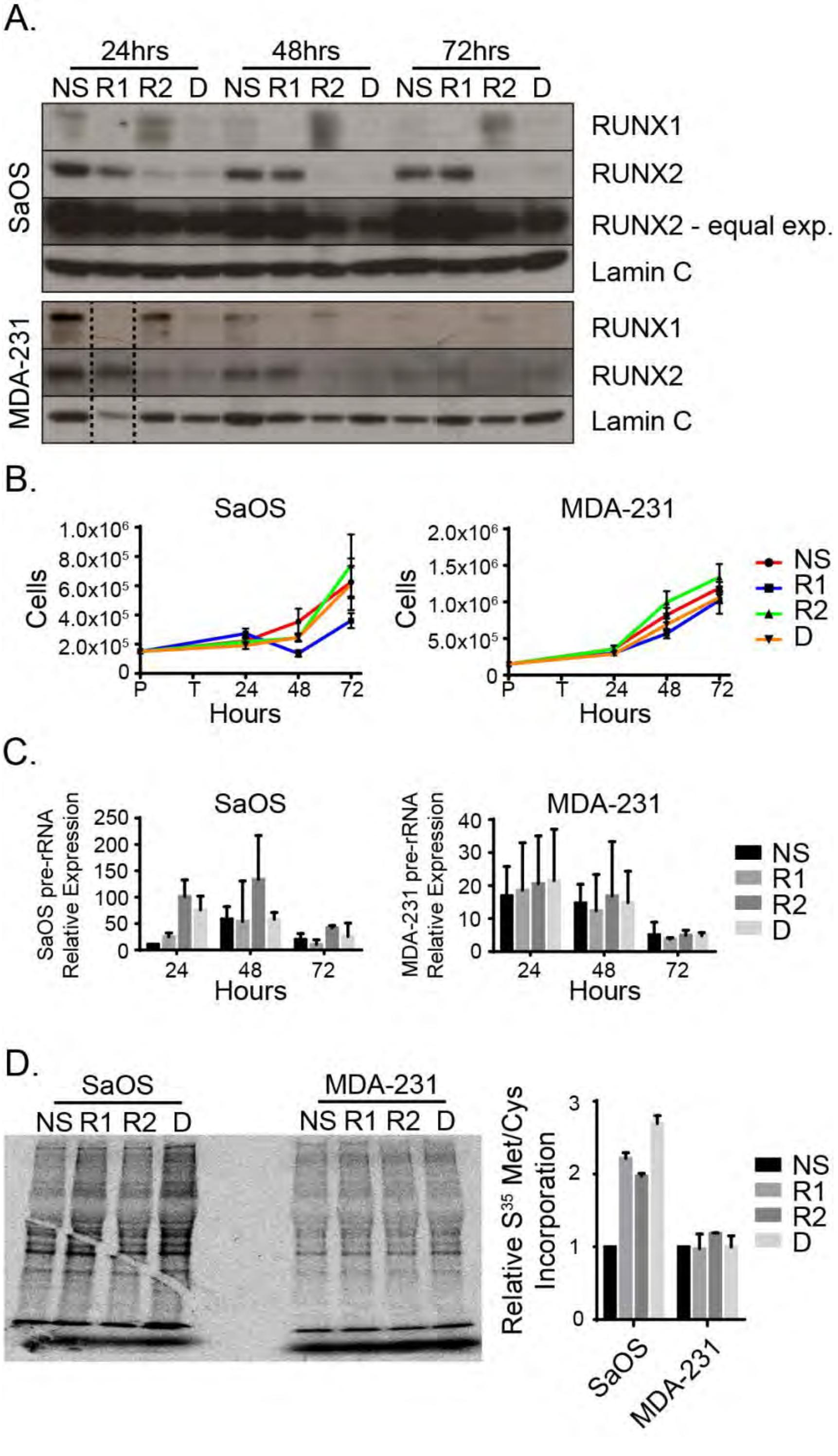


Figure 3.2: RUNX1 and RUNX2 in MDA-MB-231 cells does not appear to affect cell growth, ribosomal RNA transcription or protein synthesis.

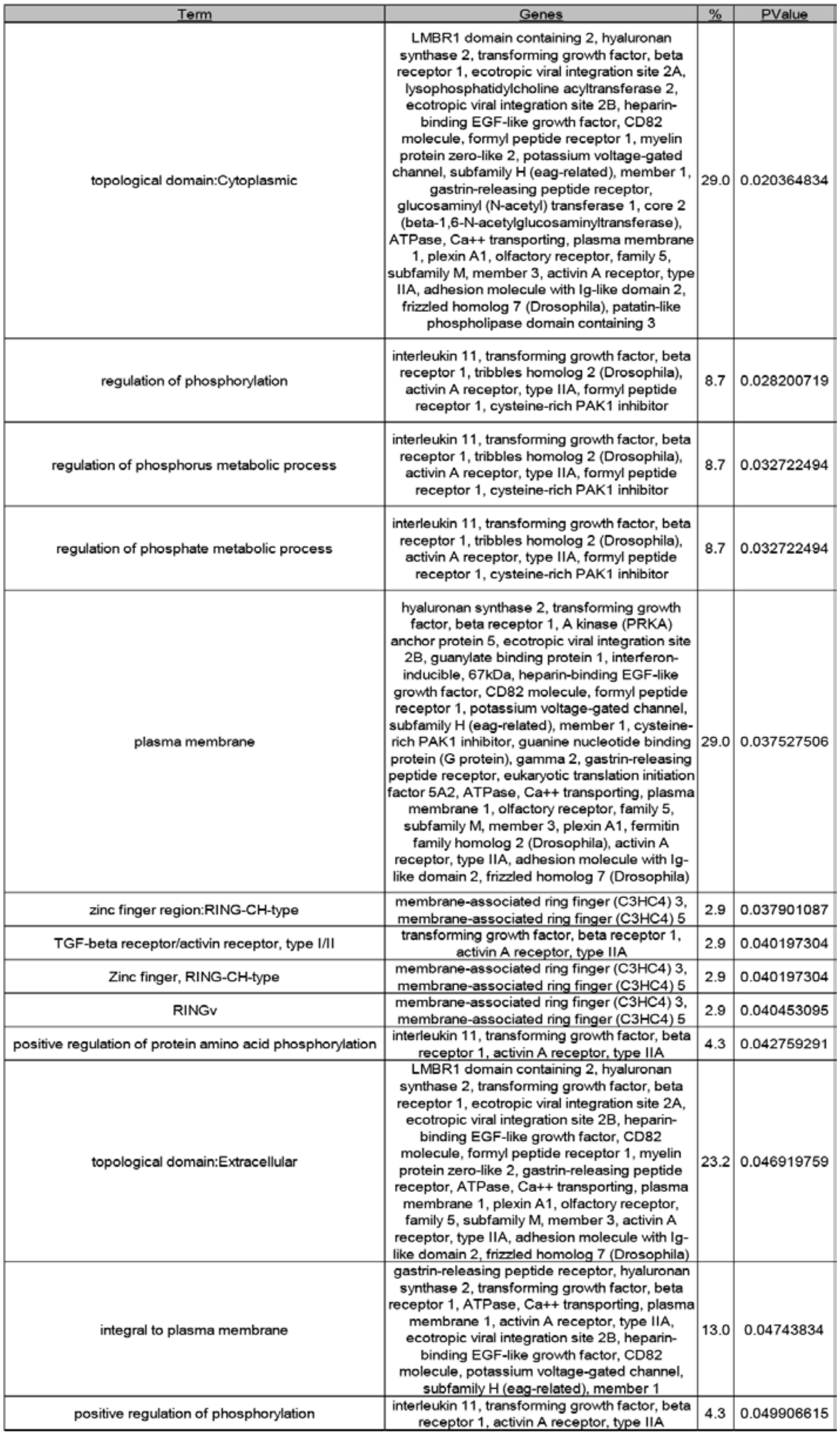
(A) Representative western blot in SaOS-2 (top panel) and MDA-MB-231 (bottom panel) cells detecting levels of RUNX1 (top blots), RUNX2 (middle blots), and Lamin C (bottom blots) in response to siRNA 24, 48, and 72 hours (x-axis) post-transfection. RUNX2 blots in SaOS-2 panel are for linear detection range (upper) and equal exposure to MDA-MB-231 cells (lower). **(B)** Cell counts for MDA-MB-231 (left panel) and SaOS-2 (right panel) at 24, 48, and 72 hours following transfection with siRNA. Symbol represents mean and error bars represent SEM for 3 transfection replicates with two counting replicates each for 2 biological replicates. **(C)** Time course of knockdown of RUNX proteins in MDA-MB-231 and SaOS-2 cells. NS = Non-silencing siRNA, R1 = RUNX1 siRNA, R2 = RUNX2 siRNA, and D = RUNX1 and RUNX2 combined siRNAs. X-axis is time post transfection, at which point RNA was isolated, cDNA amplified and qPCR for pre-rRNA was executed. Relative expression values are plotted against ACTB internal control using delta-Ct method (Bustin et al. 2009). Bars represent mean relative expression and error bars equal SEM for two technical replicates each of two biological replicates. **(D)** Representative radiograph of ^{35}S-Met/Cys pulse-labeled SaOS-2 (left half of blot) or MDA-MB-231 (right half of blot) protein lysates run on an SDS-PAGE gel and exposed to film following 48 hours of transfection (left panel). Densitometric mean and SEM (error bars) of ^{35}S-

Met/Cys incorporation for SaOS-2 and MDA-MB-231 cells for 2 biological replicates (right panel). **(All)** NS = Non-silencing siRNA, R1 = RUNX1 siRNA, R2 = RUNX2 siRNA, and D = RUNX1 and RUNX2 combined siRNAs.

Expression of RUNX proteins in MDA-MB-231 cells promotes an invasive phenotype

Given that we observed such disparate effects of RUNX knockdown on rRNA expression, protein synthesis and cell growth between breast cancer and osteosarcoma cells, we sought to understand the potentially novel transcriptional functions of RUNX proteins in MDA-MB-231 breast cancer cells. To this end, we performed Affymetrix transcriptome analysis under conditions where cells were transfected with either a non-silencing control or a combined pool of RUNX1 and RUNX2 siRNAs. We identified 66 protein coding genes whose transcript levels were changed more than 1.5 fold and whose fold changes were reproducible across biological replicates ($p < 0.05$). Using DAVID (Huang et al. 2009a, 2009b) to understand the biological consequences of changes in the levels of these transcripts, we found these genes to be ontologically associated with terms related to adhesion, migration, and invasion **(Table 3.1)**. One striking observation found in the genes responsive to RUNX RNAi in MDA-MB-231 cells was the lack of genes that are regulated during hematopoiesis and osteoblastogenesis by RUNX1 and RUNX2 respectively (Vradii et al. 2005, Young et al. 2007b, van der Deen et al. 2012). This observation combined with the non-regulation of rRNA transcription suggests that endogenously-expressed RUNX1 and RUNX2 in MDA-MB-231 cells have distinct regulatory roles as compared to RUNX1 and RUNX2 in the hematopoietic and osseous lineages, respectively.

Table 3.1: Ontological terms enriched in genes responsive to RUNX-siRNA.



Term	Genes	%	PValue
topological domain:Cytoplasmic	LMBR1 domain containing 2, hyaluronan synthase 2, transforming growth factor, beta receptor 1, ecotropic viral integration site 2A, lysophosphatidylcholine acyltransferase 2, ecotropic viral integration site 2B, heparin-binding EGF-like growth factor, CD82 molecule, formyl peptide receptor 1, myelin protein zero-like 2, potassium voltage-gated channel, subfamily H (eag-related), member 1, gastrin-releasing peptide receptor, glucosaminyl (N-acetyl) transferase 1, core 2 (beta-1,6-N-acetylglucosaminyltransferase), ATPase, Ca++ transporting, plasma membrane 1, plexin A1, olfactory receptor, family 5, subfamily M, member 3, activin A receptor, type IIA, adhesion molecule with Ig-like domain 2, frizzled homolog 7 (Drosophila), patatin-like phospholipase domain containing 3	29.0	0.020364834
regulation of phosphorylation	interleukin 11, transforming growth factor, beta receptor 1, tribbles homolog 2 (Drosophila), activin A receptor, type IIA, formyl peptide receptor 1, cysteine-rich PAK1 inhibitor	8.7	0.028200719
regulation of phosphorus metabolic process	interleukin 11, transforming growth factor, beta receptor 1, tribbles homolog 2 (Drosophila), activin A receptor, type IIA, formyl peptide receptor 1, cysteine-rich PAK1 inhibitor	8.7	0.032722494
regulation of phosphate metabolic process	interleukin 11, transforming growth factor, beta receptor 1, tribbles homolog 2 (Drosophila), activin A receptor, type IIA, formyl peptide receptor 1, cysteine-rich PAK1 inhibitor	8.7	0.032722494
plasma membrane	hyaluronan synthase 2, transforming growth factor, beta receptor 1, A kinase (PRKA) anchor protein 5, ecotropic viral integration site 2B, guanylate binding protein 1, interferon-inducible, 67kDa, heparin-binding EGF-like growth factor, CD82 molecule, formyl peptide receptor 1, potassium voltage-gated channel, subfamily H (eag-related), member 1, cysteine-rich PAK1 inhibitor, guanine nucleotide binding protein (G protein), gamma 2, gastrin-releasing peptide receptor, eukaryotic translation initiation factor 5A2, ATPase, Ca++ transporting, plasma membrane 1, olfactory receptor, family 5, subfamily M, member 3, plexin A1, fermitin family homolog 2 (Drosophila), activin A receptor, type IIA, adhesion molecule with Ig-like domain 2, frizzled homolog 7 (Drosophila)	29.0	0.037527506
zinc finger region:RING-CH-type	membrane-associated ring finger (C3HC4) 3, membrane-associated ring finger (C3HC4) 5	2.9	0.037901087
TGF-beta receptor/activin receptor, type I/II	transforming growth factor, beta receptor 1, activin A receptor, type IIA	2.9	0.040197304
Zinc finger, RING-CH-type	membrane-associated ring finger (C3HC4) 3, membrane-associated ring finger (C3HC4) 5	2.9	0.040197304
RINGv	membrane-associated ring finger (C3HC4) 3, membrane-associated ring finger (C3HC4) 5	2.9	0.040453095
positive regulation of protein amino acid phosphorylation	interleukin 11, transforming growth factor, beta receptor 1, activin A receptor, type IIA	4.3	0.042759291
topological domain:Extracellular	LMBR1 domain containing 2, hyaluronan synthase 2, transforming growth factor, beta receptor 1, ecotropic viral integration site 2A, ecotropic viral integration site 2B, heparin-binding EGF-like growth factor, CD82 molecule, formyl peptide receptor 1, myelin protein zero-like 2, gastrin-releasing peptide receptor, ATPase, Ca++ transporting, plasma membrane 1, plexin A1, olfactory receptor, family 5, subfamily M, member 3, activin A receptor, type IIA, adhesion molecule with Ig-like domain 2, frizzled homolog 7 (Drosophila)	23.2	0.046919759
integral to plasma membrane	gastrin-releasing peptide receptor, hyaluronan synthase 2, transforming growth factor, beta receptor 1, ATPase, Ca++ transporting, plasma membrane 1, activin A receptor, type IIA, ecotropic viral integration site 2B, heparin-binding EGF-like growth factor, CD82 molecule, potassium voltage-gated channel, subfamily H (eag-related), member 1	13.0	0.04743834
positive regulation of phosphorylation	interleukin 11, transforming growth factor, beta receptor 1, activin A receptor, type IIA	4.3	0.049906615

Table 3.1: Ontological terms enriched in genes responsive to RUNX-siRNA.

Using DAVID, genes responsive (fold-change >= 1.5, p-value < 0.05) to RUNX-siRNA were subjected to ontological term clustering to search for over-represented GO terms associated with Runx-siRNA-responsive genes.

Given the ontological associations of the RUNX-siRNA-responsive genes **(Table 3.1)**, we investigated the extent to which RUNX expression might affect the invasiveness of MDA-MB-231 cells. One way to measure the invasiveness of a cell line is to perform a Matrigel invasion assay. In these experiments, cells are cultured in the top of a transwell insert and a chemoattractant is cultured under and surrounding the insert. The bottom of the insert has small holes through which the cells can migrate towards the chemoattractant. The control experiment is just naked plastic, while the invasion experiment has a layer of Matrigel between the cells and bottom of the insert through which the cells have to invade to migrate towards the chemoattractant. After the cells have been allowed to migrate towards to chemoattractant for a fixed period of time, cells are fixed in ethanol and cells that did not migrate through the pores in the insert are removed. Cells that have moved through the pores of the insert are now adhered to the bottom of the insert, these are stained and the numbers of cells which have migrated or invaded are quantified. While it is known that overexpression of *Mus musculus* Runx2 in MDA-MB-231 cells causes increased invasiveness (Barnes et al. 2004, Javed et al. 2005), it is not known how RUNX1 contributes to the invasiveness of MDA-MB-231 cells. Using Matrigel invasion assays, we observe that knockdown of RUNX1 or RUNX2 **(Figure 3.3 A)** reduces the invasiveness of MDA-MB-231 cells **(Figure 3.3 B)**. These results are in accordance with what was expected based on the DAVID analysis.

Figure 3.3 RUNX proteins promote the invasiveness of MDA-MB-231 cells

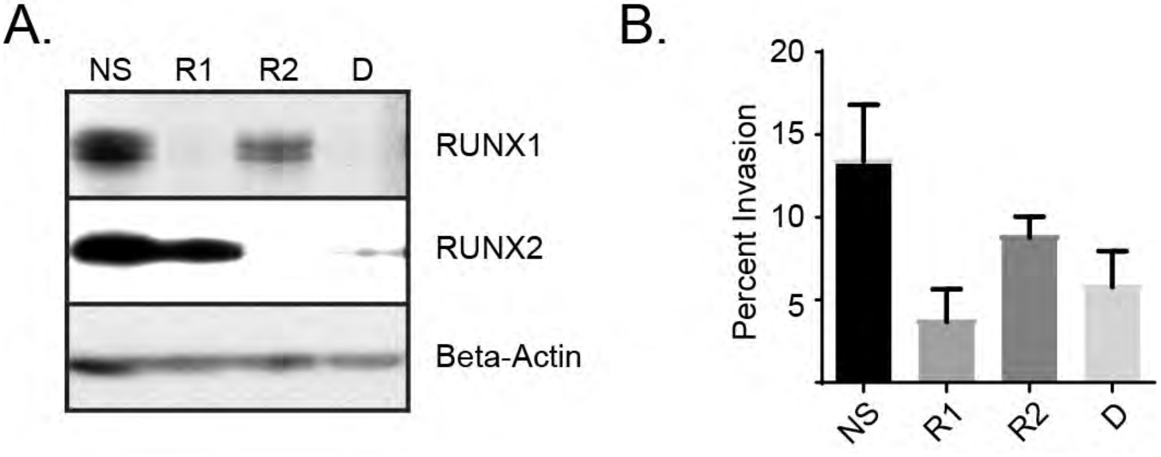


Figure 3.3 RUNX proteins promote the invasiveness of MDA-MB-231 cells

(A) Representative Western blot showing protein levels for RUNX1 (top blot), RUNX2 (middle blot), and Beta-Actin (lower blot) following 48 hours of siRNA transfection. **(B)** Column plot depicting mean and SEM (error bars) percent of invasion for each siRNA transfection (N=2). **(All)** NS = Non-silencing siRNA, R1 = RUNX1 siRNA, R2 = RUNX2 siRNA, and D = RUNX1 + RUNX2 siRNAs.

RUNX proteins share a highly conserved Runt homology domain, which binds the same consensus sequence (Van Wijnen et al. 2004); however, it is not known the extent to which these two proteins overlap in function when endogenously expressed in the same cell line. To address this, we performed gene-specific Real-Time qPCR on genes identified by Affymetrix using primer pairs designed by FoxPrimer (see Chapter 4 for methods). Of the 66 total genes, we were able to validate 44 primer pairs that were greater than 80% efficient as measured by standard curve (Bustin et al. 2009). Using these 44 genes, we performed qPCR on cDNA amplified from RNA isolated from MDA-MB-231 cells transfected with non-targeting, RUNX1, RUNX2, or RUNX1+RUNX2 siRNA and found that the expression of 41 genes were significantly changed compared to control in response to at least one siRNA transfection. Categorizing these transcripts by responsiveness to RUNX siRNA resulted in three major classes **(Figure 3.4)**: 1) responsive to either RUNX1 or RUNX2 siRNA **(Figure 3.5 A)**, 2) primarily responsive to RUNX1 siRNA **(Figure 3.5 B)**, and 3) primarily responsive to RUNX2 siRNA **(Figure 3.5 C)**. Five genes for which we were able to design efficient primers for Real Time qPCR did not show a significant change in expression in the confirmation experiments (data not shown). One gene, myelin protein zero-line 2 (MPZL2), showed an increase in detection in response to RUNX siRNAs, and therefore did not fall into the three major categories (data not shown). The observation that so many of the genes (approximately 2/3) are primarily responsive to only one RUNX protein is quite unexpected. This suggests that while the phenotypic response of RUNX protein knockdown is similar in MDA-MB-231 cells, the mechanisms in use by RUNX1 and RUNX2 may be different.

Figure 3.4 Genes responsive to RUNX-siRNAs fall into three major categories

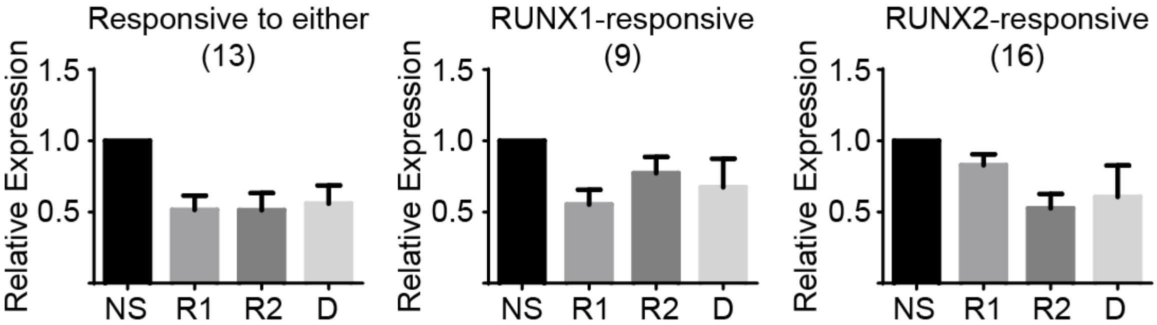


Figure 3.4 Genes responsive to RUNX-siRNAs fall into three major categories

Mean and SEM (error bars) plots of three major classes of genes identified and validated by Affymetrix Human Gene 1.0ST array based on response to 48 hours of siRNA transfection. NS = Non-silencing siRNA, R1 = RUNX1 siRNA, R2 = RUNX2 siRNA, and D = RUNX1 + RUNX2 siRNAs.

Figure 3.5 Real Time qPCR validation of Affymetrix-identified genes

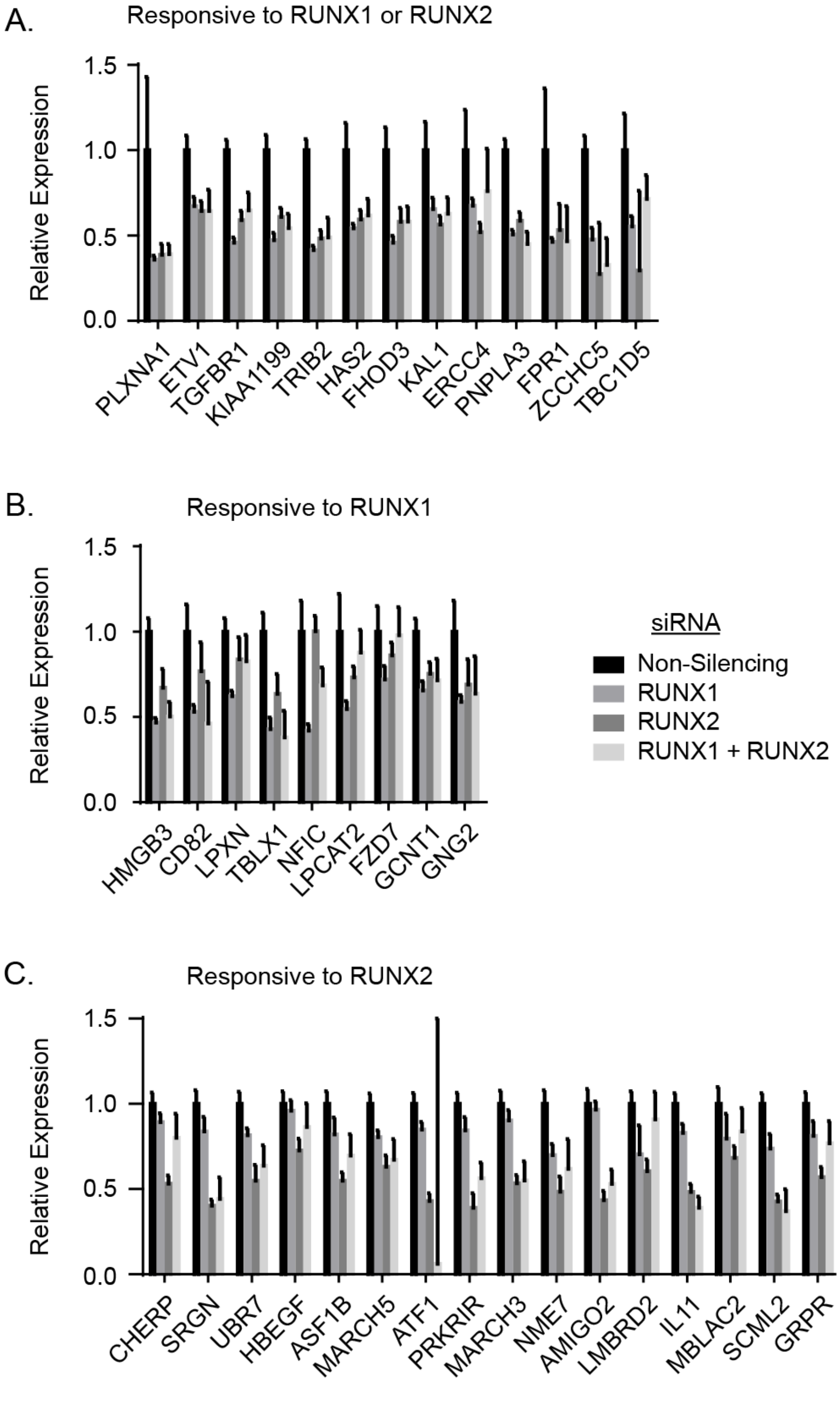


Figure 3.5 Real Time qPCR validation of Affymetrix-identifed targets

Real Time qPCR validation of genes responsive to RUNX-siRNA ordered by response group: **(A)** responsive to either, **(B)** reponsive to RUNX1, and **(C)** reponsive to RUNX2 . Primers were designed using FoxPrimer (Chapter 4). Bars represent mean and error bars are SEM for two biological replicates for each gene normalized by delta-delta-Ct to ACTB.

Alongside the Affymetrix experiments, we performed ChIP-seq for endogenous RUNX1 and RUNX2 in MDA-MB-231 cells. As mentioned before, the Affymetrix experiment was done using a double-knockdown system. Previous models of RUNX-mediated transcriptional functions based on single gene regulation suggest that we should be able to use RUNX1 and RUNX2 binding sites to determine which genes identified by Affymetrix can be uniquely regulated by RUNX1 or RUNX2 in MDA-MB-231 cells (Lian et al. 2006). Accordingly, binding within the promoter region of a gene should be sufficient for RUNX to transcriptionally regulate gene expression in trans. In our ChIP-seq of endogenous RUNX1 and RUNX2 in MDA-MB-231 cells, we found that RUNX1 and RUNX2 did not share many of the same binding sites **(Figure 3.6 A)**, which may explain the large number of genes that were uniquely responsive to RUNX1 or RUNX2 siRNA **(Figure 3.4 & Figure 3.5)**. Contrary to previous models, where RUNX proteins exclusively bind to promoter regions, we also found that the vast majority of RUNX binding sites defined by MACS (Zhang et al. 2008) were not located within promoter regions. When defining a promoter region as 1Kb 5’ of the transcriptional start site (TSS) we observed only about 10% and less than 1% of all RUNX1 and RUNX2 peaks, respectively, were bound within the promoters of RefSeq (Pruitt et al. 2009) genes in MDA-MB-231 cells **(Figure 3.6 B)**. These observations are particularly interesting because all RUNX proteins share the same common DNA sequence binding motif **(Figure 3.6 C),** which was defined *in vitro* on naked DNA (Melnikova et al. 1993), however, in MDA-MB-231 cells there was not a lot of overlap between the binding of RUNX1 and RUNX2. This suggests that the *in vivo* binding of RUNX proteins is more complicated than sequence motif recognition alone.

To investigate a potential mechanism as to how RUNX1 and RUNX2 are not occupying many of the same binding sites, we performed discriminative *de novo* motif discovery (Heinz et al. 2010) on the enriched intervals defined by MACS and found that the RUNX1 ChIP-Seq was enriched for a motif very similar to a previously identified RUNX1 motif, while the motif enriched in the RUNX2 ChIP-Seq was most similar to the motif discovered by a PU.1 ChIP-seq **(Figure 3.6 D)**. PU.1 is a member of the E-twenty six (ETS) transcription factor family, all of which bind a similar sequence motif (Gutierrez-Hartmann et al. 2007). Runx2 has been shown to interact with Ets1 and synergistically regulate gene expression in murine cells (Sato et al. 1998), so it is possible that in MDA-MB-231 cells much of the DNA-binding activity of RUNX2 is derived through an interaction with an ETS factor. However, it is not known which ETS factor RUNX2 may be interacting with in MDA-MB-231 cells. Runx2 has also been observed to synergistically cooperate with CCAAT/enhancer binding proteins alpha and delta (C/EBPα & C/EBPδ) through a protein-protein interaction in the osteocalcin promoter (Gutierrez et al. 2002). In this study, Runx2 was still recruited to the osteocalcin promoter even in when the RUNX-binding site was mutated. This suggests that RUNX2 in MDA-MB-231 cells has the potential to be recruited to

DNA independently of the RUNX-binding sequence by protein-protein interactions with co-factors. The difference in motifs found within the enriched regions along with the significantly reduced number of RUNX2 peaks in promoters suggests that in the context of MDA-MB-231 cells, RUNX2 transcriptional activities may be quite different than those of RUNX1.

Figure 3.6 RUNX proteins do not bind similar regions in MDA-MB-231 cells

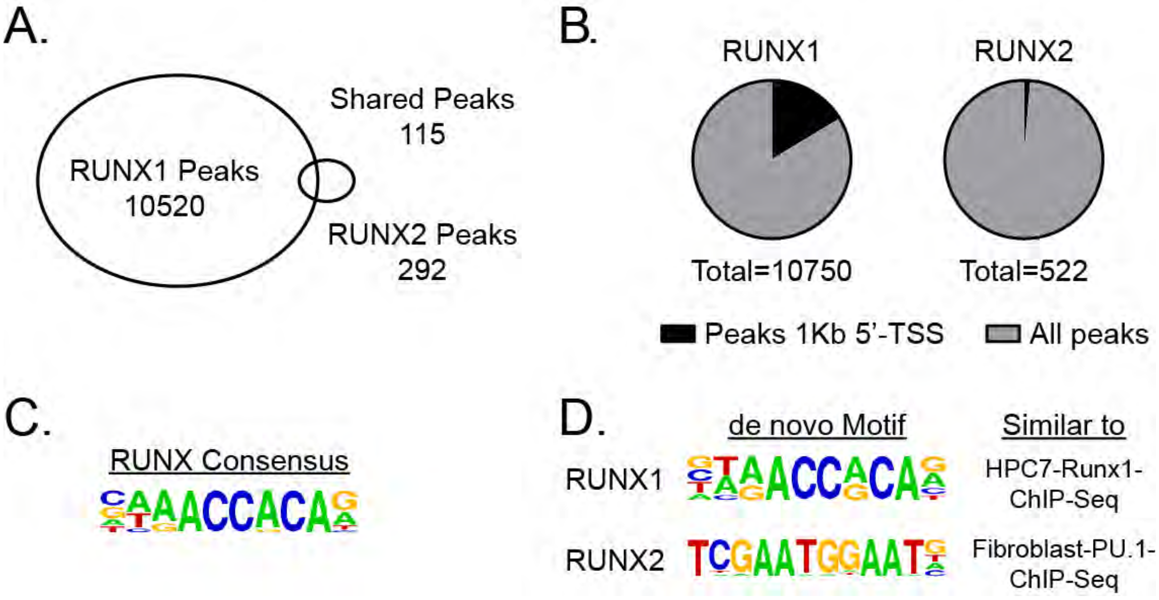


Figure 3.6 RUNX proteins do not bind similar regions in MDA-MB-231 cells

(A) Venn diagram depicting the number of overlapping RUNX1 and RUNX2 ChIP-seq peak regions as identified by MACS (Zhang et al. 2008). **(B)** Pie chart of the genomic locations of RUNX1 (left) and RUNX2 (right) MACS peaks, which are found 2Kb 5'-TSS to the TSS (black) as compared to anywhere else in the genome (grey). **(C)** SeqLogo of the consensus RUNX motif (Heinemeyer et al. 1998, Bryne et al. 2008). **(D)** Top results of discriminative motif discovery for the sequences within the genomic intervals defined as peaks for RUNX1 (top) and RUNX2 (bottom).

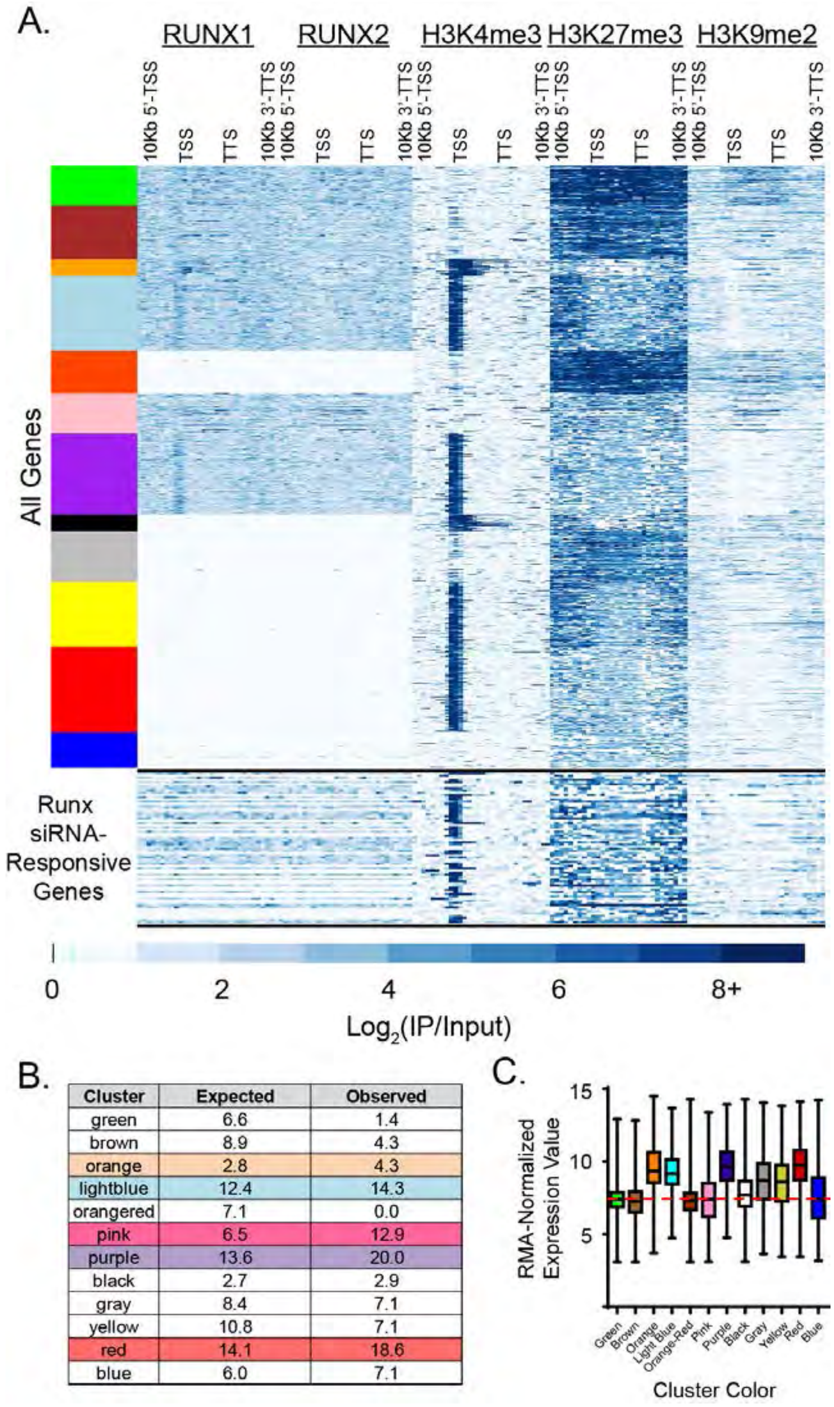
Functional RUNX1 peaks are associated with binding near the TSS and enrichment of H3K4me3

To understand what role chromatin may be playing in the binding of RUNX proteins to DNA, we compared RUNX1 and RUNX2 ChIP-seq data with ChIP-seq data for H3K4me3, H3K27me3, and H3K9me2 in MDA-MB-231 cells (from Chapter 2). We focused on the discovery of proximal (within 10Kb of genes) patterns in RUNX1 or RUNX2 binding as based on previous models of RUNX protein functions we had high confidence these regions would give insight into potential regulatory roles of RUNX proteins (Lian et al. 2006). Our goal was to understand if a pattern of RUNX1, RUNX2, H3K4me3, H3K27me3, and H3K9me2 binding was found near the genes that are responsive to RUNX siRNA. To identify these patterns, we created a matrix of enrichment data where each relative genomic region[1] for each mark is treated as a dimension and for each gene measured using the Affymetrix arrays a point based on these dimensions is defined in high-dimensional space using the corresponding enrichment ratio (IP reads / input reads). Next, we applied k-means clustering to partition these points into clusters based on Euclidean distance between all points and randomly generated cluster centers. The centers are randomly regenerated until a best fit for cluster assignment is reached. Based on these

[1] See Chapter 5 for more detailed methods regarding the definition of relative genomic regions.

cluster definitions, we plotted a heat map of the enrichment (IP reads / input reads) for relative genomic regions for each gene for each factor **(Figure 3.7 A)**. When attempting to cluster with only RUNX proteins, we were not able to identify any meaningful clusters in relation to RUNX siRNA-responsive genes (data not shown). The genes responsive to RUNX siRNA were plotted both within the main (clustered) figure and on their own **(Figure 3.7 A – bottom)**. Visually, the binding profile of RUNX1, RUNX2, H3K4me3, H3K27me3, and H3K9me2 near RUNX siRNA-responsive genes looked most similar to the purple and light blue clusters. Using a chi-squared contingency table to identify the clusters in which RUNX siRNA-responsive genes are most enriched or depleted, we found that RUNX siRNA-responsive genes were overrepresented in the purple, pink, light blue and orange clusters **(Figure 3.7 B)**. Based on the patterns of H3K4me3, which is associated with transcriptional activation and the transcriptionally repressive marks H3K27me3 and H3K9me2, we expected that the gene expression patterns would be similar (Li et al. 2007). Using gene expression measured by Affymetrix arrays, we plotted the distribution of robust means average (RMA) (Irizarry et al. 2003) transcript detection levels for each gene in each cluster as a box and whiskers plot **(Figure 3.7 C)**. Within the orange-red, black, grey, yellow, red, and blue clusters, the transcriptional initiation-associated chromatin modification, H3K4me3, was strongly enriched near the TSS, and these clusters contained some of the highest expressed genes as defined by Affymetrix arrays **(Figure 3.7 C)**. Four of the clusters (purple, pink, light blue and orange) which were overrepresented for genes responsive to RUNX siRNA were similarly enriched for H3K4me3 near the TSS **(Figure 3.7 A)** and contained genes that are highly expressed **(Figure 3.7 C)**. This suggests that the genes regulated by RUNX proteins in MDA-MB-231 cells are among the most highly expressed transcripts. We observed a clear enrichment of RUNX1 binding near the TSS of genes within these clusters, while RUNX2 binding did not appear to be strongly focused anywhere **(Figure 3.7 A)**. However, considering only the genes responsive to RUNX siRNA **(Figure 3.7 A – bottom)** this pattern of RUNX1 binding near the TSS was less apparent. This suggested that the strongest RUNX1 binding events near the TSS may not be associated with a functional response in gene expression when RUNX1 is knocked down using siRNA.

Figure 3.7 Gene-centric clustering of RUNX1, RUNX2, H3K4me3, H3K27me3, and H3K9me2 show that RUNX1 is primarily upstream of genes marked by H3K4me3 near the TSS and are strongly expressed.



Cluster	Expected	Observed
green	6.6	1.4
brown	8.9	4.3
orange	2.8	4.3
lightblue	12.4	14.3
orangered	7.1	0.0
pink	6.5	12.9
purple	13.6	20.0
black	2.7	2.9
gray	8.4	7.1
yellow	10.8	7.1
red	14.1	18.6
blue	6.0	7.1

Figure 3.7 Gene-centric clustering of RUNX1, RUNX2, H3K4me3, H3K27me3, and H3K9me2 show that RUNX1 is primarily upstream of genes marked by H3K4me3 near the TSS and are strongly expressed.

(A) Heatmap of signal ratios (IP/Input) for each RefSeq (Pruitt et al. 2009) gene (row) measured by Affymetrix Human Gene 1.0ST array. For each gene, 30 relative genomic regions were defined: 10Kb 5'-TSS to TSS in 1Kb steps, TTS to 10Kb 3'-TTS in 1Kb steps, and TSS to TTS in 10 approximately equal steps (normalized to 1Kb per step). The matrix of signal ratios per relative genomic region per transcription factor / histone modification per gene were then subjected to k-means clustering to identify 12 clusters (marked by colors on left). Scale is presented below heat map. **(B)** Chi-squared table showing expected (distribution of all genes) versus observed (distribution of genes responsive to RUNX siRNA) per cluster color. **(C)** Mean RMA-normalized expression values (Irizarry et al. 2003) for the genes in each cluster in box and whisker plot. Horizontal red dotted line represents the mean RMA-normalized detection level of all negative control probes, which is later used to define expressed versus non-expressed genes.

RUNX1 is associated with actively expressed genes and weakly associated with RUNX siRNA responsive genes.

To understand further how RUNX binding is associated with gene expression in the MDA-MB-231 cells, we used PeaksToGenes to perform statistical tests on the enrichment of each factor in relative genomic regions. As we saw in **Figure 3.7**, strong RUNX1 binding was most concentrated on the TSS of genes that are likely to be expressed based on the enrichment of H3K4me3 in the same region and the cluster of genes having a generally high expression distribution **(Figure 3.7 C).** However, this is not direct evidence that RUNX1 preferentially binds to expressed genes as many genes in these k-means-identified clusters are not responsive to RUNX siRNA. To directly test whether RUNX protein binding in MDA-MB-231 cells is associated with actively expressed genes, we first bisected the list of genes whose transcript levels are measured by Affymetrix Human Gene 1.0ST arrays into two groups. Then, using internal spike-in negative control probe sets, expressed genes and non-expressed genes were respectively defined as transcripts whose mean detection level across all biological replicates were greater than or less than or equal to the mean of the negative control probe sets.

To determine the validity of our division of transcripts into expressed genes or non-expressed genes, we first tested whether binding of H3K4me3 (activating), H3K27me3 (silencing) and H3K9me2 (silencing) were appropriately enriched in binding near expressed or non-expressed genes (Li et al. 2007). Using PeaksToGenes, we observed that H3K4me3 **(Figure 3.8 – top)** binding was significantly enriched near the TSS of expressed genes (green line), while binding of H3K27me3 **(Figure 3.8 – middle)** and H3K9me2 **(Figure 3.8 – bottom)** was significantly enriched across the entire region. Based on the associations of these chromatin modifications with gene expression (Li et al. 2007), these results suggest that our use of the Affymetrix spike-in controls allowed for reasonably appropriate definitions of expressed and non-expressed genes. There is a caveat with our bisection, which is that there are likely to be some incorrectly-classified genes, however, the robust associations we observed between the positional enrichment histone modifications and expression category suggests that these misclassifications were infrequent.

Figure 3.8 Chromatin marks are appropriately associated with expressed versus non-expressed genes in MDA-MB-231 cells.

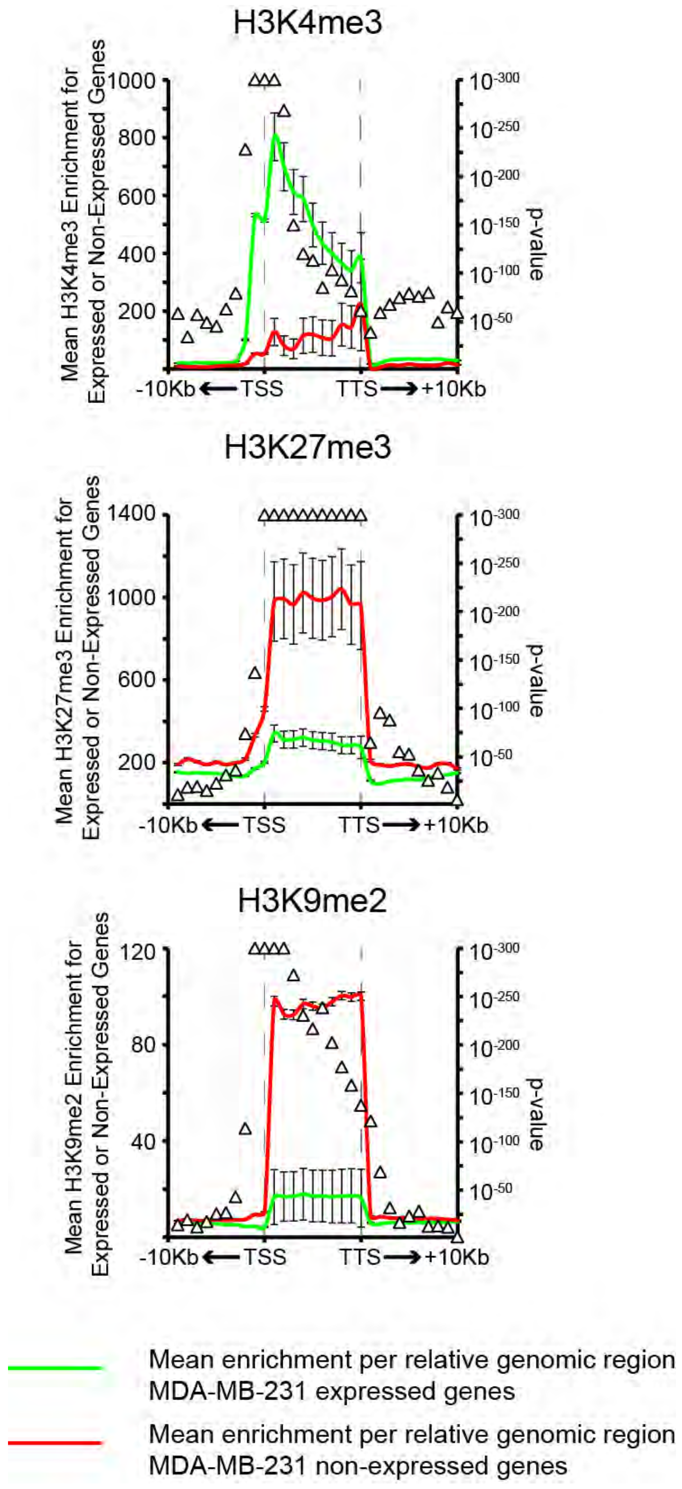


Figure 3.8 Chromatin marks are appropriately associated with expressed versus non-expressed genes in MDA-MB-231 cells.

PeaksToGenes analysis of H3K4me3 enrichment (upper panel), H3K27me3 enrichment (middle panel) and H3K9me2 enrichment (lower panel) for expressed genes (green line) and non-expressed genes (red line). Triangles are Wilcoxon Rank-Sum Test p-values (right y-axis) and error bars are SEM. For more detailed methods, please see Chapter 5.

Extending this PeaksToGenes analysis to the binding of RUNX1 and RUNX2 near expressed (green line) and non-expressed genes (red line), we found that RUNX1 binding is significantly enriched within the promoters (-2Kb of the TSS) of expressed genes and weakly so in regions flanking expressed genes **(Figure 3.9 A – left)**. RUNX2 binding was only weakly associated in regions flanking the gene bodies of expressed genes and not enriched near the promoters **(Figure 3.9 A – right)**. Interestingly, both RUNX1 and RUNX2 binding within the gene body (TSS to TTS) was strongly associated with genes defined as non-expressed **(Figure 3.9 A)**. Due to the length-based normalization done within gene bodies, there is consistently more signal found within gene bodies. Because of this technicality, statistical tests were only done within a single relative genomic region not across relative genomic regions[2]. It is difficult to know how statistically significant each individual binding event is when considering binding ratios. Specifically, a binding ratio of two-fold could be derived from two reads in the IP and one read in the Input, or the same two-fold ration could have been derived from 50 reads in the IP and 25 reads in the Input. It is more likely that the two-fold enrichment measured in the second case is more reproducible. However, instead of setting some kind of threshold for the minimum number of reads within a region to calculate a binding ratio, we relied upon the peak calling algorithm employed by MACS (Zhang et al. 2008). Peak-calling algorithms, such as MACS, measure signal ratios and then using both global and local parameters

[2] Discussed further in Chapter 5.

of variations in read numbers to determine whether a certain region is statistically significantly enriched for reads in the IP sample. To determine whether there were differences in the associations of the most significant regions (MACS peaks) of RUNX1 and RUNX2 with expressed or non-expressed genes, we applied PeaksToGenes analysis. We employed a non-parametric Wilcoxon Rank Sum test, which allows for non-normally distributed data and for unequal sample sizes (Wilcoxon 1945); the binding data for RUNX1 and RUNX2 in MDA-MB-231 cells fit these parameters. Using RUNX1 and RUNX2 peak regions, we observed a non-significant association of RUNX1 peaks with regions near the TSS of expressed genes (blue line) **(Figure 3.9 B – left)** while RUNX2 peaks were not differentially associated with either expressed (blue line) or non-expressed genes (orange line) **(Figure 3.9 B - right)**. While RUNX1 binding appeared to be increased near the TSS of expressed genes (blue line) as compared to non-expressed genes (orange line), the result of the Wilcoxon Rank Sum test was not statistically significant. As we observed in **(Figure 3.6 B)**, RUNX2 peaks were generally not found near the promoters of either expressed or non-expressed genes **(Figure 3.9 B – right)**. Using MACS peaks instead of raw signal ratios as a measure of RUNX protein binding, we observed that the binding of RUNX proteins within the gene bodies (TSS to TTS) was no longer preferentially associated with non-expressed genes (red line in **Figure 3.9 A**, orange line in **Figure 3.9 B**). This suggests that the binding of RUNX proteins within gene bodies of non-expressed genes is weak compared to other regions (therefore not defined as peaks), yet this binding is highly significant when the same relative genomic regions are compared to expressed genes **(Figure 3.9 A)**. What this PeaksToGenes analysis of RUNX2 peaks also reveals is that RUNX2 peaks were very poorly associated with genes in general as compared to RUNX1; while RUNX1 had approximately 20 times more peaks in MDA-MB-231 cells as compared to RUNX2 **(Figure 3.6 A)**, the mean enrichment of RUNX1 and RUNX2 in genic regions (within genes and 10Kb regions flanking gene bodies) was approximately 100-fold different ($3.4x10^{-2}$ peaks per Kb and $3.1x10^{-4}$ peaks per Kb respectively) **(Figure 3.9 B)**. Therefore, although there are differences in the number of highly-enriched peak regions for RUNX1 and RUNX2, the distribution of these peaks was not similar with respect to genic versus intergenic regions. These positional relationships observed for RUNX2 binding are quite different than the results obtained for the binding of endogenous RUNX proteins in osteosarcoma (Van der Deen et al. 2012), leukemia, and osteoblasts (unpublished findings – see Appendix). RUNX1, binding in genic regions appeared to be concentrated near actively expressed genes, which is consistent with previous studies of Runx1 in leukemic and hematopoietic cells (Pencovich et al. 2011, Tijssen et al. 2011, Wang et al. 2011b). Genes responsive to RUNX siRNA were overrepresented in clusters whose transcript levels were generally high, indicating that they are actively expressed genes **(Figure 3.7)**. Therefore, we conclude that RUNX siRNA-responsive genes are likely to be expressed. Through PeaksToGenes analysis, we discovered that RUNX1 is preferentially bound near the TSS and promoter regions of actively-expressed genes, while RUNX2 was not significantly associated with genic regions. Based on these results, we can hypothesize that RUNX1 binding within these regions is likely to be most associated with genes responsive to RUNX siRNA.

Figure 3.9 PeaksToGenes analysis of RUNX1 and RUNX2 binding showed that RUNX1 promoter binding was strongly associated with expressed genes

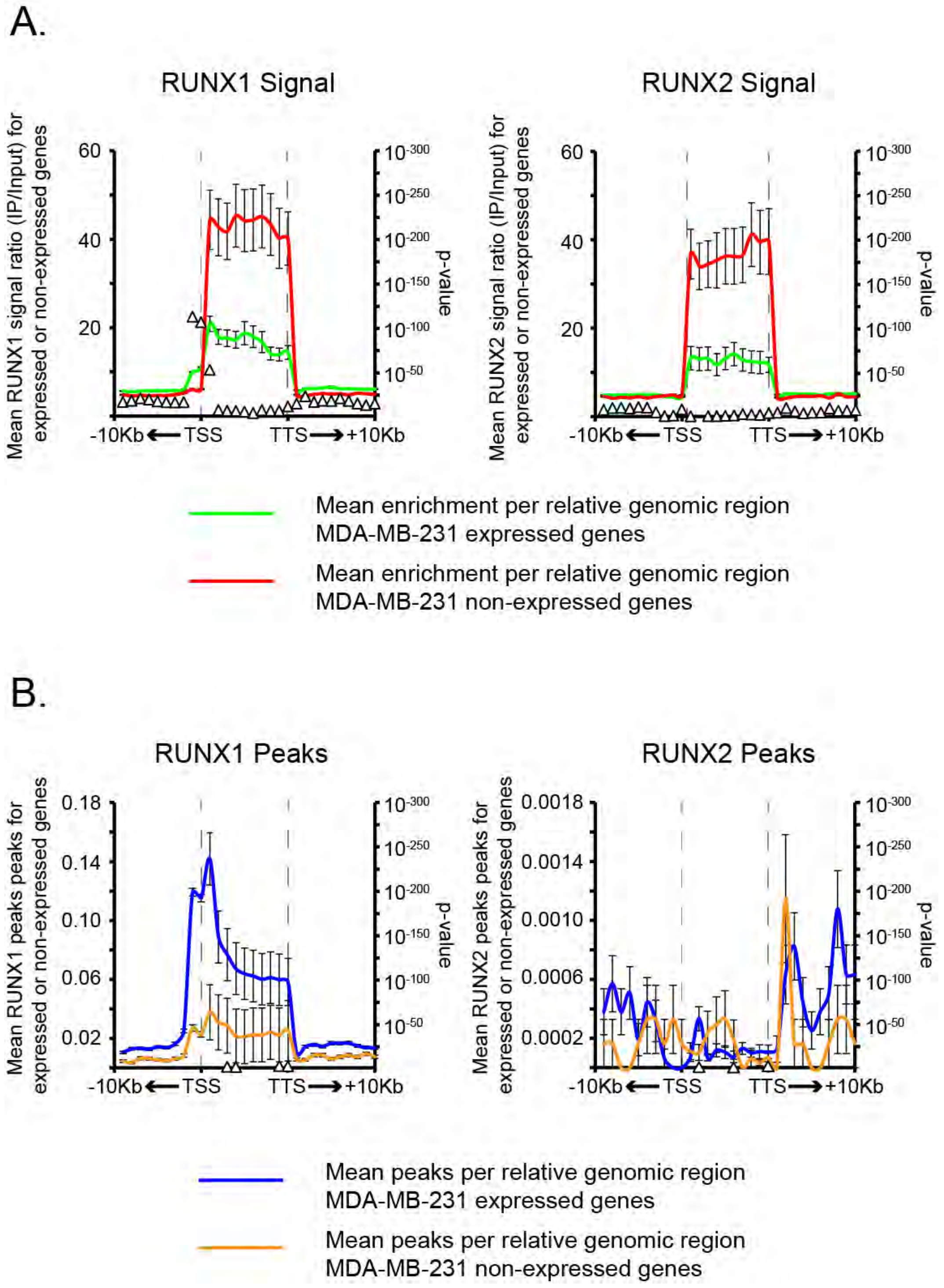


Figure 3.9 PeaksToGenes analysis of RUNX1 and RUNX2 binding showed that RUNX1 promoter binding was strongly associated with expressed genes

(A) RUNX1 signal ratios (IP/Input) (left) and RUNX2 signal ratios (IP/Input) (right) associations with expressed (green) versus non-expressed (red) genes. **(B)** Associations of RUNX1 peaks (left) and RUNX2 peaks (right) with expressed (blue) versus non-expressed (orange) genes. **(A & B)** For each relative genomic region (as shown in **Figure 3.7 A**), the series of values corresponding to the number of peaks or the signal ratios (IP/Input) or number of peaks was taken for expressed or non-expressed genes and a Wilcoxon Rank-Sum Test was performed. Plotted is the mean value of each set (as a line – left y-axis) and the resultant p-value at each relative genomic region (as the triangles – right y-axis). Error bars are SEM. For more detailed methods please see Chapter 5.

Genes responsive to RUNX siRNA are associated with H3K4me3 and RUNX1 binding near the TSS

While we observed that RUNX1 binding is preferentially located within the promoters of expressed genes, we wanted to address whether the binding patterns of RUNX1 or RUNX2 in MDA-MB-231 cells were preferentially associated with genes responsive to RUNX siRNA. If we observe that RUNX1 or RUNX2 is generally bound with certain region relative to genes responsive to RUNX siRNA, we can conclude that this binding of RUNX1 or RUNX2 is likely to be responsible for the functional responses observed after RUNX siRNA transfection. As we had observed that RUNX1 binding is preferentially associated with the promoters and TSS regions of actively-expressed genes **(Figure 3.9)**, and that genes responsive to RUNX siRNA are likely to be highly expressed **(Figure 3.7)**, we therefore used PeaksToGenes to directly test whether RUNX1 binding is more enriched near genes responsive to RUNX siRNA.

Before looking at RUNX1 and RUNX2 binding, we first examined whether binding of H3K4me3, H3K27me3, or H3K9me2 is enriched near genes responsive to RUNX siRNA. We observed that genes responsive to RUNX siRNA were overrepresented in clusters that have strong H3K4me3 signal near the TSS compared to other clusters and weak H3K27me3 and H3K9me2 signal throughout the measured genic region compared to other clusters **(Figure 3.7 A & B)**. Therefore, we used PeaksToGenes to contrast the binding of these chromatin marks near genes that are either responsive (purple line) or non-responsive (brown line) to RUNX siRNA **(Figure 3.10)**. We observed that the activating histone modification H3K4me3 was weakly associated with the TSS regions of RUNX siRNA-responsive genes **(Figure 3.10 – top)**, which indicates that in general, these RUNX siRNA-responsive genes (purple line) have H3K4me3 near the TSS, but not significantly more than other genes (brown line). When analyzing the binding of the repressive histone modifications H3K27me3 **(Figure 3.10 – middle)** and H3K9me2 **(Figure 3.10 – bottom)** we observed that the binding of these marks were generally more associated with RUNX siRNA non-responsive genes (brown line) than with RUNX siRNA responsive genes (purple line). Thus, the RUNX siRNA-responsive genes are characterized by enrichment of H3K4me3 binding near the TSS and a paucity of H3K27me3 or H3K9me2 binding, which suggests that RUNX proteins are regulating the transcript levels of euchromatic genes that are likely to be actively expressed.

Figure 3.10 Genes responsive to RUNX siRNA are not associated with H3K27me3 or H3K9me2, and are associated with H3K4me3 near the TSS

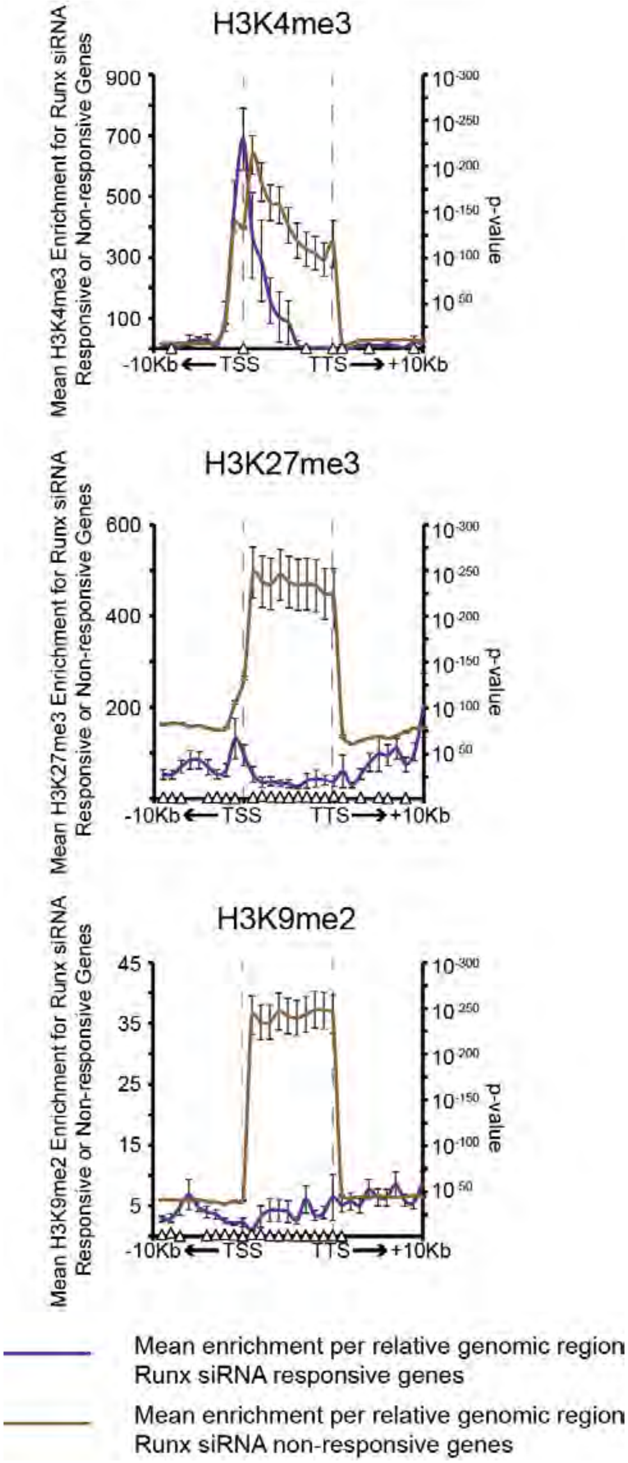


Figure 3.10 Genes responsive to RUNX siRNA are not particularly associated with any of the histone marks examined.

PeaksToGenes analysis of H3K4me3 signal ratios (IP/Input) (upper panel), H3K27me3 signal ratios (IP/Input) (middle panel) and H3K9me2 signal ratios (IP/Input) (lower panel) for genes responsive to RUNX siRNA (purple line) and genes non-responsive to RUNX siRNA (brown line). Triangles are Wilcoxon Rank-Sum Test p-values and error bars are SEM. See Chapter 5 for more detailed methods.

We next examined the relationships between the binding patterns of RUNX1 and RUNX2 in MDA-MB-231 cells and genes responsive to RUNX siRNA. Comparing the enrichment of binding (signal ratio) near RUNX siRNA-responsive genes to RUNX siRNA non-responsive genes, we observed that RUNX1 binding was non-significantly more associated with the upstream and promoter regions of genes responsive to RUNX siRNA (purple line) **(Figure 3.11 A - left)**. RUNX2 binding does not appear to have a preferential enrichment near the RUNX siRNA-responsive genes or the RUNX siRNA non-responsive genes **(Figure 3.11 A - right)**. The same rationale used for the inclusion of MACS peaks for this analysis applies here when comparing the binding of RUNX proteins near RUNX siRNA-responsive genes or RUNX siRNA non-responsive genes. Here, we observed that RUNX1 peaks were non-significantly enriched near the TSS of RUNX siRNA responsive genes (green line) **(Figure 3.11 B – left)**. Unexpectedly, there was not a single RUNX2 peak within the gene body or within 10Kb of any genes responsive to RUNX siRNA (green line) **(Figure 3.11 B – right)**. Neither RUNX2 signal ratio analysis **(Figure 3.11 A – right)** nor RUNX2 MACS peaks analysis **(Figure 3.11 B – right)** provided any indication of how RUNX2 may be functioning to regulate gene expression in MDA-MB-231 cells. We observed that in MDA-MB-231 cells RUNX2 is not binding to the RUNX motif **(Figure 3.6 C)**, has very few peaks **(Figure 3.6 A)** and has both nuclear and cytoplasmic localization **(Figure 3.1 B)**. Transfection of RUNX2 siRNA in MDA-MB-231 cells affected the transcript levels of many genes that are unrelated to bone **(Table 3.1**

& Figure 3.5 C) and were not observed to be changed when RUNX2 is knocked down in an osteoblastic cell such as SaOS-2 (Young et al. 2007b, van der Deen et al. 2012). Given the deregulated sub-cellular localization and genomic binding properties of RUNX2 we cannot determine the mechanisms by which RUNX2 is regulating the transcript levels of RUNX2 siRNA-responsive genes. Considering the preferential association of RUNX1 binding **(Figure 3.7 A & Figure 3.11 A – left & Figure 3.11 B – left)** and of H3K4me3 binding **(Figure 3.10 – top)** near the TSS of RUNX siRNA-responsive genes, leads us to conclude that this binding is functionally important for transcriptional regulation of genes responsive to RUNX siRNA. In our analysis of RUNX1 and RUNX2 binding patterns near genes responsive to RUNX siRNA, we chose to consolidate the genes responsive to RUNX1 siRNA, responsive RUNX2 siRNA and responsive to either into one group. There are several reasons for doing this: 1) Only 66 genes were identified to be responsive to combined RUNX1 and RUNX2 siRNA transfection, which is much less than we had expected given similar RUNX2 studies in SaOS-2 cells (Young et al. 2007b, van der Deen et al. 2012), 2) 66 genes is a small enough list to make statistical tests less robust, separating the genes into smaller groups would make these tests even less informative, and 3) of the 66 genes identified to be responsive to RUNX siRNA, we were only able to experimentally validate 44 of them using separate RUNX1 and RUNX2 siRNAs **(Figure 3.4 & Figure 3.5)**, which would make point number 2 an even larger concern once the genes are separated by response pattern. We observed that RUNX1 binding is preferentially associated with the TSS of RUNX siRNA-responsive genes, and if we extend the trend observed in **Figure 3.4**, approximately 1/2 of these genes would not responsive to RUNX1 siRNA. Therefore, we expect this association to be stronger when only considering RUNX1 siRNA-responsive genes. A technical issue, which can be observed in **Figure 3.5**, is that the response to transfection of both RUNX1 and RUNX2 siRNA is typically less robust than for a single siRNA. This leads us to conclude that the samples used for Affymetrix analysis may be missing a great deal of information regarding genes that are uniquely responsive to RUNX1 or RUNX2. Extrapolating this idea to the binding pattern of RUNX1 in MDA-MB-231 cells, it is possible that many of the binding events proximal to the TSS were actually responsible for the regulation of these genes but were masked in the Affymetrix experiment due to our use of combined siRNAs.

Figure 3.11 Genes responsive to RUNX siRNA are are slightly more associated with RUNX1 binding near the TSS

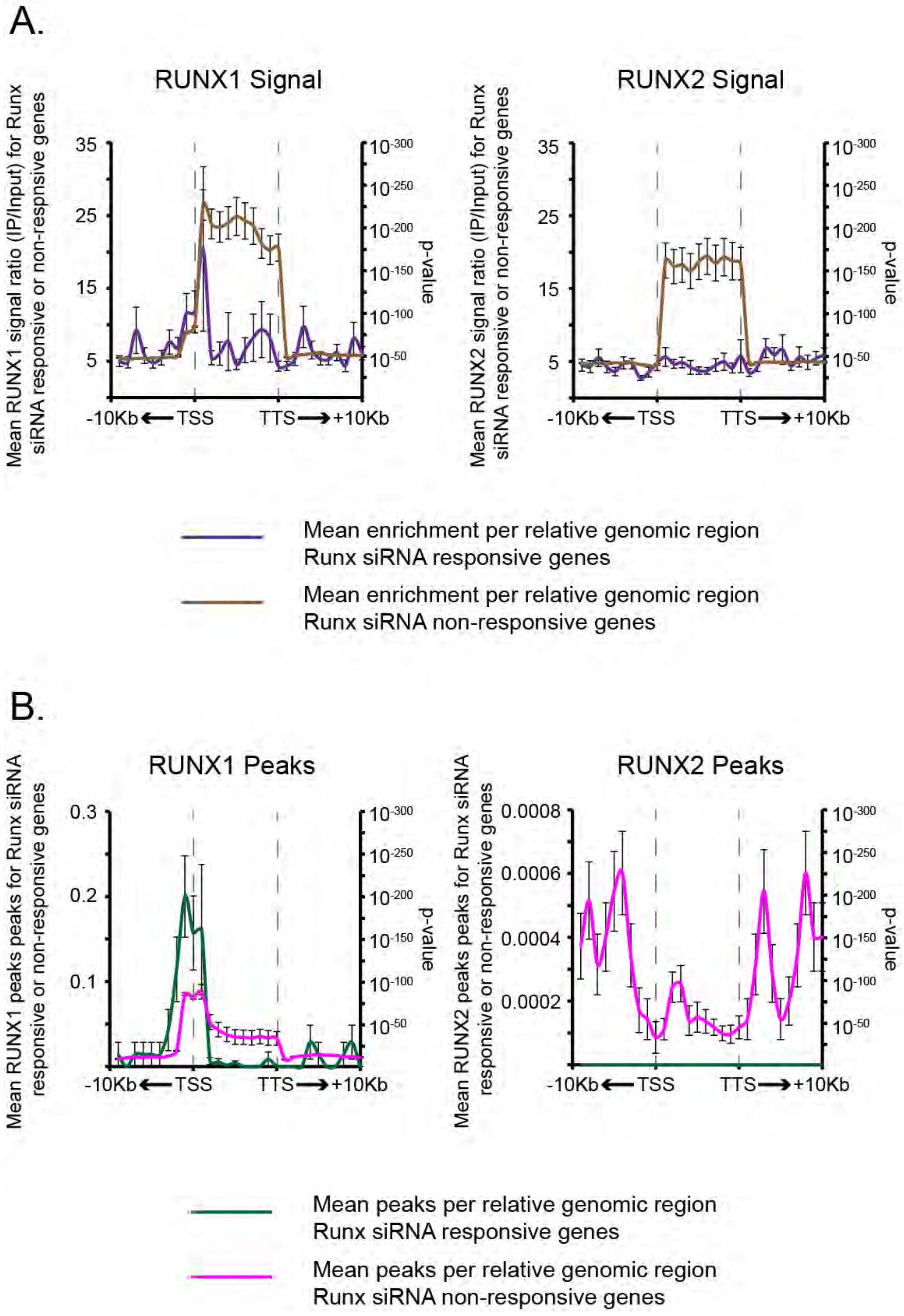


Figure 3.11 Genes responsive to RUNX siRNA are slightly more associated with RUNX1 binding near the TSS.

PeaksToGenes analysis of RUNX1 (left) and RUNX2 (right) **(A)** enrichment (IP/Input) and **(B)** MACS peaks. Purple line: mean signal ratio (IP/Input) for binding near RUNX siRNA responsive genes. Brown line: mean signal ratio (IP/Input) for binding near RUNX siRNA non-responsive genes. Green line: mean peaks per relative genomic region for RUNX siRNA responsive genes. Pink line: mean peaks per relative genomic region for RUNX siRNA non-responsive genes. No p-values were significant (<= 0.05) for Wilcoxon Rank-Sum Test and were not plotted. Error bars are SEM. See Chapter 5 for more detailed methods.

RUNX proteins are expressed in samples isolated from human patients of breast cancer

We have examined the phenotypic contributions and functions of RUNX1 and RUNX2 in MDA-MB-231 cells. These *in vitro* results suggest possible mechanisms or roles for RUNX1 and RUNX2 in human breast cancer patients, but do not allow for insight into the clinical relevance of RUNX1 and RUNX2 expression in breast cancer cells. The MDA-MB-231 cell line is a model of advanced bone, brain and lung metastatic breast cancer (Yoneda et al. 2001), however, it does not allow for insight into whether RUNX proteins are expressed during tumor initiation or progression to the metastatic state. Recent *in vivo* immunohistochemical evidence suggests that RUNX1 protein is detected in the epithelial cells of both normal breast tissue and breast cancer tissue (Uhlén et al. 2005, Pontèn et al. 2008, Uhlen et al. 2010), and that increased RUNX2 expression is associated with the disease state (Das et al. 2009, Onodera et al. 2010). It is therefore important to correlate *in vivo* expression of RUNX1 and RUNX2 with prognostic and diagnostic markers of breast cancer in various stages leading up to and including metastatic disease.

Using formalin-fixed paraffin-embedded (FFPE) human breast cancer samples from the UMMS tissue bank and FFPE human breast cancer tissue microarrays (TMAs) from US BioMax, we stained tissue samples from more than 125 patients with antibodies for RUNX1 or RUNX2. The TMAs from US BioMax were BR1503a and BR10010, which are breast cancer progression and breast cancer metastasis to lymph node arrays, respectively. BR10010 is a matched metastatic breast cancer array of samples taken from 25 patients; for each patient BR10010 contains two samples each of tissue from primary breast tissue and lymph tissue to which the primary breast tumor has metastasized. Primary tumor samples and metastatic tumor samples analyzed had not metastasized to or been taken from distal tissues such as brain, lung, liver or bone. Typically, once patients present with distal metastases, biopsies or fine needle aspirates (FNAs) are no longer taken for analysis, so it is difficult to obtain samples from advanced metastatic disease.

Using a semi-quantitative scoring system, two researchers blindly scored TMAs stained with RUNX1 and RUNX2 to determine whether RUNX protein expression was associated with any prognostic or diagnostic markers of breast cancer progression **(Figure 3.12 A)**. Strong expression of RUNX1 was observed in almost all mammary epithelial cells of any pathology, while RUNX2 positive tissues were less frequent **(Figure 3.12 B-D)**. Using this scoring system, we examined the extent to which the expression levels of RUNX proteins are significantly more or less associated with the expression of: HER2, PR, ER, and AR; as well as the cytopathologically-defined: TNM (tumor size, node status, and distal metastases status), breast cancer grade, breast cancer stage, pathology, and tissue type. We did not observe RUNX1 expression to be preferentially associated with any particular subtype of breast cancer **(Figure 3.12 C – left)**, while RUNX2 expression was primarily observed in invasive ductal carcinoma compared to normal adjacent tissue, fibroadenoma, cystosarcoma phyllodes, ductal carcinoma *in situ*, and metastatic breast tissue to the lymph nodes **(Figure 3.12 C - right)**. The column graphs presented are the mean RUNX protein detection score associated with the subtype. While these graphs may give the impression that few or no samples were judged to be greater than '++' for RUNX2, the low mean score is actually due to the high number of samples for which RUNX2 was not detected. Interestingly, while RUNX proteins appear to play a role in the invasiveness of MDA-MB-231 cells, we observe that breast cancer cells that have metastasized to the lymph node have significantly less RUNX1 **(Figure 3.10 D – left)** and RUNX2 **(Figure 3.10 – right)** expression than in the primary tumor site. This comparison is derived from the BR10010 array in which primary and metastatic patient tissues were matched and from the BR1502a in which metastatic and non-metastatic tissues primary tissues were compared. This observation is similar to a previous study in which a significant reduction in RUNX1 transcript levels in metastatic breast tissue was observed compared to non-metastatic breast cancer tissue (Dairkee et al. 2004), and suggests that the invasive phenotype *in vitro* associated with RUNX1 and RUNX2 is unlikely to have participated metastasis to primary lymph nodes.

Figure 3.12 RUNX1 is expressed in normal breast and multiple breast cancer subtypes, while RUNX2 is primarily expressed in invasive ductal carcinoma. Both RUNX1 and RUNX2 expression is significantly lower in breast cancer cells that have metastasized to lymph tissue.

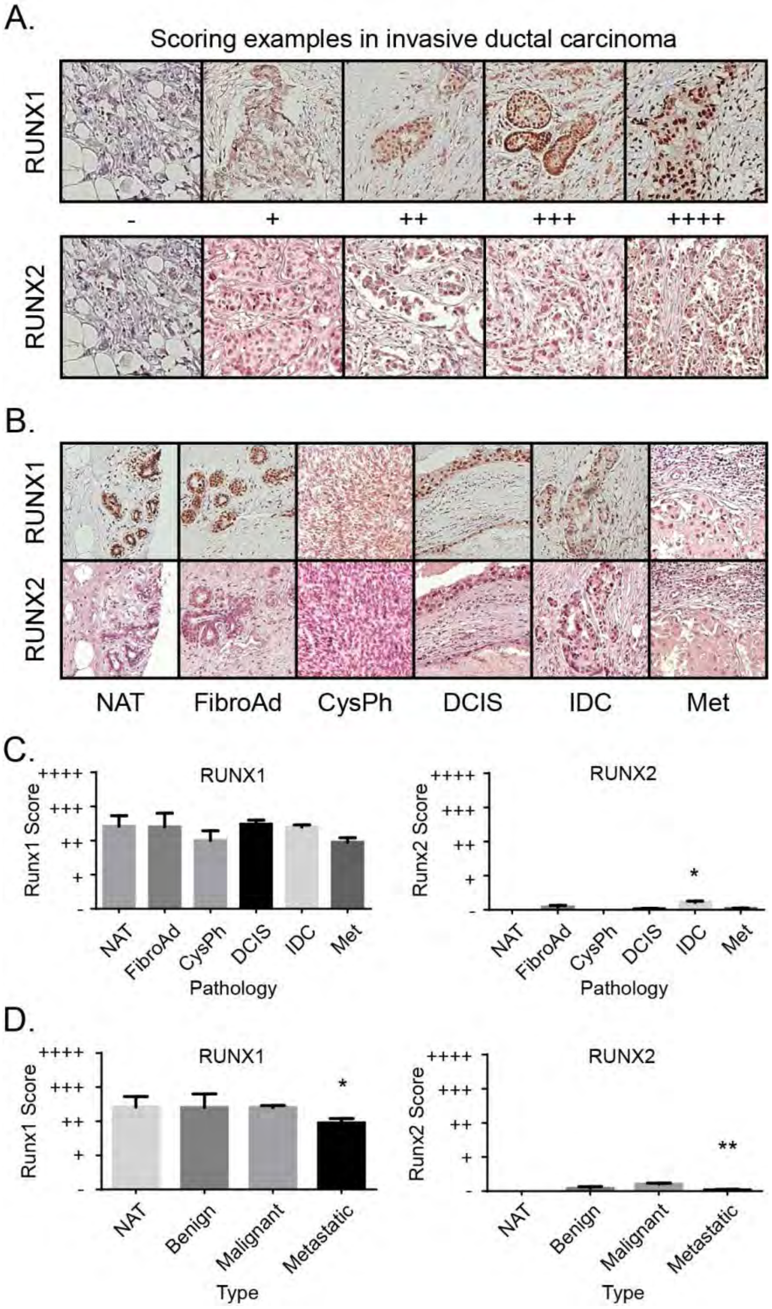


Figure 3.12 RUNX1 is expressed in normal breast and multiple breast cancer subtypes, while RUNX2 is primarily expressed in invasive ductal carcinoma. Both RUNX1 and RUNX2 expression is significantly lower in breast cancer cells that have metastasized to lymph tissue.

(A) Representative images of scoring rubric taken from invasive ductal carcinoma samples for RUNX1 (top row) and RUNX2 (bottom row). **(B)** Representative images of RUNX1 (upper panel) and RUNX2 (lower panel) in NAT (normal adjacent tissue), FibroAd (fibroadenoma), CysPh (cystosarcoma phyllodes), DCIS (ductal carcinoma *in situ*), IDC (invasive ductal carcinoma), and Met (breast tissue metastasized to lymph). **(C)** Mean and SEM (error bars) score given to RUNX1 (left) and RUNX2 (right) staining for each pathology type. * = Kruskal-Wallis p-value < 0.05 comparing RUNX2 scores in IDC to Met (right). **(D)** Mean and SEM (error bars) score given to RUNX1 (left) and RUNX2 (right) staining for tissue type. ** = Kruskal-Wallis p-value < 0.01 comparing RUNX1 staining in Malignant to Metastatic (left). ** = Kruskal-Wallis p-value < 0.01 comparing RUNX2 staining in Malignant to Metastatic (right).

RUNX1 is primarily expressed in normal tissue and in early smaller tumors, while RUNX2 is primarily expressed in middle to late-stage larger tumors

Next we examined whether the expression intensity of RUNX1 or RUNX2 in the malignant cells was associated with diagnostic markers of breast cancer (tumor node metastasis (TNM), grade, and stage). This analysis will allow us to understand whether RUNX proteins have a temporal relationship with disease progression.

TNM is a scoring metric used to define the size of the primary tumor as well as the extent to which the tumor has metastasized to proximal (node status) and distal sites[3]. T is for the size of the tumor, and has the following designations: 1) Tx – tumor was not able to be classified, 2) Tis – ductal carcinoma *in situ*, 3) T0 – no tumor, 4) T1 – can be broken down into the following four categories: i) T1mi – tumor diameter is less than 0.1cm, ii) T1a – tumor diameter is greater than 0.1cm but less than 0.5cm, iii) T1b – tumor diameter is greater than 0.5cm but less than 1.0cm, and iv) T1c – tumor diameter is greater than 1cm but less than 2cm, 5) T2 – tumor diameter is greater than 2cm but less than 5cm, 6) T3 – the tumor diameter is greater than 5cm, and 7) T4 – can be broken down into four categories: i) T4a – tumor has invaded into chest wall, ii) T4b – tumor has

[3] http://www.cancerresearchuk.org/cancer-help/type/breast-cancer/treatment/tnm-breast-cancer-staging - last accessed March 31st, 2013

invaded into skin and patient presents with local swelling, iii) T4c – both T4a and T4b, and iv) T4d – tumor is classified as inflammatory carcinoma, local area is red, swollen and painful to the touch. N is for the characterization of the node stages, and the relevant categorization is as follows: 1) NX – samples cannot be assessed, 2) N0 – nodes are negative for metastases, 3) N1 – positive metastases in nodes of upper armpit and nodes are not stuck to surrounding tissues, 4) N2 – positive metastases in nodes of upper armpit, which are stuck to each other and surrounding tissues or positive metastases in internal mammary nodes. There are further designations for node status, but none of these were present in the US BioMax arrays or samples from UMMS analyzed. Similarly, the M designation is for distal metastases and, as discussed above, because we did not analyze any samples with distal metastases the classifications of this parameter will not be discussed.

Breast cancer staging is a method of grouping the TNM classifications. These groupings have been observed to have similar clinical outcomes, and are treated likewise[4]. There are more stages defined than were present in the arrays analyzed from US BioMax. Only the stages represented on these arrays and subsequently immunostained for RUNX proteins will be discussed here. Stage 0 is cancer *in situ* or TisN0M0. Stage I is small tumors (less than 2cm in diameter).

[4] http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-staging - Last accessed March 31st, 2013.

Stage I is bisected into Ia and Ib, however, the staging data from US BioMax did not have this distinction. Stage Ia tumors are non-metastatic (T1N0M0), while stage Ib tumors have micrometastases to the axillary lymph nodes (T1N1miM0). The US BioMax data for stage two is defined as Stage II, Stage IIa, and Stage IIb. While there is not a clear definition for what Stage II is, stage IIb and Stage IIb are well-defined. It is likely that tumors classified as Stage II only had characteristics of both IIa and IIb such that a distinction could not be made. Stage IIa is small tumors that have metastasized to the lymph nodes (T1N1M0), but not to distal organs. Stage IIb is larger tumors that may or may not have metastasized to the lymph nodes (T2N1M0 or T3N0M0), and the tumor has not metastasized to distal organs. Stage IIIa is tumors no larger than 5cm in diameter that have metastasized to many axillary lymph nodes or internal mammary lymph nodes (T0-2N2M0), or tumors that are large (greater than 5cm in diameter) that have not invaded the skin or chest wall and have spread to multiple lymph nodes (T3N1-2M0). Stage IIIa tumors have not metastasized to distal organs. Stage IIIb tumors have invaded the chest wall or skin and may or may not have metastasized to the lymph (T4N0-2M0), but have not metastasized to distal organs. No samples of Stage IIIc or Stage IV were present in this study.

Tumor grade is a measure of how abnormal the cells and tissue structure are in a patient sample[5]. These grades correlate with how likely a tumor is to proliferate and affect diagnosis and prognosis. The scoring system employed by the US BioMax was not breast cancer-specific; rather US BioMax used the general scoring system that can be applied to all tumors. Grade 1 is low grade and well differentiated. Grade 2 is intermediate grade and moderately differentiated. Grade 3 is high grade and poorly differentiated.

We observed RUNX1 expression to be significantly associated with small, non-metastatic tumors (T1N0M0) **(Figure 3.13 A – left)** and RUNX2 expression was significantly associated with large, non-metastatic tumors that have grown into the chest wall or skin, but have not metastasized (T4N0M0) **(Figure 3.13 A - right)**. Again, we observed that RUNX1 was significantly associated with grade 1 (low grade), well differentiated breast cancer **(Figure 3.13 B – left)**, while RUNX2 was associated with grade 2 (intermediate grade), moderately differentiated breast cancer **(Figure 3.13 B - right)**. Contrasting RUNX expression against breast cancer stage, we observed that RUNX1 expression was most associated with stage I breast cancer **(Figure 3.13 C – left)**, while RUNX2 expression was observed to be most associated with stage IIIb breast cancer **(Figure 3.13 C - right)**.

[5] http://www.cancer.gov/cancertopics/factsheet/detection/tumor-grade - Last accessed March 31st, 2013

Taken together, these associations of RUNX1 and RUNX2 staining intensity with early and late disease states, respectively, we hypothesize that RUNX proteins are likely participating in a defined event of breast cancer progression within the primary tumor. Unfortunately, our *in vitro* cell model is a very late-stage model, so we do not know what processes may be regulated as a consequence of RUNX expression at a these earlier time points. It is quite interesting that the expression of RUNX2, which has been shown to promote the invasive and osteolytic properties of MDA-MB-231 cells (Barnes et al. 2004, Javed et al. 2005, Pratap et al. 2008), is so specifically associated with non-metastatic disease (T4N0M0) **(Figure 3.13 A – right)**. We can, however, make some correlation between the invasive phenotypes observed in association with RUNX2 *in vitro* with the observation that RUNX2 expression was significantly associated with breast tumors that have invaded into the chest wall or skin. It is, however, important to note that most metastatic tumors examined had not invaded into the chest wall or skin, so it appears that RUNX2 is not playing a role in breast cancer metastasis. Previously, *RUNX1* was shown to have a statistically significant rate of mutations in breast cancer patients (The Cancer Genome Atlas Network 2012), which suggests that RUNX1 may function as a tumor suppressor. Here, we observed that the expression of RUNX1 was primarily associated with early, malignant tumors (T1N0M0) **(Figure 3.13 A – left)** that are low grade and well differentiated (Grade 1) **(Figure 3.13 B – left)** as compared to normal or advanced disease tissue. It is unknown whether *RUNX1* is mutant in these early tumors we have examined, so we cannot determine the extent to which the RUNX1 protein detected is functioning normally. Therefore, it is difficult to speculate what role RUNX1 may be playing in breast disease progression, and further highlights the need to investigate the properties of RUNX1 in breast cancer (Janes 2011).

Figure 3.13 RUNX1 is primarily associated with early, smaller tumors while RUNX2 is most associated with late, larger tumors.

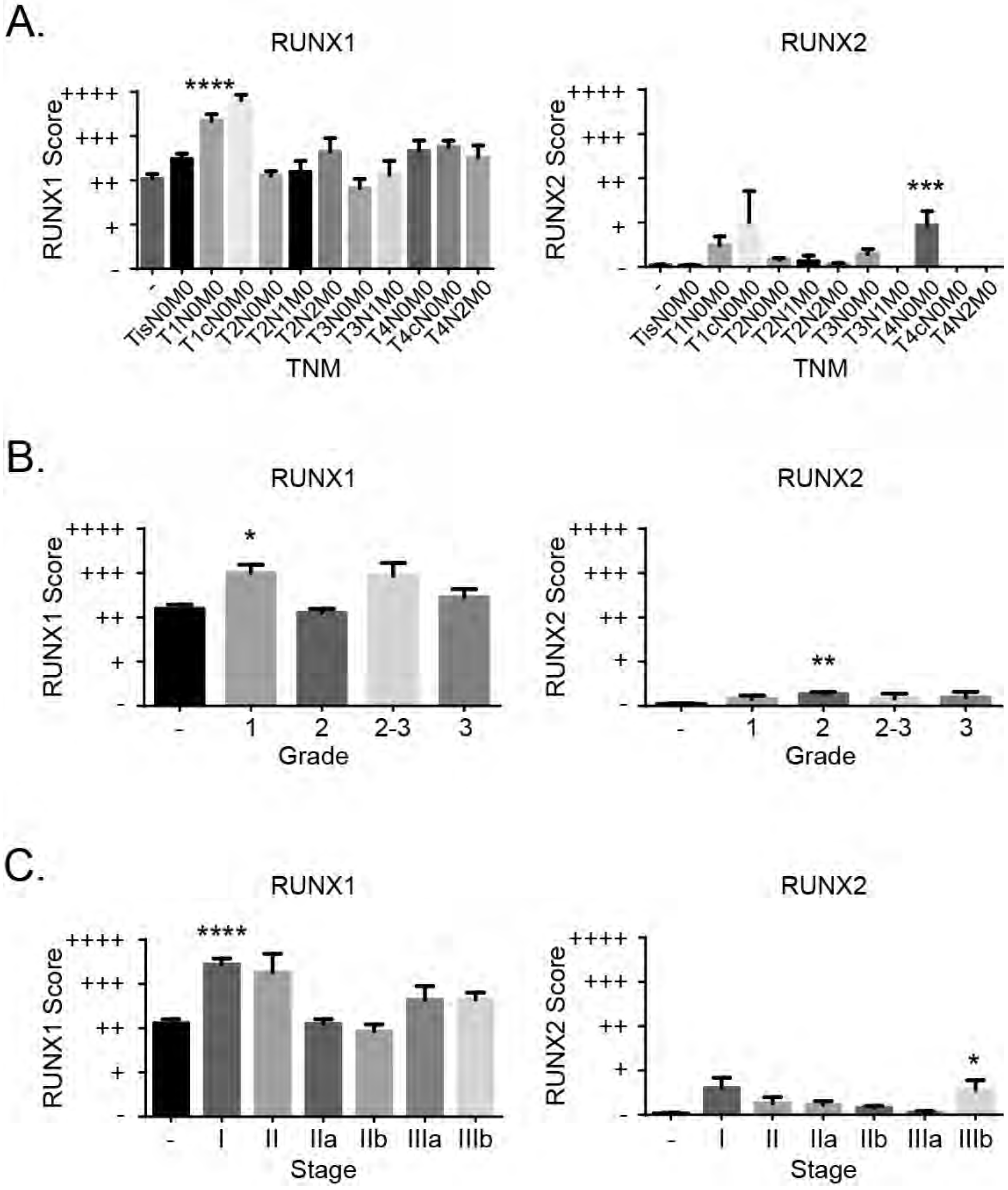


Figure 3.13 RUNX1 is primarily associated with early, smaller tumors while RUNX2 is most associated with late, larger tumors.

(A) Mean and SEM (error bars) RUNX1 (left) and RUNX2 (right) scores associated with TNM (tumor node metastases) stages. **** = Kruskal-Wallis p-value < 0.0001 comparing RUNX1 staining in T1N0M0 to – (normal) (left). *** = Kruskal-Wallis p-value < 0.0001 comparing RUNX2 staining in T4N0M0 to – (normal) (right). **(B)** Mean and SEM (error bars) RUNX1 (left) and RUNX2 (right) scores associated with breast cancer grades. * = Kruskal-Wallis p-value < 0.05 comparing RUNX1 staining in Grade 1 versus Grade 2 (left). ** = Kruskal-Wallis p-value < 0.01 comparing RUNX2 staining in Grade 2 versus – (normal) (right). **(C)** Mean and SEM (error bars) RUNX1 (left) and RUNX2 (right) scores associated with breast stage. **** = Kruskal-Wallis p-value < 0.05 comparing RUNX1 staining in Stage I versus – (normal) / Stage IIb / IIa (left). * = Kruskal-Wallis p-value < 0.05 comparing RUNX2 staining in Stage IIIb versus – (normal) (right).

RUNX1 expression is correlated with AR expression and has a strong association with ER+ breast cancer, which is dependent on AR status

As RUNX2 has been previously reported to be primarily expressed in ER- Asian breast cancer patients (Das et al. 2009), and RUNX1 has been reported to function as a tethering factor for ER-alpha in vitro (Stender et al. 2010), it is of particular interest to define the associations between RUNX protein expression and ER in North American breast cancer patients. We observed that RUNX1 expression was significantly associated with ER+ breast cancer **(Figure 3.14 A – left)**, while RUNX2 expression was not dependent on ER status **(Figure 3.14 A - right)**. This observation is particularly interesting because RUNX2 expression has previously been shown to be inversely correlated with ER *in vitro* and in Asian breast cancer patients (Das et al. 2009, Chimge et al. 2011), which is contradictory to what we have observed in North American patients. A further contradictory result is the observation that RUNX2 was an independent predictor of breast disease, progression and metastasis in Asian patients (Onodera et al. 2010), while in North American patients we observed RUNX2 is not an independent prognostic factor and is primarily expressed in large, invasive, non-metastatic (T4N0M0) ductal carcinoma **(Figure 3.12 B & Figure 3.13 A – right)**. Breast cancer incidence and death is much higher in North America as compared to Asia, and it is hypothesized that environmental factors (such as diet and lifestyle) rather than genetics contribute to these observed differences in breast cancer epidemiology (Jemal et al. 2010). It is unknown which factor or combination of factors may be driving the differences observed in the expression patterns of RUNX2 between North American and Asian patients. Further, there are no studies that provide a strong link between RUNX2 expression levels and environmental factors.

Extending our analysis to HER2, we found that the expression of both RUNX1 and RUNX2 were associated with HER2+ breast cancer **(Figure 3.14 B)**, an observation which, in the case of RUNX1, was dependent on ER status **(Figure 3.14 C)**. Neither RUNX1 nor RUNX2 expression were observed to be significantly associated with PR status **(Figure 3.14 D)**, however, RUNX1 expression was significantly associated with ER+/PR- breast cancer **(Figure 3.14 E)**. We further observed that RUNX2 expression was not altered by combined ER/PR status (data not shown).

In breast cancer, androgen receptor (AR) is a prognostic indicator, known to inhibit the activity of ER-alpha (Hickey et al. 2012). We observed that RUNX1 expression levels were correlated with AR levels **(Figure 3.14 F)**, and the association of RUNX1 expression with ER+ breast cancer was highly dependent on AR status **(Figure 3.14 G)**. RUNX2 expression was not observed to be affected by AR status or AR/ER combined status (data not shown). This observation represents a potentially novel subtype of breast cancer, which is

RUNX1, ER, and AR positive. Based on our observations **(Figure 3.13 A – left)**, this subtype of cancer is likely to be a low grade, small tumor.

A common means of describing breast cancers in both prognostic and diagnostic terms is to look at the combined statuses of HER2, ER and PR (Allred et al. 1998). We therefore investigated the extent to which combining these growth and hormone receptor statuses affected the observed expression levels of RUNX proteins. We observed no statistically significant associations between these triple statuses with RUNX1 **(Figure 3.14 H – left)** or RUNX2 **(Figure 3.14 H - right)**. However, the previously observed trend for RUNX2 association with HER+ breast cancer **(Figure 3.14 B – right)** was observed when looking at all three receptors combined **(Figure 3.14 H – right)**.

These results demonstrate that the expression of RUNX proteins, while not correlated with disease progression, are sensitive to growth and hormone receptor expression patterns. Given the observed functional differences for RUNX proteins between osteosarcoma and breast cancer cells, we hypothesize that RUNX proteins in the presence or absence of these hormone receptors may have quite distinct functional roles. A key finding in our comparison of RUNX1 and RUNX2 expression levels with growth/hormone receptor status was that triple-negative (HER2-/PR-/ER-) breast cancer was associated with the lowest RUNX1 levels **(Figure 3.14 H – left)** and nearly the lowest RUNX2 levels **(Figure**

3.14 – right). Triple-negative breast cancer is clinically associated with poor patient outcome (Dent et al. 2007), and the observation that the expression of RUNX proteins in human breast cancer was more associated with at least one growth/hormone receptor **(Figure 3.14)** leads us to hypothesize that RUNX-positive breast cancer is associated with positive patient outcome. The presence of HER2, ER, or PR is used to determine treatment regimens targeted towards these receptors and generally patients have higher survival rates (Dent et al. 2007). Unfortunately, our hypothesis was not testable with these samples as we did not have access to patient outcome data such as survival, disease-free survival or metastatic events.

Figure 3.14 RUNX1 expression correlates with AR expression in breast cancer. RUNX1 is primarily associated with HER2+/PR-/ER+ breast cancers, and RUNX1 association with ER+ breast cancer is dependent on HER2+ and AR+ status. RUNX2 expression is primarily associated with HER+ breast cancers.

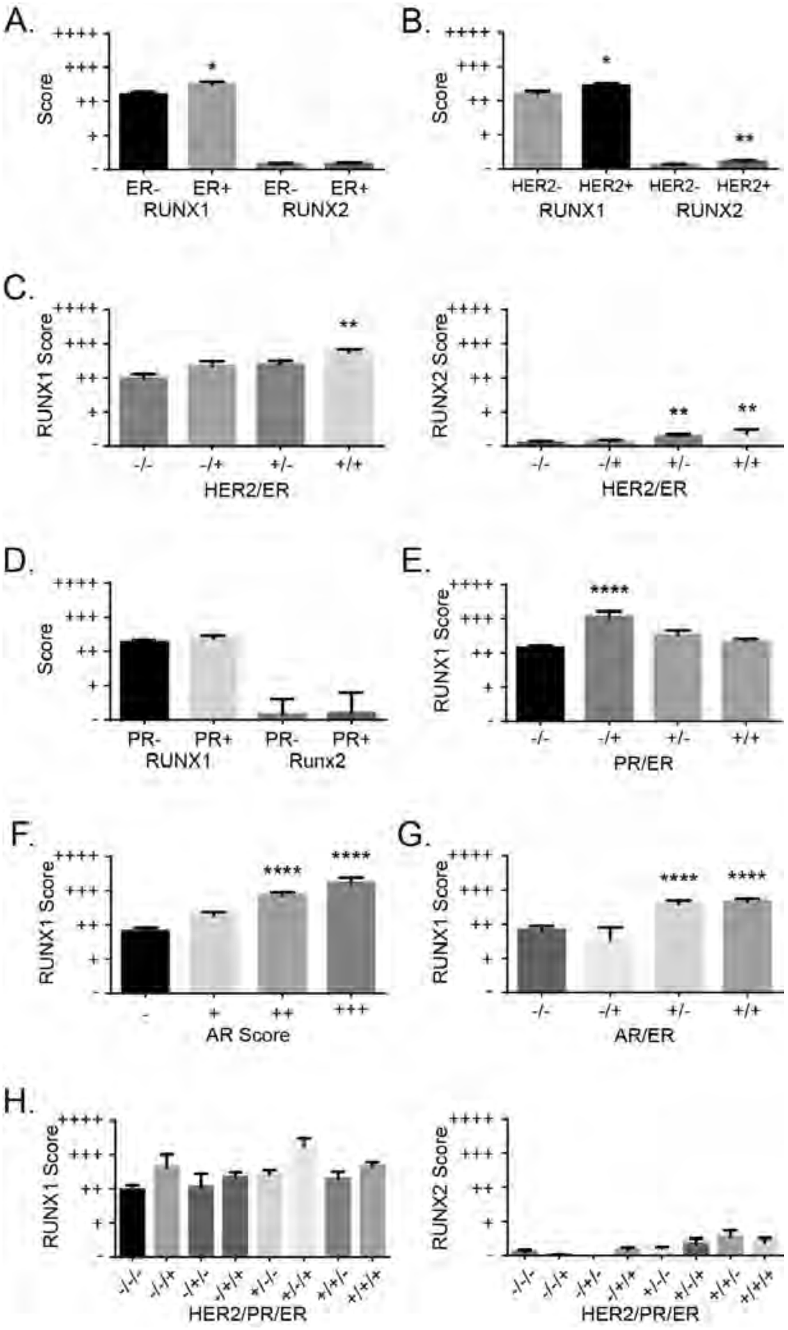


Figure 3.14 RUNX1 expression correlates with AR expression in breast cancer. RUNX1 is primarily associated with HER2+/PR-/ER+ breast cancers, and RUNX1 association with ER+ breast cancer is dependent on HER2+ and AR+ status. RUNX2 expression is primarily associated with HER+ breast cancers.

(A) Mean and SEM (error bars) scores for RUNX1 (left two columns) and RUNX2 (right two columns) in ER+ and ER- breast cancers. * = Kruskal-Wallis p-value < 0.05 comparing RUNX1 scores in ER+ versus ER-. **(B)** Mean and SEM (error bars) scores for RUNX1 (left two columns) and RUNX2 (right two columns) in HER2+ and HER2- breast cancers. * = Kruskal-Wallis p-value < 0.05 comparing RUNX1 scores in HER2+ versus HER2-. ** = Kruskal-Wallis p-value < 0.01 comparing RUNX2 scores in HER2+ versus HER2-. **(C)** Mean and SEM (error bars) scores for RUNX1 (left panel) and RUNX2 (right panel) combining HER2 and ER statuses in breast cancer samples. ** = Kruskal-Wallis p-value < 0.01 comparing RUNX1 scores in HER2+/ER+ versus HER2-/ER-. ** = Kruskal-Wallis p-value < 0.01 comparing RUNX2 scores in HER2+/ER- or HER2+/ER+ versus HER2-/ER-. **(D)** Mean and SEM (error bars) scores for RUNX1 (left two columns) and RUNX2 (right two columns) in PR+ and PR2- breast cancers. **(E)** Mean and SEM (error bars) scores for RUNX1 combining PR and ER statuses in breast cancer samples. **** = Kruskal-Wallis p-value < 0.0001 comparing RUNX1 scores in PR-/ER+ versus PR-/ER-. **(F)** Mean and SEM (error bars) of RUNX1 score associated with AR score (x-axis). **** = Kruskal-Wallis p-value < 0.0001 comparing RUNX1 scores in AR- to AR++ or AR+++ **(F)**. Mean and SEM (error bars) scores for RUNX1 combining AR and ER statuses in breast cancer samples. **** = Kruskal-Wallis p-value < 0.0001 comparing RUNX1 scores in AR+/ER- or AR+/ER+ versus AR-/ER- **(G)**. Mean and SEM (error bars) scores for RUNX1 (left panel) and RUNX2 (right panel) in all permutations of HER2, PR and ER. RUNX1 (left panel) non-parametric (Kruskal-Wallis) ANOVA p-value < 0.05. RUNX2 (right panel) non-parametric (Kruskal-Wallis) ANOVA p-value < 0.01 **(H)**.

Discussion

Here we present an unbiased analysis of the functions of RUNX1 and RUNX2 in MDA-MB-231 cells. We demonstrate that RUNX proteins did not affect the rate of protein synthesis MDA-MB-231 cells through the transcriptional regulation of ribosomal RNA as they did in SaOS-2 cells. Using transcriptome-wide profiling of gene expression following knockdown of RUNX proteins, we found that the majority of genes responsive to RUNX siRNA were uniquely responsive to either RUNX1 or RUNX2. This observation is particularly interesting as this is the first analysis of the extent to which endogenous RUNX1 and RUNX2 functionally overlap on a genome-wide scale. This study is also the first time the genome-wide binding profile of RUNX proteins has been characterized using endogenous protein in a breast cancer cell line. Similar to the observations made for transcriptome functions, RUNX binding positions had very little overlap. Given that RUNX proteins share a highly-conserved Runt-homology domain responsible for DNA-binding, this is a very interesting result. Because this is an *in vivo* result and not an EMSA or other *in vitro* approach, this suggests that RUNX proteins are sensitive to more than just local genomic sequence when interacting with chromatinized DNA. As discussed earlier, previous studies have demonstrated that RUNX proteins can be recruited to genomic regions through protein-protein interactions with co-factors that are independent of a RUNX-binding site (Gutierrez et al. 2002). Our results suggest that the genomic locations of RUNX proteins are sensitive to more than just the local sequence motifs. The binding of other transcription factors has similarly been shown to be sensitive to co-factors and chromatin states (Wang et al. 2012), so our observations we not unexpected. Given that RUNX proteins appear to be sensitive to context, we propose that understanding the contextual factors that affect the binding and regulatory roles of RUNX proteins are equally as important as studying the temporal expression patterns of RUNX proteins in breast cancer progression and metastasis.

We also observed that while the majority of RUNX1 binding loci are not in the “promoter” of genes, those near the TSS were highly likely to be functional binding sites as these locations were enriched for H3K4me3 and near genes that were responsive to RUNX siRNA. We observed strong RUNX “foci” when staining nuclei for RUNX proteins, which leads us to believe that RUNX proteins are in close three-dimensional proximity with one another in intact nuclei. A limitation of this current study is the ability to understand how these non-promoter intergenic RUNX binding loci are related to one another in three-dimensional context, especially given evidence suggesting that RUNX1 mediates chromatin looping to activate gene expression (Jiang and Peterlin 2008, Levantini et al. 2011).

This study is also the first study to analyze the expression of RUNX1 in human breast cancer patients via immunohistochemistry and the first study to examine RUNX2 expression in North American patients. Our results demonstrate that in North American patients RUNX proteins were not independent prognostic factors of disease. The observation that RUNX1 and RUNX2 were associated with early and late disease states, respectively, and regulated a novel cohort of genes *in vitro* suggests that RUNX proteins may play important temporal roles during tumor progression. This is an important point given the many observations of the context-specific roles for RUNX proteins (Cameron and Neil 2004). While RUNX1 may function as a tumor-suppressor protein in normal mammary epithelial cells (Kadota et al. 2010, Wang et al. 2011a), we observed RUNX1 functions to promote the invasive/metastatic potential of malignant metastatic MDA-MB-231 cells. Extending these *in vitro* observations to our observations in human patients suggest that it will be quite interesting to investigate the functions of RUNX proteins when expressed in breast cancer cells of varying hormone receptor statuses. However, as discussed before, *RUNX1* is frequently mutated in human breast cancer (The Cancer Genome Atlas Network 2012), and we have observed that RUNX1 expression is low in late disease and significantly lower in metastatic diease. These previously observations, combined with our study suggest that RUNX1 may play an important tumor suppressor role in breast cancer. As cancers progress, traits that give the cells a growth advantage for the cancer cells are selected. Therefore, the observations that RUNX1 expression is reduced in late-stage and metastatic disease, combined with the high rate of mutations observed in the *RUNX1* gene suggests that it is advantageous for the cell to have either less RUNX1 or a mutant form of RUNX1.

These results show a strong correlation between RUNX1 expression and androgen receptor (AR) expression. This is especially interesting as AR expression is an emergent marker associated with disease progression and treatment (Hickey et al. 2012). This study further shows that RUNX2 expression is not associated with estrogen receptor (ER) status, while one study in Asian patients found that RUNX2 expression is inversely correlated with ER expression (Das et al. 2009) and another study in Asian patients defined RUNX2 as an independent prognostic marker of breast cancer (Onodera et al. 2010). The differences observed between RUNX2 clinical associations in North American patients and Asian patients reinforce the idea that the expression and functions of RUNX proteins are highly sensitive to context. A recent study in North American breast cancer patients by The Cancer Genome Atlas Network showed that neither RUNX1 nor RUNX2 expression is significantly associated with disease progression (The Cancer Genome Atlas Network 2012). This study also found that in ER+ breast tumors, RUNX1 has a significant frequency of mutations that would cause early translational termination. The endogenous RUNX1 in the MDA-MB-231 is the correct length, so our *in vitro* studies did not address the functions of these truncated RUNX1 proteins; however, previous work has shown that shortened RUNX transcripts lack the ability to interact with the nuclear matrix and therefore cannot function correctly (Zeng et al. 1997). This may be a potential mechanism by which the tumor suppressor functions observed for RUNX in normal mammary epithelial cells can be bypassed (Kadota et al. 2010, Wang et al. 2011a).

There are several technical limitations of this study that are important to discuss. First, our RNAi study was not as well-controlled as it could be. We did not use multiple siRNAs targeting RUNX mRNAs to demonstrate that the majority of the genes responsive to the RUNX siRNA transfections were specific and not off-target effects. We also did not perform a compensation experiment in which we express transgenic RUNX protein following an siRNA-mediated knockdown of the endogenous protein to rescue the transcriptional effect of the knockdown. With these caveats, we recognize that an unknown number of genes identified by Affymetrix could be the result of off-target effects. This is hopefully not the case as the number of genes (66) was quite small as compared to the several hundred genes identified in knockdowns done in SaOS-2 cells (Young et al. 2007b, van der Deen et al. 2012) or by overexpression of Runx2 in MCF-7 cells (Chimge et al. 2011). Another important consideration to make is our use of the MDA-MB-231 cells as a model cell line for RUNX protein functions in breast cancer. MDA-

MB-231 cells are isolated via pleural effusion from a patient who had a recurrent disease following mastectomy and treatment with 5-fluorouracil followed by combined treatment with Adriamycin, Cytoxan, and methotrexate (Brinkley et al. 1980). The patient was again treated with methotrexate 5 days prior to isolation of MDA-MB-231 cells, and the patient died a few months after isolation of the cells. Therefore, these MDA-MB-231 cells represent an extremely aggressive form of breast cancer. We observed much weaker expression of RUNX proteins in more advanced breast cancers and significantly less in metastatic breast disease. Further, we observed that RUNX proteins were more strongly expressed when the breast cancer was positive for at least one growth or hormone receptor. MDA-MB-231 cells are triple-negative breast cancer cells, which further make them a less appropriate cell-based model for studying the functions of RUNX proteins in breast cancer. This is especially critical given that we are proposing that the functions of RUNX proteins are sensitive to cellular context. A further complication, for which we have not addressed, is the mutation status of *RUNX1* and *RUNX2* in MDA-MB-231 cells. As mentioned before, we observed that the RUNX1 and RUNX2 Western blot bands migrated at the appropriate molecular weight in an SDS-PAGE gel, so we do not suspect that RUNX proteins were truncated in MDA-MB-231 cells. However, this does not rule out the possibility that these *RUNX* genes may have point mutations that cause aberrant functions. We have operated under the assumption that *RUNX1* and *RUNX2* were wild-type in MDA-MB-231 cells, but we should not rule out the possibility that many of the breast cancer-specific functional roles for RUNX1 and RUNX2 we have observed may be a consequence of mutations.

Materials and Methods

Tissue Microarrays

TMAs (BR1503a & BR10010) were obtained from US BioMax. Information pertaining to Grade, Stage, TNM, Type, ER, PR, HER2, AR, p53, and ki67 were provided by US BioMax.

BR1503a is a primary breast tissue array of 150 samples of 75 patient cases: 3 cases of adjacent normal breast tissue, 3 cases of breast fibroadenoma, 2 cases of breast cystosarcoma phyllodes, 7 cases of breast intraductal carcinoma, and 60 cases of breast invasive ductal carcinoma. Duplicate cores per case.

BR10010 is a breast carcinoma and matched metastatic carcinoma array of 100 samples of 50 patient cases: 46 cases of invasive ductal carcinoma, 1 case of micropapillary carcinoma, 2 cases of invasive lobular carcinoma, and 1 case of neuroendocrine carcinoma. Duplicate cores per case.

Immunohistochemistry

RUNX2 staining was done using a lab stock of Mouse Monoclonal hybridoma (IgG purified) clone 8G5 as previously (Das et al. 2009) with the following modification: antibody concentration was reduced to 1:500 dilution.

RUNX1 staining was done as previously described (Liu et al. 2011) using RUNX1 Rabbit Polyclonal 4334 from Cell Signaling.

Histology Quantification

Each tissue section was imaged and independent researchers blindly scored the sections based on the metric in Figure 3.12 A, which is based on the nuclear intensity of DAB.

Histological Statistics

Statistical analyses were performed by converting histological scores (-, +, ++, +++, and ++++) into scalar variables (0, 100, 200, 300, and 400). Scores such as +/++ were converted to the halfway point i.e. 150. Statistical testing was performed in GraphPad Prism using the Kruskal-Wallis non-parametrc ANOVA or (for 3 or more groups) or the non-parametric t-tests (Wilcoxon Rank-Sum). Non-parametric tests were used as the data values are not normally distributed. Kruskal-Wallis ANOVA tests were followed by multiple comparisons for each group.

siRNA Oligos

Smart Pool: ON-TARGETplus Non-silencing siRNA (D-001810-0X) Dharmacon.

SMARTpool: ON-TARGETplus RUNX1 siRNA (L-003926-00-0005) Dharmacon

SMARTpool: ON-TARGETplus RUNX2 siRNA (L-012665-00-0005) Dharmacon

siRNA Transfection

siRNA transfection was done using Oligofectamine (Invitrogen) and Opti-MEM (Invitrogen) using 50nM siRNA according to manufacturer's instructions.

Affymetrix Arrays and Analysis

Human Gene 1.0ST arrays from Affymetrix were used to measure gene expression levels in MDA-MB-231 cells following Runx siRNA transfection. cRNA amplification and hybridization to the array was performed by the UMASS Medical School Genomic Core as previously described (Dowdy et al. 2010).

Analysis was performed in R to execute normalization, quality control, transcript-level reporting, annotation, and contrast tests. The "affy" (Gautier et al. 2004) package was used to read in raw fluorescent values from arrays; values were normalized across all arrays using quantile normalization (Bolstad et al. 2003), robust means average (RMA) background correction and median polish (Irizarry et al. 2003). Quality control plots were generated to ensure the arrays did not have any artifacts and post-processing values were in similar ranges. Transcripts were annotated using the "annotate" package, the "hugene10sttranscriptcluster.db" package. Contrast tests were generated and performed using the "limma" package.

Matrigel Invasion and Migration Assays

Proliferating MDA-MB-231 cells were trypsinized and counted using Cellometer Auto T4 Cell Counter. A cell suspension of 100,000 cells/mL in growth medium was prepared and 100µL of the suspension was loaded into each BD Matrigel 24-well 8.0 µm PET Membrane Invasion Chamber (#354483). Matrigel coated plates, and control insert plates had 500µL NIH3T3-conditioned medium loaded in the bottom as the chemoattractant. Plates and chemoattractant medium were incubated at 37°C for 3-4 hours prior to loading MDA-MB-231 cells. Cells were incubated for 16 hours at 37°C in 5% CO_2 and then fixed and stained using the Fisher HealthCare PROTOCOL Hema 3 Manual Staining System (#22-122-911) according to the manufacturer's instructions. Matrigel and cells that did not invade were eliminated by cotton swabs. Cells that had migrated or invaded to the other side of the inserts were counted using an inverted light microscope.

Immunofluorescence

Cells were grown on gelatin-coated coverslips, fixed, stained, and imaged for RUNX1 (Cell Signaling Rabbit Polyclonal RUNX1 4334), RUNX2 (Santa Cruz

Biotechnologies Rabbit Polyclonal RUNX2 M-70), and UBF (Santa Cruz Biotechnologies Mouse Monoclonal UBF F-9) as previously described (Ali et al. 2010, 2012).

Growth Curve

For each biological replicate, cells were plated at equal density (150,000 cells per well on day 1). On day 2, three wells of a 6-well plate per time point (24, 48, and 72 hours post-transfection) were transfected with siRNA. At each time point, three wells per siRNA were trypsinized, spun down, and resuspended in equal volume growth media. The cell suspension from each well was used on each side of a Cellometer counting slide and counted using a Cellometer T4 Auto Cell Counter (making 6 technical replicates per time point, per biological replicate).

^{35}S-Protein Synthesis Labeling

Protein synthesis was measured as described in (Ali et al. 2010, 2012). Briefly, cells were grown in Met/Cys free medium prior to pulse labeling with EasyTag EXPRESS^{35}S Protein Labeling Mix (Perkin Elmer). Protein was lysed and run on an SDS-PAGE gel, then dried on Whatman paper. Blot was exposed to film and densitometry was used to quantify differential amino acid incorporation.

Chromatin Immunoprecipitation

ChIP conditions for RUNX1, RUNX2 and UBF in MDA-MB-231 cells as described (Lee et al. 2006) with the following modifications:

Crosslinking

1. Cells were grown to near-confluence in 100mm dishes. 10 plates per ChIP antibody are used.
2. Cells were washed in 37°C serum-free medium twice, and then placed in 10mL 37°C serum-free medium before addition of 1mL crosslinking buffer (50mM HEPES-KOH, 100mM sodium chloride, 1mM EDTA, 0.5mM EGTA, 2.75% w/v formaldehyde) and incubated for 5 minutes at room temperature.
3. 500µL freshly-prepared 2.5M glycine was added to quench crosslinking reaction, and incubated for 5-10 minutes at room temperature.
4. Plates were then placed on ice such that the plates were surrounded by ice on all sides and washed twice with ice-cold PBS. Then, 500µL of ice-

cold PBS supplemented with Roche cOmplete EDTA-Free Protease Inhibitor and 25nM MG132 is added.

5. Cells were scraped and placed in nuclease-free 1.75mL tubes (Axygen MCT-175-C 1.7) and spun at 500g for 5 minutes at 4°C.
6. Supernatant was removed and tube was dropped in liquid nitrogen and stored at -80°C until sonication.

Nuclear Isolation & Sonication

1. Cells were removed from -80°C and placed on ice in 1mL freshly-prepared ice-cold Buffer A (50mM HEPES-KOH pH 7.5, 140mM sodium chloride, 1mM EDTA, 10% v/v glycerol, 0.5% v/v NP-40, 0.25% v/v Triton X-100, 1x cOmplete EDTA-Free Protease Inhibitor, 25μM MG-132 in nuclease-free water).
2. Once cells had thawed, tubes were placed on upright rotator at 4°C for 10 minutes.
3. Nuclei were spun down at 700g for 5 minutes at 4°C and Buffer A supernatant was removed.
4. Nuclei were resuspended in 1mL freshly-prepared ice-cold Buffer B (10mM Tris-HCl pH 8.0, 200mM sodium chloride, 1mM EDTA, 1mM EGTA, 1x cOmplete EDTA-Free Protease Inhibitor, 25μM MG-231 in

nuclease-free water) and placed on upright rotator at room temperature for 10 minutes.

5. Nuclei were spun down at 700g for 5 minutes at 4°C and Buffer B supernatant was removed.

6. Nuclei were resuspended in freshly-prepared ice-cold Buffer C (10mM tris-HCL pH 8.0, 100mM sodium chloride, 1mM EDTA, 1mM EGTA, 0.1% w/v sodium deoxycholate, 0.5%w/v N-lauroylsarcosine, 1x cOmplete EDTA-Free Protease Inhibitor, 25μM MG-132 in nuclease-free water) and incubated on ice for at least 20 minutes prior to sonication.

7. Sonication was done on a Misonix Sonic Dismembrator S-4000 using a 1.6mm microtip adapter. Tube was positioned in an ice water bath such that the probe tip was a few millimeters from the bottom of the tube (clear tube is helpful here) and perfectly centered. Sonication program was as follows:

    a. 8 cycles, with 30 second rest periods between each cycle.

    b. Each cycle was 20 seconds total pulse time, oscillating between 1 second "on" and 2 seconds "off".

8. Following sonication, an aliquot of lysate was taken and processed for DNA purification to verify by agarose gel that the median fragment size is approximately 400bp. Lysates were frozen until validation was completed.

Immunoprecipitation

1. Once fragment size had been validated, lysates were thawed and 100µL 10% v/v Triton X-100 was added to each tube and tubes were spun at full speed at 4°C for 20 minutes. Soluble fractions from all tubes (from same cell line / condition) were combined into one nuclease-free tube on ice.

2. For each IP, approximately 10mL lysate was added to a nuclease-free 15mL conical. 50µL of lysate was set aside as “input” and stored at 4°C.

3. 50µg of each antibody was added to appropriate 15mL conical tubes, and tubes were placed on vertical rotator at 4°C for 16 hours.

4. Invitrogen Protein A and Protein G Dynalbeads were thoroughly resuspended.

5. 50µL each Protein A and Protein G Dynalbeads were added to each 15mL tube. Tubes were placed back on vertical rotator at 4°C and incubated for 4 hours.

6. 1 1.75mL nuclease-free tube per IP was placed on MagnaRack (Invitrogen). Lysates were added to tubes, when beads were fully adhered to magnet, cleared lysate was removed. This was repeated until all beads had adhered to magnets.

7. Beads were gently washed with the following freshly-prepared ice-cold buffers in order: Low Salt Wash Buffer (0.1% w/v SDS, 1% v/v Triton X-

100, 2mM EDTA, 150mM Sodium Chloride, 20mM Tris-HCl pH 8.0 in Nuclease-Free Water), High Salt Wash Buffer (0.1%w/v SDS, 1% v/v Triton X-100, 2mM EDTA, 500mM Sodium Chloride, 20mM Tris-HCl pH 8.0 in Nuclease-Free Water), TEN (1mM ETDA, 10mM Tris-HCl pH 8.0, 50mM Sodium Chloride).

8. Beads were placed in 100uL freshly-prepared room temperature Elution Buffer (1% w/v SDS, 100mM Sodium Bicarbonate, 10mM EDTA in Nuclease Free Water) and tubes were placed on tube shaker for 30 minutes at room temperature. Tubes were then moved to MangaRack, Cleared solution was transferred to fresh tube. These steps was repeated twice.

9. 200µL TE Buffer was added to lysates to bring the final volume to 400uL. 150µL Elution Buffer and 200µL TE Buffer was added to 50 µL Input sample set aside earlier. Genomic DNA isolation proceeds as described (Lee et al. 2006).

ChIP-qPCR

For ChIP-DNA isolated as above, ChIP-qPCR was performed as described (Ali et al. 2010, 2012).

ChIP-seq Library Preparation

For ChIP-DNA isolated as above, libraries were prepared for paired-end multiplexed Illumina/Solexa sequencing using the Invitrogen TruSeq DNA Sample Prep v2.0 Kit according to manufacturer's instructions.

Read Mapping

Reads were mapped to hg19 using Bowtie2 for paired-end reads (Langmead and Salzberg 2012).

Peak Calling

Peaks were called on aligned sequences using MACS (Zhang et al. 2008).

Conversion of SAM to BED files

Mapped reads were converted to BED-format reads for analysis using SAMTools (Li et al. 2009) and BEDtools (Quinlan and Hall 2010).

Discriminative Motif Discovery

Homer (Heinz et al. 2010) motif discovery software was used to discover enriched motifs. For each set of test intervals, a GC/CpG/length-matched set of intervals was generated by Homer and used as background.

CHAPTER 4 FOXPRIMER: A QPCR PRIMER DESIGN PROGRAM

Authors and contributions

Jason R. Dobson, Ricardo Medina, Andre J. van Wijnen, Janet L. Stein, Jane B. Lian, Gary S. Stein.

Workflow for manual primer design and validation was optimized and established by RM.

FoxPrimer was designed, written and tested by JRD.

Background

The advent of high-throughput sequencing technologies has allowed for the application of systems biology-type approaches to a wide range of cell biological models. Often, it is necessary to validate genome-wide findings at single gene or genomic locus resolution, and this is commonly done through the design of Real-Time qPCR (quantitative polymerase chain reaction) primers for the amplification of either cDNA (complimentary DNA) or genomic DNA. We propose that the tools and methods used to design Real-Time qPCR primers should be both rigorous in design parameters and accommodate batch design. A major hurdle in the design of Real-Time qPCR primers is the investment of time required to make good primers. While there are many available programs to automate the design of primer pairs, they do not provide the user with enough information to decide which primers have the best chance of being efficient for qPCR. Further, we have not found a program that provides sufficient detail about the positions of primers to determine whether they can be used for a circumstance such as amplification of a specific isoform, or detection of cDNA in samples with genomic DNA contamination. To facilitate the process of designing qPCR primers in a rapid manner, we have designed the FoxPrimer suite of primer design algorithms. Through the use of a web interface, FoxPrimer mechanizes the design workflow via a number of robust command-line tools and stores the primer results in a rapidly-searchable local database. These design choices provide a simple interface, which can be shared among a group of users.

Purpose

The design of Real Time qPCR primers is a critical step in any experimental approach where quantitative analysis is essential. Biological interpretation of results is often highly dependent on the use of primers that are specific, highly efficient, and well-annotated in position. In our experience, with the wealth of tools available to end-users for primer design, creating primers that fit the specificity requirement is generally successful. However, a single open-source program that meets all three requirements does not exist; instead users must rely on multiple programs to design and annotate primer pairs. FoxPrimer is designed to create and store Real-Time qPCR primers that meet all three of these requirements (specific, efficient, and well-annotated), while providing a simple web-based interface for the end-user.

A non-experimental hurdle for primer design is the management of primer sequences within a research group or among collaborators. Commonly, researchers maintain a large spreadsheet of primers and their cognate information. While this may work well for individual use, this can easily become quite unwieldy when a single document is shared among many users and may become a problem if all users are allowed to read and write to the same file. Further complicating the issue with spreadsheet-style storage is the implicit requirement that researchers must manually enter the primer information, which in many cases is very time-consuming. This is especially problematic when implementing a policy in which primer location is required for entry into the spreadsheet, as many primer design programs do not provide this information and would therefore become a barrier for users to easily store their primer information in a consistent format that is useful to others.

FoxPrimer is designed to rapidly annotate primer pairs that have been experimentally validated, storing the primer information in a searchable database. To store validated primers, users only need to enter the sequences and a few other pieces of information, while FoxPrimer handles the annotation and storage into the primer databases. FoxPrimer therefore offers a solution for management of a database of primer information indexed with uniform information and with minimal time and effort requirements on the end-user for entry.

Real Time qPCR Primer Design Rules

General considerations

Primer pairs designed for Real Time qPCR typically have the following characteristics: primer oligo length between 18 and 24 nucleotides, melting temperature between 58°C and 62°C, and primer product size of approximately 100bp. There are other considerations as well, such as no self-complementarity and low probability of recognizing repetitive sequences. Real-Time qPCR primers should amplify a single product as measured experimentally by a single band on a gel (after reaction) or a single sharp peak on a dissociation curve. One of the most critical, and often overlooked, aspects of Real Time qPCR primer usage is validation of primer efficiency by standard curve. Two common means of determining relative levels of a target sequence rely on the use of one or more internal controls or reference genes; both of these methods require all primer pairs (target and control) to amplify products at similar rates (Pfaffl 2001, Hellemans et al. 2007). Primer pair efficiency is measured by the rate at which the target is amplified in response to template concentration; across a wide range of concentrations, each time the concentration doubles, the detection value (C_t) should increase by one giving an efficiency of 100% (Bustin et al. 2009). Primer pairs that are not highly sensitive to template concentration are not suitable for quantitative use, and should therefore not be used for Real Time qPCR. While many algorithms are designed to give the best chance at having highly efficient primer pairs, experimental validation of primer pair efficiency is absolutely required.

cDNA Primer Design Considerations

When designing Real Time qPCR primers with the intent of amplifying specific cDNA targets, there are several considerations that must be kept in mind. The first is the ability to recognize the mature mRNA sequence with greater efficiency than any genomic DNA contamination. The amount of contaminating genomic DNA in an RNA sample can be reduced by DNAse I digestion prior to cDNA amplification. Experimentally, taking an equal amount of RNA and performing the cDNA amplification reaction without the addition of reverse transcriptase creates a negative control template that can be used in the Real Time qPCR reaction to quantify the extent of genomic DNA contamination in the sample.

In eukaryotes, intronic sequences that are part of the pre-mRNA are spliced out of the transcript to form the mature mRNA. The junction points at which the exons are joined represent sequences that are typically not found at the genomic level and are unique to the mature transcript. When designing Real Time qPCR primers for the amplification of specific cDNA templates, using primer pairs for which one or both oligos span these exon-junction points should create a primer pair that is highly specific for the mature form of the mRNA. Alternatively, designing a primer pair for which the primers recognize different exons with a large intron between the exons creates a bias towards the mature mRNA, if the Real Time qPCR elongation time is quite short. Lastly, primer pairs within the same exon will have the same ability to amplify genomic DNA as cDNA. Often, a primer pair that falls within the same exon is the only primer pair that meets the requirements for efficiency, and it is therefore critical to include experimental controls as described above. This last consideration is especially important when designing primers for genes that do not contain introns.

Manual Primer Design

Before describing how FoxPrimer works, we describe our workflow for manual primer design, which has proven to be quite successful in meeting our three requirements for primer design. FoxPrimer mechanizes each of these steps, saving time and effort on the part of the researcher.

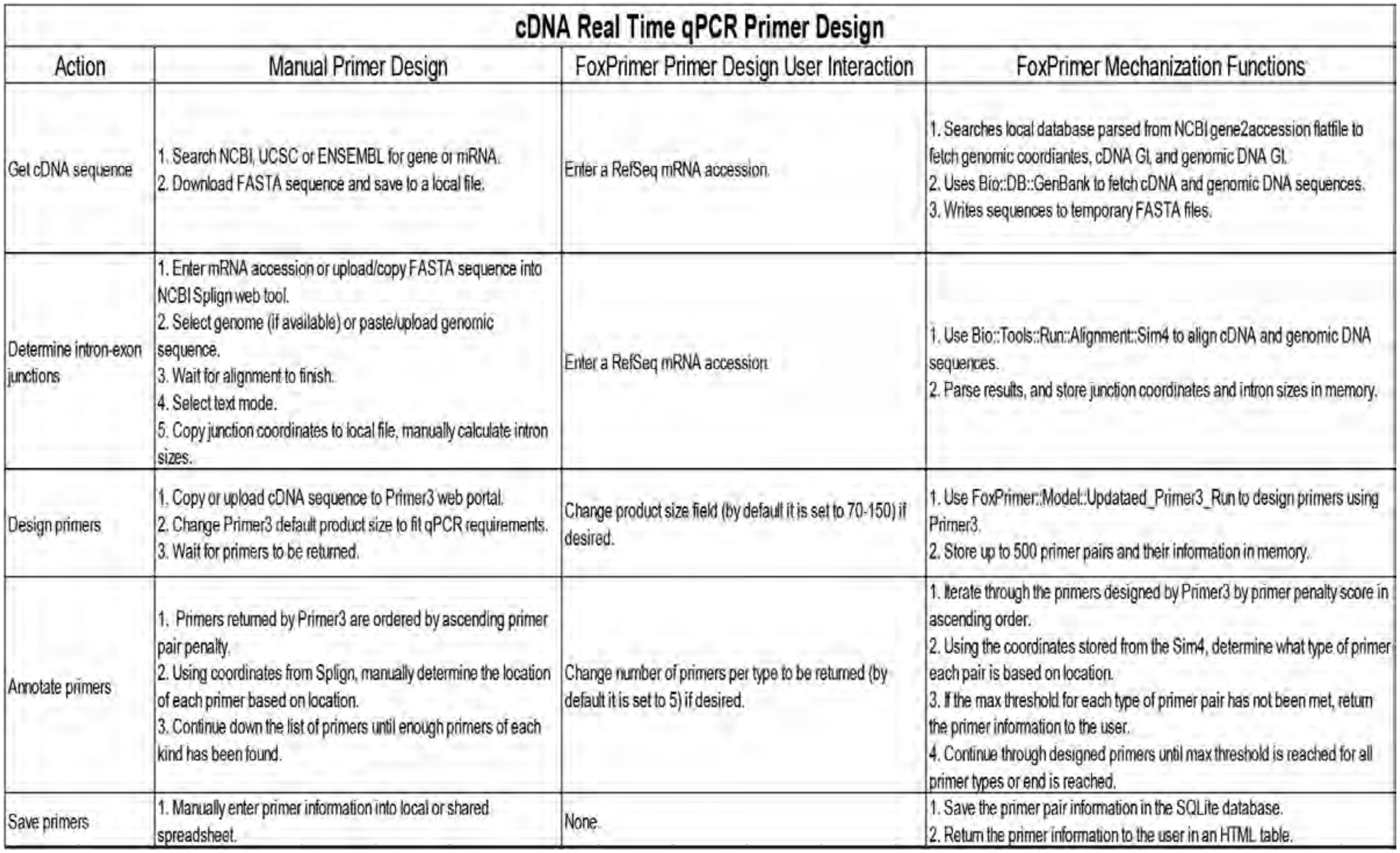
cDNA Primer Design

Our typical workflow for manual design of cDNA primers for a single transcript has been empirically defined to have a very high rate of success in terms of target specificity and primer pair efficiency. The design and positional-annotation of Real-Time qPCR primers for the amplification of cDNA has two parts. First, the cDNA sequence is aligned to the reference genome to define intron-exon junctions and intron sizes. Then, primers are designed for the cDNA sequence and locations are manually annotated based on cDNA-genomic DNA alignment coordinates. To accomplish these tasks, we use two web-based tools: NCBI Splign <http://www.ncbi.nlm.nih.gov/sutils/splign/splign.cgi> (Kapustin et al. 2008) to define transcript splice junctions and Primer3-Web <http://frodo.wi.mit.edu/> (Rozen and Skaletsky 2000) for primer design. We then manually determine the locations of primer pairs relative to exon/intron coordinates. Comparing this workflow to the one executed by FoxPrimer, we can see that while the underlying tools for this workflow are quite similar, FoxPrimer requires much less investment of time and effort from the user **(Table 4.1)**.

Genomic DNA (ChIP) Primer Design

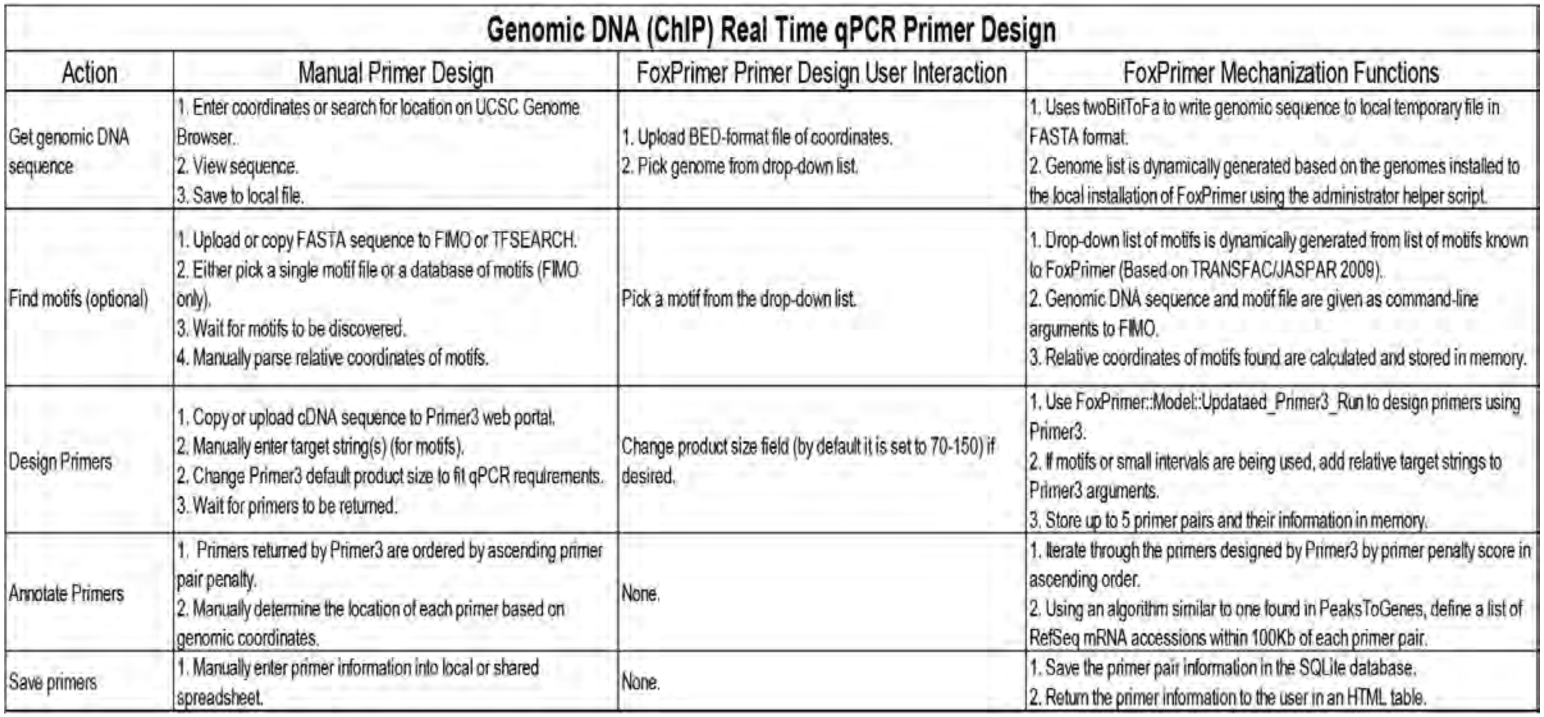
When designing primers for amplification of genomic DNA, we first determine the genomic region in which we would like the primers to be designed and then download the genomic sequence. FoxPrimer only requires a BED-format file of genomic coordinates, which is, conveniently, the type of file generated by many popular peak-calling algorithms (MACS (Zhang et al. 2008), SPP (Kharchenko et al. 2008), PeakSeq (Rozowsky et al. 2009)). The next consideration is whether we want primers to amplify a defined sub-sequence such as a motif or whether the goal is to amplify a broader region. If the objective is to design primers flanking a specific motif, either TFSEARCH <http://www.cbrc.jp/research/db/TFSEARCH.html> (Heinemeyer et al. 1998) or FIMO (Find Individual Motif Occurences) <http://meme.nbcr.net/meme/cgi-bin/fimo.cgi> (Grant et al. 2011) is used to define the location of the motif of interest in a genomic sequence. Primer3-Web <http://frodo.wi.mit.edu/> (Rozen and Skaletsky 2000) is then used for primer design (setting the subsequence string if making primers for a specific motif). The positions of the primers within the genomic sequence are then manually calculated primers relative to a gene of interest. Again, comparison of this workflow to the one executed by FoxPrimer, shows that while the underlying tools are quite similar, FoxPrimer requires much less investment of time and effort from the user and provides more information about the relative location of the primer pair **(Table 4.2)**.

Table 4.1 cDNA Primer Design



cDNA Real Time qPCR Primer Design			
Action	Manual Primer Design	FoxPrimer Primer Design User Interaction	FoxPrimer Mechanization Functions
Get cDNA sequence	1. Search NCBI, UCSC or ENSEMBL for gene or mRNA. 2. Download FASTA sequence and save to a local file.	Enter a RefSeq mRNA accession.	1. Searches local database parsed from NCBI gene2accession flatfile to fetch genomic coordiantes, cDNA GI, and genomic DNA GI. 2. Uses Bio::DB::GenBank to fetch cDNA and genomic DNA sequences. 3. Writes sequences to temporary FASTA files.
Determine intron-exon junctions	1. Enter mRNA accession or upload/copy FASTA sequence into NCBI Splign web tool. 2. Select genome (if available) or paste/upload genomic sequence. 3. Wait for alignment to finish. 4. Select text mode. 5. Copy junction coordinates to local file, manually calculate intron sizes.	Enter a RefSeq mRNA accession.	1. Use Bio::Tools::Run::Alignment::Sim4 to align cDNA and genomic DNA sequences. 2. Parse results, and store junction coordinates and intron sizes in memory.
Design primers	1. Copy or upload cDNA sequence to Primer3 web portal. 2. Change Primer3 default product size to fit qPCR requirements. 3. Wait for primers to be returned.	Change product size field (by default it is set to 70-150) if desired.	1. Use FoxPrimer::Model::Updataed_Primer3_Run to design primers using Primer3. 2. Store up to 500 primer pairs and their information in memory.
Annotate primers	1. Primers returned by Primer3 are ordered by ascending primer pair penalty. 2. Using coordinates from Splign, manually determine the location of each primer based on location. 3. Continue down the list of primers until enough primers of each kind has been found.	Change number of primers per type to be returned (by default it is set to 5) if desired.	1. Iterate through the primers designed by Primer3 by primer penalty score in ascending order. 2. Using the coordinates stored from the Sim4, determine what type of primer each pair is based on location. 3. If the max threshold for each type of primer pair has not been met, return the primer information to the user. 4. Continue through designed primers until max threshold is reached for all primer types or end is reached.
Save primers	1. Manually enter primer information into local or shared spreadsheet.	None.	1. Save the primer pair information in the SQLite database. 2. Return the primer information to the user in an HTML table.

Table 4.2 Genomic DNA (ChIP) Primer Design



Genomic DNA (ChIP) Real Time qPCR Primer Design			
Action	Manual Primer Design	FoxPrimer Primer Design User Interaction	FoxPrimer Mechanization Functions
Get genomic DNA sequence	1. Enter coordinates or search for location on UCSC Genome Browser. 2. View sequence. 3. Save to local file.	1. Upload BED-format file of coordinates. 2. Pick genome from drop-down list.	1. Uses twoBitToFa to write genomic sequence to local temporary file in FASTA format. 2. Genome list is dynamically generated based on the genomes installed to the local installation of FoxPrimer using the administrator helper script.
Find motifs (optional)	1. Upload or copy FASTA sequence to FIMO or TFSEARCH. 2. Either pick a single motif file or a database of motifs (FIMO only). 3. Wait for motifs to be discovered. 4. Manually parse relative coordinates of motifs.	Pick a motif from the drop-down list.	1. Drop-down list of motifs is dynamically generated from list of motifs known to FoxPrimer (Based on TRANSFAC/JASPAR 2009). 2. Genomic DNA sequence and motif file are given as command-line arguments to FIMO. 3. Relative coordinates of motifs found are calculated and stored in memory.
Design Primers	1. Copy or upload cDNA sequence to Primer3 web portal. 2. Manually enter target string(s) (for motifs). 2. Change Primer3 default product size to fit qPCR requirements. 3. Wait for primers to be returned.	Change product size field (by default it is set to 70-150) if desired.	1. Use FoxPrimer::Model::Updataed_Primer3_Run to design primers using Primer3. 2. If motifs or small intervals are being used, add relative target strings to Primer3 arguments. 3. Store up to 5 primer pairs and their information in memory.
Annotate Primers	1. Primers returned by Primer3 are ordered by ascending primer pair penalty. 2. Manually determine the location of each primer based on genomic coordinates.	None.	1. Iterate through the primers designed by Primer3 by primer penalty score in ascending order. 2. Using an algorithm similar to one found in PeaksToGenes, define a list of RefSeq mRNA accessions within 100Kb of each primer pair.
Save primers	1. Manually enter primer information into local or shared spreadsheet.	None.	1. Save the primer pair information in the SQLite database. 2. Return the primer information to the user in an HTML table.

Program Structure - Model, View, Controller

FoxPrimer is written in Perl on the Catalyst web framework <http://www.catalystframework.org/>. The program is structured such that the Model, View and Controller (MVC) (GuangChun et al. 2003) architecture can be easily separated for future development of FoxPrimer.

Model

FoxPrimer is structured where the majority of program logic and execution occurs within the Model. The Perl modules that constitute the Model of the FoxPrimer MVC are self-contained and are capable of executing primer design functions independent of the Catalyst Controller module. This design approach allows for easier maintenance and testing of the FoxPrimer suite of primer design functions should interactions with external utilities become unstable in the future.

Controller

The FoxPrimer Controller module is only responsible for relaying information from the user to the Model where all of the business logic takes place. Then, the Controller returns error messages or final data to the View (in this case the web page). The FoxPrimer controller ensures that the data entered by the user are valid and relays error messages as needed.

View

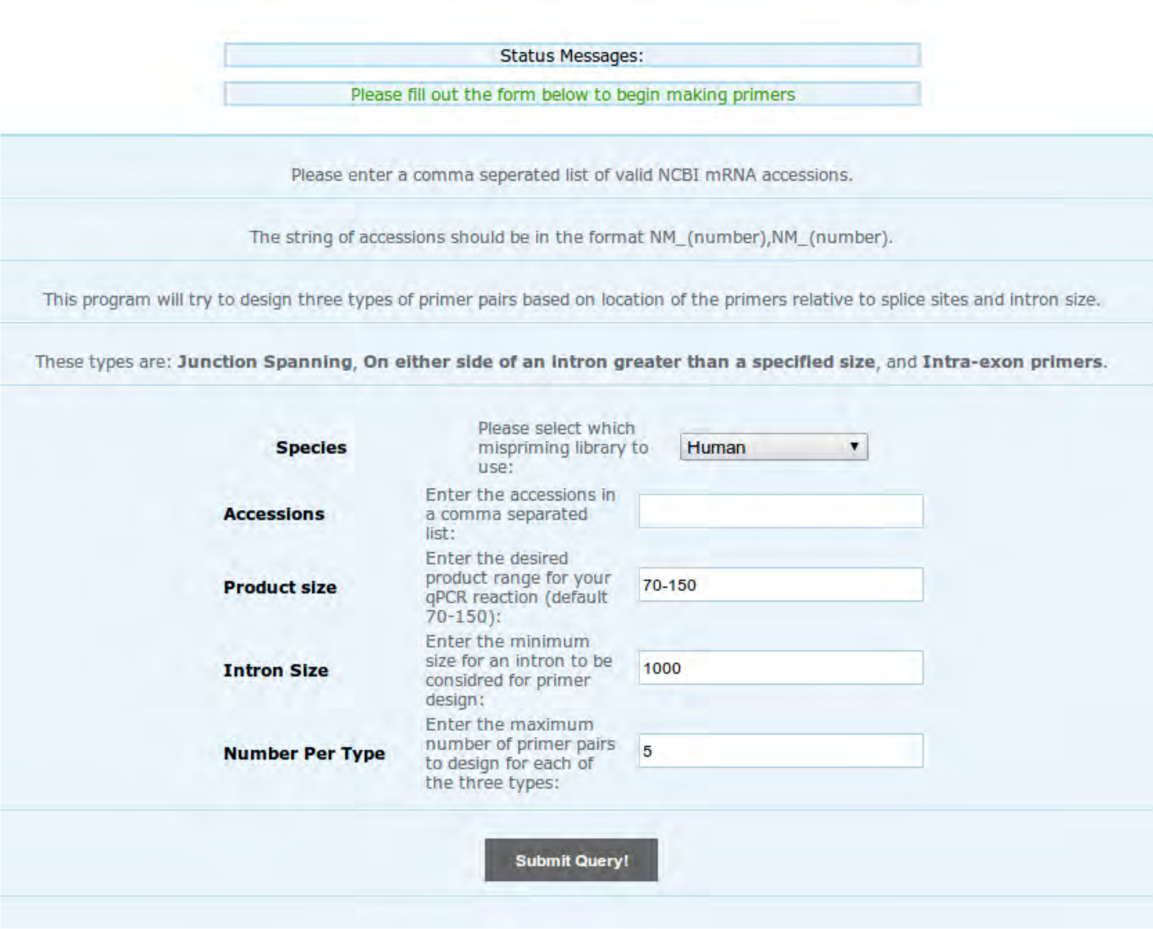
The FoxPrimer View module is the web page where the user can interact with the settings and view output from the program. All code related to the display of the web pages is separate from any code involved in the business logic of the program. This design allows editing of the aesthetics without affecting the Model and Controller elements of the program. The FoxPrimer View module uses Template Toolkit <http://www.template-toolkit.org/> to render HTML code, and allows the Controller to pass Perl data structures to the View and have the Template Toolkit iterate through the information and properly display data to the user.

Program Function

cDNA Primer Design

Upon entering the program, the View module (or the web page) renders and presents the user with a prompt for a comma-delimited list of RefSeq (Pruitt et al. 2009) RNA accessions **(Figure 4.1)**. By default, the fields for product size, minimum intron size, and number per type are already filled, but can be changed by the user. The number per type field determines how many of each type of cDNA primer type will be returned. Primer types are defined by the locations of the primer pairs relative to the intron-exon junctions of the transcript. There are four types of primers that are returned to the user: junction spanning primers (one or both olgios span the intron-exon junction), exon primer pair (primers target different exons that have one or more introns between the exons whose combined length is greater than the user-defined minimum intron size), smaller exon primer pair (primers target different exons that have one or more introns between the exons whose combined length is less than the user-defined minimum intron size), and intra-exon primers (primers that map to the same exon). Also by default, the mispriming library for Primer3 is set to 'human', but can be changed based on the organism for which the user is designing qPCR primers. Once the FoxPrimer Controller has determined that all of the fields are valid, it will begin to check the RefSeq accessions entered by the user.

Figure 4.1 cDNA primer design form



Status Messages:

Please fill out the form below to begin making primers

Please enter a comma seperated list of valid NCBI mRNA accessions.

The string of accessions should be in the format NM_(number),NM_(number).

This program will try to design three types of primer pairs based on location of the primers relative to splice sites and intron size.

These types are: **Junction Spanning**, **On either side of an intron greater than a specified size**, and **Intra-exon primers**.

Species	Please select which mispriming library to use:	Human ▾
Accessions	Enter the accessions in a comma separated list:	
Product size	Enter the desired product range for your qPCR reaction (default 70-150):	70-150
Intron Size	Enter the minimum size for an intron to be considred for primer design:	1000
Number Per Type	Enter the maximum number of primer pairs to design for each of the three types:	5

Gene2Accession Database

FoxPrimer relies on RefSeq accession and NCBI "GenInfo Identifier" or GI to determine the location of the sequence on genomic DNA, and to rapidly fetch the cDNA and genomic sequences from GenBank. In order to implement this large database of information, FoxPrimer provides a helper script that interacts with the NCBI FTP server to download the gene2accession flat file <ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2accession.gz> containing the requisite information, parses the file, and then stores the relevant information in a local SQLite <www.sqlite.org> database. Storage in the SQLite database allows FoxPrimer to rapidly determine if a user-defined RefSeq accession is valid before attempting to interact with GenBank.

FoxPrimer will design primers for each valid accession, while each accession that is not found by FoxPrimer in the gene2accession database will be returned to the user in an error message. If the user has not entered any valid accessions, then the program will exit and inform the user which accession(s) were not found in the gene2accession database in an error message.

BioPerl

FoxPrimer uses BioPerl (Stajich et al. 2002) to interact with NCBI and fetch a sequence object for each cDNA and genomic DNA pair found in the gene2accession database. FoxPrimer stores in memory the GI for the cDNA, and the GI, start position, stop position and strand for the genomic DNA. Then, using the BioPerl Bio::DB::GenBank module FoxPrimer fetches sequence objects from NCBI and extracts a description of the mRNA from the cDNA object. Sequences from both cDNA and genomic DNA objects are written to temporary files in FASTA format.

Sim4

Once the sequences have been fetched from NCBI and written to file, FoxPrimer uses Sim4 (Florea et al. 1998) to align the cDNA sequence to the genomic DNA sequence to determine the coordinates of intron-exon junctions and lengths (if any) of each intron. Sim4 is an algorithm written to rapidly determine a best-fit alignment for a cDNA sequence to a genomic DNA sequence. Sometimes, multiple alignments are found, but this is rare. In such an instance, FoxPrimer will consider each alignment as valid, and proceed with primer design algorithm treating each alignment as a different pair of cDNA and genomic DNA.

BioPerl provides an interface to call Sim4 from the FoxPrimer Model modules and fetches the results for each alignment determined by Sim4 (usually only one per cDNA-gDNA pair). FoxPrimer then determines the coordinates of the 5' and 3' end of each exon, as well as the length of each intron, and stores the information in memory.

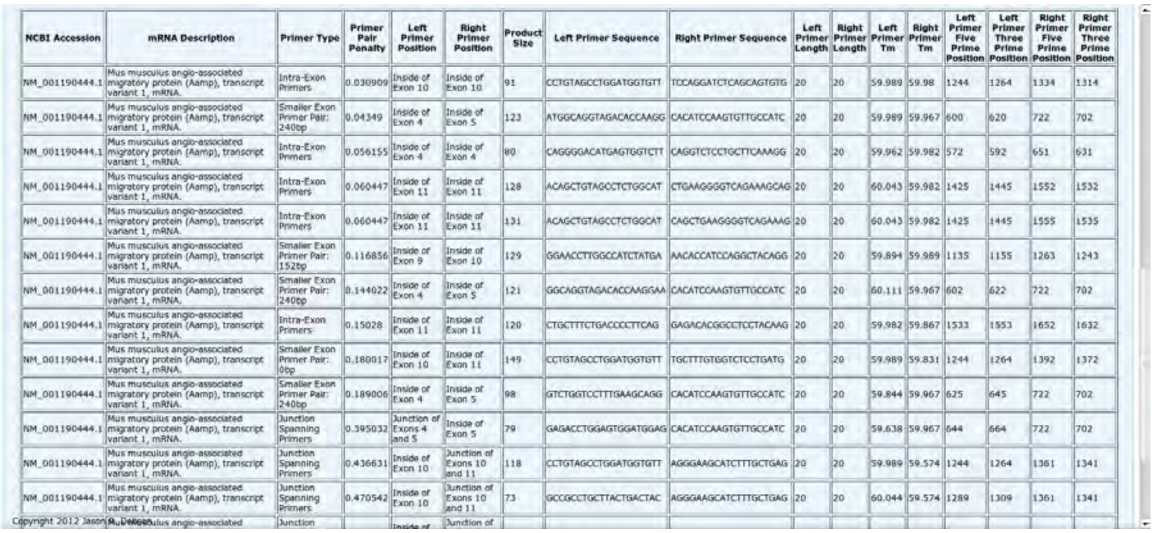
Updated_Primer3_Run

FoxPrimer then uses Primer3 to design several hundred primers using the cDNA sequence as the template and the user-defined product size boundaries as the only constraint. BioPerl provides a module to interact with Primer3, however, this module was written for a previous release of Primer3. The algorithm for Primer3 (current version is 2.2.3) has since been updated and so have the inputs for Primer3. Therefore, FoxPrimer features an updated version of BioPerl::Module:Bio::Tools::Run::Primer3, which we will submit to the BioPerl group when FoxPrimer is made public.

Based on the constraints defined by the user for how many of each type of primer to be designed, FoxPrimer will iterate through the primers designed by Primer3 in increasing primer pair penalty score. For each primer pair, FoxPrimer defines the primer pair type. FoxPrimer will continue through the list of primers until the maximum number of each type of primer pair has been reached, or the primers designed by Primer3 have been exhausted.

Once FoxPrimer has finished designing primers for each accession entered by the user, the primers are stored in a local SQLite database for designed cDNA primer pairs. The primer information is then returned the user in a table via the FoxPrimer View module, together with any error messages **(Figure 4.2)**. Once the primers have been stored in the created primers database, they can be searched and retrieved rapidly using the FoxPrimer search functions described later.

Figure 4.2 Example output from FoxPrimer for cDNA primer design



NCBI Accession	mRNA Description	Primer Type	Primer Pair Penalty	Left Primer Position	Right Primer Position	Product Size	Left Primer Sequence	Right Primer Sequence	Left Primer Length	Right Primer Length	Left Primer Tm	Right Primer Tm	Left Primer Five Prime Position	Left Primer Three Prime Position	Right Primer Five Prime Position	Right Primer Three Prime Position
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Intra-Exon Primers	0.030909	Inside of Exon 10	Inside of Exon 10	91	CCTGTAGCCTGGATGGTGTT	TCCAGGATCTCAGCAGTGTG	20	20	59.989	59.98	1244	1264	1334	1314
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Smaller Exon Primer Pair: 240bp	0.04349	Inside of Exon 4	Inside of Exon 5	123	ATGGCAGGTAGACACCAAGG	CACATCCAAGTGTTGCCATC	20	20	59.989	59.967	600	620	722	702
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Intra-Exon Primers	0.056155	Inside of Exon 4	Inside of Exon 4	80	CAGGGGACATGAGTGGTCTT	CAGGTCTCCTGCTTCAAAGG	20	20	59.962	59.982	572	592	651	631
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Intra-Exon Primers	0.060447	Inside of Exon 11	Inside of Exon 11	128	ACAGCTGTAGCCTCTGGCAT	CTGAAGGGGTCAGAAAGCAG	20	20	60.043	59.982	1425	1445	1552	1532
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Intra-Exon Primers	0.060447	Inside of Exon 11	Inside of Exon 11	131	ACAGCTGTAGCCTCTGGCAT	CAGCTGAAGGGGTCAGAAAG	20	20	60.043	59.982	1425	1445	1555	1535
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Smaller Exon Primer Pair: 152bp	0.116856	Inside of Exon 9	Inside of Exon 10	129	GGAACCTTGGCCATCTATGA	AACACCATCCAGGCTACAGG	20	20	59.894	59.989	1135	1155	1263	1243
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Smaller Exon Primer Pair: 240bp	0.144022	Inside of Exon 4	Inside of Exon 5	121	GGCAGGTAGACACCAAGGAA	CACATCCAAGTGTTGCCATC	20	20	60.111	59.967	602	622	722	702
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Intra-Exon Primers	0.15028	Inside of Exon 11	Inside of Exon 11	120	CTGCTTTCTGACCCCTTCAG	GAGACACGGCCTCCTACAAG	20	20	59.982	59.867	1533	1553	1652	1632
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Smaller Exon Primer Pair: 0bp	0.180017	Inside of Exon 10	Inside of Exon 11	149	CCTGTAGCCTGGATGGTGTT	TGCTTTGTGGTCTCCTGATG	20	20	59.989	59.831	1244	1264	1392	1372
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Smaller Exon Primer Pair: 240bp	0.189006	Inside of Exon 4	Inside of Exon 5	98	GTCTGGTCCTTTGAAGCAGG	CACATCCAAGTGTTGCCATC	20	20	59.844	59.967	625	645	722	702
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Junction Spanning Primers	0.395032	Junction of Exons 4 and 5	Inside of Exon 5	79	GAGACCTGGAGTGGATGGAG	CACATCCAAGTGTTGCCATC	20	20	59.638	59.967	644	664	722	702
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Junction Spanning Primers	0.436631	Inside of Exon 10	Junction of Exons 10 and 11	118	CCTGTAGCCTGGATGGTGTT	AGGGAAGCATCTTTGCTGAG	20	20	59.989	59.574	1244	1264	1361	1341
NM_001190444.1	Mus musculus angio-associated migratory protein (Aamp), transcript variant 1, mRNA.	Junction Spanning Primers	0.470542	Inside of Exon 10	Junction of Exons 10 and 11	73	GCCGCCTGCTTACTGACTAC	AGGGAAGCATCTTTGCTGAG	20	20	60.044	59.574	1289	1309	1361	1341
	Mus musculus angio-associated	Junction		Inside of	Junction of											

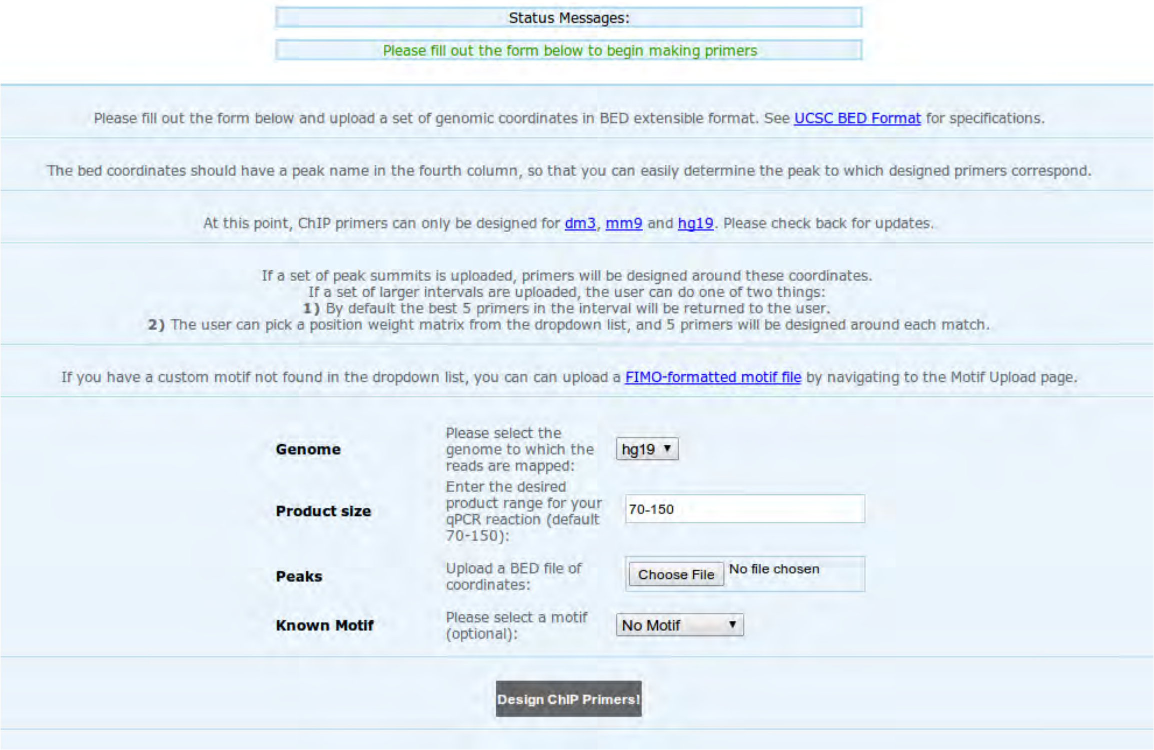
Figure 4.2 Example output from FoxPrimer for cDNA primer design

Screenshot of example output from FoxPrimer cDNA primer design. Tables of primers for each transcript are ordered by ascending primer pair penalty score from Primer3.

Genomic DNA Primer Design

For design of genomic DNA qPCR primers, the FoxPrimer View module renders a form for the user to complete, along with a widget allowing the user to upload a BED-file of genomic interval coordinates **(Figure 4.3)**. FoxPrimer will check to ensure that a file has been uploaded, and if not, the Controller will inform the View to prompt the user to upload a BED file. By default, the View has entered a product size for the primer pairs to be created, and the drop down menus for the genome and motif have been set to hg19 and "no motif", respectively. Each of these parameters can be changed by the user.

Figure 4.3 Genomic DNA primer design form



Status Messages:

Please fill out the form below to begin making primers

Please fill out the form below and upload a set of genomic coordinates in BED extensible format. See UCSC BED Format for specifications.

The bed coordinates should have a peak name in the fourth column, so that you can easily determine the peak to which designed primers correspond.

At this point, ChIP primers can only be designed for dm3, mm9 and hg19. Please check back for updates.

If a set of peak summits is uploaded, primers will be designed around these coordinates.
If a set of larger intervals are uploaded, the user can do one of two things:
1) By default the best 5 primers in the interval will be returned to the user.
2) The user can pick a position weight matrix from the dropdown list, and 5 primers will be designed around each match.

If you have a custom motif not found in the dropdown list, you can can upload a FIMO-formatted motif file by navigating to the Motif Upload page.

Genome	Please select the genome to which the reads are mapped:	hg19 ▾
Product size	Enter the desired product range for your qPCR reaction (default 70-150):	70-150
Peaks	Upload a BED file of coordinates:	Choose File No file chosen
Known Motif	Please select a motif (optional):	No Motif ▾

Dynamic Genome Addition

FoxPrimer is capable of designing genomic DNA qPCR primers for all RefSeq-annotated genomes. To extract the genomic sequences for primer design, FoxPrimer requires large (hundreds of megabytes to several gigabytes, depending on genome size) UCSC 2bit-format files for each genome. It is unreasonable to expect users of FoxPrimer to need a sequence file for all RefSeq genomes or to have the requisite disk space to store these files. FoxPrimer therefore offers dynamic genome addition for each reference genome the user would like to design genomic DNA qPCR primers for. A helper script is provided, which interacts with the UCSC MySQL server <http://genome.ucsc.edu/goldenPath/help/mysql.html> to extract the necessary files and genomic coordinates for genomic DNA primer design. Once a genome has been installed using this helper script, the genome will appear in the dropdown menu on the website and users will be able to design genomic DNA primers for this genome.

Available Motifs

Often, users will want to design their genomic DNA primers so that they amplify a genomic region that matches the DNA-binding motif for a specific protein. To assist in the design of primers, which are targeted around motifs, FoxPrimer provides a set of 454 motifs from the 2009 JASPAR/TRANSFAC position weight motif matrices (Bryne et al. 2008). FoxPrimer provides helper scripts to update the motif list or add custom position weight matrices. The FoxPrimer Controller makes a call to the FoxPrimer Model modules to dynamically fetch a list of motifs available to the user. This list is sent to the FoxPrimer View module and displayed in the form of a drop down menu.

By default, FoxPrimer will only read and design primers for the first 10 lines of a BED file. This design limits the time between execution and results. This variable is clearly defined in the FoxPrimer Model modules, and can be changed by the administrator.

If any interval in the uploaded BED-file is 30bp or less, FoxPrimer will assume the user is requesting primer pairs to flank this region and will mark them as such. FoxPrimer will extend the coordinates of such an interval in both the 5' and 3' directions by twice the length of the maximum user-defined product size. The coordinates of the original interval are stored and defined as the target sequence for Primer3. This ensures that any short interval supplied by the user will be included in the product amplified by the designed primer pairs.

TwoBitToFa

For each set of valid coordinates, FoxPrimer will use these coordinates to write the FASTA format sequence to a temporary file. TwoBitToFa is provided as part of the source code written by Jim Kent for the UCSC genome browser (Meyer et al. 2013). This program provides very rapid retrieval of genomic sequence from a highly compressed 2bit file containing all genomic sequences. A FASTA format file is needed to find motifs (if required) using FIMO, and to design primers using Primer3.

FIMO

If the user elects to search for a motif within the genomic intervals, FoxPrimer will search for the user-defined motif using FIMO (Find Individual Motif Occurences) from the MEME (Multiple Em for Motif Elicitation) suite of motif-oriented programs (Grant et al. 2011). FoxPrimer makes a call to FIMO, informing FIMO of the sequence to search and the motif for which to search. If the motif is not discovered in an interval, that interval is returned to the user in an error message after primers have been designed around all other motifs discovered. For each motif discovered, FoxPrimer will treat these short intervals as described above (coordinates are extended, and used as target sequence for primer design).

For each motif found in the target sequences, FoxPrimer will use the aforementioned Updated_Primer3_Run Model module to create 5 primer pairs.

FoxPrimer will store the necessary information about the primer oligos, and then use the relative template coordinates to extrapolate these coordinates into genomic coordinates.

Primer Position Determination

Designing a pair of primers to amplify genomic DNA is not a difficult task. It is, however, time-consuming to determine where these primers are relative to nearby genes. This information is important and should be stored for each primer pair. For some users, simply knowing where the primers are in raw genomic coordinates will satisfy their research needs, and FoxPrimer provides this information for each primer pair. However, knowing where a primer pair is located relative to nearby genes is a critical piece of information. To define the positions of primers relative to nearby genes, FoxPrimer implements a version of the algorithm used in PeaksToGenes (see Chapter 5). FoxPrimer creates a temporary BED-file of coordinates using the 5'-end of each primer as the genomic start and genomic end of the positive and negative strand primers, respectively. Then, using intersectBed from the BEDTools suite of programs (Quinlan and Hall 2010), FoxPrimer defines a list of genes that are within 100Kb of the primer pair. FoxPrimer then iterates through the list of genes, and mathematically determines the relative positions of the primer pair to each gene in the list.

Once FoxPrimer has determined the relative locations of all designed primers, the information for each primer pair is entered into the created primers database for genomic DNA primers. Like the created cDNA primers database, this database allows all created primers to be rapidly searchable by the user. Finally, the primer information is sent from the FoxPrimer Controller to the FoxPrimer View, and a table of primer information as well as messages, describing any errors that that may have occurred during primer design, is returned to the user.

Validated Primer Entry

Given the investments of time and reagent cost to validate a primer pair, it is critical that validated primers be stored in a location where they are both secure and accessible. By implementing an SQLite database for primer information storage, FoxPrimer accomplishes both of these goals.

To facilitate rapid addition of validated primers to the SQLite database, FoxPrimer allows batch entry of primers. FoxPrimer requires the user to enter primer information in the form of a tab-delimited file containing the following information: “Primer Type”, “Left Primer Sequence”, “Right Primer Sequence”, “Accession”, “Your Name”. “Efficiency”, “Left Primer Location”, “Right Primer Location”, and “Genome”. The last three parameters are only required if the primers are designed for genomic DNA amplification, and informs FoxPrimer about where it should make the template DNA for *in silico* validation.

The FoxPrimer Controller will make calls to the FoxPrimer Model modules to ensure that the fields entered in the file are valid. The user will be informed of any errors via an error message from the FoxPrimer View. For each primer pair, FoxPrimer will use Primer3 to calculate the Tm and primer pair penalty, and use the aforementioned algorithms to determine exonic or genomic positions of the primers for cDNA and genomic primers, respectively. After these parameters have been defined, the primer pair is entered into the validated primer database, which can be searched by the user.

Discussion

FoxPrimer provides a simple interface to a series of robust programs for the design of Real Time qPCR primers that amplify either cDNA or genomic DNA sequences. For each primer pair designed, FoxPrimer provides detailed information about the location relative to either exons or nearby genes. Because FoxPrimer also reports the primer pair penalty score, the user can make informed decisions about which primer pairs should be experimentally tested.

FoxPrimer can be used to make primer pairs for a single mRNA or genomic interval quickly. However, the real strength of this program lies in its ability to accomplish these tasks in batch. A limited version of FoxPrimer is available at <http://www.foxprimer.org>, and is running on a micro-instance of Amazon EC2, therefore, strict limits have been set to reduce compute cycles and costs. On more powerful private servers, FoxPrimer has been tested to design hundreds of primer pairs within a few minutes. High-throughput design of primers with a strong probability of meeting the requirements of Real Time qPCR will be a useful time-saving tool for the validation of next-generation sequencing projects.

FoxPrimer has been a valuable tool for many researchers within our research group over the course of its development. It is our hope this program will see widespread adoption due to its simple interface and open source nature.

Program Requirements

- Perl >= 5.14.2
- Perl module dependencies (see ‘INSTALL’ file for specifics)
- Primer3 command-line version

- Sim4
- twoBitToFa (Kent Source)
- FIMO (From MEME)
- BEDtools
- SQLite3
- MySQL
- OS X or Linux

CHAPTER 5 PEAKSTOGENES: AVERAGE GENE PLOT GENERATION AND STATISTICAL TESTING

Authors and contributions

Jason R. Dobson, Andre J. van Wijnen, Jane B. Lian, Gary S. Stein, Janet L. Stein.

PeaksToGenes was designed, written and tested by JRD.

Introduction

Gene expression is a highly regulated process that ultimately leads to the determination of cellular phenotype and functions. Characterization of the patterns and relationships in the regulatory networks controlling gene expression at the transcriptional level is much sought after information. This will allow for a greater understanding of the mechanism contributing to processes such as development, tissue turnover, response to external stimuli, and disease.

Chromatin features including nucleosome positioning, histone modifications, hypersensitive regions, and DNA-binding proteins like transcription factors contribute to transcriptional regulation. Many of these epigenetic marks have been assigned a particular role in transcriptional regulation based on observed functions related to candidate genes. Approaches utilizing next-generation sequencing to experimentally understand the genome-wide functions of these chromatin features have allowed an unprecedented level of insight into and understanding of the regulatory roles of many proteins.

Examination of the patterns of chromatin feature locations in relation to gene expression from next-generation sequencing data has dramatically increased our understanding of chromatin-mediated regulation of transcription. However, a major hurdle in measuring and defining these relationships has been the researchers' ability to manage the data according to their experimental needs. To assist in the process of defining patterns and relationships between chromatin features and gene expression, we will describe the programming framework PeaksToGenes.

PeaksToGenes is designed to help the user identify patterns in sequencing results for the purpose of making predictions about the potential transcriptional functions of the chromatin features being investigated. A common means to begin identifying patterns in chromatin features is to construct an average gene profile of the mark across the entire genome. These average gene profiles show where a particular mark is most commonly found relative to gene bodies and can be used to perform statistical tests to determine how significant a given observation is. At the core, PeaksToGenes constructs these average gene profiles and then allows the user to query these profiles without manually designing scripts to create or extract information.

PeaksToGenes is written in Modern Perl, and relies only on open-source dependencies. This design approach allows for a completely exposed API, allowing the user to easily write their own extensions for PeaksToGenes should they have more specific experimentally-driven questions. Here, we describe the major functions and API of PeaksToGenes.

Average Gene Profiles

PeaksToGenes is designed to test whether the binding profile of a given protein is different near one subset of genes versus another subset of genes. These gene lists should be large enough so that the group sizes will be sufficient for robust statistical testing. Although there are many instances where proteins are known to bind distal to gene bodies and have enhancer-type functions, PeaksToGenes is focused on proximal binding and therefore focuses on genomic regions 10Kb upstream of the transcriptional start site (TSS), within the gene body, and 10Kb downstream of the transcriptional termination site (TTS). For each gene, PeaksToGenes stores either the number of peaks or the ratio of “IP” over “input” reads per relative genomic region (described below), making it relatively quick to contrast two lists of genes against one another.

The 10Kb regions upstream of the TSS and downstream of the TTS are split into 1Kb non-overlapping intervals. If used in “contrast” mode, PeaksToGenes will look in each one of these intervals and calculate the ratio of “IP” reads to “input” reads. PeaksToGenes can optionally use a scaling factor, such as sequence enrichment scaling (SES) (Diaz et al. 2012) or RPKM (reads per kilobasepair per million reads) (Landt et al. 2012), which is multiplicatively applied to the number of “input” reads found in the interval. Alternatively, if PeaksToGenes is run in “annotate” mode, the number of peak regions will be counted for each relative genomic region.

PeaksToGenes looks at the gene body in two ways: based on transcriptional regions and based on relative genomic coordinates. The transcriptional regions are defined by the function of the transcribed sequences, which fall into four categories: five prime untranslated regions (5’-UTRs), exons, introns, and three prime untranslated regions (3’-UTRs). The genomic regions are defined by relative coordinates between the TSS and TTS and are divided into ten approximately equal length intervals, which we call “gene body deciles”. Because the relative coordinates within the gene bodies are not the same length as the relative genomic coordinates, the number of peaks or reads found in each coordinate are linearly scaled to 1Kb prior to applying the same ratio or counting functions described above for flanking regions.

Once PeaksToGenes has defined either the number of peaks or the signal ratio for each relative genomic region for each gene in the given RefSeq genome, these scalar numbers are stored in a local SQLite database indexed by genome, RefSeq mRNA accession and the name the user has chosen for the experimental sample. The user can then specify multiple gene lists to be utilized for statistical testing or just simple averaging of the signal ratios or numbers of peaks near the list of genes.

Matrix Files

The average genome profile (generated as described above) is stored in the local SQLite database. The data can be accessed by the user several ways. Often, users will want to create some kind of heat map, or manipulate the data using functions separate from the statistical contrast tests provided by PeaksToGenes. To that end PeaksToGenes provides the “matrix” function, which allows the user to specify a list of RefSeq accessions and a series of datasets (from the same genome) for which they would like a tab-delimited file of binding information printed to file. Each row represents a particular RefSeq gene, and each column represents a relative genomic region for a particular data set. The relative regions included in this matrix file are ordered from 10Kb 5’-TSS, gene body deciles, to 10Kb 3’-TTS for each dataset (protein) in the order (left to right) in which the datasets were defined as arguments for the “matrix” function. We often find that the files exported from this function are especially useful for the creation of heat maps and for identifying patterns in protein binding using expectation maximization algorithms such as k-means clustering.

Statistical Tests

PeaksToGenes is designed to help the user test the hypothesis “is the binding of this protein different near one set of genes versus another?” By defining the binding profile of the protein in terms of peaks or signal ratios within gene bodies and in the flanking genomic regions, PeaksToGenes creates “populations” of scalar data points to which we can apply statistical approaches. In the current version of PeaksToGenes, the statistical tests are one-way tests contrasting the binding profile of a protein on one set of genes versus another set of genes within each relative genomic region. More specifically, PeaksToGenes will take two lists of genes and extract all of the binding data in the “1Kb upstream” region (or “promoter”) and determine whether the binding data within the two lists is significantly different. This is the repeated for each relative genomic region defined above.

By default, PeaksToGenes will not run any statistical tests because it is important for the user to understand which tests are appropriate for their data. The two tests included in this implementation of PeaksToGenes make very different assumptions about the nature of the data in each list, and ignoring these assumptions can lead to both false positive and false negative results; referred to as type I and type II statistical errors, respectively. Regardless of which combination of tests or no tests is chosen by the user, PeaksToGenes will create a tab-delimited file containing the sum, mean, and standard error of the mean (SEM) of binding data for each relative genomic region for each list of genes.

Fisher-ANOVA

The one-way parametric ANOVA test (Fisher ANOVA) provided by PeaksToGenes tests whether the means of two lists are different (Field 2007). This tests whether, on average, the binding profile is significantly different in a given genomic region near one set of genes versus another set of genes. This test may not be appropriate in all cases as it assumes that the data are normally distributed. In the case of data derived from peak intervals, this test is never appropriate as there are more values of zero than any other, and in the case of signal ratio data this must be examined on a case-by-case basis, as often the data are not normally distributed.

Wilcoxon Rank-Sum (Mann-Whitney U)

The Wilcoxon Rank-Sum (or Mann-Whitney U) Test (Wilcoxon 1945) is used to determine whether the values in one list tend to be greater than the values in the other list. In the case of PeaksToGenes, this test determines whether the binding profiles in a particular relative genomic region near one set of genes tend to have higher values than the binding profiles near another set of genes. Instead of using the scalar value from the binding profile, this test assigns a rank to each value and then determines if the sum of the ranks in each list meet the expected values under the assumption that the lists have similar data. This test is most appropriate for binding profiles derived from signal ratios and can be used for binding profiles derived from peaks data. It should be noted that using binding profiles derived from peaks data can be computationally difficult as the number of genomic regions with zero peaks is quite high and the algorithm used to rank the data will take quite a long time dealing with all the ties found with a zero value.

Implementation

PeaksToGenes is written in Modern Perl, using objective-oriented program function encapsulation style. The objective-oriented set of Perl modules imported by Moose is used to provide the syntactic sugar for the creation of classes and objects. User-defined experimental information and corresponding binding data is stored in a local SQLite database, using the Perl module DBIx::Class to form relationships between tables and form the interface between the database and the user. These design choices allow for additional Perl modules to be added by the user or in future development for the rapid addition of complementary functions.

Interface

The primary interface to PeaksToGenes is the script'`peaksToGenes.pl`'. This script is a command-line interface, which allows the user to perform all the requisite functions described herein: installing databases, annotating datasets, deleting datasets, and performing statistical contrast tests. The '`peaksToGenes.pl`' script uses command-line arguments such as, "`--annotate`", "`--contrast`", "`--processors`", or "`--test_genes`", which are used to define input files and settings.

Each of the functions used by PeaksToGenes is encapsulated as an object class by Moose, allowing the user to write their own script or Perl module to utilize the

PeaksToGenes functions and database utilizing object-oriented programming paradigms.

Input Files

When using the “signal_ratio” function, PeaksToGenes expects BED-format files of reads for both the IP sample and the input sample. We use the following BASH pseudocode to generate these BED-format files from raw reads:

```
for sample in IP Input
do
     # Map to reference genome
     bowtie [bowtie options] [ebwt_files] –S
     ${sample}.fastq > ${sample}.sam

     # Convert from SAM to BAM
     samtools view –bS ${sample}.sam > ${sample}.bam

     # Sort the BAM file
     samtools sort ${sample}.bam ${sample}_sorted

     # Remove the unsorted BAM file
     rm ${sample}.bam
```

```
        # Convert the sorted BAM file to BED format
        bamToBed –i ${sample}_sorted.bam > ${sample}.bed

        # Remove the sorted BAM file
        Rm ${sample}_sorted.bam

    done
```

Alternatively, when using the “annotate” function, PeaksToGenes expects a BED-format file of intervals an external program or algorithm has defined as “peaks”.

Error Handling

There are a multitude of safeguards written into the current implementation of PeaksToGenes designed to prevent wasted time and unexpected program behavior. The first line of defense against these problems occurs in the main PeaksToGenes class ‘`PeaksToGenes.pm`’, which ensures that all required command-line arguments are defined for a given function and are properly formatted. If there any of these parameters are in error, PeaksToGenes will return an error message to the user, and terminate execution. The second error-checking function occurs via checking the formatting of each peaks or reads file defined by the user to ensure the file adheres to proper BED-format specifications. For each error found in a given file, PeaksToGenes informs the user of the type of error and the line at which the error(s) has occurred and will then cease execution. Another error-checking function implemented in PeaksToGenes is the checking of RefSeq accessions defined for contrast testing. If there is an invalid or deprecated accession for the given genome, PeaksToGenes returns these accessions to the user, and continues execution with the valid accessions found. If too many accessions are invalid, the user can choose to terminate execution; the program will still function correctly without user intervention. If there are no valid accessions entered, PeaksToGenes will exit before trying to extract genomic annotation data and run statistical contrast tests.

Parallel Processing

CPU manufacturers are beginning to hit limits in individual processor speeds and have begun pushing towards using many cores in parallel. PeaksToGenes is designed to take advantage of the ability to utilize multiple processors anywhere parallel operations were safe and tested to increase speed of operation. Parallelization is accomplished through the Perl module Parallel::ForkManager, which relies on the Unix fork operation to carry out instructions in parallel. The ability to utilize multiple processors in parallel is especially powerful when

PeaksToGenes is used on a computing cluster. Utilizing as many as 24 (a limitation of our cluster, not PeaksToGenes) processors in parallel allows for rapid execution of the PeaksToGenes functions. Many of the signal ratio annotation functions take quite a long time to complete, so parallelization of these functions allows for faster generation of results.

Genomic Indexes

In order to understand spatial relationships between the user's dataset and genes, a meta-gene profile must be created. To facilitate this process, PeaksToGenes uses a set of BED-format coordinate files, which contain genomic intervals whose positions are relative to a given RefSeq accession. PeaksToGenes does not come with any of these indexes installed, rather it provides a function to dynamically add genomic information based on the user's needs. This allows a great deal of flexibility, and at the time of writing PeaksToGenes is capable of dynamically adding 62 RefSeq genome definitions.

To accomplish these functions, PeaksToGenes::Update::UCSC uses DBIx::Class to interact with the UCSC MySQL server to fetch the genomic coordinates of each RefSeq transcript as well as the sizes of each chromosome in the genome of interest. These coordinates are then used to define 34 distinct genomic locations relative to RefSeq transcript genomic coordinates within the limits of the chromosome sizes. Each location is written in a separate BED-format file, and the full path to the file is stored in the PeaksToGenes SQLite database by PeaksToGenes::Update.

For each relative location, PeaksToGenes creates a BED-format file with the coordinates for that relative location for every gene whose coordinates are valid. The path to these BED-format files are stored in the PeaksToGenes database, so that when the user chooses to annotate a dataset, the location of the files will be statically stored within the program and need not be known to the user. If for some reason, the user moves or deletes the BED-format relative location files, PeaksToGenes will quit during the annotation process, and inform the user that the files could not be found prompting the user to run the update function again. This will prevent unpredictable behavior and errors in interpretation downstream.

Annotation / Average Gene Profile Creation

With a set of genome-defined indexes in place, PeaksToGenes uses the intersectBed utility from the BedTools suite of command-line tools to assign information from the user's dataset into the intervals defined by PeaksToGenes::Update. This operation is primarily performed on BED-formatted reads files (generated using the pseudocode above). Alternatively, PeaksToGenes can use BED-format files corresponding to externally defined “peak” intervals.

For each relative genomic index, PeaksToGenes normalizes the number of peaks/peak scores/reads per 1Kb. This calculation is done as needed during the parsing of the results of the intersectBed command, and allows for bias to be reduced from larger sub-genomic regions such as gene body deciles or introns. This calculation further allows the raw values for each relative genomic region to be interpreted as the aggregate number of peaks/peak scores/reads per Kb.

For each peak (or read interval) found within a particular relative genomic location, PeaksToGenes::Annotate::BedTools (for peaks) or PeaksToGenes::SignalRatio::BedTools (for reads) stores in memory both the number of peaks/reads as well as the aggregate peak scores (from MACS, SPP, etc.). These large hash references are then converted into a DBIx::Class insert statement by PeaksToGenes::Annotate::Database (for peaks) or PeaksToGenes::SignalRatio (for reads), and then inserted into the PeaksToGenes database using a bulk insert.

SQLite Database

The process of generating an average gene profile is relatively quick. The issue is how to store this information in a way that is consistent, so that strict functions can be written to interpret and parse the files, and that is less prone to end-user tampering, so that PeaksToGenes will have an easier time recalling information from annotation/signal_ratio functions. We did not want to come up with an additional file format for the rapid storage and retrieval of this type of information, so we chose to use a local SQLite database to store the results of the meta-genome profiles.

Storing the data in a relational database management system (RDBMS) allows for strict data types in each field, adding an extra layer of error checking to ensure that data is properly handled and that results can be properly interpreted. Further, the use of an RDBMS allows for other means of extracting meta-genome information from PeaksToGenes database should a user decide to write their own implementation to access the database in another language (C++, Python, etc.) or through the SQLite command-line.

There are many different implementations of SQL-mediated control of RDBMS, however, we chose SQLite as we feel it the easiest to implement for end-users, especially those who do not have root/admin privileges (such as on a computing cluster). SQLite does, however, have some drawbacks in performance as compared to MySQL or PostgreSQL, which causes the time for insertion of the meta-genome profile data into the database to be quite long (SQLite, at this time, does not allow asynchronous transactions). However, we feel that the lack in performance is outweighed by the flexibility permitted by SQLite.

Results

Using PeaksToGenes to Identify Patterns in Chromatin Context Prior to Estrogen Stimulation in MCF-7 Breast Cancer Cells

To demonstrate the functions of PeaksToGenes, we chose to examine the relationships between the chromatin context of MCF-7 breast cancer cells and their response to estrogen stimulation. Of particular interest are the positions of transcription factors such as ER-alpha and CTCF relative to genes responsive to estrogen stimulation. Positional binding of ER-alpha in relation to ER-alpha-responsive genes has been investigated using other means (Carroll et al. 2006, Hurtado et al. 2011, Zwart et al. 2011) as well as the cooperative functions of

ER-alpha and CTCF on the promoters of ER-responsive genes (Ross-Innes et al. 2011). To understand what role the chromatin context may be playing in ER-response, the positions and binding intensities of the chromatin modifications histone 3 lysine 4 trimethylation (H3K4me3), histone 3 lysine 27 acetylation (H3K27ac), and histone 3 lysine 27 trimethylation (H3K27me3) were examined as well. Functional ER-alpha binding sites are associated with H3K4me3, slightly associated with H3K27ac and not significantly associated with H3K27me3 (Joseph et al. 2010). Data for ER-responsive genes, ER binding, CTCF binding, and histone marks were extracted from publicly-available data for each of the aforementioned studies.

Using PeaksToGenes, we find that genes responsive to ER-stimulation in MCF-7 breast cancer cells are significantly enriched in binding of both ER-alpha and CTCF near the TSS. Further, we observe that the chromatin near the TSS of ER-responsive genes is in a predominately open conformation, as H3K4me3 is highly enriched at these loci. PeaksToGenes provides a rapid means to perform a common form of epigenomics analysis, which produces results similar to those produced by home-grown scripting and code.

ER-responsive genes are significantly enriched in association with ER-alpha in proliferating MCF-7 cells

The binding profile of ER-alpha in MCF-7 cells was defined using the PeaksToGenes “signal_ratio” function with input scaling using the sequence enrichment scaling algorithm. RefSeq mRNAs defined as responsive to ER-stimulation after 3 hours were used as the list of “test_genes”, while the remaining RefSeq mRNAs on the Affymetrix HGU133A Plus 2 array were used as the “background_genes”. Here we observe that regions near the TSS of genes responsive to ER-stimulation are significantly enriched in binding of ER-alpha **(Figure 5.1)**. This is consistent with similar analyses done with ER-alpha (Carroll et al. 2006, Hurtado et al. 2011).

Figure 5.1 ER-responsive genes are significantly enriched in ER-alpha binding near the TSS

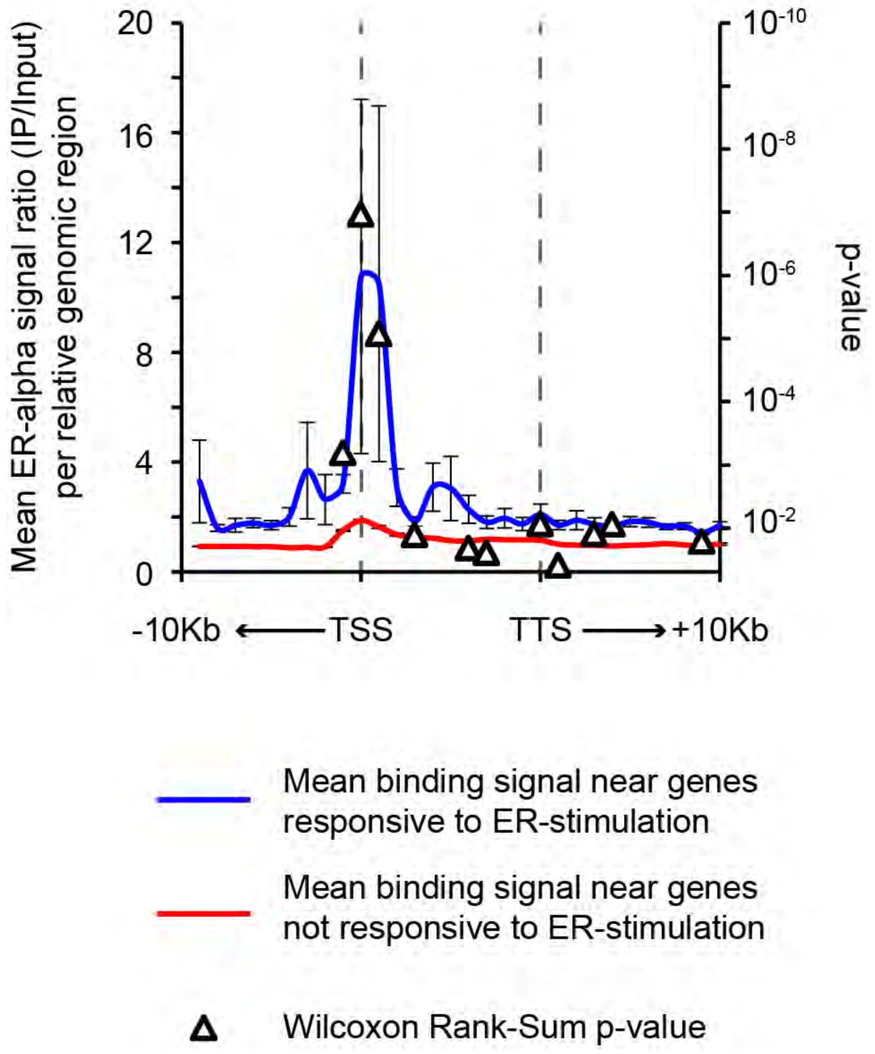


Figure 5.1 ER-responsive genes are significantly enriched in ER-alpha binding near the TSS

PeaksToGenes average gene profile of the binding of ER-alpha as measured by the ratio of IP reads over input reads. Blue line is the mean ER-alpha signal ratio per relative genomic region near genes responsive to ER-stimulation, while the red line is the same data near genes defined as non-responsive to ER-stimulation. Triangles represent the Wilcoxon Rank Sum p-value generated by using this test to compare the binding of ER for ER-responsive genes versus the non-responsive genes in each relative genomic region. Error bars are SEM.

Genes responsive to ER stimulation are enriched in CTCF binding near the TSS

ER-alpha and CTCF bind many of the same genomic regions, and these co-binding events are highly functional in response to ER stimulation (Ross-Innes et al. 2011). Using PeaksToGenes, we can observe that the same genes, which have significant binding of ER-alpha near the TSS are also significantly associated with CTCF binding near the TSS **(Figure 5.2)**. It is interesting that some of the genes that are not responsive to ER stimulation (red line) show an increase in CTCF binding near the TSS, which suggests that CTCF may have TSS-centric functions independent of ER-alpha binding in MCF-7 cells. CTCF typically binds to open regions of chromatin that are DNAse I-sensitive, and it is thought that these CTCF binding events promote the interactions of distal enhancer regions with promoter regions (Sanyal et al. 2012, Merkenschlager and Odom 2013). Therefore, the ER-independent binding of CTCF within promoter regions is not unexpected.

Genes Responsive to Estrogen Signaling are Associated with Open Chromatin Marks

To further characterize the pre-estrogen-stimulation chromatin state of genes that are responsive to estrogen, we looked the chromatin modifications H3K4me3 and H3K27ac, which are associated with transcriptional activation, and H3K27me3, commonly associated with transcriptional repression. We used PeaksToGenes to address the extent to which more binding of these histones was differentially associated with genes responsive to estrogen signaling. Functional ER-alpha binding near the TSS of genes is associated with H3K4me3 (Joseph et al. 2010); using PeaksToGenes, we observe the same association between H3K4me3 and genes responsive to ER-stimulation (which has strong ER-binding in near the TSS) **(Figure 5.3)**. Because we have chosen to plot these three marks on the same scale, it is difficult to resolve some of the finer differences in H3K27ac and H3K27me3 binding between the ER-responsive and non-responsive genes. However, while the magnitudes of the mean differences are not great, we do observe statistically significantly increased levels of H3K27ac and reduced levels of H3K27me3 near the TSS of ER-responsive genes. These results, combined with the H3K4me3 results, suggest that ER-responsive genes have an open chromatin configuration near the TSS.

Figure 5.2 Genes responsive to ER stimulation are strongly associated with CTCF binding near the TSS.

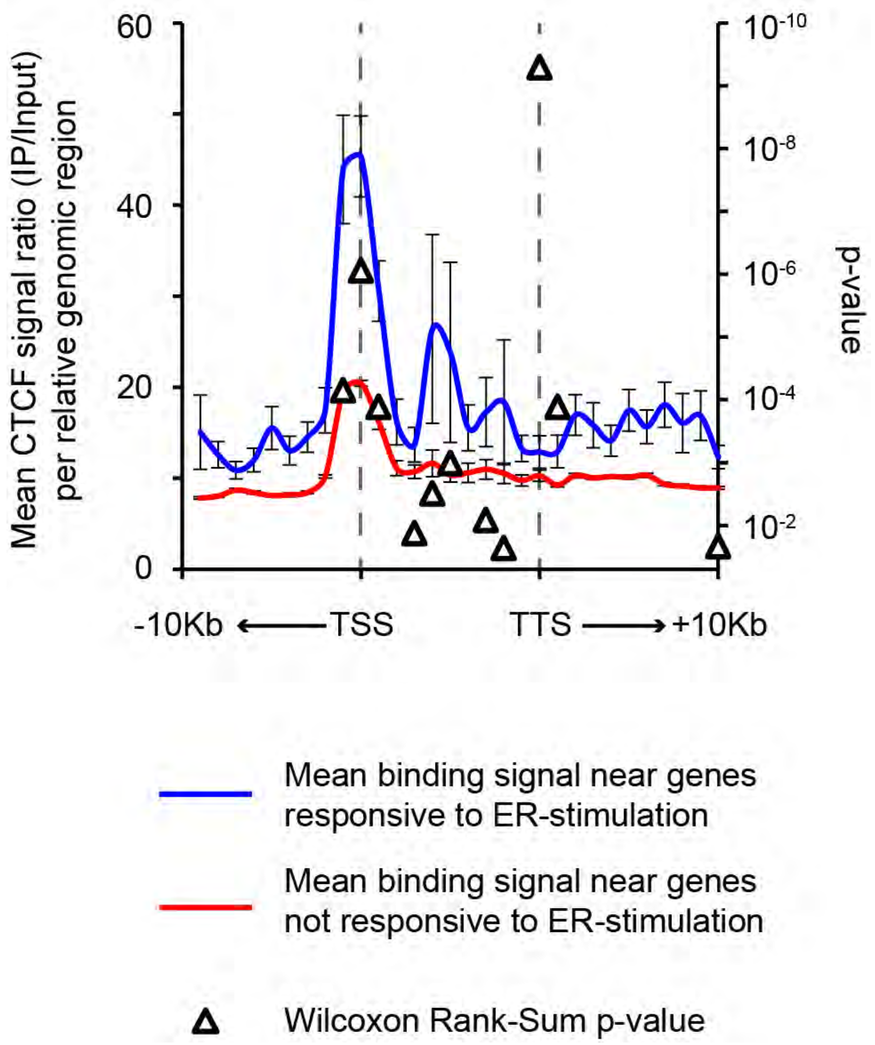


Figure 5.2 Genes responsive to ER stimulation are strongly associated with CTCF binding near the TSS.

PeaksToGenes average gene profile of the binding of CTCF as measured by the ratio of IP reads over input reads. Blue line is the mean CTCF signal ratio per relative genomic region near genes responsive to ER stimulation, while the red line is the same data near genes defined as non-responsive to ER stimulation. Triangles represent the Wilcoxon Rank Sum p-value generated by using this test to compare the binding of ER for ER-responsive genes versus the non-responsive genes in each relative genomic region. Error bars are SEM.

Figure 5.3 The TSS of genes responsive to ER stimulation are enriched in open chromatin marks

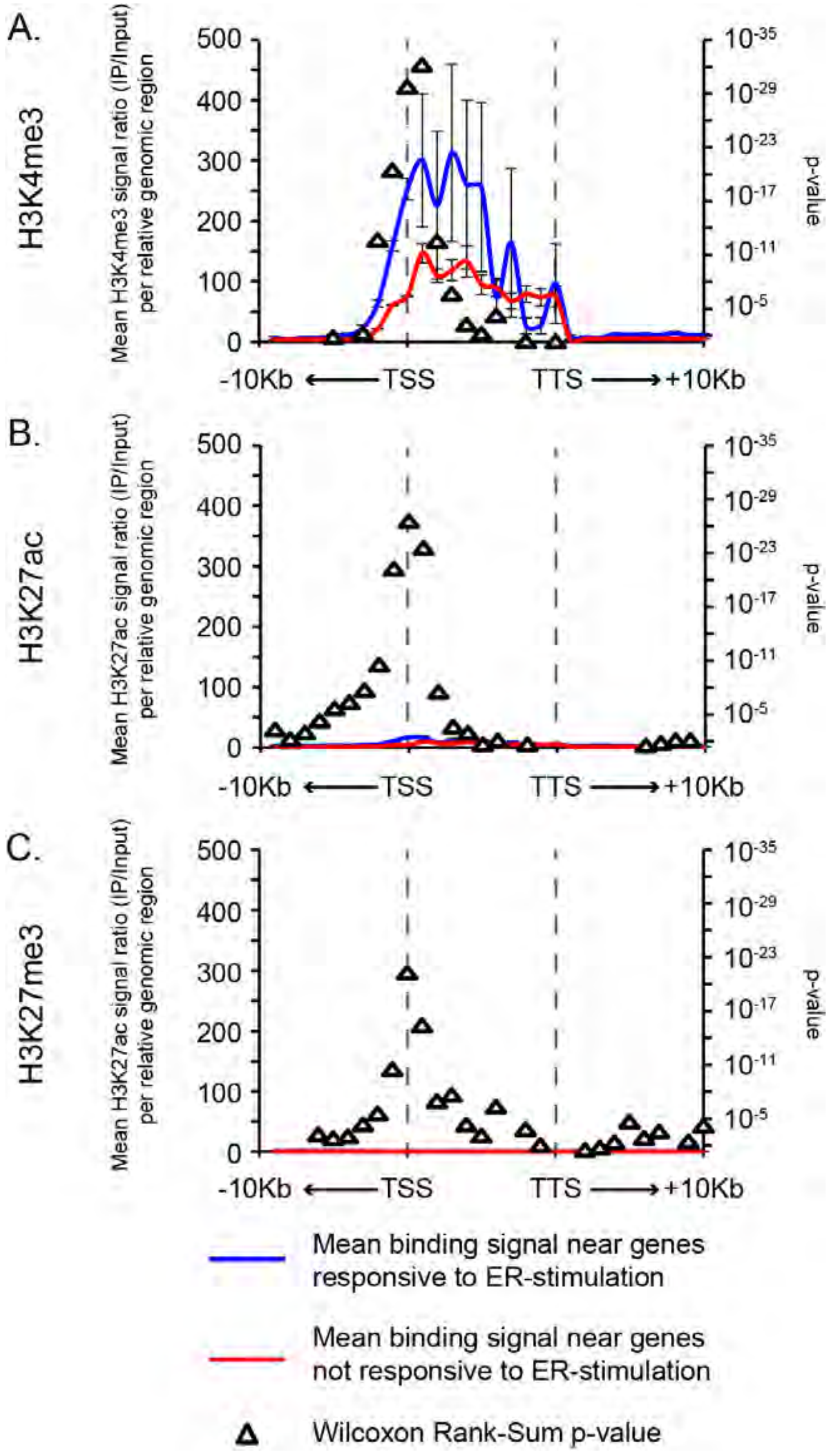


Figure 5.3 The TSS of genes responsive to ER stimulation are enriched in open chromatin marks

PeaksToGenes average gene profile of the binding of **(A)** H3K4me3, **(B)** H3K27ac and **(C)** H3K27me3 as measured by the ratio of IP reads over input reads. Blue line is the mean histone mark signal ratio per relative genomic region near genes responsive to ER-stimulation, while the red line is the same data near genes defined as non-responsive to ER-stimulation. Triangles represent the Wilcoxon Rank Sum p-value generated by using this test to compare the binding of ER for ER-responsive genes versus the non-responsive genes in each relative genomic region. Error bars are SEM.

Discussion

PeaksToGenes provides an open source framework for the analysis of epigenomic data in a gene-centric manner. The addition of statistical testing provides access to robust statistical methods without having to restructure the data. PeaksToGenes also provides convenient functions to write the average gene profiles to file so the user may perform their own analyses. We further demonstrate the ability of PeaksToGenes to reproduce previous results using ER stimulation in MCF-7 cells as a model.

Using publicly available datasets we demonstrate a potential use case for PeaksToGenes. The analysis provided by PeaksToGenes allows for rapid insight into the binding events near a subset of genes. In this case, we examined the average profile of genes that are defined as responsive to ER-stimulation (Carroll et al. 2006). ER-alpha transcriptional activity occurs near the TSS of actively expressed genes marked by H3K4me3, and cooperatively functions with CTCF (Joseph et al. 2010, Ross-Innes et al. 2011).

PeaksToGenes is not designed to be the only form of analysis done with epigenomic data; rather, PeaksToGenes can be considered as a complementary analysis approach. In no way is PeaksToGenes a replacement for such approaches as chromosome segmentation or hidden Markov modeling (Ernst and Kellis 2012, Hoffman et al. 2012); however, PeaksToGenes requires minimal computational aptitude to find more general patterns and associations between protein binding and genes.

Program Requirements

- Perl >= 5.14.2
- Perl Module dependencies (see ‘INSTALL’ file for details)
- BEDtools
- SQLite3
- MySQL
- OS X or Linux

Data sources

ER-responsive genes

ER-responsive genes in MCF-7 cells are genes defined as ER-responsive after 3 hours of ER stimulation using Affymetrix arrays (Carroll et al. 2006).

ER-alpha ChIP-seq

ER-alpha binding profile in MCF-7 cells was extracted from (Hurtado et al. 2011).

CTCF and H3K4me3 ChIP-seq

CTCF and H3K4me3 binding profiles in MCF-7 cells were downloaded from ENCODE, and produced by the University of Washington ENCODE group (ENCODE Project Consortium 2011).

H3K27ac and H3K27me3 ChIP-seq

H3K27ac and H3K27me3 binding profiles in MCF-7 cells were downloaded from ENCODE and produced by the Stanford / Yale / USC / Hardvard (SYDH) (ENCODE Project Consortium 2011).

CHAPTER 6: DISCUSSION

Isolation of the in situ nuclear matrix

The nuclear matrix and previous methods of biochemical fractionation has been quite a controversial concept, and has resulted in the nuclear matrix being maligned as “operationally defined”. Much of the scientific disagreement stems from the development of a myriad of protocols to isolate the nuclear matrix and thereby questions the functional relevance of this insoluble fraction. While compositions of the “matrix” isolated in most procedures are similar in terms of the relative amounts of RNA, DNA and proteins, the ultrastructural organization of the *in situ* nuclear matrix is only preserved under specific circumstances (Belgrader et al. 1991).

The physical appearance of many nuclear bodies is largely tethered to functional activity, suggesting that organizational parameters are synergistically derived from physiological demands. Nuclear bodies form *de novo*, utilize RNA as a structural scaffold, and are self-organizing (reviewed in (Dundr and Misteli 2010)). Given the dynamic relationship between functional demand and structural organization, it seems counterintuitive to expect that extracting these nuclear bodies and the nuclear matrix while destroying native structure will result in a biochemically faithful isolate.

The nuclear matrix is enriched in newly transcribed RNA, and in close physical proximity to open, active chromatin (Bachellerie et al. 1975, Fakan and Nobis 1978, Fakan and Hughes 1989). The Laemli group developed a protocol to isolate matrix-associated DNA, which relies on LiS buffer for extraction (Mirkovitch et al. 1984, Izaurralde et al. 1988). Due to differences in the structural appearance of the residual structures from the “high salt” and “LiS” matrix preparations, DNA isolated from these preparations were defined as matrix associated regions “MARs” and scaffold associated regions “SARs” respectively (Belgrader et al. 1991). This marked one of the major splits in the nuclear matrix field in terms of the functional characterization of NM-DNA.

For both “high salt” and “LiS” matrix preparations, without some form of stabilization, many matrix-DNA interactions are lost. Somehow, the experiments demonstrating loss of matrix-DNA attachments were not considered in the majority of attempts to understand the role of NM-DNA in gene expression regulation; most of these protocols are reliant upon methods that destabilize matrix-DNA interactions in the preparation of non-stabilized “nuclear halos” prior to nuclease digestion and NM-DNA isolation (Maya-Mendoza and Aranda-

Anzaldo 2003, Linnemann et al. 2008, Keaton et al. 2011). Given the functionally-derived organizational properties of nuclear bodies, and the dynamic matrix-DNA interactions (Misteli 2010, Dundr and Misteli 2010), the question is then, how can native DNA-matrix interactions be extracted using a protocol that does not preserve structural integrity or matrix-DNA interactions? As we are interested in understanding how malignant nuclear disorganization and nuclear matrix transcriptional functions are related, we realized how critical it would be to ensure that the methods used to isolate matrix-associated DNA would preserve the structural integrity of the nuclear matrix and associated nuclear bodies.

As mentioned above, there are two approaches that are commonly used to isolate matrix DNA, both of which rely on the intermediate “nuclear halo” preparation prior to nuclease digestion. Using the high salt extraction protocol, MARs are enriched, which are experimentally defined as transcriptionally repressive (Maya-Mendoza and Aranda-Anzaldo 2003, Rivera-Mulia and Aranda-Anzaldo 2010, Trevilla-García and Aranda-Anzaldo 2011). In contrast, using the “LiS” extraction method, SARs are enriched, which are associated with transcriptionally-active genes (Heng et al. 2004, Keaton et al. 2011). One study has shown that using both methods side-by-side in the same cell line to extract MARs and SARs results in the isolation of DNA sequences enriched near transcriptionally silent or active genes respectively (Linnemann et al. 2008). The aforementioned studies used either tiling-PCR or hybridization arrays of a subset of the genome to measure the positional enrichment of DNA isolated from the nuclear matrix; however, we wanted to generate completely unbiased, genome-wide measurements of matrix-associated DNA positional enrichment. The investment of time, energy, and money to perform a genome-wide screen using both the high salt and LiS methods in parallel seemed excessive, and the intermediate “nuclear halo” step seemed antithetical to our goal of preserving the ultrastructural organization of the *in situ* nuclear matrix. We therefore began looking at alternative approaches to isolating DNA associated with the nuclear matrix.

Two approaches became immediately attractive to us in terms of their ability to preserve nuclear integrity. One is the use of agarose-suspended digestion and electroelution-mediated extraction developed by Jackson and Cook (Jackson and Cook 1988). The other uses formaldehyde to stabilize matrix-DNA interactions prior to nuclease digestion and salt extraction (Nickerson et al. 1997). The formaldehyde-stabilized method was more practical in terms of our ability to execute this procedure on a large enough scale to collect sufficient material for deep-sequencing library preparation, so we used this approach as a starting point for the optimization of the methods described above. This method of matrix isolation results in the isolation of a fibrogranular network of RNPs that is structurally indistinguishable from the that which is observed in an intact nucleus via regressive EDTA staining and electron microscopy (Nickerson et al. 1997). This level of structural preservation, combined with reversibly cross-linked stabilization, satisfied our requirements for an experimental approach to isolate native DNA-matrix interactions.

Roles of RUNX1 and RUNX2 in recruitment of DNA to the nuclear matrix

RUNX proteins have a conserved amino acid sequence in the C-terminal domain, the nuclear matrix targeting signal (NMTS), which is required for interaction with the nuclear matrix and transcriptional activities (Zeng et al. 1997, 1998, Zaidi et al. 2001, 2006). The DNA-binding domain of RUNX proteins is located in the N-terminal half of the protein and is not required for interaction with the nuclear matrix (Zeng et al. 1997, van Wijnen et al. 2004). It is not known whether RUNX proteins recruit genomic DNA to the nuclear matrix, however, it has been hypothesized that this is a function of RUNX proteins.

Given that RUNX transcriptional activity was observed to be coupled to association with the nuclear matrix, and the strong associations we observed between actively expressed gene promoters and RUNX1 binding, we expect to see cooperative enrichment of matrix-associated DNA and RUNX1 in the MDA-MB-231 cells. However, we did not see any relationships between the two datasets nor with RUNX2 and NM-DNA (data not shown), which would lead us to believe that RUNX proteins are not involved in the recruitment of DNA to the nuclear matrix in MDA-MB-231 cells.

The MDA-MB-231 cells are perhaps not the best cell line to try to understand whether RUNX proteins recruit DNA to the nuclear matrix, as the normal associations between transcriptional activity and matrix-association are not well represented in the malignant MDA-MB-231 cells. This may, in part, explain why RUNX functions in breast cancer cells are so different from what is observed in hematopoietic and osteoblastic cells. It would be interesting to execute a similar study in the context of a blood or bone cell line to understand how related NM-DNA and RUNX binding are, as a normal organizational context may be more representative of RUNX-matrix functions.

Computational tools for analysis of nuclear organization

In this dissertation, we describe the development and application of FoxPrimer and PeaksToGenes, which are used to interrogate experimentally derived nuclear organization data. While the specific applications of these tools may differ, FoxPrimer and PeaksToGenes are designed to provide results to the user in a rapid yet thorough manner and are open source software packages. We suggest that the shared design philosophy of FoxPrimer and PeaksToGenes will allow groups utilizing next-generation type sequencing approaches a means to address some of their specific questions using free, well-documented software.

Appendices

CHAPTER A1 HSA-MIR-30C PROMOTES THE INVASIVE PHENOTYPE OF MDA-MB-231 CELLS

Authors and contributions

Jason R. Dobson, Hanna Taipaleenmäki, Yu-Jie Hu, Deli Hong, Andre J. van Wijnen, Janet L. Stein, Gary S. Stein, Jane B. Lian, Jitesh Pratap.

qPCR for miRNAs performed by JRD and HT.

qPCR for RUNX2 and NOV performed and analyzed by JRD.

siRNA and miRNA transfections performed by JRD.

Invasion assays performed by JRD and HT, quantified by JRD.

Western blotting performed by JRD.

MDA-MB-231 cells stably overexpressing Runx2 and Runx2-RY engineered by DH.

Initial cancer-centric screens for hsa-mir-30c targets in MDA-MB-231 cells performed by YH and JP (data not shown).

Ontological and qPCR screen designed and executed by JRD.

Introduction

Breast cancer is the most commonly diagnosed disease among women. Aggressive breast cancers have high potential to become metastatic, a transition that makes clinical intervention very difficult. Therefore, understanding the molecular mechanisms that have the potential to facilitate a more invasive or metastatic state are critical to understand the progression to metastatic breast cancer.

A frequent site of breast cancer metastasis is bone. Upon metastasizing to bone, breast cancer cells typically participate in what is known as the "vicious cycle" of osteolysis. During this process, breast cancer cells push the normally homeostatic signals of bone resorption and bone mineralization towards a more resorptive state, thereby causing osteolysis, bone density loss and ultimately pathological fractures (Guise et al. 2006). At a molecular level, many of the genes upregulated in bone metastatic breast cancer cells are typically involved in the process of bone differentiation, such as MMP9, MMP13, VEGF, and PTHLH (Pratap et al. 2006). High expression of these genes is clinically associated with poor outcome and prognosis for patients (Kingsley et al. 2007).

Runx2 is a transcription factor required for the development of ossified bone (Komori et al. 1997). In breast cancer cell lines, Runx2 transcriptionally regulates the expression of many genes known to be important for breast cancer metastasis and bone osteolysis and is involved in promoting the osteolytic properties of MDA-MB-231 cells as observed in intra-tibial orthotopic injection studies (Barnes et al. 2004, Javed et al. 2005, Pratap et al. 2008).

During osteoblast differentiation, a number of small non-coding RNAs, microRNAs (including mir30c), have been characterized to target the 3'-UTR of the Runx2 transcript as a mechanism of lineage specification. When osteoblast progenitor cells commit to the osteoblast lineage, the levels of these Runx2-targeting miRNAs are significantly reduced, thereby facilitating the developmental requirement for increased Runx2 levels as a function of osteoblast differentiation (Zhang et al. 2011).

One of the many challenges in cancer is discovering intervention therapies that will more specifically target the cancer cells rather than normal cells. As has been previously described, mir30c is highly capable to target the Runx2 mRNA and reduce Runx2 protein levels (Zhang et al. 2011). We investigated the extent to which mir30c down-regulation of RUNX2 protein could reduce the invasive phenotype of MDA-MB-231 breast cancer cells. We observed that while mir30c does reduce the levels of RUNX2 protein, mir30c significantly increases the invasiveness of MDA-MB-231 cells. We used a qPCR screen of in silico predicted targets of mir30c whose ontological associations are related to the invasive phenotype. We identify NOV as the target of mir30c, which is likely playing a role in the increased invasiveness of MDA-MB-231 cells. The functional consequences of altering NOV through siRNA are consistent with the predicted function of repressing the invasiveness of MDA-MB-231 cells. These studies demonstrate the potential for miRNAs to have quite undesirable effects *in vivo* due to the multiplicity of potential targets and affected pathways.

Results

hsa-mir-30c promotes the invasiveness of MDA-MB-231 breast cancer cells

In vitro, MCF-7 breast cancer cells are much less invasive than MDA-MB-231 cells (Morini et al. 2000). We observed statistically significant higher levels of endogenous hsa-mir-30c in the more invasive breast cancer cell line MDA-MB-231 compared to the MCF7 cell line **(Figure A.1 A)**.

In osteoblasts, mmu-mir-30c post-transcriptionally regulates the levels of Runx2 (Zhang et al. 2011). In the MDA-MB-231 cell line, trans-expression of hsa-mir-30c or an anti-mir sequence designed to inhibit the function of endogenous hsa-mir-30c causes the expected changes in the RUNX2 proteins levels **(Figure A.1 B)**. Therefore, as in osteoblasts (Zhang et al. 2011), RUNX2 is regulated by hsa-mir-30c.

To assess the effect of hsa-mir-30c or the anti-mir on the invasive potential of MDA-MB-231 cells, we transfected the cells for 48 hours and then loaded the cells into transwell culture plates with or without a layer of MatriGel. Cells were placed in the top of the transwell, while conditioned osteoblast media was placed in the bottom well as a chemoattractant for the cells to mimic a bone environment. After the cells were permitted to either migrate or invade towards the osteoblastic medium, cells were fixed and stained with HEMA3 **(Figure A.1 C)**. We observe that hsa-mir-30c expression levels are linked with the invasive potential of MDA-MB-231 breast cancer cells. We observe that the anti-mir control affects the endogenous levels of RUNX2, so the anti-mir reagents were not used for the rest of this study as we were concerned about off-target effects.

Figure A.1 hsa-mir-30c promotes the invasiveness of MDA-MB-231 cells and regulates RUNX2 levels

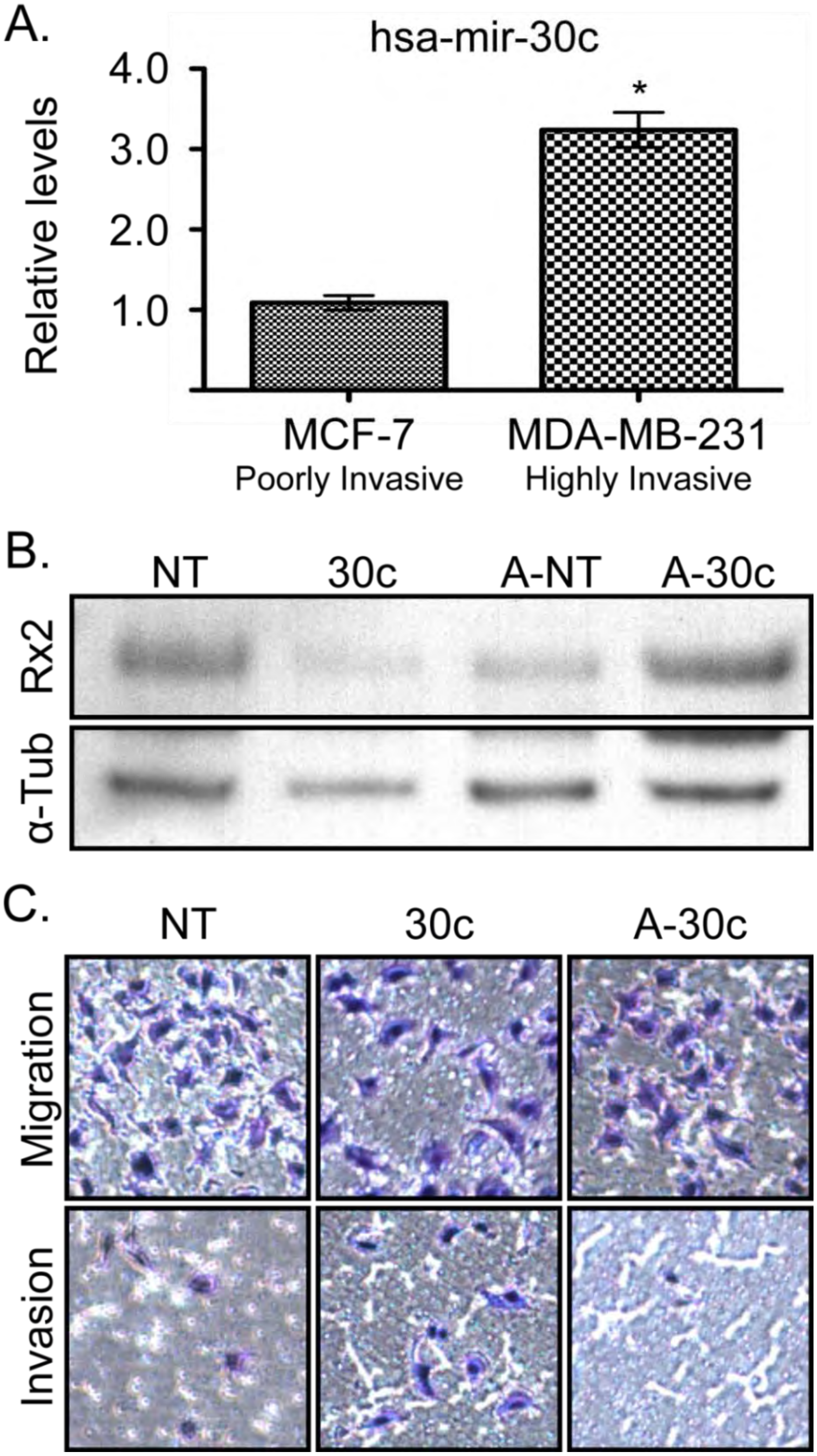


Figure A.1 hsa-mir-30c promotes the invasiveness of MDA-MB-231 cells and regulates RUNX2 levels

(A) qPCR for hsa-mir-30c in MCF-7 and MDA-MB-231 cells, normalized to U6 RNA. Mean and SEM (error bars) for two technical replicates of two biological replicates. * = Student's t-test p-value < 0.05. **(B)** Representative Western blots of lysates of MDA-MB-231 following 48 hours of transient transfection of non-targeting miRNA (NT), hsa-mir-30c (30c), non-targeting anti-miRNA (A-NT) and anti-hsa-mir-30c (A-30c). Top blot: RUNX2, bottom blot: α-Tubulin.**(C)** Representative Matrigel invasion assay in MDA-MB-231 cells following 48 hours of transient transfection of non-targeting miRNA (NT), hsa-mir-30c (30c) and anti-hsa-mir-30c (A-30c), cells are stained with HEMA3.

The canonical 5-prime end of hsa-mir-30c, not the "star" or 3-prime end of hsa-mir-30c, promotes the invasiveness of MDA-MB-231 cells

During normal miRNA biogenesis, the stem-loop structure that is termed the pre-miRNA is cleaved and only one end of the longer pre-miRNA is integrated into the Dicer complex (Kim 2005). In the case of hsa-mir-30c, the 5'-end of the stem-loop is most commonly detected as the mature form of the miRNA (Griffiths-Jones et al. 2006). When the "star" strand or in the case of hsa-mir-30c the 3'-end of the stem-loop is utilized, a unique set of mRNAs can be targeted by the Dicer complex. The phenomenon of alternate utilization of the non-canonical part of the pre-miRNA has been observed in leukemic cells (Kuchenbauer et al. 2011), and we investigated the extent to which this kind of transformation may be occurring in MDA-MB-231 metastatic breast cancer cells.

To test whether the hsa-mir-30c-mediated effect on the invasiveness of MDA-MB-231 cells was due to the canonical miRNA strand, we transfected either hsa-mir-30c-5p (canonical) or hsa-mir-30c-3p (star) to evaluate potential changes in the invasive properties of MDA-MB-231 cells. While hsa-mir-30c-3p does inhibit the migration of MDA-MB-231 cells, it does not affect the invasiveness **(Figure A.2 A-C]**. Further, using primers specifically designed to detect either the 5'- or 3'-form of hsa-mir-30c, we observe that the ratio of the endogenous levels of 5'-form are several hundred fold higher than the 3'-form in both MDA-MB-231 and

MCF-7 breast cancer cells **(Figure A.2 D)**. These results suggest that the phenomenon of “star”-strand miRNA activity does not have a major contribution to the invasiveness of MDA-MB-231 cells.

Figure A.2 The 5'-end of the hsa-mir-30c hairpin is the predominant mature miRNA detected in and promoting the invasiveness of MDA-MB-231 cells

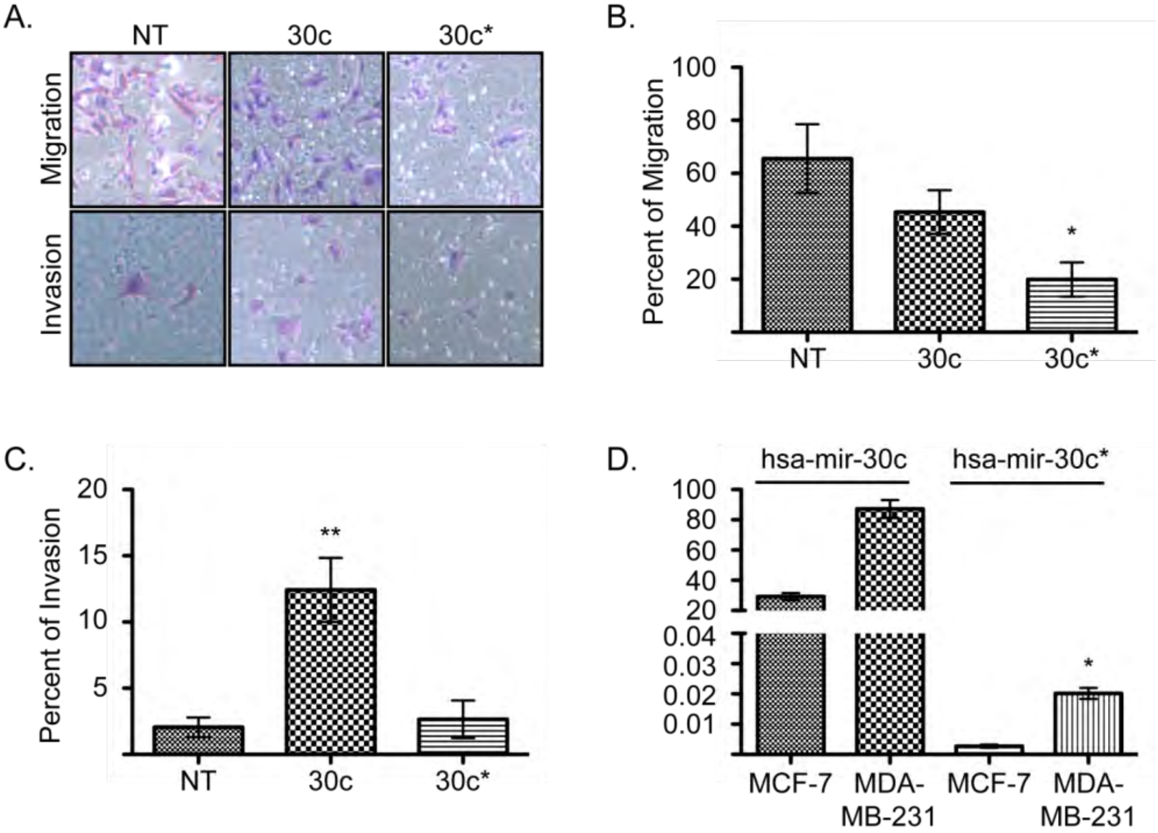


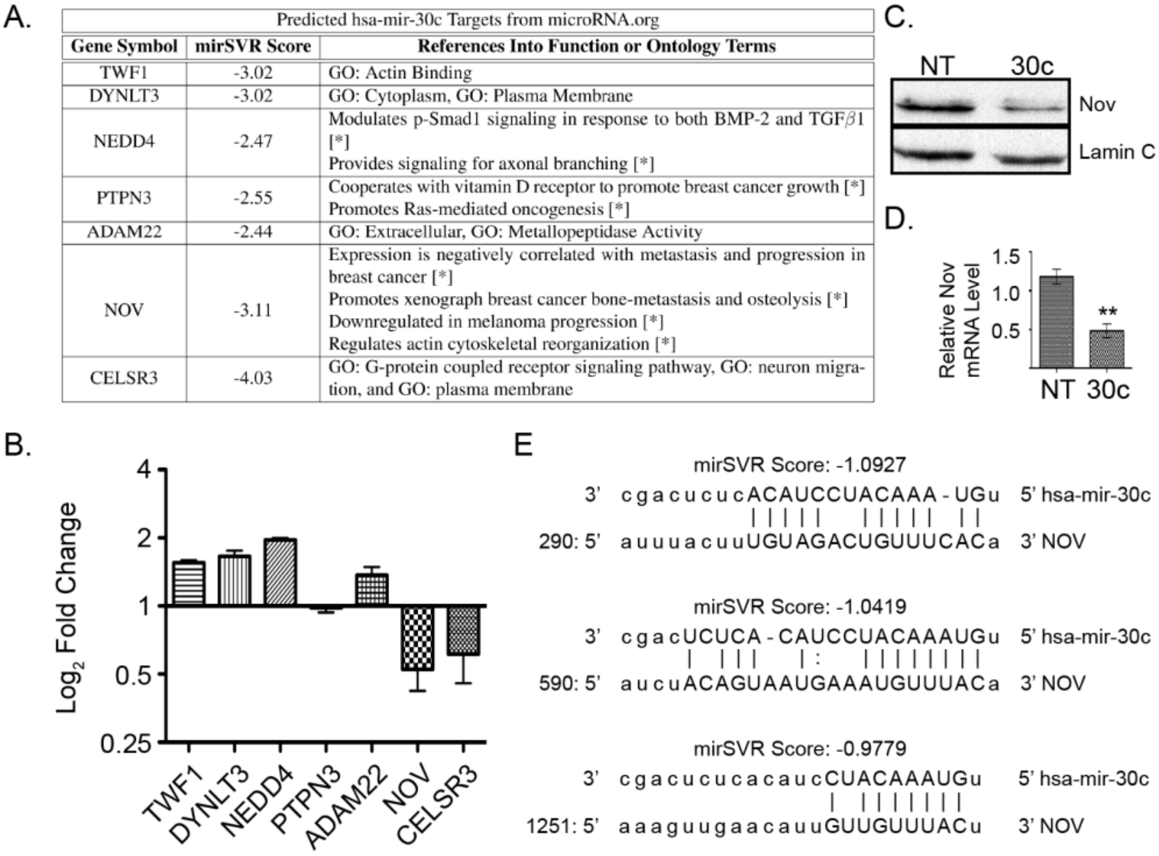
Figure A.2 The 5'-end of the hsa-mir-30c hairpin is the predominant mature miRNA detected in and promoting the invasiveness of MDA-MB-231 cells

(A) Representative images Matrigel invasion assay following 48 hours of transient transfection of siRNA of cells stained with HEMA3. **(B)** Percent of cells that migrated through the control inserts, defined by count of stained cells on the bottom of the inserts as a percent of cells loaded into inserts. Mean and SEM (error bars) for four technical replicates each of two biological replicates. * = p-value from one-way ANOVA followed by paired t-tests < 0.05. **(C)** Percent of cells that invaded through the Matrigel inserts, normalized to the number of cells that migrated through the control inserts. Mean with SEM (error bars) for four technical replicates each of two biological replicates. ** = p-value from ANOVA with paired follow-up t-tests < 0.01. **(A-C)** NT = non-targeting miRNA, 30c = hsa-mir-30c, and 30c* = hsa-mir-30c-3p. **(D)** qPCR detection of endogenous hsa-mir-30c (hsa-mir-30c-5p) and hsa-mir-30c* (hsa-mir-30c-3p) in MCF-7 and MDA-MB-231 cells. Mean with SEM (error bars) for two technical replicates each of two biological replicates normalized to U6 snRNA using the delta-Ct method. * = p-value from Student's t-test comparing hsa-mir-30c (hsa-mir-30c-5p) and hsa-mir-30c* (hsa-mir-30c-3p) in MDA-MB-231 cells < 0.05.

Screening via ontological terms and qPCR reveals Nov to be a target of hsa-mir-30c in MDA-MB-231 breast cancer cells

To identify potential targets of hsa-mir-30c participating in the invasive phenotype associated with hsa-mir-30c expression in MDA-MB-231 cells, we performed a screen in three steps: 1) generate a list of potential targets based on seed sequence targeting potential using the top 300 human mRNA targets from microRNA.org; 2) Filter the list based on known functions and ontological terms for mRNAs that code for proteins associated with invasion, adhesion, or migration, as well as mRNAs that code for transcription factors **(Figure A.3 A)**; 3) qPCR screen for functional targets measuring the relative mRNA levels in MDA-MB-231 cells after being transfected with either non-targeting miRNA or hsa-mir-30c. We observe NOV to be the most downregulated upon transfection of hsa-mir-30c **(Figure A.3 B)**. Furthermore, the protein levels of NOV were reduced by hsa-mir-30c transfection **(Figure A.3 C)**. The observation that NOV mRNA is reduced upon transfection of hsa-mir-30c is highly reproducible as observed in multiple biological replicates **(Figure A.3 D)**. Aligning the sequence of hsa-mir-30c with the sequence of the NOV 3’-UTR, mirSVR predicts three potential binding sites for hsa-mir-30c **(Figure A.3 E)**. Further, hsa-mir-30c is unique among the mir-30 family members, as a unique site is predicted for hsa-mir-30c that is not shared by the other family members **(Figure A.4)**. Based on ontological terms associated with Nov, and the reduction of NOV protein levels in response to hsa-mir-30c levels, we hypothesize that NOV may be involved in the regulation of MDA-MB-231 invasiveness.

Figure A.3 Predictive, ontological, and qPCR screen for hsa-mir-30c reveals NOV as a target of hsa-mir-30c



Predicted hsa-mir-30c Targets from microRNA.org

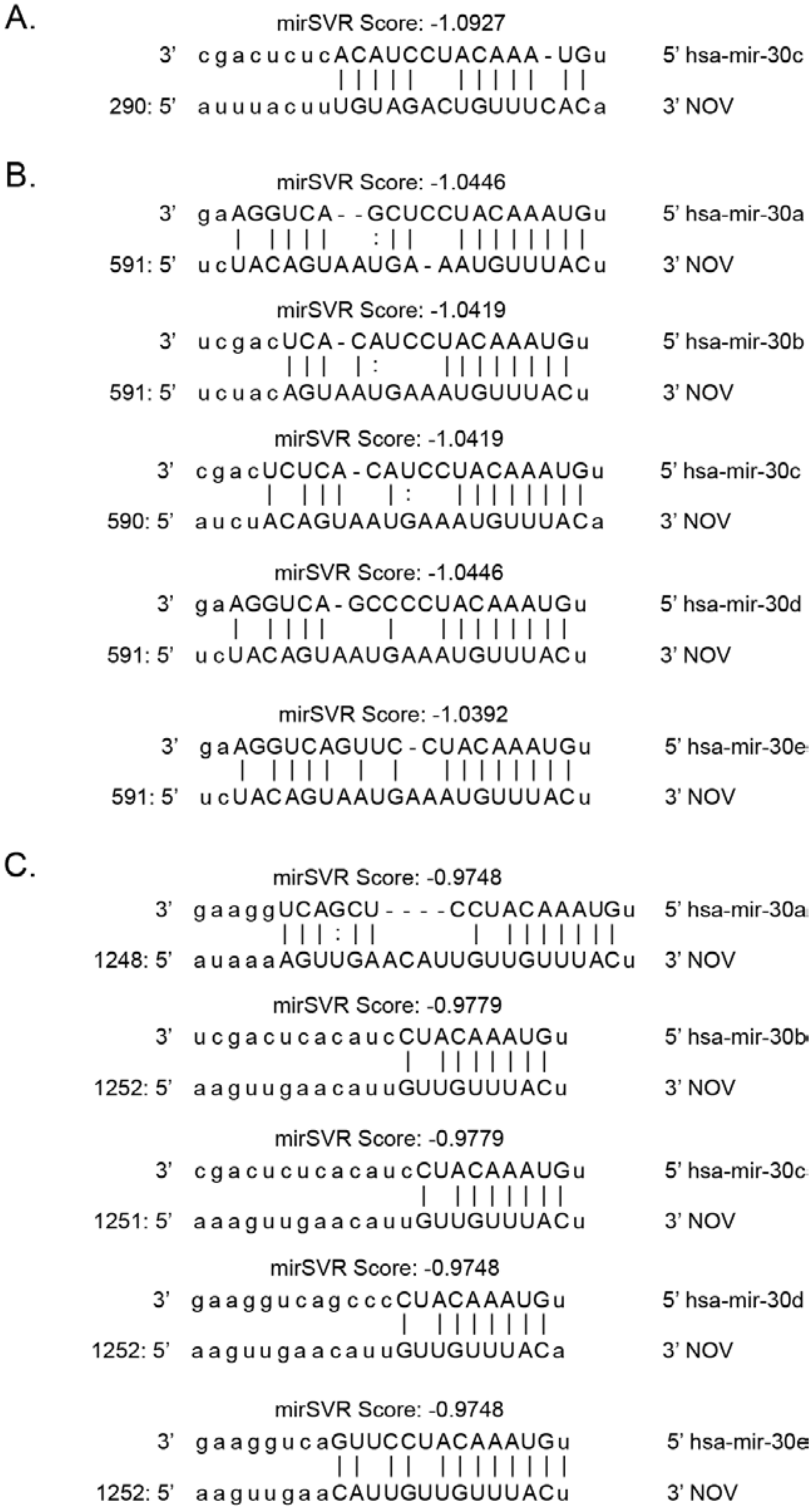
Gene Symbol	mirSVR Score	References Into Function or Ontology Terms
TWF1	-3.02	GO: Actin Binding
DYNLT3	-3.02	GO: Cytoplasm, GO: Plasma Membrane
NEDD4	-2.47	Modulates p-Smad1 signaling in response to both BMP-2 and TGFβ1 [*] Provides signaling for axonal branching [*]
PTPN3	-2.55	Cooperates with vitamin D receptor to promote breast cancer growth [*] Promotes Ras-mediated oncogenesis [*]
ADAM22	-2.44	GO: Extracellular, GO: Metallopeptidase Activity
NOV	-3.11	Expression is negatively correlated with metastasis and progression in breast cancer [*] Promotes xenograph breast cancer bone-metastasis and osteolysis [*] Downregulated in melanoma progression [*] Regulates actin cytoskeletal reorganization [*]
CELSR3	-4.03	GO: G-protein coupled receptor signaling pathway, GO: neuron migration, and GO: plasma membrane

A.
B.
Log2 Fold Change
4
2
1
0.5
0.25
TWF1
DYNLT3
NEDD4
PTPN3
ADAM22
NOV
CELSR3
C.
NT
30c
Nov
Lamin C
D.
Relative Nov mRNA Level
1.5
1.0
0.5
**
NT
30c
E
mirSVR Score: -1.0927
3' cgacucucACAUCCUACAAA-UGu 5' hsa-mir-30c
290: 5' auuuacuuUGUAGACUGUUUCACa 3' NOV
mirSVR Score: -1.0419
3' cgacUCUCA-CAUCCUACAAAUGu 5' hsa-mir-30c
590: 5' aucuACAGUAAUGAAAUGUUUACa 3' NOV
mirSVR Score: -0.9779
3' cgacucucacaucCUACAAAUGu 5' hsa-mir-30c
1251: 5' aaaguugaacauuGUUGUUUACu 3' NOV

Figure A.3 Predictive, ontological, and qPCR screen for hsa-mir-30c reveals NOV as a target of hsa-mir-30c

(A) Summary table of genes chosen for qPCR screen based on: mirSVR score <http://www.microrna.org> (Betel et al. 2008), NCBI GeneRIF, and ontological terms. **(B)** The $\log_2$ of the fold change in mean with SEM detection levels (hsa-mir-30c / non-targeting miRNA) is plotted for each set of primers for each transcript normalized to HPRT using delta-delta Ct method. **(C)** Representative Western blots for lysates of MDA-MB-231 cells following 48 hours of transient transfection with non-targeting miRNA (NT) or hsa-mir-30c (30c). Top blot: NOV, bottom blot: Lamin C. **(D)** qPCR for NOV levels for two technical replicates each for three biological replicates following 48 hour transient transfections of non-targeting miRNA (NT) and hsa-mir-30c (30c) normalized to HPRT using delta-delta Ct method. Mean with SEM. ** = p-value from Student's t-test < 0.01. **(E)** Alignment of hsa-mir-30c (top sequences) with the 3'-UTR of NOV (bottom sequences) with the 5'-positions within the NOV 3'-UTR being relative to the 5'-start of the 3'-UTR for each of the three predicted targeting sites. Target scores are provided by mirSVR. Uppercase letters linked with a "|" character indicates a perfect match, while uppercase letters linked with a ":" character indicates a wobble pair.

Figure A.4 Alignment of mir-30 family members on the NOV 3'-UTR



```
A.
                mirSVR Score: -1.0927
      3'  cgacucucACAUCCUACAAA-UGu   5' hsa-mir-30c
                  |||||  ||||| ||
 290: 5'  auuuacuuUGUAGACUGUUUCACa   3' NOV

B.
                mirSVR Score: -1.0446
      3'  gaAGGUCA--GCUCCUACAAAUGu   5' hsa-mir-30a
            | ||||  :||  ||||||||
 591: 5'  ucUACAGUAAUGA-AAUGUUUACu   3' NOV

                mirSVR Score: -1.0419
      3'  ucgacUCA-CAUCCUACAAAUGu    5' hsa-mir-30b
               |||  |:  ||||||||
 591: 5'  ucuacAGUAAUGAAAUGUUUACu    3' NOV

                mirSVR Score: -1.0419
      3'  cgacUCUCA-CAUCCUACAAAUGu   5' hsa-mir-30c
              | |||  |:  ||||||||
 590: 5'  aucuACAGUAAUGAAAUGUUUACa   3' NOV

                mirSVR Score: -1.0446
      3'  gaAGGUCA-GCCCCUACAAAUGu    5' hsa-mir-30d
            | ||||   |  ||||||||
 591: 5'  ucUACAGUAAUGAAAUGUUUACu    3' NOV

                mirSVR Score: -1.0392
      3'  gaAGGUCAGUUC-CUACAAAUGu    5' hsa-mir-30e
            | ||||  | |  ||||||||
 591: 5'  ucUACAGUAAUGAAAUGUUUACu    3' NOV

C.
                mirSVR Score: -0.9748
      3'  gaaggUCAGCU----CCUACAAAUGu   5' hsa-mir-30a
               |||:||     | |||||||
1248: 5'  auaaaAGUUGAACAUUGUUGUUUACu   3' NOV

                mirSVR Score: -0.9779
      3'  ucgacucacaucCUACAAAUGu     5' hsa-mir-30b
                      | |||||||
1252: 5'  aaguugaacauuGUUGUUUACu     3' NOV

                mirSVR Score: -0.9779
      3'  cgacucucacaucCUACAAAUGu    5' hsa-mir-30c
                       | |||||||
1251: 5'  aaaguugaacauuGUUGUUUACu    3' NOV

                mirSVR Score: -0.9748
      3'  gaaggucagcccCUACAAAUGu     5' hsa-mir-30d
                      | |||||||
1252: 5'  aaguugaacauuGUUGUUUACa     3' NOV

                mirSVR Score: -0.9748
      3'  gaaggucaGUUCCUACAAAUGu     5' hsa-mir-30e
                  || || ||||||||
1252: 5'  aaguugaaCAUUGUUGUUUACu     3' NOV
```

Figure A.4 Alignment of mir-30 family members on the NOV 3'-UTR

For each of the predicted sites of targeting by hsa-mir-30c on the NOV 3'-UTR **(A-C)**, if an alignment is possible, the alignment of each mir-30 member is presented. Alignment of hsa-mir-30 family members (top sequences) with the 3'-UTR of NOV (bottom sequences); the 5'-positions within the NOV 3'-UTR are relative to the 5'-start of the 3'-UTR for each of the three predicted targeting sites. Target scores are provided by mirSVR. Uppercase letters linked with a "|" character indicates a perfect match, while uppercase letters linked with a ":" indicate a wobble pair.

hsa-mir-30c regulation of NOV and invasion is independent of RUNX2

To understand whether there is cross-talk between RUNX2/hsa-mir-30c/NOV, we modulated RUNX2 levels in MDA-MB-231 cells using lentivirus-mediated stable cell lines expressing either empty vector, wild-type *Mus musculus* Runx2 or a subnuclear-targeting-deficient mutant form of *Mus musculus* Runx2, which inhibits the invasiveness of MDA-MB-231(Barnes et al. 2004, Javed et al. 2005). In overexpression conditions, we observe that neither Runx2 nor the RY-mutant form of Runx2 altered the protein levels of Nov **(Figure A.5 A)** or the levels of hsa-mir-30c **(Figure A.5 B)**. Further, we tested whether a reduction in the endogenous levels of RUNX2 in MDA-MB-231 cells via siRNA, which reduces invasiveness of MDA-MB-231 cells (Pratap et al. 2008), affects the levels of NOV protein or hsa-mir-30c. Here, we observe that Nov protein levels are not affected by RUNX2 knockdown **(Figure A.5 C)**, and that hsa-mir-30c levels, while slightly reduced, are not statistically significantly changed by RUNX2 siRNA **(Figure A.5 D)**. These results suggest that hsa-mir-30c/NOV-mediated regulation of the invasiveness of MDA-MB-231 cells is occurring through a RUNX2-independent pathway.

Figure A.5 RUNX2 does not significantly regulate the expression levels of either hsa-mir-30c or NOV

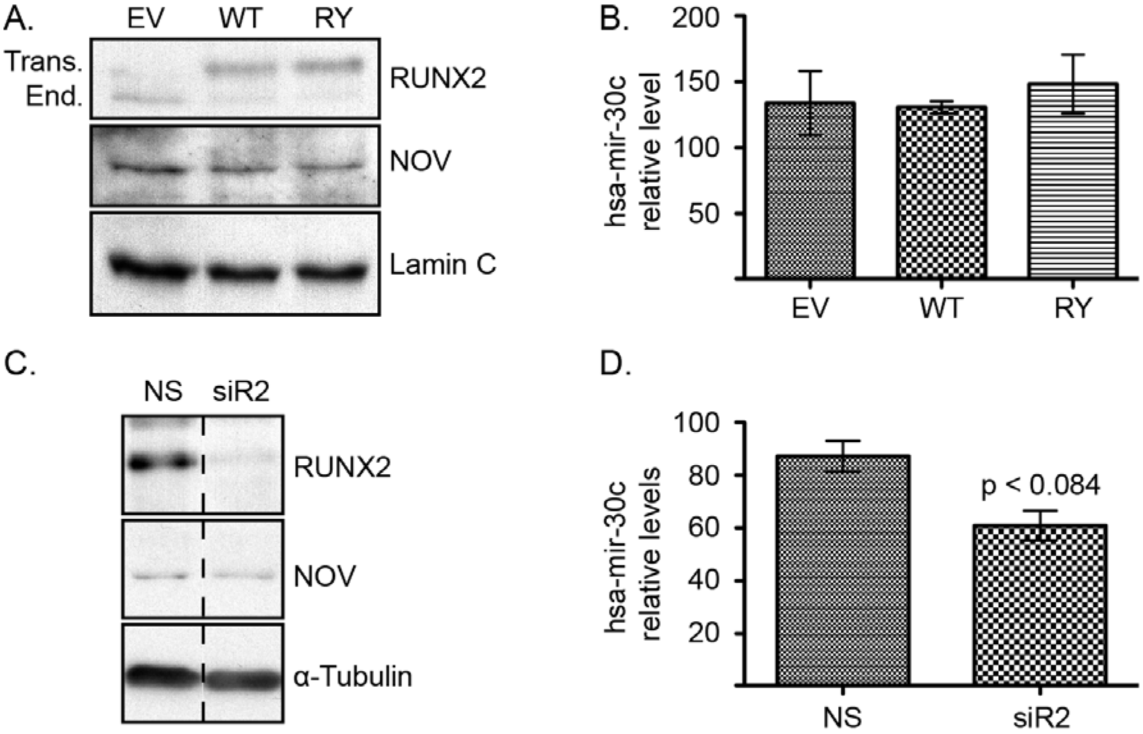


Figure A.5 RUNX2 does not significantly regulate the expression levels of either hsa-mir-30c or NOV

(A & B) Detection of NOV and hsa-mir-30c levels in MDA-MB-231 stably expressing empty vector (EV), wild-type Runx2 (WT), or R398A/Y428A mutant Runx2 (RY). **(A)** Representative Western blots of MDA-MB-231 stable cell lysates. Top blot: Runx2 (top band: transgenic murine Runx2, lower band: endogenous human RUNX2). Middle blot: NOV. Lower blot: Lamin C. B) qPCR detection for hsa-mir-30c in consecutive (N=2) passages of stable MDA-MB-231 cells normalized to U6 snRNA. **(C & D)** Detection of Nov and hsa-mir-30c following 48 hours of transient transfection of non-targeting siRNA (NS) and RUNX2 siRNA (siR2). **(C)** Representative Western blots of MDA-MB-231 lysates following 48 hours of siRNA transfection. Vertical dashed line indicates that the image of the blot was cut for figure. Top blot: RUNX2. Middle blot: NOV. Lower Blot: α-Tubulin. **(D)** qPCR detection of hsa-mir-30c levels of two technical replicates each of four biological replicates following 48 hour transfection of siRNA. p-value from Student's t-test: approaching statistical significance. B,D) Bars equal mean, error bars equal SEM.

NOV inhibits the invasiveness of MDA-MB-231 cells

To determine the involvement of NOV in the invasive phenotype of MDA-MB-231 cells, we used siRNA specific for Nov to knock down the protein **(Figure A.6 A)** and observed the effects of reduced NOV on MDA-MB-231 invasiveness. When NOV protein is significantly reduced, similar to hsa-mir-30c overexpression, the invasiveness of MDA-MB-231 cells is significantly increased **(Figure A.6 B-D)**. These results suggest that the targeting of NOV by hsa-mir-30c is a contributing factor in the invasive phenotype imparted to the MDA-MB-231 cells by hsa-mir-30c.

Figure A.6 NOV inhibits the invasiveness of MDA-MB-231 cells

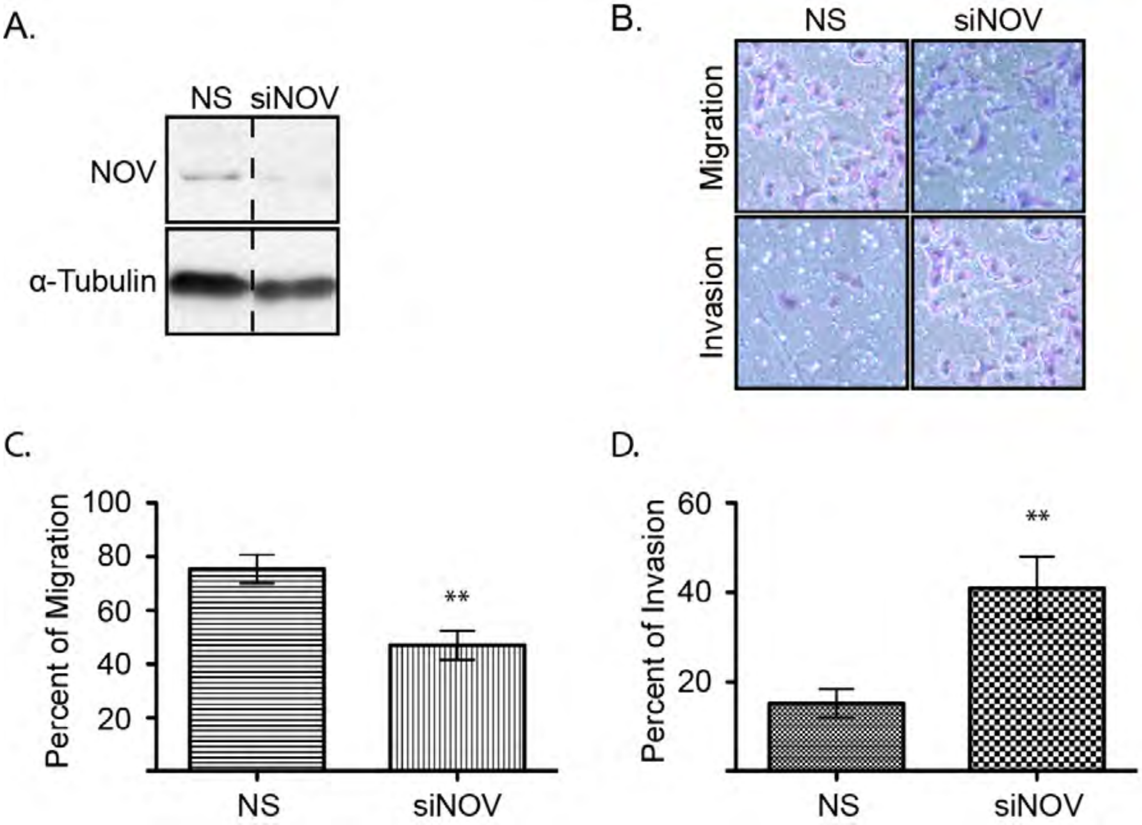


Figure A.6 NOV inhibits the invasiveness of MDA-MB-231 cells.

(A) Representative Western blot for NOV (upper blot) and tubulin (lower blot) 48 hours post-transfection with siRNA. Vertical dashed line indicates where image of gel was cut for figure. **(B)** Representative image of HEMA-3 stained cells, which migrated through either the control inserts (upper row) or Matrigel inserts (lower row) after 48 hours of transfection with siRNA. **(C)** Quantification of 4 technical replicates of 2 biological replicates measuring the percent of cells that migrated through the control inserts (100% being the number of cells loaded into the inserts). **(D)** Quantification of 4 technical replicates of 2 biological replicates measuring the percent of cells that invaded through the Matrigel normalized by the number of cells migrated through the control inserts. **(A-D)** NS = Non-silencing siRNA, siNOV = NOV siRNA. **(C & D)** Bars equal mean, error bars equal SEM. ** = Student's t-test p-value < 0.01.

Discussion

Here we identify a novel pathway by which hsa-mir-30c promotes the invasiveness of the MDA-MB-231 cell line through targeting of NOV. Concomitant reductions in the levels of both NOV and RUNX2 upon increased levels of hsa-mir-30c causes increased invasiveness of the MDA-MB-231 cells. Demonstrating the specificity of NOV's involvement in the invasive phenotype observed, transfection of siRNA targeting NOV results in significant increases in the invasiveness of MDA-MB-231 cells.

While it is clear that RUNX2 plays a major role in promoting the osteomimetic and osteolytic properties of MDA-MB-231 cells (Barnes et al. 2004, Javed et al. 2005, Pratap et al. 2008), it remains unclear how RUNX2 levels are regulated at both the transcriptional and post-transcriptional levels in breast cancer cells. mir-30c targets Runx2 in osteoblasts, and during osteoblast differentiation hsa-mir-30c levels are reduced in concert with increased levels of Runx2, a process that is required for proper mineralization of osteoblasts (Zhang et al. 2011).

The mir-30 family shares a conserved seed sequence. However, our *in silico* research suggests that seed sequence differences may give rise to selectivity of targeting among mir-30 members. This is particularly interesting taking into account a recent study showing that hsa-mir-30a targets the 3’-UTR of VIM, causing reduced VIM protein levels and invasiveness of MDA-MB-231 cells (Cheng et al. 2012). While levels of VIM and hsa-mir-30a were outside the scope of this study, we do observe that hsa-mir-30a does not appear to be a strong *in silico*-predicted target of NOV, whereas hsa-mir-30c both *in silico* and *in vitro* appears likely to target NOV. We did not examine the extent to which hsa-mir-30c may be targeting VIM in our system; however, the functional consequence of increased invasiveness of MDA-MB-231 cells following the reduction of NOV through either hsa-mir-30c or NOV-siRNA suggests that the invasiveness of MDA-MB-231 cells is quite sensitive to NOV levels. It is also quite interesting that these mir-30 miRNAs appear on many chromosomes rather than co-regulated in a cluster, and are involved in the regulation of a myriad of pathways such as tumor suppression (p53) (Li et al. 2010), apoptosis (BCL) (Jia et al. 2011), and epithelial to mesenchymal transition (VIM) (Cheng et al. 2012). Individual members of the mir-30 family have been implicated in both tumor suppression and oncogenesis; it is therefore difficult to define the family as a “tumor suppressive” or “oncogenic”. This ambiguity makes studying the functions of the mir30 family members on a case-by-case basis critical for understanding the basis of post-transcriptional molecular mechanisms of disease. The relative levels of the mir-30 family members and their temporal expression may play a critical role in disease progression.

NOV appears to play a context-sensitive role in oncogenesis/tumor suppression (Brigstock 2003). In several cancers that develop from mesenchymal tissues, NOV has been shown to promote tumor growth and metastasis (Manara et al. 2002, Benini et al. 2005, Vallacchi et al. 2008). By contrast, in the context of brain cancer NOV appears to inhibit tumor progression (Fu et al. 2004, Sin et al. 2008). While it is unclear what the function of NOV is when looking at breast cancer cell line data (Ghayad et al. 2009, Sin et al. 2009), NOV expression in human tissue samples of breast cancer shows a clear negative association between NOV expression and late-stage and metastatic disease (Jiang et al. 2004). These histological results strongly suggest that in breast cancer, a disease that develops from epithelial tissue, NOV functions to inhibit disease progression. We demonstrate that a major function of NOV is inhibition of the invasive phenotype of a metastatic breast cancer cell line (MDA-MB-231), similar to the associations of NOV expression with disease progression and metastasis in patients.

Materials and Methods

RNA isolation

RNA was isolated using Qiagen miRNAeasy Mini Kit (217004) following the manufacturer's recommended protocol with optional in-column DNAse I digestion of genomic DNA (Qiagen RNase-Free DNase Set 79254).

miRNA amplification and detection

Complimentary miRNA-specific cDNA was amplified and detected using Applied Biosystems TaqMan MicroRNA Assays for hsa-mir-30c (#4427975) hsa-mir-30c-2* (#4427975) and RNU6B (#4427975).

cDNA amplification and detection

cDNA was amplified from equal quantities of total cellular RNA for each treatment or cell line. cDNA was amplified using the Invitrogen SuperScript First-Strand Synthesis System for RT-PCR (#11904-018) according to the manufacturer's protocol. Reactions were volumetrically diluted, and reaction products were used as templates for Real Time qPCR using Bio-Rad iQ SYBR Green Supermix (#170-8880).

cDNA qPCR primers

Real Time qPCR primers were designed using FoxPrimer (www.foxprimer.org) and validated for efficiency by standard curve using cDNA amplified from untreated MDA-MB-231 cells.

Protein isolation and Western blotting

Cells grown on tissue culture plates were placed directly on ice, and washed twice with PBS supplemented with Roche cOmplete, EDTA-free Protease Inhibitor Cocktail (#11873580001) and 25µM MG132 (Calbiochem (EMD Millipore) CAS 133407-82-6). Cells were scraped into screw-top microcentrifuge tubes, gently spun down to pellet cells and excess PBS was aspirated and discarded. Cells were snap-frozen in liquid nitrogen. Protein lysates were prepared by the addition of RIPA buffer (50mM Tris pH 7.4, 150mM NaCl, 2mM EDTA, 1% v/v NP-40, 0.1% w/v SDS, 1x Roche cOmplete, EDTA-free Protease Inhibitor Cocktail and 25µM MG132) and placing tubes on a 100°C heat block for 10 minutes. Protein lysates were quantified using Pierce BCA Protein Assay Kit (#23225) according to manufacturer's instructions. 50µg protein per sample was loaded onto an SDS-PAGE gel. SDS-PAGE was performed as described (Jitesh Cancer Research Paper). Briefly, lysates were run through an 8.5% acrylamide gel, and then transferred to a PVDF Transfer Membrane (Thermo Scientific #88518).

Membranes were blocked with 5% (w/v) milk (BioRad #170-6404XTU) in PBS and then subjected to immunodetection using the following primary antibodies and dilution factors in 1% (w/v) milk in PBS: Nov (Santa Cruz Biotechnology H-71 sc-50304 1:1000), Lamin A/C (Santa Cruz Biotechnology N-18 sc-6215 1:5000), α-Tubulin (Santa Cruz Biotechnology H-300 sc-5546 1:2000), Runx2 (Lab hybridoma clone 8G5 1:1000). Secondary antibodies used were from Santa Cruz Biotechnology and were diluted 1:5000 in 1% (w/v) milk in PBS: donkey anti-goat IgG-HRP (sc-2020), goat anti-mouse IgG-HRP (sc-2005), and goat anti-rabbit IgG-HRP (sc-2004). After incubation with primary and secondary antibodies, the membranes were washed three times for thirty minutes each with 0.1% (v/v) Tween-20 in PBS. HRP reaction was achieved by one minute incubation with Perkin Elmer Western Lightning ECL (NEL102001EA). Membranes were exposed to Kodak BioMax Light File for Chemiluminescent Imaging (#868-9358) in serial exposure times to empirically determine the exposure time at which signal is most linear.

Matrigel invasion and migration assays

Proliferating MDA-MB-231 cells were trypsinized and counted using Cellometer Auto T4 Cell Counter. A cell suspension of 100,000 cells/mL in growth medium was prepared and 100μL of the suspension was loaded into each BD Matrigel 24-well 8.0 μm PET Membrane Invasion Chamber (#354483). Matrigel coated plates, and control insert plates had 500μL NIH3T3-conditioned medium loaded in the bottom as the chemoattractant. Plates and chemoattractant medium were incubated at 37°C for 3-4 hours prior to loading MDA-MB-231 cells. Cells were incubated for 16 hours at 37°C in 5% CO_2 and then fixed and stained using the Fisher HealthCare PROTOCOL Hema 3 Manual Staining System (#22-122-911) according to the manufacturer's instructions. Cotton swabs were used to eliminate cells which did not migrate/invade as well as Matrigel. Cells were counted using an inverted light microscope.

Transient transfection

Proliferating MDA-MB-231 cells were transfected with 50nM of siRNA/miRNA using Oligofectamine (Invitrogen #12252-011) accoding to the Oligofectamine protocol.

siRNAs

Dharmacon SMARTpool: ON-TARGETplus RUNX2 siRNA (L-012665-00-0005)

Dharmacon SMARTpool: ON-TARGETplus NOV siRNA (L-010527-00-0005)

Dharmacon ON-TARGETplus Non-targeting Pool (D-001810-10-05) *miRNAs and anti-miRNAs*

Dharmacon miRIDIAN microRNA hsa-mir-30c-1 mimic (C-300542-03-0005)

Dharmacon miRIDIAN microRNA hsa-mir-30c-1* mimic (C-301199-01-0005)

Dharmacon miRIDIAN microRNA hsa-mir-30c-1 haripin inhibitor (IH-300542-07-0005)

Dharmacon miRIDIAN microRNA Mimic Negative Control #1 (CN-001000-01-05)

Dharmacon miRIDIAN microRNA Hairpin Inhibitor Negative Control #1 (IN-001005-01-05)

Screen for hsa-mir-30c targets

The top 300 targets of hsa-mir-30c based on mirSVR were downloaded from microrna.org in January, 2011. Gene symbols were used to access gene ontology (GO) terms from DAVID (http://david.abcc.ncifcrf.gov/) and gene reference into function (GeneRIF) from NCBI (http://www.ncbi.nlm.nih.gov/gene/about-generif). Genes whose GO terms or GeneRIFs were associated with invasion, migration, extracellular matrix, or transcription factors were selected and qPCR primers were designed. After 48 hours of transfection, RNA was isolated, cDNA was amplified and Real Time qPCR was carried out to detect the relative levels of mRNAs following transfection with hsa-mir-30c.

Cell lines

Cell lines were grown and maintained as previously described (Jitesh Cancer Research paper).

Stable cell lines

Constructs and stable cell lines were generated as previously described (Pande et al. 2013).

CHAPTER A2: ABSTRACTS FOR MANUSCRIPTS IN PREPARATION OUTSIDE THE SCOPE OF THIS THESIS

Global genomic analysis of AML1-ETO and transcriptional co-regulators in t(8;21) leukemia

Trombly, D.J., Whitfield, T.W., Padmanabahn, S., **Dobson, J.R.**, Gordon, J.A., Lian, J.B., van Wijnen, A.J., Stein, J.L., and Stein, G.S.*

Abstract

The acute myeloid leukemia-related t(8;21) fusion protein AML1-ETO impairs the function of AML1 and other myeloid transcription factors, which results in differentiation arrest and increased self-renewal properties. The oncogenic phenotype caused by AML1-ETO has been primarily studied at single gene resolution, therefore we utilized chromatin immunoprecipitation-sequencing (ChIP-seq) to understand the global contributions of AML1-ETO and associated co-regulatory proteins to the leukemic properties of the t(8;21) model cell line Kasumi-1. We find that both AML1 and AML1-ETO are more associated with the nuclear co-repressor protein (N-CoR) as compared to the histone acetyltransferase p300, which indicates a bias towards transcriptional repression as a mechanism for self-renewal and differentiation arrest. Terms identified by independent gene ontology analyses show significant overlap between the AML1, AML1-ETO, and N-CoR, datasets but not for p300, further suggesting that these proteins function cooperatively. To understand the alternative functions of

AML1-ETO, *de novo* motif discovery was used to identify potential co-regulatory DNA-binding proteins. Genomic regions co-occupied by AML1-ETO and N-CoR in Kasumi-1 cells are enriched in PU.1, RUNX1 and CEBPβ motifs indicating that PU.1 may be important for the repressive functions of AML1-ETO and N-CoR. In summary, our study identified a RUNX1/AML1-ETO/N-CoR gene regulatory network that can be used to interrogate the molecular mechanisms of differentiation arrest in t(8;21) leukemia.

Genomic occupancy of RUNX2 with global expression profiling identifies novel mechanisms regulating osteoblastogenesis

Wu, H., Whitfield, T.W., Gordon, J.A., **Dobson, J.R.**, Moore, J., van Wijnen, A.J., Stein, J.L., Lian, J.B., and Stein, G.S.*

Abstract

Proliferation and differentiation of osteoblasts are highly regulated during bone development and formation, as well as normal turnover and repair in the adult skeleton. Runx2, the master regulator of osteoblastogenesis, directs a transcription program necessary for bone formation through both transcriptional and epigenetic mechanisms. While individual Runx2 gene targets have been identified, insight into the broad spectrum of Runx2 functions is obtainable by global analysis of Runx2 binding. Here, we performed genome-wide characterization of Runx2 occupancy at three major stages of osteoblast differentiation: proliferation, matrix deposition and mineralization. Novel findings include: 1) distinct patterns of Runx2 distribution in genomic regions including upstream, proximal promoters, introns, and exons, and intergenic regions; 2) greater than 80% of all Runx2 binding occurs in non-proximal promoter regions indicating that Runx2 function extends beyond classical transcriptional control; 3) Runx2 binding profiles during osteoblast differentiation that result in functional

changes in gene expression; and 4) novel biological targets and validated functional cis-elements regulated by Runx2. These data comprise a comprehensive map of Runx2 interactions with chromatin revealing gene regulatory roles, as well as potential involvement in other nuclear functions and chromatin organization in developing osteoblast.

Disruption of the RUNX2 response to SMAD signaling affects bone turnover in adult mice

Lou, Y., **Dobson, J.R.**, Wu, H., Frederick, D., Hussain, S., van Wijnen, A.J., Stein, G.S., Lian, J.B., Stein, J.L.*

Abstract

Bone morphogenetic protein (BMP) and transforming growth factor-β (TGFβ) are required for bone formation and bone turnover *in vivo*. Previous studies have shown that three critical residues (HTY426-428) of the transcription factor RUNX2 are required for its interaction with SMAD proteins. Mutation of HTY426-428 to AAA426-428 can abolish the activity of RUNX2 to execute and complete BMP2/ TGFβ signaling for osteoblasogenesis *in vitro*. Here, we describe a mouse model with this triple amino acid mutation inserted into the endogenous RUNX2 locus to test the consequences of disruption of a Runx2-Smad transcriptional complex in vivo. The RUNX2HTY426-428AAA mice have grossly normal skeleton at the birth. Histological and μCT imaging analysis of RUNX2HTY426-428AAA mice beyond three months of age revealed similar bone lengths but extended lengths of trabecular area of tibiae and femurs ($P<0.01$) as compared to wild-type mice of matched ages. To define pathways affected by the HTY mutation that may be causing the observed trabecular phenotype, genome-

wide transcriptome analysis was performed. Proliferating primary osteoblasts isolated from newborn RUNX2HTY426-428AAA mice or wild-type mice were cultured with or without BMP2 treatment and gene expression was measured using Affymetrix GeneChip Mouse Genome 2.0 arrays. The analyzed result has revealed that disruption of Runx2-Smad interaction can alter 860 genes' response to the BMP2 signaling pathway. Ontological terms associated with these genes were examined to identify biological themes associated with each group of genes. This analysis revealed an increase in proliferation-related and bone-related genes combined with reduced expression of genes involved in adipocyte and chondrocyte differentiation. Thus, coordinated changes in the expression of these genes can cause a cell autonomous defect. By *ex vivo* and *in vivo* studies, we have also found that the RUNX2HTY426-428AAA mice exhibited a slightly enhanced osteogenesis differentiation of calvarial osteoblast and bone marrow stromal cells (BMSCs), an inhibition of osteoclast and adipocyte lineage differentiation of BMSCs, and accelerated bone fracture-healing process. These observations were mirrored in our analysis of proliferating primary osteoblasts. These findings indicate that the bone resorption is compromised by the Runx2HTY mutation, while bone formation is slightly increased in RUNX2HTY426-428AAA mice, which caused increased trabecular bone in our RUNX2HTY426-428AAA knock-in mice. Taken together, our findings suggest that a RUNX2-SMAD functional complex may be dispensable for normal skeletal development, but is required for the balance between bone formation and bone resorption in vivo.

Regulation of breast cancer cell proliferation and metabolism by SWI/SNF chromatin remodeling enzyme ATPases

Wu, Q., Madany, P., Akech, J., **Dobson, J.R.**, Douthwright, S., Underwood, J.M., Colby, J.L., van Wijnen, A.J., Stein, J.L., Chiosea, S., Lian, J.B., Stein, G.S., Imbalzano, A.N., Nickerson, J.A.*

Abstract

The mammalian SWI/SNF complexes mediate ATP-dependent chromatin remodeling that functions as a master regulator of gene expression in a broad range of biological function. Dysregulation of this process has been shown to play an important role in oncogenic transformation. Human SWI/SNF consists of at least nine subunits, including one of two mutual exclusive ATPases hBRM (human Brahma) and BRG1 (Brahma-Related Gene 1). Several subunits of this complex, SNF5, BAF57 and BAF180, have been documented as tumor suppressors in human and mice BRG1 is commonly considered as a tumor suppressor based on the observations that it is frequently deleted in lung cancer, and that ~10% of Brg1 heterozygous mice developed mammary tumors. However, there is an increasing body of evidences indicated that BRG1 may function differently in melanoma, prostate cancer and glioma. BRG1 and BRM

are expressed in nearly all breast cancer cell lines, and their role in breast cancer has yet not been fully characterized.

We reported here that BRG1 and BRM expression is up-regulated in primary human breast cancers. Knockdown of these enzymes in metastatic human breast cancer MDA-MB-231 cells decreased cell proliferation by extending cell cycle without cell cycle arrest in a specific phase. Loss of BRG1/BRM had a more profound effect on anchorage-independent growth and tumor-initiation progenitor population. When injected into mammary fat pad, those knockdown cells were unable to or formed much smaller tumors. We link the proliferation defect to the observation that BRG1/BRM regulates fatty acid synthesis by binding to the promoters of and controlling expression of key metabolic enzymes controlling fatty acid synthesis such as ACC and FASN. An additional mechanism of regulation occurs via BRG1/BRM directly interacting with AMPK to sequester it from inactivate ACC. As a consequence, inhibition of BRG1/BRM expression sensitized cancer cells to chemotherapeutic agents, suggesting that the SWI/SNF ATPases may have potential as a therapeutic target for malignant breast cancer.

CHAPTER A3: CONTRIBUTING AUTHOR MANUSCRIPTS PUBLISHED DURING THESIS OUTSIDE THE SCOPE OF THIS DISSERTATION

Runx2 transcriptional activation of Indian Hedgehog and a downstream bone metastatic pathway in breast cancer cells.

Pratap, J., Wixted, J.J., Gaur, T., Zaidi, S.K., **Dobson, J.R.**, Gokul, K., Hussain, S., van Wijnen AJ., Stein, J.L., Stein, G.S., Lian, J.B. *Cancer Res.* 2008 October; 68(19): 7795.

Abstract

Runx2, required for bone formation, is ectopically expressed in breast cancer cells. To address the mechanism by which Runx2 contributes to the osteolytic disease induced by MDA-MB-231 cells, we investigated the effect of Runx2 on key components of the "vicious cycle" of transforming growth factor beta (TGFbeta)-mediated tumor growth and osteolysis. We find that Runx2 directly up-regulates Indian Hedgehog (IHH) and colocalizes with Gli2, a Hedgehog signaling molecule. These events further activate parathyroid hormone-related protein (PTHrP). Furthermore, Runx2 directly regulates the TGFbeta-induced PTHrP levels. A subnuclear targeting deficient mutant Runx2, which disrupts TGFbeta-induced Runx2-Smad interactions, failed to induce IHH and downstream events. In addition, Runx2 knockdown in MDA-MB-231 inhibited IHH and PTHrP expression in the presence of TGFbeta. In vivo blockade of the Runx2-IHH pathway in MDA-MB-231 cells by Runx2 short hairpin RNA inhibition

prevented the osteolytic disease. Thus, our studies define a novel role of Runx2 in up-regulating the vicious cycle of metastatic bone disease, in addition to Runx2 regulation of genes related to progression of tumor metastasis.

Transcriptional corepressor TLE1 functions with Runx2 in epigenetic repression of ribosomal RNA genes.

Ali, S.A., Zaidi, S.K., **Dobson, J.R.**, Shakoori, A.R., Lian, J.B., Stein, J.L., van Wijnen, A.J., Stein, G.S.* *Proc. Natl. Acad. Sci. U.S.A.* 2010 March; 107(9): 4165.

Abstract

Epigenetic control of ribosomal RNA (rRNA) gene transcription by cell type-specific regulators, such as the osteogenic transcription factor Runx2, conveys cellular memory of growth and differentiation to progeny cells during mitosis. Here, we examined whether coregulatory proteins contribute to epigenetic functions that are mitotically transmitted by Runx2 in osteoblastic cells. We show that the transcriptional corepressor Transducin Like Enhancer-1 (TLE1) associates with rRNA genes during mitosis and interphase through interaction with Runx2. Mechanistically, depletion of TLE1 relieves Runx2-mediated repression of rRNA genes transcription and selectively increases histone modifications linked to active transcription. Biologically, loss of TLE-dependent rRNA gene repression coincides with increased global protein synthesis and enhanced cell proliferation. Our findings reinforce the epigenetic marking target genes by phenotypic transcription factors in mitosis and demonstrate a

requirement for retention of coregulatory factors to sustain physiological control of gene expression during proliferation of lineage committed cells.

Cancer-related ectopic expression of the bone-related transcription factor RUNX2 in non-osseous metastatic tumor cells is linked to cell proliferation and motility.

Leong, D.T.*, Lim, J., Goh, X., Pratap, J., Pereira, B.P., Kwok, H., Nathan, S., **Dobson, J.R.**, Lian, J.B., Ito, Y., Voorhoeve, P.M., Stein, G.S., Salto-Tellez, M., Cool, S.M., van Wijnen, A.J.* *Breast Cancer Res*. 2010 October; 12(5):R89.

Abstract

INTRODUCTION:

Metastatic breast cancer cells frequently and ectopically express the transcription factor RUNX2, which normally attenuates proliferation and promotes maturation of osteoblasts. RUNX2 expression is inversely regulated with respect to cell growth in osteoblasts and deregulated in osteosarcoma cells.

METHODS:

Here, we addressed whether the functional relationship between cell growth and RUNX2 gene expression is maintained in breast cancer cells. We also investigated whether the aberrant expression of RUNX2 is linked to phenotypic parameters that could provide a selective advantage to cells during breast cancer progression.

RESULTS:

We find that, similar to its regulation in osteoblasts, RUNX2 expression in MDA-MB-231 breast adenocarcinoma cells is enhanced upon growth factor deprivation, as well as upon deactivation of the mitogen-dependent MEK-Erk pathway or EGFR signaling. Reduction of RUNX2 levels by RNAi has only marginal effects on cell growth and expression of proliferation markers in MDA-MB-231 breast cancer cells. Thus, RUNX2 is not a critical regulator of cell proliferation in this cell type. However, siRNA depletion of RUNX2 in MDA-MB-231 cells reduces cell motility, while forced exogenous expression of RUNX2 in MCF7 cells increases cell motility.

CONCLUSIONS:

Our results support the emerging concept that the osteogenic transcription factor RUNX2 functions as a metastasis-related oncoprotein in non-osseous cancer cells.

A RUNX2-HDAC1 co-repressor complex regulates rRNA gene expression by modulating UBF acetylation.

Ali, S.A., **Dobson, J.R.**, Lian, J.B., Stein, J.L, van Wijnen, A.J., Zaidi, S.K., Stein, G.S.* *J. Cell Sci.* June 2012; 125(Pt 11): 2732.

Abstract

The osteogenic and oncogenic transcription factor RUNX2 downregulates the RNA polymerase I (RNA Pol I)-mediated transcription of rRNAs and changes histone modifications associated with the rDNA repeat. However, the mechanisms by which RUNX2 suppresses rRNA transcription are not well understood. RUNX2 cofactors such as histone deacetylases (HDACs) play a key role in chromatin remodeling and regulation of gene transcription. Here, we show that RUNX2 recruits HDAC1 to the rDNA repeats in osseous cells. This recruitment alters the histone modifications associated with active rRNA-encoding genes and causes deacetylation of the protein upstream binding factor (UBF, also known as UBTF). Downregulation of RUNX2 expression reduces the localization of HDAC1 to the nucleolar periphery and also decreases the association between HDAC1 and UBF. Functionally, depletion of HDAC1 relieves the RUNX2-mediated repression of rRNA-encoding genes and concomitantly increases cell proliferation and global protein synthesis in osseous cells. Our

findings collectively identify a RUNX2-HDAC1-dependent mechanism for the regulation of rRNA-encoding genes and suggest that there is plasticity to RUNX2-mediated epigenetic control, which is mediated through selective mitotic exclusion of co-regulatory factors.

Bibliography

Abal, M., J. Planaguma, A. Gil-Moreno, M. Monge, M. Gonzalez, T. Baro, A. Garcia, J. Castellvi, S. Ramon Y Cajal, J. Xercavins, F. Alameda, and J. Reventos. 2006. Molecular pathology of endometrial carcinoma: transcriptional signature in endometrioid tumors. Histology and histopathology 21:197–204.

Adachi, Y., E. Käs, and U. K. Laemmli. 1989. Preferential, cooperative binding of DNA topoisomerase II to scaffold-associated regions. The EMBO Journal 8:3997–4006.

Akech, J., J. J. Wixted, K. Bedard, M. Deen, S. Hussain, T. A. Guise, A. Wijnen, J. L. Stein, L. R. Languino, D. C. Altieri, J. Pratap, E. Keller, G. S. Stein, J. B. Lian, M. van der Deen, and A. J. van Wijnen. 2010. Runx2 association with progression of prostate cancer in patients: mechanisms mediating bone osteolysis and osteoblastic metastatic lesions. Oncogene 29:811–821.

Alberts, B., A. Johnson, J. Lewis, P. Walter, M. Raff, and K. Roberts. 2002. Molecular Biology of the Cell 4th Edition.

Ali, S. A., J. R. Dobson, J. B. Lian, J. L. Stein, A. J. van Wijnen, S. K. Zaidi, and G. S. Stein. 2012. A RUNX2-HDAC1 co-repressor complex regulates rRNA gene expression by modulating UBF acetylation. Journal of Cell Science 125:2732–9.

Ali, S. A., S. K. Zaidi, J. R. Dobson, A. R. Shakoori, J. B. Lian, J. L. Stein, A. J. van Wijnen, and G. S. Stein. 2010. Transcriptional corepressor TLE1 functions with Runx2 in epigenetic repression of ribosomal RNA genes. Proceedings of the National Academy of Sciences of the United States of America 107:4165–9.

Allred, D. C., J. M. Harvey, M. Berardo, and G. M. Clark. 1998. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 11:155–168.

Alsheimer, M., B. Liebe, L. Sewell, C. L. Stewart, H. Scherthan, and R. Benavente. 2004. Disruption of spermatogenesis in mice lacking A-type lamins. Journal of cell science 117:1173–8.

Alva-Medina, J., A. Maya-Mendoza, M. Dent, and A. Aranda-Anzaldo. 2011. Continued Stabilization of the Nuclear Higher-Order Structure of Post-Mitotic Neurons In Vivo. PLoS ONE 6.

Alvarez, J. D., D. H. Yasui, H. Niida, T. Joh, D. Y. Loh, and T. Kohwi-Shigematsu. 2000. The MAR-binding protein SATB1 orchestrates temporal and spatial expression of multiple genes during T-cell development. Genes & Development 14:521–535.

Andrulis, E. D., A. M. Neiman, D. C. Zappulla, and R. Sternglanz. 1998. Perinuclear localization of chromatin facilitates transcriptional silencing. Nature 394:592–5.

Bachellerie, J. P., E. Puvion, and J. P. Zalta. 1975. Ultrastructural Organization and Biochemical Characterization of Chromatin· RNA· Protein Complexes Isolated from Mammalian Cell Nuclei. European journal of biochemistry / FEBS 58:327–337.

Bakshi, R., S. K. Zaidi, S. Pande, M. Q. Hassan, D. W. Young, M. Montecino, J. B. Lian, A. J. van Wijnen, J. L. Stein, and G. S. Stein. 2008. The leukemogenic t(8;21) fusion protein AML1-ETO controls rRNA genes and associates with nucleolar-organizing regions at mitotic chromosomes. Journal of Cell Science 121:3981–3990.

Barnes, G. L., K. E. Hebert, M. Kamal, A. Javed, T. A. Einhorn, J. B. Lian, G. S. Stein, and L. C. Gerstenfeld. 2004. Fidelity of Runx2 activity in breast cancer cells is required for the generation of metastases-associated osteolytic disease. Cancer Research 64:4506–4513.

Barseguian, K., B. Lutterbach, S. W. Hiebert, J. Nickerson, J. B. Lian, J. L. Stein, A. J. van Wijnen, and G. S. Stein. 2002. Multiple subnuclear targeting signals of the leukemia-related AML1/ETO and ETO repressor proteins. Proceedings of the National Academy of Sciences of the United States of America 99:15434–15439.

Belgrader, P., A. J. Siegel, and R. Berezney. 1991. A comprehensive study on the isolation and characterization of the HeLa S3 nuclear matrix. Journal of Cell Science 98 (Pt 3):281–291.

Belle, I., S. Cai, T. Kohwi-Shigematsu, and I. de Belle. 1998. The genomic sequences bound to special AT-rich sequence-binding protein 1 (SATB1) in vivo in Jurkat T cells are tightly associated with the nuclear matrix at the bases of the chromatin loops. The Journal of cell biology 141:335–348.

Benini, S., B. Perbal, D. Zambelli, M. P. Colombo, M. C. Manara, M. Serra, M. Parenza, V. Martinez, P. Picci, and K. Scotlandi. 2005. In Ewing’s sarcoma CCN3(NOV) inhibits proliferation while promoting migration and invasion of the same cell type. Oncogene 24:4349–4361.

Benyajati, C., and A. Worcel. 1976. Isolation, characterization, and structure of the folded interphase genome of Drosophila melanogaster. Cell 9:393–407.

Berezney, R. 1984. Organization and functions of the nuclear matrix. Pages 119–180 *in* L. S. Hnilica, editor. The Cell NucleusVolume IV. CRC Press, Boca Raton, FL.

Berezney, R., and D. S. Coffey. 1974. Identification of a nuclear protein matrix. Biochemical and biophysical research communications 60:1410–1417.

Berezney, R., and D. S. Coffey. 1977. Nuclear matrix. Isolation and characterization of a framework structure from rat liver nuclei. Journal of Cell Biology 73:616–637.

Berezney, R., M. J. Mortillaro, H. Ma, X. Wei, and J. Samarabandu. 1995. The Nuclear Matrix: A Structural Mileu for Nuclear Genomic Function. Pages 2–66 *in* R. Berezney and K. W. Jeon, editors. International Review of Cytology, A Survey of Cell Biology: Structural and functional organization of the nuclear matrix.Volume 162. Academic Press, San Diego, CA.

Bernhard, W. 1969. A new staining procedure for electron microscopical cytology. Journal of Ultrastructure Research 27:250–265.

Betel, D., M. Wilson, A. Gabow, D. S. Marks, and C. Sander. 2008. The microRNA.org resource: targets and expression. Nucleic Acids Research 36:D149–53.

Bidwell, J., E. Fey, A. Wijnen, S. Penman, J. Stein, J. Lian, and G. Stein. 1994. Nuclear matrix proteins distinguish normal diploid osteoblasts from osteosarcoma cells. Cancer Research 54:28–32.

Bidwell, J. P., A. Wijnen, E. G. Fey, S. Dworetzky, S. Penman, J. L. Stein, J. B. Lian, G. S. Stein, and A. J. Van Wijnen. 1993. Osteocalcin gene promoter-binding factors are tissue-specific nuclear matrix components. Proceedings of the National Academy of Sciences of the United States of America 90:3162–3166.

Blyth, K., F. Vaillant, L. Hanlon, N. Mackay, M. Bell, A. Jenkins, J. C. Neil, and E. R. Cameron. 2006. Runx2 and MYC collaborate in lymphoma development by suppressing apoptotic and growth arrest pathways in vivo. Cancer research 66:2195–201.

Blyth, K., F. Vaillant, A. Jenkins, L. McDonald, M. Pringle, C. Huser, T. Stein, J. Neil, and E. Cameron. 2010. Runx2 in normal tissues and cancer cells: A developing story. Blood Cells Mol Dis 45:117–123.

Bode, J., Y. Kohwi, L. Dickinson, T. Joh, D. Klehr, C. Mielke, and T. Kohwi-Shigematsu. 1992. Biological significance of unwinding capability of nuclear matrix-associating DNAs. Science (New York, NY) 255:195–197.

Bolstad, B. M., R. a Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (Oxford, England) 19:185–93.

Boulikas, T. 1995. Chromatin Domains and Prediction of MAR Sequences. Pages 279–388 *in* R. Berezney and K. W. Jeon, editors. International Review of Cytology, A Survey of Cell Biology: Structural and functional organization of the nuclear matrix., 162nd edition. Academic Press, San Diego, CA.

De Braekeleer, E., C. Férec, and M. De Braekeleer. 2009. RUNX1 translocations in malignant hemopathies. Anticancer research 29:1031–7.

Brigstock, D. R. 2003. The CCN family: a new stimulus package. The Journal of endocrinology 178:169–175.

Brinkley, B. R., P. T. Beall, L. J. Wible, M. L. Mace, D. S. Turner, and R. M. Cailleau. 1980. Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. Cancer research 40:3118–29.

Bryne, J. C., E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, A. Sandelin, and I. Piedade. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research 36:D102–6.

Bustin, S. A., V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele, and C. T. Wittwer. 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clinical Chemistry 55:611–622.

Cai, S., C. C. Lee, and T. Kohwi-Shigematsu. 2006. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Nature genetics 38:1278–1288.

Cameron, E. R., and J. C. Neil. 2004. The Runx genes: lineage-specific oncogenes and tumor suppressors. Oncogene 23:4308–4314.

Carroll, J. S., C. A. Meyer, J. Song, W. Li, T. R. Geistlinger, J. Eeckhoute, A. S. Brodsky, E. K. Keeton, K. C. Fertuck, G. F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E. A. Fox, P. A. Silver, T. R. Gingeras, X. S. Liu, and M. Brown. 2006. Genome-wide analysis of estrogen receptor binding sites. Nature genetics 38:1289–1297.

Cheng, C.-W., H.-W. Wang, C.-W. Chang, H.-W. Chu, C.-Y. Chen, J.-C. Yu, J.-I. Chao, H.-F. Liu, S.-L. Ding, and C.-Y. Shen. 2012. MicroRNA-30a inhibits cell migration and invasion by downregulating vimentin expression and is a potential prognostic marker in breast cancer. Breast Cancer Research and Treatment 134:1081–1093.

Chimge, N.-O., S. K. Baniwal, G. H. Little, Y. Chen, M. Kahn, D. Tripathy, Z. Borok, and B. Frenkel. 2011. Regulation of breast cancer metastasis by Runx2 and estrogen signaling: the role of SNAI2. Breast cancer research : BCR 13:R127.

Chimge, N.-O., and B. Frenkel. 2012. The RUNX family in breast cancer: relationships with estrogen signaling. Oncogene 32:2121–30.

Choi, J. Y., J. Pratap, A. Javed, S. K. Zaidi, L. Xing, E. Balint, S. Dalamangas, B. Boyce, A. J. van Wijnen, J. B. Lian, J. L. Stein, S. N. Jones, and G. S. Stein. 2001. Subnuclear targeting of Runx/Cbfa/AML factors is essential for tissue-specific differentiation during embryonic development. Proceedings of the National Academy of Sciences of the United States of America 98:8650–5.

Ciejek, E. M., M. J. Tsai, and B. W. O’Malley. 1983. Actively transcribed genes are associated with the nuclear matrix. Nature 306:607–9.

Cook, P. R. 1989. The nucleoskeleton and the topology of transcription. European journal of biochemistry / FEBS 185:487–501.

Cook, P. R. 1999. The organization of replication and transcription. Science (New York, N.Y.) 284:1790–5.

Cook, P. R., and I. A. Brazell. 1975. Supercoils in human DNA. Journal of cell science 19:261–79.

Cook, P. R., and I. A. Brazell. 1980. Mapping sequences in loops of nuclear DNA by their progressive detachment from the nuclear cage. Nucleic acids research 8:2895–906.

Craig, J. M., S. Boyle, P. Perry, and W. A. Bickmore. 1997. Scaffold attachments within the human genome. Journal of Cell Science 110 (Pt 2:2673–2682.

Dairkee, S., Y. Ji, Y. Ben, D. Moore, Z. Meng, and S. Jeffrey. 2004. A molecular “signature” of primary breast cancer cultures; patterns resembling tumor tissue. BMC Genomics 5:47.

Das, K., D. T. Leong, A. Gupta, L. Shen, T. Putti, G. S. Stein, A. J. van Wijnen, and M. Salto-Tellez. 2009. Positive association between nuclear Runx2 and oestrogen-progesterone receptor gene expression characterises a biological subtype of breast cancer. European Journal of Cancer 45:2239–2248.

Davie, J. R. 1995. The Nuclear matrix and the Regulation of Chromatin Organization and Function. Pages 191–250 *in* R. Berezney and K. W. Jeon, editors. International Review of Cytology, A Survey of Cell Biology: Structural and functional organization of the nuclear matrix., 162nd edition. Academic Press, San Diego, CA.

Debnath, J., S. K. Muthuswamy, and J. S. Brugge. 2003. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. Methods (San Diego, Calif.) 30:256–68.

Van der Deen, M., J. Akech, D. Lapointe, S. Gupta, D. W. Young, M. A. Montecino, M. Galindo, J. B. Lian, J. L. Stein, G. S. Stein, and A. J. van Wijnen. 2012. Genomic promoter occupancy of runt-related transcription factor RUNX2 in osteosarcoma cells identifies genes involved in cell adhesion and motility. The Journal of biological chemistry 287:4503–4517.

Deng, B., S. Melnik, and P. R. Cook. 2012. Transcription factories, chromatin loops, and the dysregulation of gene expression in malignancy. Seminars in cancer biology 23:65–71.

Dent, R., M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. a Sawka, L. a Lickley, E. Rawlinson, P. Sun, and S. a Narod. 2007. Triple-negative breast cancer: clinical features and patterns of recurrence. Clinical cancer research : an official journal of the American Association for Cancer Research 13:4429–34.

Diaz, A., K. Park, D. A. Lim, and J. S. Song. 2012. Normalization, bias correction, and peak calling for ChIP-seq. Statistical applications in genetics and molecular biology 11:Article 9.

Dickinson, L. A., T. Joh, Y. Kohwi, and T. Kohwi-Shigematsu. 1992. A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. Cell 70:631–645.

Dijkwel, P. A., P. W. Wenink, and J. Poddighe. 1986. Permanent attachment of replication origins to the nuclear matrix in BHK-cells. Nucleic acids research 14:3241–9.

Doll, A., M. Gonzalez, M. Abal, M. Llaurado, M. Rigau, E. Colas, M. Monge, J. Xercavins, G. Capella, B. Diaz, A. Gil-Moreno, F. Alameda, and J. Reventos. 2009. An orthotopic endometrial cancer mouse model demonstrates a role for RUNX1 in distant metastasis. International journal of cancer. Journal international du cancer 125:257–63.

Dowdy, C. R., R. Xie, D. Frederick, S. Hussain, S. K. Zaidi, D. Vradii, A. Javed, X. Li, S. N. Jones, J. B. Lian, A. Wijnen, J. L. Stein, G. S. Stein, and A. J. van Wijnen. 2010. Definitive hematopoiesis requires Runx1 C-terminal-mediated subnuclear targeting and transactivation. Human Molecular Genetics 19:1048–57.

Dundr, M., and T. Misteli. 2010. Biogenesis of nuclear bodies. Cold Spring Harbor Perspectives in Biology 2:a000711.

Dworetzky, S., K. Wright, E. Fey, S. Penman, J. Lian, J. Stein, and G. Stein. 1992. Sequence-specific DNA-binding proteins are components of a nuclear matrix-attachment site. Proceedings of the National Academy of Sciences of the United States of America 89:4178–4182.

ENCODE Project Consortium. 2011. A user’s guide to the encyclopedia of DNA elements (ENCODE). PLoS biology 9:e1001046.

Ernst, J., and M. Kellis. 2012. ChromHMM: automating chromatin-state discovery and characterization. Nature methods 9:215–6.

Fackelmayer, F. O., K. Dahm, A. Renz, U. Ramsperger, and A. Richter. 1994. Nucleic-acid-binding properties of hnRNP-U/SAF-A, a nuclear-matrix protein which binds DNA and RNA in vivo and in vitro. European journal of biochemistry / FEBS 221:749–57.

Fakan, S., and W. Bernhard. 1971. Localisation of rapidly and slowly labelled nuclear RNA as visualized by high resolution autoradiography. Experimental cell research 67:129–41.

Fakan, S., and M. E. Hughes. 1989. Fine structural ribonucleoprotein components of the cell nucleus visualized after spreading and high resolution autoradiography. Chromosoma 98:242–249.

Fakan, S., and P. Nobis. 1978. Ultrastructural localization of transcription sites and of RNA distribution during the cell cycle of synchronized CHO cells. Experimental cell research 113:327–337.

Fawcett, D. W. 1966. An Atlas of Fine Structure. The Cell. Its Organelles and Inclusions. Annals of Internal Medicine 64:968.

Feldman, L. T., and J. R. Nevins. 1983. Localization of the adenovirus E1Aa protein, a positive-acting transcriptional factor, in infected cells infected cells. Molecular and cellular biology 3:829–38.

Fey, E. G., G. Krochmalnic, and S. Penman. 1986. The nonchromatin substructures of the nucleus: the ribonucleoprotein (RNP)-containing and RNP-depleted matrices analyzed by sequential fractionation and resinless section electron microscopy. The Journal of cell biology 102:1654–65.

Field, A. P. 2007. Analysis of Variance (ANOVA). Pages 33–36 *in* N. J. Salkind, editor. Encyclopedia of Measurement and Statistics. . SAGE Publications, Inc., Thousand Oaks, CA.

Florea, L., G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Research 8:967–974.

Frock, R. L., B. A. Kudlow, A. M. Evans, S. A. Jameson, S. D. Hauschka, and B. K. Kennedy. 2006. Lamin A/C and emerin are critical for skeletal muscle satellite cell differentiation. Genes & development 20:486–500.

Fu, C. T., J. F. Bechberger, M. A. Ozog, B. Perbal, and C. C. Naus. 2004. CCN3 (NOV) interacts with connexin43 in C6 glioma cells: possible mechanism of connexin-mediated growth suppression. The Journal of biological chemistry 279:36943–36950.

Galindo, M., J. Pratap, D. Young, H. Hovhannisyan, H.-J. Im, J.-Y. Choi, J. Lian, J. Stein, G. Stein, and A. Wijnen. 2005. The bone-specific expression of Runx2 oscillates during the cell cycle to support a G1-related antiproliferative function in osteoblasts. The Journal of biological chemistry 280:20274–20285.

Gasser, S. M., and U. K. Laemmli. 1987. A glimpse at chromosomal order. Trends in Genetics 3:16–22.

Gautier, L., L. Cope, B. M. Bolstad, and R. a Irizarry. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics (Oxford, England) 20:307–15.

Georgiev, G. P., Y. S. Vassetzky, A. N. Luchnik, V. V Chernokhvostov, and S. V Razin. 1991. A. E. Braunstein Plenary Lecture. Nuclear skeleton, DNA domains and control of replication and transcription. European journal of biochemistry / FEBS 200:613–24.

Gerdes, M. G., K. C. Carter, P. T. Moen, and J. B. Lawrence. 1994. Dynamic changes in the higher-level chromatin organization of specific sequences revealed by in situ hybridization to nuclear halos. The Journal of Cell Biology 126:289–304.

Germann, S., and V. Gaudin. 2011. Mapping in vivo protein-DNA interactions in plants by DamID, a DNA adenine methylation-based method. Methods in molecular biology (Clifton, N.J.) 754:307–21.

Getzenberg, R. H., B. R. Konety, T. A. Oeler, M. M. Quigley, A. Hakam, M. J. Becich, and R. R. Bahnson. 1996. Bladder Cancer-associated Nuclear Matrix Proteins Bladder Cancer-associated:1690–1694.

Getzenberg, R., K. Pienta, E. Huang, and D. Coffey. 1991. Identification of nuclear matrix proteins in the cancer and normal rat prostate. Cancer research 51:6514–6520.

Geyer, P. K., M. W. Vitalini, and L. L. Wallrath. 2011. Nuclear organization: taking a position on gene expression. Current Opinion in Cell Biology 23:354–359.

Ghamari, A., M. P. C. van de Corput, S. Thongjuea, W. A. van Cappellen, W. van IJcken, J. van Haren, E. Soler, D. Eick, B. Lenhard, and F. G. Grosveld. 2013. In vivo live imaging of RNA polymerase II transcription factories in primary cells. Genes & Development 27:767–777.

Ghayad, S. E., J. A. Vendrell, I. Bieche, F. F. Spyratos, C. Dumontet, I. Treilleux, R. Lidereau, and P. A. Cohen. 2009. Identification of TACC1, NOV, and PTTG1 as new candidate genes associated with endocrine therapy resistance in breast cancer. Journal of Molecular Endocrinology 42:87.

Ghule, P. N., Z. Dominski, J. B. Lian, J. L. Stein, A. J. van Wijnen, and G. S. Stein. 2009. The subnuclear organization of histone gene regulatory proteins and 3’ end processing factors of normal somatic and embryonic stem cells is compromised in selected human cancer cell types. Journal of cellular physiology 220:129–35.

Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics (Oxford, England) 27:1017–1018.

Greer, E. L., and Y. Shi. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. Nature Reviews Genetics 13:343–357.

Griffiths-Jones, S., R. J. Grocock, S. Dongen, A. Bateman, A. J. Enright, and S. van Dongen. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research 34:D140–4.

Grummt, I. 2010. Wisely chosen paths - regulation of rRNA synthesis: Delivered on 30 June 2010 at the 35th FEBS Congress in Gothenburg, Sweden. The FEBS journal 277:4626–4639.

Grummt, I., and R. Voit. 2010. Linking rDNA transcription to the cellular energy supply. Cell cycle (Georgetown, Tex.) 9:225–6.

GuangChun, L., W. Lu, and X. Hanhong. 2003. A novel web application frame developed by MVC. ACM SIGSOFT Software Engineering Notes 28:7.

Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. Faza, W. Talhout, B. Eussen, A. Klein, L. Wessels, W. Laat, and B. Steensel. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453:948–951.

Guise, T. A., K. S. Mohammad, G. Clines, E. G. Stebbins, D. H. Wong, L. S. Higgins, R. Vessella, E. Corey, S. Padalecki, L. Suva, and J. M. Chirgwin. 2006. Basic mechanisms responsible for osteolytic and osteoblastic bone metastases. Clinical cancer research : an official journal of the American Association for Cancer Research 12:6213s–6216s.

Gutierrez, S., A. Javed, D. K. Tennant, M. van Rees, M. Montecino, G. S. Stein, J. L. Stein, and J. B. Lian. 2002. CCAAT/enhancer-binding proteins (C/EBP) beta and delta activate osteocalcin gene transcription and synergize with Runx2 at the C/EBP element to regulate bone-specific expression. The Journal of biological chemistry 277:1316–23.

Gutierrez-Hartmann, A., D. L. Duval, and A. P. Bradford. 2007. ETS transcription factors in endocrine systems. Trends in endocrinology and metabolism: TEM 18:150–8.

Hager, G., J. Mcnally, and T. Misteli. 2009. Transcription Dynamics. Molecular Cell 35:741–753.

Han, H.-J., J. Russo, Y. Kohwi, and T. Kohwi-Shigematsu. 2008. SATB1 reprogrammes gene expression to promote breast tumour growth and metastasis. Nature 452:187–193.

Hanahan, D., and R. A. Weinberg. 2000. The Hallmarks of Cancer. Cell 100:57–70.

Harewood, L., H. M. Robinson, R. L. Harris, M. J. Al-Obaidi, G. R. Jalali, M. Martineau, A. V Moorman, N. Sumption, S. Richards, C. Mitchell, C. J. Harrison, Z. J. Broadfield, K. L. Cheung, L. Harewood, R. L. Harris, G. R. Jalali, M. Martineau, A. V Moorman, K. E. Taylor, S. Richards, C. Mitchell, and C. J. Harrison. 2003. Amplification of AML1 in acute lymphoblastic leukemia is associated with a poor outcome. Leukemia 17:547–53.

Hawkins, R. D., G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J. A. Thomson, J. R. Ecker, and B. Ren. 2010. Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells. Cell Stem Cell 6:479–491.

He, D. C., J. A. Nickerson, and S. Penman. 1990. Core filaments of the nuclear matrix. The Journal of Cell Biology 110:569–580.

He, S., K. L. Dunn, P. S. Espino, B. Drobic, L. Li, J. Yu, J.-M. Sun, H. Y. Chen, S. Pritchard, and J. R. Davie. 2008. Chromatin organization and nuclear microenvironments in cancer cells. Journal of Cellular Biochemistry 104:2004–2015.

Heinemeyer, T., E. Wingender, I. Reuter, H. Hermjakob, A. E. Kel, O. V Kel, E. V Ignatieva, E. A. Ananko, O. A. Podkolodnaya, F. A. Kolpakov, N. L. Podkolodny, and N. A. Kolchanov. 1998. Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Research 26:362–367.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. 2010. Simple Combinations of Lineage-

Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 38:576–589.

Hellemans, J., G. Mortier, A. Paepe, F. Speleman, J. Vandesompele, and A. De Paepe. 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biology 8:R19.

Heng, H. H. Q., S. Goetze, C. J. Ye, G. Liu, J. B. Stevens, S. W. Bremer, S. M. Wykes, J. Bode, and S. A. Krawetz. 2004. Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. Journal of cell science 117:999–1008.

Herman, R., L. Weymouth, and S. Penman. 1978. Heterogeneous nuclear RNA-protein fibers in chromatin-depleted nuclei. The Journal of cell biology 78:663–74.

Hewitt, S. L., F. A. High, S. L. Reiner, A. G. Fisher, and M. Merkenschlager. 2004. Nuclear repositioning marks the selective exclusion of lineage-inappropriate transcription factor loci during T helper cell differentiation. European journal of immunology 34:3604–13.

Hickey, T. E., J. L. L. Robinson, J. S. Carroll, and W. D. Tilley. 2012. Minireview: The androgen receptor in breast tissues: growth inhibitor, tumor suppressor, oncogene? Molecular endocrinology (Baltimore, Md.) 26:1252–1267.

Higgins, G., K. M. Roper, I. J. Watson, F. H. Blackhall, W. N. Rom, H. I. Pass, J. F. X. Ainscough, and D. Coverley. 2012. Variant Ciz1 is a circulating biomarker for early-stage lung cancer. Proceedings of the National Academy of Sciences of the United States of America 109:E3128–35.

Hoffman, M. M., O. O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods 9:473–476.

Hoi, C. S. L., S. E. Lee, S.-Y. Lu, D. J. McDermitt, K. M. Osorio, C. M. Piskun, R. M. Peters, R. Paus, and T. Tumbar. 2010. Runx1 directly promotes proliferation of hair follicle stem cells and epithelial tumor formation in mouse skin. Molecular and cellular biology 30:2518–36.

Hon, G. C., R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, M. Pelizzola, A. Valsesia, Z. Ye, S. Kuan, L. E. Edsall, A. A. Camargo, B. J. Stevenson, J. R. Ecker, V. Bafna, R. L. Strausberg, A. J. Simpson, and B. Ren. 2012. Global

DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome research 22:246–58.

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

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009b. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research 37:1–13.

Huber, L., and L. Chodosh. 2005. Dynamics of DNA repair suggested by the subcellular localization of Brca1 and Brca2 proteins. Journal of cellular biochemistry 96:47–55.

Hurtado, A., K. a Holmes, C. S. Ross-Innes, D. Schmidt, and J. S. Carroll. 2011. FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nature genetics 43:27–33.

Inman, C., and P. Shore. 2003. The osteoblast transcription factor Runx2 is expressed in mammary epithelial cells and mediates osteopontin expression. The Journal of biological chemistry 278:48684–48689.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) 4:249–264.

Izaurralde, E., J. Mirkovitch, and U. K. Laemmli. 1988. Interaction of DNA with nuclear scaffolds in vitro. Journal of molecular biology 200:111–125.

Jackson, D. A., and P. R. Cook. 1985. Transcription occurs at a nucleoskeleton. The EMBO journal 4:919–25.

Jackson, D. A., and P. R. Cook. 1988. Visualization of a filamentous nucleoskeleton with a 23 nm axial repeat. The EMBO Journal 7:3667–3677.

Jackson, D. A., and P. R. Cook. 1995. The Structural Basis of Nuclear Function. Pages 125–150 *in* R. Berezney and K. W. Jeon, editors. International Review of Cytology, A Survey of Cell Biology: Structural and functional organization of the nuclear matrix., 162nd edition. Academic Press, San Diego, CA.

Jackson, D. A., A. B. Hassan, R. J. Errington, and P. R. Cook. 1993. Visualization of focal sites of transcription within human nuclei. The EMBO journal 12:1059–65.

Jackson, D. A., F. J. Iborra, E. M. Manders, and P. R. Cook. 1998. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. Molecular biology of the cell 9:1523–36.

Janes, K. A. 2011. RUNX1 and its understudied role in breast cancer. Cell cycle (Georgetown, Tex) 10:3461–5.

Javed, A., G. L. Barnes, J. Pratap, T. Antkowiak, L. C. Gerstenfeld, A. J. van Wijnen, J. L. Stein, J. B. Lian, G. S. Stein, and A. Wijnen. 2005. Impaired intranuclear trafficking of Runx2 (AML3/CBFA1) transcription factors in breast cancer cells inhibits osteolysis in vivo. Proceedings of the National Academy of Sciences of the United States of America 102:1454–1459.

Jemal, A., M. M. Center, C. DeSantis, and E. M. Ward. 2010. Global patterns of cancer incidence and mortality rates and trends. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 19:1893–907.

Jia, W., J. O. Eneh, S. Ratnaparkhe, M. K. Altman, and M. M. Murph. 2011. MicroRNA-30c-2* expressed in ovarian cancer cells suppresses growth factor-induced cellular proliferation and downregulates the oncogene BCL9. Molecular cancer research : MCR 9:1732–1745.

Jiang, H., and B. M. Peterlin. 2008. Differential chromatin looping regulates CD4 expression in immature thymocytes. Molecular and Cellular Biology 28:907–912.

Jiang, W. G., G. Watkins, O. Fodstad, A. Douglas-Jones, K. Mokbel, and R. E. Mansel. 2004. Differential expression of the CCN family members Cyr61, CTGF and Nov in human breast cancer. Endocrine-related cancer 11:781–791.

Joseph, R., Y. L. Orlov, M. Huss, W. Sun, S. L. Kong, L. Ukil, Y. F. Pan, G. Li, M. Lim, J. S. Thomsen, Y. Ruan, N. D. Clarke, S. Prabhakar, E. Cheung, and E. T. Liu. 2010. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. Molecular systems biology 6:456.

Jost, J. P., and M. Seldran. 1984. Association of transcriptionally active vitellogenin II gene with the nuclear matrix of chicken liver. The EMBO journal 3:2005–8.

Kadota, M., H. H. Yang, B. Gomez, M. Sato, R. J. Clifford, D. Meerzaman, B. K. Dunn, L. M. Wakefield, and M. P. Lee. 2010. Delineating genetic alterations for tumor progression in the MCF10A series of breast cancer cell lines. PLoS ONE 5:e9201.

Kagoshima, H., K. Shigesada, M. Satake, Y. Ito, H. Miyoshi, M. Ohki, M. Pepling, and P. Gergen. 1993. The Runt domain identifies a new family of heteromeric transcriptional regulators. Trends in genetics : TIG 9:338–41.

Kapustin, Y., A. Souvorov, T. Tatusova, and D. Lipman. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. Biology direct 3:20.

Kayed, H., X. Jiang, S. Keleg, R. Jesnowski, T. Giese, M. R. Berger, I. Esposito, M. Löhr, H. Friess, and J. Kleeff. 2007. Regulation and functional role of the Runt-related transcription factor-2 in pancreatic cancer. British journal of cancer 97:1106–15.

Keaton, M. A., C. M. Taylor, R. M. Layer, and A. Dutta. 2011. Nuclear Scaffold Attachment Sites within ENCODE Regions Associate with Actively Transcribed Genes. PLoS ONE 6:e17912.

Kharchenko, P. V, M. Y. Tolstorukov, and P. J. Park. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology 26:1351–1359.

Kim, V. N. 2005. MicroRNA biogenesis: coordinated cropping and dicing. Nature reviews. Molecular cell biology 6:376–385.

Kingsley, L. A., P. G. J. Fournier, J. M. Chirgwin, and T. A. Guise. 2007. Molecular biology of bone metastasis. Molecular cancer therapeutics 6:2609–2617.

Komori, T., H. Yagi, S. Nomura, A. Yamaguchi, K. Sasaki, K. Deguchi, Y. Shimizu, R. T. Bronson, Y. H. Gao, M. Inada, M. Sato, R. Okamoto, Y. Kitamura, S. Yoshiki, and T. Kishimoto. 1997. Targeted disruption of Cbfa1 results in a complete lack of bone formation owing to maturational arrest of osteoblasts. Cell 89:755–64.

Kosak, S. T., J. A. Skok, K. L. Medina, R. Riblet, M. M. Le Beau, A. G. Fisher, and H. Singh. 2002. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. Science (New York, N.Y.) 296:158–62.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. a Marra. 2009. Circos: an information aesthetic for comparative genomics. Genome research 19:1639–45.

Kuchenbauer, F., S. M. Mah, M. Heuser, A. McPherson, J. Rüschmann, A. Rouhi, T. Berg, L. Bullinger, B. Argiropoulos, R. D. Morin, D. Lai, D. T. Starczynowski, A. Karsan, C. J. Eaves, A. Watahiki, Y. Wang, S. A. Aparicio, A. Ganser, J. Krauter, H. Döhner, K. Döhner, M. A. Marra, F. D. Camargo, L. Palmqvist, C. Buske, and R. K. Humphries. 2011. Comprehensive analysis of mammalian miRNA* species and their role in myeloid cells. Blood 118:3350–3358.

Kuo, Y.-H., S. K. Zaidi, S. Gornostaeva, T. Komori, G. S. Stein, and L. H. Castilla. 2009. Runx2 induces acute myeloid leukemia in cooperation with Cbfbeta-SMMHC in mice. Blood 113:3323–32.

Lanctôt, C., T. Cheutin, M. Cremer, G. Cavalli, and T. Cremer. 2007. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nature reviews. Genetics 8:104–15.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Landt, S. G., G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22:1813–1831.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9:357–9.

Larkin, J. D., P. R. Cook, and A. Papantonis. 2012. Dynamic reconfiguration of long human genes during one transcription cycle. Molecular and cellular biology 32:2738–47.

Lee, T. I., S. E. Johnstone, and R. A. Young. 2006. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nature protocols 1:729–748.

Levanon, D., O. Brenner, F. Otto, and Y. Groner. 2003. Runx3 knockouts and stomach cancer. EMBO reports 4:560–4.

Levantini, E., S. Lee, H. S. Radomska, C. J. Hetherington, M. Alberich-Jorda, G. Amabile, P. Zhang, D. A. Gonzalez, J. Zhang, D. S. Basseres, N. K. Wilson, S. Koschmieder, G. Huang, D.-E. Zhang, A. K. Ebralidze, C. Bonifer, Y. Okuno, B. Gottgens, and D. G. Tenen. 2011. RUNX1 regulates the CD34 gene in haematopoietic stem cells by mediating interactions with a distal regulatory element. The EMBO Journal 30:4059–4070.

Lewis, C., and J. Lebkowski. 1984. Interphase nuclear matrix and metaphase scaffolding structures. Journal of cell science. Supplement 1:103–22.

Li, B., M. Carey, and J. L. Workman. 2007. The role of chromatin during transcription. Cell 128:707–719.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li, J., S. Donath, Y. Li, D. Qin, B. S. Prabhakar, and P. Li. 2010. miR-30 regulates mitochondrial fission through targeting p53 and the dynamin-related protein-1 pathway. PLoS genetics 6:e1000795.

Li, J., H. Shen, K. L. Himmel, A. J. Dupuy, D. A. Largaespada, T. Nakamura, J. D. Shaughnessy, N. A. Jenkins, and N. G. Copeland. 1999. Leukaemia disease genes: large-scale cloning and pathway predictions. Nature genetics 23:348–53.

Lian, J. B., G. S. Stein, A. Javed, A. J. van Wijnen, J. L. Stein, M. Montecino, M. Q. Hassan, T. Gaur, C. J. Lengner, and D. W. Young. 2006. Networks and hubs for the transcriptional control of osteoblastogenesis. Reviews in endocrine & metabolic disorders 7:1–16.

Linnemann, A. K., A. E. Platts, N. Doggett, A. Gluch, J. Bode, and S. A. Krawetz. 2007. Genomewide identification of nuclear matrix attachment regions: an analysis of methods. Biochemical Society transactions 35:612–617.

Linnemann, A. K., A. E. Platts, and S. A. Krawetz. 2008. Differential nuclear scaffold/matrix attachment marks expressed genes. Human Molecular Genetics 18:645–654.

Liu, J. C., C. J. Lengner, T. Gaur, Y. Lou, S. Hussain, M. D. Jones, B. Borodic, J. L. Colby, H. a Steinman, A. J. van Wijnen, J. L. Stein, S. N. Jones, G. S. Stein, and J. B. Lian. 2011. Runx2 protein expression utilizes the Runx2 P1 promoter to establish osteoprogenitor cell number for normal bone formation. The Journal of biological chemistry 286:30057–70.

Liu, W. M., F. K. Guerra-Vladusic, S. Kurakata, R. Lupu, and T. Kohwi-Shigematsu. 1999. HMG-I(Y) recognizes base-unpairing regions of matrix attachment sequences and its increased expression is directly linked to metastatic breast cancer phenotype. Cancer Research 59:5695–5703.

Ludérus, M., A. Graaf, E. Mattia, J. Blaauwen, M. Grande, L. Jong, and R. Driel. 1992. Binding of matrix attachment regions to lamin B1. Cell 70:949–959.

Manara, M. C., B. Perbal, S. Benini, R. Strammiello, V. Cerisano, S. Perdichizzi, M. Serra, A. Astolfi, F. Bertoni, J. Alami, H. Yeger, P. Picci, and K. Scotlandi. 2002. The expression of ccn3(nov) gene in musculoskeletal tumors. The American Journal of Pathology 160:849–59.

Mangan, J. K., and N. A. Speck. 2011. RUNX1 mutations in clonal myeloid disorders: from conventional cytogenetics to next generation sequencing, a story 40 years in the making. Critical reviews in oncogenesis 16:77–91.

Marsden, M. P., and U. K. Laemmli. 1979. Metaphase chromosome structure: evidence for a radial loop model. Cell 17:849–58.

Mateos-Langerak, J., S. Goetze, H. Leonhardt, T. Cremer, R. Driel, and C. Lanctôt. 2007. Nuclear architecture: Is it important for genome function and can we prove it? Journal of Cellular Biochemistry 102:1067–1075.

Maya-Mendoza, A., and A. Aranda-Anzaldo. 2003. Positional mapping of specific DNA sequences relative to the nuclear substructure by direct polymerase chain reaction on nuclear matrix-bound templates. Analytical biochemistry 313:196–207.

McStay, B., and I. Grummt. 2008. The epigenetics of rRNA genes: from molecular to chromosome biology. Annual review of cell and developmental biology 24:131–57.

Meaburn, K., P. Gudla, S. Khan, S. Lockett, and T. Misteli. 2009. Disease-specific gene repositioning in breast cancer. J Cell Biol.

Melillo, R. M., G. M. Pierantoni, S. Scala, S. Battista, M. Fedele, A. Stella, M. C. De Biasio, G. Chiappetta, V. Fidanza, G. Condorelli, M. Santoro, C. M. Croce, G. Viglietto, and A. Fusco. 2001. Critical role of the HMGI(Y) proteins in adipocytic cell growth and differentiation. Molecular and cellular biology 21:2485–95.

Melnikova, I. N., B. E. Crute, S. Wang, and N. a Speck. 1993. Sequence specificity of the core-binding factor. Journal of virology 67:2408–11.

Merkenschlager, M., and D. T. Odom. 2013. CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. Cell 152:1285–1297.

Merriman, H. L., A. Wijnen, S. Hiebert, J. P. Bidwell, E. Fey, J. Lian, J. Stein, G. S. Stein, and A. J. van Wijnen. 1995. The tissue-specific nuclear matrix protein, NMP-2, is a member of the AML/CBF/PEBP2/runt domain transcription factor family: interactions with the osteocalcin gene promoter. Biochemistry 34:13125–13132.

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

Miething, C., R. Grundler, C. Mugler, S. Brero, J. Hoepfl, J. Geigl, M. R. Speicher, O. Ottmann, C. Peschel, and J. Duyster. 2007. Retroviral insertional mutagenesis identifies RUNX genes involved in chronic myeloid leukemia disease persistence under imatinib treatment. Proceedings of the National Academy of Sciences of the United States of America 104:4594–9.

Mirkovitch, J., S. M. Gasser, and U. K. Laemmli. 1987. Relation of chromosome structure and gene expression. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 317:563–74.

Mirkovitch, J., M. E. Mirault, and U. K. Laemmli. 1984. Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. Cell 39:223–232.

Misteli, T. 2007. Beyond the sequence: cellular organization of genome function. Cell 128:787–800.

Misteli, T. 2010. Higher-order Genome Organization in Human Disease. Cold Spring Harbor Perspectives in Biology 2:a000794–a000794.

modENCODE Consortium, S. Roy, J. Ernst, P. Kharchenko, P. Kheradpour, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science (New York, NY) 330:1787–1797.

Monneron, A., and W. Bernhard. 1969. Fine structural organization of the interphase nucleus in some mammalian cells. Journal of ultrastructure research 27:266–288.

Morini, M., M. Mottolese, N. Ferrari, F. Ghiorzo, S. Buglioni, R. Mortarini, D. M. Noonan, P. G. Natali, and A. Albini. 2000. The alpha 3 beta 1 integrin is associated with mammary carcinoma cell metastasis, invasion, and gelatinase B (MMP-9) activity. International journal of cancer. Journal international du cancer 87:336–342.

Nickerson, J. A. 2001. Experimental observations of a nuclear matrix. Journal of cell science 114:463–474.

Nickerson, J. A., B. J. Blencowe, and S. Penman. 1995. The Architectural Organization of Nuclear Metabolism. Pages 67–124 *in* R. Berezney and K. W. Jeon, editors. International Review of Cytology, A Survey of Cell Biology: Structural and functional organization of the nuclear matrix., 162nd edition. Academic Press, San Diego, CA.

Nickerson, J. A., G. Krockmalnic, K. M. Wan, and S. Penman. 1997. The nuclear matrix revealed by eluting chromatin from a cross-linked nucleus. Proceedings of the National Academy of Sciences of the United States of America 94:4446–4450.

Niedojadlo, J., C. Perret-Vivancos, K.-H. Kalland, D. Cmarko, T. Cremer, R. van Driel, and S. Fakan. 2011. Transcribed DNA is preferentially located in the perichromatin region of mammalian cell nuclei. Experimental cell research 317:433–44.

Niini, T., J. Kanerva, K. Vettenranta, U. M. Saarinen-Pihkala, and S. Knuutila. 2000. AML1 gene amplification: a novel finding in childhood acute lymphoblastic leukemia. Haematologica 85:362–6.

Ogata, N. 1990. Preferential association of a transcriptionally active gene with the nuclear matrix of rat fibroblasts transformed by a simian-virus-40-pBR322 recombinant plasmid. The Biochemical journal 267:385–90.

Ong, C.-T., and V. G. Corces. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature reviews. Genetics 12:283–93.

Onodera, Y., Y. Miki, T. Suzuki, K. Takagi, J.-I. Akahira, T. Sakyu, M. Watanabe, S. Inoue, T. Ishida, N. Ohuchi, and H. Sasano. 2010. Runx2 in human breast carcinoma: its potential roles in cancer progression. Cancer science 101:2670–2675.

Ostermeier, G. C., Z. Liu, R. P. Martins, R. R. Bharadwaj, J. Ellis, S. Draghici, and S. A. Krawetz. 2003. Nuclear matrix association of the human beta-globin locus utilizing a novel approach to quantitative real-time PCR. Nucleic Acids Research 31:3257–3266.

Pande, S., G. Browne, S. Padmanabhan, S. K. Zaidi, J. B. Lian, A. J. van Wijnen, J. L. Stein, and G. S. Stein. 2013. Oncogenic cooperation between PI3K/Akt signaling and transcription factor Runx2 promotes the invasive properties of metastatic breast cancer cells. Journal of Cellular Physiology:n/a–n/a.

Papanicolaou, G. N., and H. F. Traut. 1997. The diagnostic value of vaginal smears in carcinoma of the uterus. 1941. Archives of pathology & laboratory medicine 121:211–24.

Papantonis, A., and P. R. Cook. 2010. Genome architecture and the role of transcription. Current opinion in cell biology 22:271–6.

Partin, A. W., R. H. Getzenberg, M. J. CarMichael, D. Vindivich, J. Yoo, J. I. Epstein, and D. S. Coffey. 1993. Nuclear matrix protein patterns in human benign prostatic hyperplasia and prostate cancer. Cancer research 53:744–746.

Paulson, J. R., and U. K. Laemmli. 1977. The structure of histone-depleted metaphase chromosomes. Cell 12:817–828.

Pencovich, N., R. Jaschek, A. Tanay, and Y. Groner. 2011. Dynamic combinatorial interactions of RUNX1 and cooperating partners regulates megakaryocytic differentiation in cell line models. Blood 117:e1–e14.

Peric-Hupkes, D., W. Meuleman, L. Pagie, S. Bruggeman, I. Solovei, W. Brugman, S. Gräf, P. Flicek, R. Kerkhoven, M. Lohuizen, M. Reinders, L.

Wessels, and B. Steensel. 2010. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Molecular Cell 38:603–613.

Petrov, P., and C. E. Sekeris. 1971. Early action of α-amanitin on extranucleolar ribonucleoproteins, as revealed by electron microscopic observation. Experimental Cell Research 69:393–401.

Pfaffl, M. W. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Research 29:e45.

Planagumà, J., M. Díaz-Fuertes, A. Gil-Moreno, M. Abal, M. Monge, A. García, T. Baró, T. M. Thomson, J. Xercavins, F. Alameda, and J. Reventós. 2004. A differential gene expression profile reveals overexpression of RUNX1/AML1 in invasive endometrioid carcinoma. Cancer research 64:8846–53.

Planaguma, J., M. Gonzalez, A. DOLL, M. Monge, A. GILMORENO, T. Baro, A. Garcia, J. Xercavins, F. Alameda, and M. Abal. 2006. The up-regulation profiles of p21WAF1/CIP1 and RUNX1/AML1 correlate with myometrial infiltration in endometrioid endometrial carcinoma☆. Human pathology 37:1050–1057.

Planagumà, J., M. Liljeström, F. Alameda, R. Bützow, I. Virtanen, J. Reventós, and M. Hukkanen. 2011. Matrix metalloproteinase-2 and matrix metalloproteinase-9 codistribute with transcription factors RUNX1/AML1 and ETV5/ERM at the invasive front of endometrial and ovarian carcinoma. Human pathology 42:57–67.

Pontèn, F., K. Jirström, and M. Uhlen. 2008. The Human Protein Atlas—a tool for pathology. The Journal of pathology:387–393.

Pratap, J., M. Galindo, S. K. Zaidi, D. Vradii, B. M. Bhat, J. A. Robinson, J.-Y. Choi, T. Komori, J. L. Stein, J. B. Lian, G. S. Stein, and A. J. van Wijnen. 2003. Cell growth regulatory role of Runx2 during proliferative expansion of preosteoblasts. Cancer research 63:5357–62.

Pratap, J., J. B. Lian, A. Javed, G. L. Barnes, A. J. van Wijnen, J. L. Stein, and G. S. Stein. 2006. Regulatory roles of Runx2 in metastatic tumor and cancer cell interactions with bone. Cancer metastasis reviews 25:589–600.

Pratap, J., J. B. Lian, and G. S. Stein. 2010. Metastatic bone disease: Role of transcription factors and future targets. Bone 48:1–7.

Pratap, J., J. J. Wixted, T. Gaur, S. K. Zaidi, J. Dobson, K. D. Gokul, S. Hussain, A. J. van Wijnen, J. L. Stein, G. S. Stein, and J. B. Lian. 2008. Runx2 transcriptional activation of Indian Hedgehog and a downstream bone metastatic pathway in breast cancer cells. Cancer research 68:7795–7802.

Pruitt, K. D., T. Tatusova, W. Klimke, and D. R. Maglott. 2009. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Research 37:D32–6.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Ramaswamy, S., K. N. Ross, E. S. Lander, and T. R. Golub. 2003. A molecular signature of metastasis in primary solid tumors. Nature genetics 33:49–54.

Rivera-Mulia, J. C., and A. Aranda-Anzaldo. 2010. Determination of the in vivo structural DNA loop organization in the genomic region of the rat albumin locus by means of a topological approach. DNA Research 17:23–35.

Ronneberger, O., D. Baddeley, F. Scheipl, P. Verveer, H. Burkhardt, C. Cremer, L. Fahrmeir, T. Cremer, and B. Joffe. 2008. Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 16:523–562.

Roshon, M. J., and H. E. Ruley. 2005. Hypomorphic mutation in hnRNP U results in post-implantation lethality. Transgenic research 14:179–92.

Ross-Innes, C. S., G. D. Brown, and J. S. Carroll. 2011. A co-ordinated interaction between CTCF and ER in breast cancer cells. BMC genomics 12:593.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods in molecular biology (Clifton, NJ) 132:365–386.

Rozowsky, J., G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature biotechnology 27:66–75.

Samuel, S. K., T. M. Minish, and J. R. Davie. 1997. Nuclear matrix proteins in well and poorly differentiated human breast cancer cell lines. Journal of Cellular Biochemistry 66:9–15.

De Sandre-Giovannoli, A., M. Chaouch, S. Kozlov, J.-M. Vallat, M. Tazir, N. Kassouri, P. Szepetowski, T. Hammadouche, A. Vandenberghe, C. L. Stewart, D. Grid, and N. Lévy. 2002. Homozygous defects in LMNA, encoding lamin A/C nuclear-envelope proteins, cause autosomal recessive axonal neuropathy in human (Charcot-Marie-Tooth disorder type 2) and mouse. American journal of human genetics 70:726–36.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker. 2012. The long-range interaction landscape of gene promoters. Nature 489:109–13.

Sase, T., T. Suzuki, K. Miura, K. Shiiba, I. Sato, Y. Nakamura, K. Takagi, Y. Onodera, Y. Miki, M. Watanabe, K. Ishida, S. Ohnuma, H. Sasaki, R. Sato, H. Karasawa, C. Shibata, M. Unno, I. Sasaki, and H. Sasano. 2012. Runt-related transcription factor 2 in human colon carcinoma: a potent prognostic factor associated with estrogen receptor. International journal of cancer. Journal international du cancer 131:2284–93.

Sato, M., E. Morii, T. Komori, H. Kawahata, M. Sugimoto, K. Terai, H. Shimizu, T. Yasui, H. Ogihara, N. Yasui, T. Ochi, Y. Kitamura, Y. Ito, and S. Nomura. 1998. Transcriptional regulation of osteopontin gene in vivo by PEBP2alphaA/CBFA1 and ETS1 in the skeletal tissues. Oncogene 17:1517–25.

Shariat, S. F., R. Casella, F. H. Wians, R. Ashfaq, J. Balko, T. Sulser, T. C. Gasser, and A. I. Sagalowsky. 2004. Risk stratification for bladder tumor recurrence, stage and grade by urinary nuclear matrix protein 22 and cytology. European urology 45:304–13; author reply 313.

Shore, P. 2005. A role for Runx2 in normal mammary gland and breast cancer bone metastasis. Journal of Cellular Biochemistry 96:484–489.

Sin, W. C., J. F. Bechberger, W. J. Rushlow, and C. C. Naus. 2008. Dose-dependent differential upregulation of CCN1/Cyr61 and CCN3/NOV by the gap junction protein Connexin43 in glioma cells. Journal of Cellular Biochemistry 103:1772–1782.

Sin, W.-C., M. Tse, N. Planque, B. Perbal, P. D. Lampe, and C. C. Naus. 2009. Matricellular protein CCN3 (NOV) regulates actin cytoskeleton reorganization. Journal of Biological Chemistry 284:29935–29944.

Singh, G. B., J. A. Kramer, and S. A. Krawetz. 1997. Mathematical model to predict regions of chromatin attachment to the nuclear matrix. Nucleic acids research 25:1419–25.

Soloway, M. S., V. Briggman, G. A. Carpinito, G. W. Chodak, P. A. Church, D. L. Lamm, P. Lange, E. Messing, R. M. Pasciak, G. B. Reservitz, D. B. Rukstalis, M. F. Sarosdy, W. M. Stadler, R. P. Thiel, and C. L. Hayden. 1996. Use of a new tumor marker, urinary NMP22, in the detection of occult or rapidly recurring transitional cell carcinoma of the urinary tract following surgical treatment. The Journal of urology 156:363–367.

Song, W. J., M. G. Sullivan, R. D. Legare, S. Hutchings, X. Tan, D. Kufrin, J. Ratajczak, I. C. Resende, C. Haworth, R. Hock, M. Loh, C. Felix, D. C. Roy, L. Busque, D. Kurnit, C. Willman, A. M. Gewirtz, N. A. Speck, J. H. Bushweller, F. P. Li, K. Gardiner, M. Poncz, J. M. Maris, and D. G. Gilliland. 1999. Haploinsufficiency of CBFA2 causes familial thrombocytopenia with propensity to develop acute myelogenous leukaemia. Nature genetics 23:166–75.

Spitz, F., and E. E. M. Furlong. 2012. Transcription factors: from enhancer binding to developmental control. Nature reviews. Genetics 13:613–26.

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. R. Gilbert, I. Korf, H. Lapp, H. Lehväslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Research 12:1611–1618.

Van Steensel, B., A. D. van Haarst, E. R. de Kloet, and R. van Driel. 1991. Binding of corticosteroid receptors to rat hippocampus nuclear matrix. FEBS letters 292:229–31.

Stein, G., J. Stein, A. Wijnen, J. Lian, S. Zaidi, J. Nickerson, M. Montecino, and D. Young. 2011. An architectural genetic and epigenetic perspective. Integr. Biol. 3:297–303.

Stein, G., S. Zaidi, C. Braastad, M. Montecino, A. Wijnen, J.-Y. Choi, J. Stein, J. Lian, and A. Javed. 2003. Functional architecture of the nucleus: organizing the regulatory machinery for gene expression, replication and repair. Trends in Cell Biology 13:584–592.

Stender, J. D., K. Kim, T. H. Charn, B. Komm, K. C. N. Chang, W. L. Kraus, C. Benner, C. K. Glass, and B. S. Katzenellenbogen. 2010. Genome-wide analysis of estrogen receptor alpha DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation. Molecular and Cellular Biology 30:3943–3955.

Stewart, M., A. Terry, M. Hu, M. O’Hara, K. Blyth, E. Baxter, E. Cameron, D. E. Onions, and J. C. Neil. 1997. Proviral insertions induce the expression of bone-specific isoforms of PEBP2alphaA (CBFA1): evidence for a new myc collaborating oncogene. Proceedings of the National Academy of Sciences of the United States of America 94:8646–51.

Subong, E. N., M. J. Shue, J. I. Epstein, J. V Briggman, P. K. Chan, and A. W. Partin. 1999. Monoclonal antibody to prostate cancer nuclear matrix protein (PRO:4-216) recognizes nucleophosmin/B23. The Prostate 39:298–304.

Sullivan, G. J., J. M. Bridger, A. P. Cuthbert, R. F. Newbold, W. A. Bickmore, and B. McStay. 2001. Human acrocentric chromosomes with transcriptionally silent nucleolar organizer regions associate with nucleoli. The EMBO journal 20:2867–74.

Tang, L., B. Guo, A. Javed, J. Y. Choi, S. Hiebert, J. B. Lian, A. J. van Wijnen, J. L. Stein, G. S. Stein, and G. W. Zhou. 1999. Crystal structure of the nuclear matrix targeting signal of the transcription factor acute myelogenous leukemia-1/polyoma enhancer-binding protein 2alphaB/core binding factor alpha2. The Journal of biological chemistry 274:33580–6.

The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. Nature 490:61–70.

Tijssen, M. R., A. Cvejic, A. Joshi, R. L. Hannah, R. Ferreira, A. Forrai, D. C. Bellissimo, S. H. Oram, P. a Smethurst, N. K. Wilson, X. Wang, K. Ottersbach, D. L. Stemple, A. R. Green, W. H. Ouwehand, and B. Göttgens. 2011. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Developmental cell 20:597–609.

Trevilla-García, C., and A. Aranda-Anzaldo. 2011. Cell-type-specific organization of nuclear DNA into structural looped domains. Journal of Cellular Biochemistry 112:531–40.

Tubo, R. A., and R. Berezney. 1987. Pre-replicative association of multiple replicative enzyme activities with the nuclear matrix during rat liver regeneration. The Journal of biological chemistry 262:1148–54.

Uhlén, M., E. Björling, C. Agaton, C. A.-K. Szigyarto, B. Amini, et al. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. Molecular & cellular proteomics : MCP 4:1920–32.

Uhlen, M., P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Björling, and F. Ponten. 2010. Towards a knowledge-based Human Protein Atlas. Nature biotechnology 28:1248–50.

Vaillant, F., K. Blyth, A. Terry, M. Bell, E. R. Cameron, J. Neil, and M. Stewart. 1999. A full-length Cbfa1 gene product perturbs T-cell development and promotes lymphomagenesis in synergy with myc. Oncogene 18:7124–34.

Vallacchi, V., M. Daniotti, F. Ratti, D. Stasi, P. Deho, A. Filippo, G. Tragni, A. Balsari, A. Carbone, L. Rivoltini, D. Di Stasi, A. De Filippo, G. Parmiani, N. Lazar, B. Perbal, and M. Rodolfo. 2008. CCN3/nephroblastoma overexpressed matricellular protein regulates integrin expression, adhesion, and dissemination in melanoma. Cancer research 68:715–723.

Velden, H., and F. Wanka. 1987. The nuclear matrix—Its role in the spatial organization and replication of eukaryotic DNA. Molecular biology reports.

Vogelstein, B., D. M. Pardoll, and D. S. Coffey. 1980. Supercoiled loops and eucaryotic DNA replicaton. Cell 22:79–85.

Voit, R., A. Kuhn, E. E. Sander, and I. Grummt. 1995. Activation of mammalian ribosomal gene transcription requires phosphorylation of the nucleolar transcription factor UBF. Nucleic acids research 23:2593–9.

Vradii, D., S. K. Zaidi, J. B. Lian, A. J. van Wijnen, J. L. Stein, and G. S. Stein. 2005. Point mutation in AML1 disrupts subnuclear targeting, prevents myeloid differentiation, and effects a transformation-like phenotype. Proceedings of the National Academy of Sciences of the United States of America 102:7174–9.

Wan, K., J. Nickerson, G. Krockmalnic, and S. Penman. 1999. The nuclear matrix prepared by amine modification. Proceedings of the National Academy of Sciences of the United States of America 96:933–938.

Wang, J., J. Zhuang, S. Iyer, X. Lin, T. Whitfield, M. Greven, B. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. Rando, E. Birney, R. Myers, W. Noble, M. Snyder, and Z. Weng. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research 22:1798–1812.

Wang, L., J. S. Brugge, and K. A. Janes. 2011a. Intersection of FOXO- and RUNX1-mediated gene expression programs in single breast epithelial cells during morphogenesis and tumor progression. Proceedings of the National Academy of Sciences of the United States of America 108:E803–12.

Wang, L., A. Gural, X.-J. Sun, X. Zhao, F. Perna, G. Huang, M. a Hatlen, L. Vu, F. Liu, H. H. Xu, T. Asai, T. Deblasio, S. Menendez, F. Voza, Y. Jiang, P. a Cole, J. Zhang, A. Melnick, R. G. Roeder, and S. D. Nimer. 2011b. The Leukemogenicity of AML1-ETO Is Dependent on Site-Specific Lysine Acetylation. Science (New York, NY) 333:765–9.

Wang, Q., T. Stacy, M. Binder, M. Marin-Padilla, A. H. Sharpe, and N. A. Speck. 1996. Disruption of the Cbfa2 gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. Proceedings of the National Academy of Sciences of the United States of America 93:3444–3449.

Wansink, D. G., W. Schul, I. van der Kraan, B. van Steensel, R. van Driel, and L. de Jong. 1993. Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. The Journal of cell biology 122:283–93.

Van Wijnen, A. J., J. P. Bidwell, E. G. Fey, S. Penman, J. B. Lian, J. L. Stein, and G. S. Stein. 1993. Nuclear matrix association of multiple sequence-specific DNA binding activities related to SP-1, ATF, CCAAT, C/EBP, OCT-1, and AP-1. Biochemistry 32:8397–402.

Van Wijnen, A. J., G. S. Stein, J. P. Gergen, Y. Groner, S. W. Hiebert, Y. Ito, P. Liu, J. C. Neil, M. Ohki, and N. Speck. 2004. Nomenclature for Runt-related (RUNX) proteins. Oncogene 23:4209–10.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1:80–83.

Wilson, R. H. C., and D. Coverley. 2013. Relationship between DNA replication and the nuclear matrix. Genes to cells : devoted to molecular & cellular mechanisms 18:17–31.

Xu, M., and P. R. Cook. 2008. Similar active genes cluster in specialized transcription factories. The Journal of cell biology 181:615–23.

Yoneda, T., P. J. Williams, T. Hiraga, M. Niewolna, and R. Nishimura. 2001. A bone-seeking clone exhibits different biological properties from the MDA-MB-231 parental human breast cancer cells and a brain-seeking clone in vivo and in vitro. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research 16:1486–1495.

Young, D. W., M. Q. Hassan, J. Pratap, M. Galindo, S. K. Zaidi, S. Lee, X. Yang, R. Xie, A. Javed, J. M. Underwood, P. Furcinitti, A. N. Imbalzano, S. Penman, J. A. Nickerson, M. A. Montecino, J. B. Lian, J. L. Stein, A. J. van Wijnen, G. S. Stein, and A. Wijnen. 2007a. Mitotic occupancy and lineage-specific transcriptional control of rRNA genes by Runx2. Nature 445:442–446.

Young, D. W., M. Q. Hassan, X.-Q. Yang, M. Galindo, A. Javed, S. K. Zaidi, P. Furcinitti, D. Lapointe, M. Montecino, J. B. Lian, J. L. Stein, A. Wijnen, G. S. Stein, and A. J. van Wijnen. 2007b. Mitotic retention of gene expression patterns by the cell fate-determining transcription factor Runx2. Proceedings of the National Academy of Sciences of the United States of America 104:3189–3194.

Zaidi, S. K., A. Javed, J. Y. Choi, A. Wijnen, J. L. Stein, J. B. Lian, G. S. Stein, and A. J. van Wijnen. 2001. A specific targeting signal directs Runx2/Cbfa1 to subnuclear domains and contributes to transactivation of the osteocalcin gene. Journal of cell science 114:3093–3102.

Zaidi, S. K., A. Javed, J. Pratap, T. M. Schroeder, J. J Westendorf, J. B. Lian, A. Wijnen, G. S. Stein, J. L. Stein, and A. J. van Wijnen. 2006. Alterations in intranuclear localization of Runx2 affect biological activity. Journal of cellular physiology 209:935–942.

Zaidi, S. K., D. W. Young, J.-Y. Choi, J. Pratap, A. Javed, M. Montecino, J. L. Stein, A. J. van Wijnen, J. B. Lian, and G. S. Stein. 2005. The dynamic organization of gene-regulatory machinery in nuclear microenvironments. EMBO Reports 6:128–133.

Zaidi, S. K., D. W. Young, A. Javed, J. Pratap, M. Montecino, A. van Wijnen, J. B. Lian, J. L. Stein, G. S. Stein, and A. Wijnen. 2007. Nuclear microenvironments in biological control and cancer. Nature Reviews Cancer 7:454–463.

Zehnbauer, B. A., and B. Vogelstein. 1985. Supercoiled loops and the organization of replication and transcription in eukaryotes. BioEssays 2:52–54.

Zeng, C., S. McNeil, S. Pockwinse, J. Nickerson, L. Shopland, J. B. Lawrence, S. Penman, S. Hiebert, J. B. Lian, A. J. van Wijnen, J. L. Stein, G. S. Stein, and A. Wijnen. 1998. Intranuclear targeting of AML/CBFalpha regulatory factors to nuclear matrix-associated transcriptional domains. Proceedings of the National Academy of Sciences of the United States of America 95:1585–1589.

Zeng, C., A. Wijnen, J. Stein, S. Meyers, W. Sun, L. Shopland, J. Lawrence, S. Penman, J. Lian, G. Stein, and S. Hiebert. 1997. Identification of a nuclear matrix targeting signal in the leukemia and bone-related AML/CBF-alpha transcription factors. Proceedings of the National Academy of Sciences of the United States of America 94:6746–51.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, and C. Nussbaum. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biology 9:R137.

Zhang, Y., R.-L. Xie, C. M. Croce, J. L. Stein, J. B. Lian, A. J. van Wijnen, G. S. Stein, and A. Wijnen. 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 108:9863–9868.

Zink, D., M. D. Amaral, A. Englmann, S. Lang, L. A. Clarke, C. Rudolph, F. Alt, K. Luther, C. Braz, N. Sadoni, J. Rosenecker, and D. Schindelhauer. 2004a. Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei. The Journal of cell biology 166:815–25.

Zink, D., A. H. Fischer, and J. A. Nickerson. 2004b. Nuclear structure in cancer cells. Nature Reviews Cancer 4:677–87.

Zwart, W., V. Theodorou, M. Kok, S. Canisius, S. Linn, and J. S. Carroll. 2011. Oestrogen receptor-co-factor-chromatin specificity in the transcriptional regulation of breast cancer. The EMBO journal 30:4764–76.