# UCLA
## UCLA Electronic Theses and Dissertations

**Title**

Systems Biology Approaches Investigating Cardiac Epigenome Dynamics in Heart Hypertrophy and Failure

**Permalink**

https://escholarship.org/uc/item/5z2964kp

**Author**

Chen, Haodong

**Publication Date**

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Systems Biology Approaches Investigating Cardiac

Epigenome Dynamics in Heart Hypertrophy and Failure

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in

Molecular, Cellular and Integrative Physiology

by

Haodong Chen

2013

ABSTRACT OF THE DISSERTATION

Systems Biology Approaches Investigating Cardiac

Epigenome Dynamics in Heart Hypertrophy and Failure

by

Haodong Chen

Doctor of Philosophy in Molecular, Cellular and Integrative Physiology

University of California, Los Angeles, 2013

Professor Thomas M. Vondriska, Chair

Heart failure is a syndrome resulting from multiple genetic and environmental factors. In response to neural/hormonal changes or hemodynamic stress, the heart can generate extra force through hypertrophic growth. Chronic hypertrophy, however, is deleterious because it leads to irreversible decrease of cardiac contractility. This process requires many molecular and cellular abnormalities in the cells of the failing heart. Cardiac hypertrophy and heart failure arise as the result of multiple biological processes acting within the context of multicomponent, interrelated cellular networks. The past decade has seen an

explosion of research using high-throughput techniques, identifying hundreds of genes involved in disease pathogenesis. These studies raise the question of how global, coordinated changes in transcription are precisely regulated within the cardiac nucleus. We reason that altered chromatin structure and DNA methylation endow distinct gene expression patterns during the development of disease.

This dissertation details several systems biology approaches aimed at defining cardiac epigenomic variations in the hypertrophied and failing heart. Chapter 1 reports the genome-wide nucleosome positioning in normal and pressure overload-induced hypertrophic cardiac myocytes. Because nucleosomes are the basic units of chromatin, their localization fundamentally affects the actions many types of regulatory machinery. In Chapter 2, we examined the non-histone chromatin structural protein High Mobility Group Protein B2, which we had previously determined to be a regulator of hypertrophic growth. We measured its genome-wide binding pattern in both normal and $\alpha$-adrenergic receptor agonist-induced hypertrophic cardiac myocytes. Our findings demonstrate that HMGB2 is reorganized away from coding regions in the genome of the hypertrophic myocyte, an action we interpret as globally coordinated relaxation of specific loci. In Chapter 3, we detail a comprehensive analysis of the DNA methylome in both normal and $\beta$-adrenergic agonist-induced failing heart. We used two different mouse strains with different susceptibility to heart failure and

identified several strain-specific DNA methylation modules. The results presented here advance our understanding of the cardiac epigenome and provided essential insights into global mechanisms of chromatin structural remodeling.

The dissertation of Haodong Chen is approved.


Aldons J. Lusis


W. Robb MacLellan


Yibin Wang


Xinshu (Grace) Xiao


Thomas M. Vondriska, Committee Chair


University of California, Los Angeles

2013

DEDICATION

This dissertation is dedicated to my parents for their love,

support, encouragement, and inspiration.


谨以此论文献给我的父母

## TABLE OF CONTENTS

# LIST OF FIGURES

## LIST OF TABLES

# ACKNOWLEDGEMENTS

I would like to first give my thanks to my mentor Dr. Thomas Vondriska for his guidance and support over the past years. Tom taught me everything I need to become a scientist, including how to write a fundable grant and how to efficiently communicate with other scientists. Furthermore, he always encouraged me to think independently and critically. I am extremely grateful to the members of the Vondriska lab for their generously support and close friendship: Mr. Doug Chapski, Dr. Sarah Franklin, Mr. Ricardo Gray, Ms. Elaheh Karbassi, Mr. Todd Kimball, Dr. Yifeng Li, Ms. Rachel Lopez, Dr. Andrea Matlock, Dr. Scherise Mitchell-Jordan, Ms. Emma Monte, Dr. Michelle Parvatiyar, Dr. Manuel Rosa Garrido, and Mr. Michael Zhang. In particular, I want to thank Sarah for her help in proteomics, Scherise and Todd for their help in mouse Echocardiography, Manuel and Ricardo for their help in chromatin immunoprecipitation, and Emma and Elaheh for the two songs they wrote for me and the Halloween party idea.

I want to thank the members of my dissertation committee for their exceptional kindness and thoughtful suggestions: Drs. Aldons J. Lusis, W. Robb MacLellan, Yibin Wang and Xinshu (Grace) Xiao. This dissertation would not have made it without my collaborators and thank you to Drs. Zugen Chen, Luz Orozco, Christoph Rau, Vincent Ren and Jessica Wang. Moreover, I wish to thank Drs. Siavash Kurdistani, Stanley Nelson, Matteo Pellegrini and Zhilin Qu for their

inspiring discussion.

I also wish to thank my brilliant fiancée Jun for her continual support and patience. She was always there to inspire me and I owe her a huge thanks. I must thank my parents for their guidance and sacrifices throughput my life.

Lastly, I want to acknowledge the permission to reuse/reproduce manuscripts as part of my dissertation from the American Society for Biochemistry and Molecular Biology and Elsevier Limited. I also want to thank the American Heart Association for the pre-doctoral fellowship award. The research in Vondriska lab was supported by the American Heart Association, National Institutes of Health/National Heart, Lung and Blood Institute and the Laubisch Endowment.

# Biographical Sketch

**Education**

| Institution & Location | Dates Attended | Degree | Conferred (mm/yyyy) | Field of Study |
|---|---|---|---|---|
| **University of Science and Technology of China** | 09/2003~06/2007 | BS | 06/2007 | Biotechnology |
| **University of California, Los Angeles** | 09/2008~ | PhD Student | | Molecular Physiology |

1. **ACADEMIC AND PROFESSIONAL HONORS**

- New Investigator Travel Award, BCVS Scientific Sessions, Las Vegas, NV, 2013

- Predoctoral Fellowship, American Heart Association, Western States Affiliate, 2011-2013 (Priority Score: 1.26)

- Student Scholarship, Chinese-American Engineers and Scientists Association of Southern California, 2012

- Abstract Travel Award, American Heart Association, Orlando, FL, 2011

- Guo Moruo Scholarship (President's Scholarship), University of Science and Technology of China, 2006

2. **PEER REVIEWED PUBLICATIONS**

**Research Papers:**

- **Chen H**, Monte E, Vondriska TM, Franklin S: Systems proteomics of healthy and diseased chromatin. *Methods Mol Biol*. 2013, 1005:77-93

- Monte E, **Chen H**, Kolmakova M, Parvatiyar M, Vondriska TM, Franklin S: Quantitative analysis of chromatin proteomes in disease. *J Vis Exp*. 2012, 70, e4294.

- **Chen H**, Monte E, Parvatiyar MS, Franklin S, Vondriska TM: Structural considerations for chromatin state models with transcription as a functional readout. *FEBS Lett*. 2012, 586(20):3548-3554.

- Mitchell-Jordan S, **Chen H**, Franklin S, Stefani E, Bentolila LA, Vondriska TM: Features of endogenous cardiomyocyte chromatin revealed by super-resolution STED microscopy. *J Mol Cell Cardiol*. 2012, 53(4): 552-558.

- Franklin S, **Chen H**, Mitchell-Jordan SA, Ren S, Wang Y, Vondriska TM: Quantitative analysis of the chromatin proteome in disease reveals remodeling principles and identifies

high mobility group protein B2 as a regulator of hypertrophic growth. *Mol Cell Proteomics* 2012, 11:M111.014258.

- Franklin S, Zhang MJ, **Chen H**, Paulsson AK, Mitchell-Jordan SA, Li Y, Ping P, Vondriska TM: Specialized compartments of cardiac nuclei exhibit distinct proteomic anatomy. *Mol Cell Proteomics* 2011, 10(1):M100.000703.

**Abstracts:**

- **Chen H,** Orozco LD, Rubbi L, Wang J, Rau CD, Chapski D, Ren S, Wang Y, Pellegrini M, Lusis AJ, Vondriska TM: Genome-wide Bisulfite Sequencing After Isoproterenol Identifies DNA Methylation Modules Influencing Heart Failure Susceptibility. BCVS Scientific Sessions 2013. Las Vegas, NV.

- **Chen H**, Franklin S, Chen Z, Nelson SF, Mitchell-Jordan SA, Ren S, Wang Y, Vondriska TM: Nucleosome Positioning is a Mechanism of Global Chromatin Structural Remodeling in Cardiac Hypertrophy. American Heart Association Scientific Sessions 2011. Orlando, FL. (**Oral presentation**)

3. **INVITED LECTURES AND SEMINARS**

- Chromatin Biology and Epigenetics in Heart Diseases: What Central Dogma Didn't Tell You. Invited Seminar for Cal Poly Pomona RISE Program. Cal State Poly Univ, Pomona. CA. 2012 Feb 3.

- Chromatin Biology and Epigenetics in Heart Diseases. MCIP Student Seminar, University of California, Los Angeles, CA. 2012 July 10.

- Dynamics of Genome-Wide DNA Binding by High Mobility Group Protein B2 Reveals Novel Role to Maintain Normal Gene Expression in Cardiomyocytes. MBI Retreat Talk, UCLA. 2012 Oct 19.

- Chromatin Biology and Epigenetics in Heart Diseases. Physiology Happy Hour and Seminar, University of California, Los Angeles, CA. 2012 Nov 30.

- Genome-Wide DNA Methylation Profile in Heart Hypertrophy. Annual MCIP Retreat, Santa Anita Park. 2013 Feb 16.

# PREFACE

**Part I. Quantitative Analysis of the Chromatin Proteome in Disease Reveals Remodeling Principles and Identifies High Mobility Group Protein B2 as a Regulator of Hypertrophic Growth.**

Sarah Franklin, Haodong Chen, Scherise Mitchell-Jordan, Shuxun Ren, Yibin Wang, Thomas M. Vondriska

Heart is the organ that supplies the main pumping force to propel blood to the whole body. Cardiac hypertrophy, which is defined as the increase of heart chamber wall mass and individual myocyte size, can be either physiological or pathological. Under the condition of exercise or pregnancy, heart function can be increased by physiological hypertrophy. This process is reversible and does not cause cardiac dysfunction. However, under pathological conditions, e.g. hypertension or myocardial infarction, heart will undergo a change into pathological hypertrophy, and finally develop into heart failure with thin chamber wall, enlarged chamber size and poor heart function (measured as decreased ejection fraction). It has been known that the gene expression profile of failing heart is similar with fetal heart, compared with adult heart. This process is called fetal gene reprogramming, namely gene expression is

reprogrammed into fetal profile. Two outcomes of fetal gene reprogramming can affect pathophysiological function of myocardium: a) glycolytic metabolic pathway re-construction and b) contractile apparatus re-organization. It was found that the source of energy generation in failure heart switches from fatty acid β-oxidation to glycolysis, which is mainly used by newborn heart. There is also a switch of cardiac myosin heavy chain (MHC), from α isoform to β isoform. Actually, mRNA microarray assay has shown that fetal gene reprogramming involves thousands of genes. Many cardiac specific transcription factors are also involved, reflecting the complexity of this pathological process.

Transcriptional regulation must be preceded by non-random structural reorganization of the genome, such that stimulus-specific transcriptional regulators are recruited to the correct genomic regions and excluded from the wrong ones. During mitosis, eukaryotic chromosomes adapt an ordered and stunningly reproducible three-dimensional structure. During interphase, however, and in cells that do not divide (such as adult cardiomyocytes and neurons) the structure of the genome is much less clear. Chromosome territories have been described[1], which are thought to facilitate co-localization of similarly regulated genes within the nucleus, and recent advances in chromosomal conformation capture techniques have provided exciting new insights into the global structure of the interphase genome, suggesting the structure resembles a fractal globule[2]. However, changes in this structure following stimulation, and the specific proteins responsible for dynamics of genomic structure, are less well understood.

From a structural standpoint, DNA (~146-147 base pairs) is packaged around the nucleosome octamer[3], itself composed of two copies each of four core histones (H2A, H2B, H3 and H4). Previous work has described a 30nm fiber-level feature[4] and it has been shown that chromatin structural molecules such as histone H1[5–7] and high mobility group proteins[8,9]

contribute to higher order structure, and that core nucleosome histone variants[10] may alter DNA packaging. The ability of histone-modifying enzymes to alter gene expression, presumably by altering chromatin compaction, has been well studied in various cell types and disease states[11–13]. These chromatin remodelers, however, are insufficient to explain global coordination of gene expression that occurs simultaneously and across distinct loci in the genome. Toward this end, several recent investigations have employed proteomic/genomic approaches to characterize large-scale changes in chromatin structure in eukaryotes[14–18]. Global principles for how the entire genome is structurally remodeled for gene expression changes are lacking. We reasoned that *modulation of the abundance and stoichiometry of the chromatin backbone itself* could be a mechanism by which the cell rapidly alters gene expression across the genome in a coordinated manner. To test this hypothesis and address the general paucity of knowledge about chromatin-bound proteins in the heart, we have carried out a fundamentally new type of analysis of chromatin remodeling -- one that quantifies the constituents of the genome's protein backbone at different stages of global transcriptional activation. Our mass spectrometry data reveal the itineraries of chromatin modifiers and chromatin structural proteins during the development of disease and provide a novel comprehensive blueprint for chromatin-associated proteins in the heart.

## Experimental Procedures

### Mouse model of cardiac hypertrophy and failure

Adult male balb/c mice aged 8-12 weeks (Charles River Laboratories) were used for this study. All protocols involving animals conform to the NIH Guide for the Care and Use of Laboratory Animals and were approved by the UCLA Animal Research Committee. The murine

model of transverse aortic banding-induced cardiac hypertrophy was performed as described previously[19–21]. Animals were anesthetized with 1.5-2.0% isoflurane, intubated, and ventilated with 1.5-2.0% isoflurane in 95% $O_2$ / 5% $CO_2$. After shaving hair from the animal, the chest was entered from the left side via the third intercostal space, the aorta identified at the T8 region and a venous vascular clamp (Fine Science Tools), outfitted with a band of silastic tubing at the distal edge of one of the clamps, placed around the vessel. The internal diameter of the resulting modified clamp was that of a 27 ½ gauge needle, a common diameter for aortic banding in the mouse; the clamp is of a size that does not impede pulmonary function. Once the aorta was clamped, distal blood flow was measured and quantified using a flow probe (Transonic Instruments). The chest was then closed using 6-0 proline suture, during which negative pressure in the thorax was reestablished by removing air with a PE-50 chest tube attached to a syringe. SHAM operated mice underwent the same procedure without placement of aortic clamp.

***Echocardiography***

ECHO was used to determine cardiac parameters in live mice as described[20,21], including the following indices: left ventricular size (end-diastolic and end-systolic dimension), wall thickness (intra-ventricular septum and posterior wall thickness), ventricular mass, ventricular function (ejection fraction), and blood flow. ECHO was performed on mice sedated with isoflurane vaporized (2.5% for induction, 1.0% for maintenance) in oxygen. The animal's chest was shaved and positioned in the left lateral decubitus position for ultrasonic imaging with Vevo 770 high-resolution ECHO system equipped with a 35 MHz transducer. The short axis view (M-mode) and the long axis view (Doppler) measurements were performed and data stored for off-line analysis. LV chamber dimensions, VST, PWT, and LV mass index (the ratio of LV mass to

body weight) were obtained from M-mode images; LV systolic function is also assessed from these measurements by calculating EF. All mice underwent ECHO analyses once before TAC or SHAM surgery, once a day after and then once every 5 days for the duration of the study. Animals were considered "hypertrophic" when their LV mass was greater than the mean of the control animal with no depression of LV function as measured by EF; animals were considered in "failure" when the EF was significantly decreased below the mean of the control animal. These phenotypes corresponded to roughly 2 and 4 weeks after TAC surgery in most animals. Data from mice appear in Figures 1-3 and Supplemental Figures 1-4 and 6.

### *Nuclear Isolation and Fractionation*

All buffers used for cell isolation or fractionation in this study contained the following protease, phosphatase and deacetylase inhibitors, respectively: 0.1mM phenylmethanesulfonylflouride, protease inhibitor cocktail pellet (Roche), 0.2mM sodium orthovanadate, 0.1 mM sodium flouride and 10mM sodium butyrate.

Cardiac nuclei were isolated as previously described[22]. Briefly, hearts were excised, washed in PBS, minced with scissors and homogenized in a glass dounce in buffer containing 10mM Tris (pH 7.4), 250 mM sucrose, 1mM EDTA, 0.15% NP-40. After homogenization the suspension was passed through a 100um nylon strainer (BD Falcon, #352360) to remove any large insoluble material. Subcellular fractionation was carried out by centrifuging at 1,000g to pellet crude nuclear fraction. The crude nuclear pellet was resuspended in homogenization buffer, layered on a 2M sucrose pad and centrifuged at 7,500g for 5 min to isolate the enriched nuclear fraction. All steps were carried out at 4°C. The initial relative purity of the nuclei fraction was evaluated via western blotting for compartment specific proteins and compared to

unfractionated cells from the heart (whole heart lysate; WHL) using antibodies to histone H3 and GAPDH. Whole heart lysate (WHL) was made by homogenizing the entire heart in 20mM Tris (pH 7.4), 150mM NaCl, 1mM EDTA, 1mM EGTA, 1% triton X-100, 2.5mM sodium pyrophosphate, 1mM beta-glycerophosphate, followed by sonication and centrifugation at 13,000 rpm to remove insoluble material. We consistently achieve ≥80% enrichment of nuclei with this method as observed by electron microscopy and western blotting analysis[22].

***Purification of cardiac chromatin and nucleoplasm and acid extraction of nuclear proteins***

Following isolation of nuclei, further fractionation was carried out to separate nucleoplasm from chromatin as previously described[22]. Briefly, isolated nuclei were resuspended in buffer (20mM HEPES (pH 7.6), 7.5 mM $MgCl_2$, 30 mM NaCl, 1M Urea, 1% NP-40) to solubilize the nuclear membrane and extract soluble proteins in the nucleoplasm. After solubilization samples were centrifuged at 13,000g for 10 min to pellet the insoluble chromatin and remove the nucleoplasm fraction. The chromatin pellet was washed with PBS, solubilized in 50mM Tris (pH 8), 10 mM EDTA, 1% SDS, sonicated to shear the DNA and centrifuged at 13,000g to extract proteins (referred to as chromatin fraction). For acid extraction, isolated chromatin was treated with 400μL of 0.4N $H_2SO_4$ and placed on a rotator at 4°C overnight. Samples were centrifuged at 16,000g to pellet acid-insoluble material. The supernatant was treated with 132μL of TCA in a drop wise manner, inverting the tube between drops, and placed on ice for 30 min to precipitate the acid extracted proteins. Precipitated proteins were collected by centrifuging at 16,000g for 15 min (4°C). The pellet was washed twice with ice-cold acetone without disturbing the pellet and allowed to air dry. The pellet was resuspended in buffer

containing 1% SDS, 50mM Tris (pH 8) and 10mM EDTA. Proteins extracted in this manner were examined in Figures 1-3 and Supplemental Figures 2-4 and 6.

*Isolated Myocytes*

Neonatal rat ventricular myocytes (NRVMs; used in Figure 4 and for the microarray studies in Figure 5 and Supplemental Figure 5) were obtained by enzymatic dissociation from 1 day old litters and plated in DMEM media (Invitrogen, #11965) containing 1% penicillin, 1% streptomycin, 1% insulin-transferrin-sodium selenite supplement and 10% fetal bovine serum for the first 24 hr after which the cells are cultured in serum- and antibiotic-free media. NRVMs were treated with 50nM siRNA targeted to HMGB1 (25nM each SI03034696 and SI03111626; Qiagen) or HMGB2 (25nM each SI02877252 and SI02877266; Qiagen) for 72 hrs. Control cells were treated with 50nM scrambled siRNA (Cat.# 1027280; Qiagen). Transfections were performed with Lipofectamine RNAiMax Transfection Reagent (Invitrogen). For cell size analysis NRVMs were fixed with 4% paraformaldehyde, incubated with rhodamine-conjugated phalloidin (Invitrogen) to stain the actin filaments and imaged on a Nikon Eclipse TE2000-U. The average size of treated and untreated NRVMs was quantified by measuring the area of actin staining in Photoshop (Adobe). For each condition at least 50 cells were quantified in three separate experiments. To induce hypertrophy in isolated NRVMs, cells were treated with one of the following three hypertrophic agonists: isoproterenol (1μM), phenylephrine (10μM) or endothelin-1(1nM) for 48 hours.

*Electrophoresis and western blotting*

Proteins were separated by standard SDS-PAGE using Laemmli buffer. Gels (12%) were stained with Oriole (BioRad). For western blotting, proteins were transferred to nitrocellulose,

membranes blocked with milk and protein signals detected by enzyme linked chemiluminescence (GE Biosciences). Ponceau staining of membranes was used to confirm transfer and protein loading. Antibodies used in this study were as follows, including sources: Histone H3 (Abcam, ab1791, 1:10,000 dilution); Histone H4 (Abcam, ab10158, 1:1000 dilution); SNRP70 (Abcam, ab51266, 1:400 dilution); HMGB1 (Abcam, ab18256, 1:1000 dilution); HMGB2 (Abcam, ab67282, 1:1000 dilution); Gapdh (Santa Cruz, sc-20357, 1:1000 dilution); Histone H3-trimethylated-K4 (Abcam, ab8898, 1:500 dilution); Histone H3-mono/di/tri-methylated-K4 (Millipore, 04-791, 1:1000 dilution); Histone H1.0 (Abcam, ab11079, 1:1000 dilution); Histone H3-trimethylated-K4 (Abcam, ab71998, 1:500 dilution); Ruvb-like 1 (Proteintech, 10210-2-AP, 1:1000 dilution). Nuclear lysate control (Nuc) was made by solubilizing intact nuclei in buffer containing 1% SDS, 50mM Tris (pH 8) and 10mM EDTA followed by centrifugation at 13,000 rpm for 15 min to remove insoluble material. Hela lysate control (He) was made by solubilizing Hela cells, isolated by scraping cells from culture dishes, in 20mM Tris (pH 7.4), 150mM NaCl, 1mM EDTA, 1mM EGTA, 1% triton X-100, 2.5mM sodium pyrophosphate, 1mM beta-glycerophosphate. The homogenate was then sonicated three times for 5 sec each followed by centrifugation at 13,000rpm to remove insoluble material.

***Enzyme digestion***

Proteins isolated from acid extraction of chromatin were separated by SDS-PAGE. Each gel lane was cut into 25 slices (~2mm each, labeled A through Y in Supplemental Figure 2) for protein identification by mass spec. Gel plugs were dehydrated in acetonitrile and dried in a speedvac. Samples were reduced and alkylated with 10 mM dithiotreitol and 10mM TCEP solution in 50 mM $NH_4HCO_3$ (30 min at 56°C) and 100 mM iodoacetamide (45 min in dark),

respectively. Gels were washed with 50 mM $NH_4HCO_3$, dehydrated with acetonitrile, and dried down in a speedvac. Gel pieces were then swollen in digestion buffer containing 50mM $NH_4HCO_3$, and 20.0 ng/µL of trypsin (37°C, overnight). Peptides were extracted with 0.1%TFA in 50% acetonitrile solution, dried down and resuspended in 0.1% formic acid. For each condition (basal, hypertrophy and failure) three biological (*de novo* preparation of samples from different animals) and two technical (multiple LC/MS/MS experiments on the same preparation) replicates were analyzed by mass spectrometry.

### *Mass spectrometry analyses and database searching via SEQUEST.*

Extracted peptides were analyzed by nano-flow LC/MS/MS on a Thermo Orbitrap with dedicated Eksigent nanopump using a reversed phase column (75 µm i.d. 10 cm, BioBasic C18 5µm particle size, New Objective) and a flow rate of 200 nL/min. For peptide separation a linear gradient was utilized from 95% Buffer A (0.1% formic acid, 2% acetonitrile) and 5% Buffer B (0.1% formic acid, 20% water in acetonitrile) to 50% Buffer A and 50% Buffer B over 60 minutes. Spectra were acquired in data-dependent mode with dynamic exclusion where the instrument selects the top six most abundant ions in the parent spectra for fragmentation. Data were searched against the mouse IPI database (version 3.76; 57, 605 entries) using the SEQUEST algorithm[23] in the BioWorks software program version 3.3.1 SP1. False discovery rate, which was calculated on several independent datasets within this study by reverse database searching, ranged from 1.4-1.7%. All spectra used for identification had deltaCN>0.1, consensus score ≥20 and met the following Xcorr criteria: >3 (+2), >4 (+3), and >5 (+4). Peptides with +1 charge were not used for identification of proteins; however, some +1 peptides were included in quantitative measurements (following manual inspection) only in the case of histone variants

(see Figure 2). Searches required full tryptic cleavage, ≤3 missed cleavages and were performed with the differential modifications of carbamidomethylation on cysteine and methionine oxidation. Mass tolerance was 2 Da for precursor and 1 Da for product ions. All proteins were identified on the basis of two or more unique peptides.

***Bioinformatics and protein annotation.***

The molecular weight, isoelectric point (pI) and the GRAVY (Grand Average of Hydropathy) score were calculated by ProtParam tool provided in the SeqUtils package of Biopython (v1.50; Supplemental Figure 3F). Protein expression plots were generated as described[24]. Redundancy in proteins was eliminated at the primary sequence level by manual inspection using CLUSTAL to compare the sequences in UniProt. Genome analysis (i.e. for determining from where in the genome mRNAs for the detected proteins were transcribed) was performed from the NCBI genome browser following conversion of IPI to ref-seq accession numbers (Supplemental Figures 3E and 4D). GO annotation enrichment analysis was performed using the DAVID Bioinformatics Resource (v6.7) developed by the NIAID (NIH; Figures 3B and 5A and Supplemental Figure 4A)[25]. The Interpro and KEGG analysis functions of DAVID were utilized to determine enrichment in protein domains and functional pathways (Supplemental Figures 4B, 4C, 5A and 5B).

Mass spectrometry raw data were searched two ways in this study: by "conventional" SEQUEST searching (which were used to generate Figure 2) and via a quantitative analysis using the Rosetta Elucidator software (from Microsoft, which led to Figures 1 and 3 and Supplemental Figures 3A-D and 4D)[26,27]. The reason for this approach was that a main component of chromatin—histones—are small molecular weight proteins that yield few peptides

following enzymatic digestion and which often differ by only a few amino acids. To accurately quantify them, we sought to confirm all peptide identifications manually, which is more easily done after conventional SEQUEST search. For Elucidator analyses, peptides across the entire chromatographic run for each sample were aligned between mass spec runs and between conditions (basal, hypertrophy, failure). The peak intensity for each eluting peptide was calculated as area under the curve. To determine protein abundance, intensity data for all peptides mapping to a protein were combined and data from 3 biological and 2 technical replicates was averaged for each of the three conditions. Proteins whose intensity changed $\geq 2$ fold between conditions with a p-value $\leq 0.01$ were considered to be statistically significant. To identify modules of proteins with similar expression behavior (Figure 3), intensity values were converted to z-scores. Proteins were then clustered using a self-organizing map with x and y nodes of of 3,3 and cosine correlation. Intensity data was coupled to peptide identification which was determined using the SEQUEST algorithm using the same criteria described above. Mass tolerance was 0.5 Da for precursor and 1 Da for product ions and all identifications were based on $\geq 2$ peptides.

***Quantitative Real Time PCR Analysis***

Total RNA was isolated from the left ventricle of the heart (Supplemental Figure 1C) or from cultured NRVMs (Figure 4C) using TRIzol (Invitrogen) according to the manufacturer's protocol. Total RNA was transcribed using SuperScript First-Strand Synthesis system for RT-PCR (Invitrogen) according to the manufacturer's protocol to produce cDNA. cDNA transcripts were amplified on the iCycler iQ real-time PCR detection system with iQ SYBR Green Supermix (Bio-Rad, Hercules, Calif). Expression levels were analyzed using the iQ5 Optical

Systems software v2.0 and normalized against GAPDH by subtracting the mean cycle number for each experimental group from the mean cycle number for GAPDH from the same group. The normalized means were then applied to the formula $((2^{(cycle\#)})^{-1})$ to calculate fold change. Primers used in this study are as follows: GAPDH F-5′-CCCACTAACATCAAATGGGG-3′ R-5′-CCTTCCACA ATGCCAAAGTT-3′; HMGB2 F-5'-AAGCCGCGGGGCAAGATGTC-3' R-5'-TGCCCTTGGCACGGTAT GCA-3'; ANF F-5'-CTGATGGATTTCAAGAACCTGCT-3' R-5'-CTCTGGGCTCCAATCCTGTC-3'; SERCA2a F5'-CCTTCTACCAGCTGAGTCATTT-3' R5'-CAGATGGAGCCCACGACCCA-3'; α-MHC F-5'-GAACAGCTGGGAGAAGGGGG-3' R-5'-GCCTCTGAGGCTATTCTATTGG-3'; and β-MHC F-5'-CTCAACTGGGAAGAGCATCCA-3' R-5'-CCTTCAGCAAACTCTGGAGGC-3'.

*Microarrays*

Total RNA was isolated from cultured NRVMs after treatment with HMGB2 or scrambled siRNA using TRIzol (Invitrogen) according to the manufacturer's protocol (Figure 4D and 5A). RNA was analyzed for genome-wide expression analysis using an Illumina rat bead chip (RatRef-12_V1_0_R2_11222119_A). The RatRef beadchip contains 22,523 probes which map to 21,791 genes in the rat genome. 200 ng of each sample was processed using the Illumina specific version of the Ambion TotalPrep 96 kit (Cat # 4393543). 750 ng of biotinylated cRNA product was added to each array hybridized according to standard Illumina protocols. Washing and signal development were performed with the aid of a SciGene model 650c microarray processor (LittleDipper). Chips were scanned on an Illumina iScan confocal scanner under standard parameters. Samples were analyzed in triplicate for each treatment group and data for each gene was averaged between the three samples. Data was subjected to background

subtraction and quantile normalization. For data analyses and figure generation, the coordinates of each probe on the microarray was extracted from the annotation file provided by Illumina. The Rat genome (Baylor 3.4/rn4) was used to generate Figure 4D, with each chromosome divided into bins of 1 megabase. Only transcripts meeting a p value cutoff of 0.001 were examined and are reported.

## Results

### *Mass spectrometric measurements of the chromatin proteome*

To investigate changes in chromatin-bound proteins commensurate with global changes in gene expression, adult mice were subjected to cardiac pressure overload by transverse aortic constriction[19], a model that recapitulates the human disease of heart failure (Supplemental Figure 1), which affects millions of people worldwide[28]. This disease is associated with a global shift in gene expression (Supplemental Figure 1C)[29] which drive a progressive increase in cardiac mass at the cell and organ level accompanied by deterioration of cardiac function (Supplemental Figure 1B). Hearts from mice in stages of cardiac hypertrophy, failure or basal state were excised. Chromatin was fractionated (Figure 1A and Supplemental Figure 2) from mouse cardiac nuclei and DNA-bound proteins were acid extracted. As with any fractionation study of a complex tissue like the heart, complete purification of individual fractions is not possible when using a non-biased technique like mass spectrometry for detection. Our previous work demonstrates that this method significantly enriches for histone and other chromatin-bound proteins[22], although additional studies will be required to determine what role each individual protein has in the chromatin fraction (i.e. to discriminate between true residents and

contaminants). Proteins were separated by SDS-PAGE, in-gel trypsin digested, analyzed by quantitative mass spectrometry on an Orbi-trap instrument, and peptide spectra examined by a combination of traditional database searching (SEQUEST), label-free quantitative analysis using the Elucidator software package[26,27], and manual inspection (please see Experimental Procedures and Supplemental Figure 3).

This study includes the analysis of 3 biological replicates (i.e. 3 different animals), two technical replicates each, in the basal, hypertrophy (moderate disease) and failing heart (severe disease) conditions, resulting in the acquisition of ~5.14 million spectra from 450 LC/MS/MS runs. Aligning the multiple features of retention time, parent ion intensity and MS/MS-confirmed sequence, this quantitative analysis allows for trends of protein abundance to be extracted from highly reproducible LC/MS/MS runs. Using this approach, combined with manual validation, we measured the association of >300 proteins with chromatin during distinct stages of heart failure in the mouse (note: only proteins detected by mass spectrometry are reported in this study). The examples of histone 1.0 and histone 1.4 represent a protein with altered abundance during the development of disease contrasted with one whose abundance is stable, respectively (Figure 1B). Progression of disease was associated with global shifts in the abundance of chromatin-associated proteins (Figure 1, C and D). Principle component analysis of all peptide intensities from all 18 datasets after ANOVA indicates good separation between the three experimental groups (and good agreement within them) (Figure 1E), supports the quantitative analysis of protein abundance and demonstrates strong reproducibility of the experimental workflow. Genes for proteins undergoing altered expression localize without bias throughout the genome (Supplemental Figure 3), suggesting genome-wide rearrangements are a prerequisite for disease. Furthermore, annotation of these proteins reveals extensive reprogramming at the level of

chromatin structure and regulation, including proteins harboring a panel of domains known to regulate transcription and to bind DNA (Supplemental Figure 4).

***Alteration in linker to core histone ratio as a mechanism for chromatin remodeling***

The protein building block of chromatin is the histone molecule. The murine genome has 80 histone genes for which mass spectrometric evidence has been observed for at least 25, based on unequivocal peptide hits, and 54, based on combinatorial induction, highlighting the complexity of functional units available for DNA packaging and genomic structure in the heart[22]. In this study, disease progression is associated with isoform-specific (i.e. different gene products) reprogramming of histone variants (Figure 2A). Strikingly, the ratio of linker (i.e. H1) to core (H2A, H2B, H3 and H4) histones was decreased during hypertrophy and returned to baseline during failure (Figure 2B and Supplemental Figure 3G top panel). This change in core to linker histone ratio was associated with a global shift from hetero- to euchromatin (i.e. from silenced to more transcriptionally active chromatin)[30,31], as evinced by the increase in histone mark H3K4me3 and decrease in H3K9me3 (Figure 2C). While this change in linker to core histone ratio is modest (~15%; Figure 2B), the functional impact could be profound, as it may be sufficient to affect ~15% of the coding and non-coding genome. This finding supports the concept that linker histones facilitate packaging of chromatin into higher order structures[7,32], the local scale effects of which are represented diagrammatically in Figure 2D.

Because these analyses are unbiased and quantitative, the opportunity exists to discover emergent properties within the data. Trend analyses reveal self-organizing modules of chromatin-bound proteins with similar behavior during disease progression (Figure 3A&B). Unsupervised clustering of z-score transformed data (which allows for analysis across ranges of

absolute abundance) in this manner allows for determination of groups of molecules(26, 27) with non-intuitive relationships that could not be revealed by annotation and would otherwise not be examined in a traditional gain/loss of function experiment. It remains unknown at what level(s) the abundance of these proteins are principally controlled (e.g. transcription, nuclear translocation, degradation, and so forth).

### *HMGB2 is a novel regulator of pathologic cardiac cell growth*

To explore the isoform-specific roles of individual chromatin structural proteins from within modules identified in Figure 3, siRNA was used to knock-down expression of two high mobility group proteins identified in our study (HMGB1 and 2) and previously implicated in chromatin structure[8,32]. Loss of HMGB2 (but not HMGB1; Figure 4A) induced genomic structural remodeling conducive to cell growth, as measured by cell size (Figure 4B) and some aspects of genomic reprogramming associated[20,29] with a fetal phenotype (Figure 4C). Although previous investigations had linked this family to cancer signaling[33,34], this is to our knowledge the first demonstration of a functional role for HMGBs in cardiac phenotype. Microarray gene expression analyses following HMGB2 knock-down indicate that the resulting growth phenotype is the product of specific cardiac growth pathways, including hypertrophic gene up-regulation and MAPK pathway activation, among others (Figure 5A and Supplemental Figures 5A & B). HMGB2 appears to be important for structural inhibition of a genome-wide growth expression program, in that genes whose expression is influenced by HMGB2 knockdown are distributed with apparent randomness across the genome (Figure 4D). Interestingly, the role these proteins play in regulating chromatin structure is very stimulus-specific, in that hypertrophic growth induced by isoproterenol and phenylephrine induces an increase in HMGB1 and a decrease in

HMGB2, whereas stimulation with endothelin-1 has the antithetic effect (Figure 5B).

Biochemical evidence was observed to support distinct populations of HMGB-associated chromatin, in that different amounts of each isoform were recovered depending on the use of either detergent or low pH in the chromatin isolation buffer (Supplemental Figure 6). Lastly, decreased HMGB2-dependent chromatin regulation was accompanied by a shift from hetero- to euchromatin, as shown by H3K4me3, H3K9me3 and heterochromatin binding protein-1 alpha abundance (Figure 5C). These findings support the concept that chromatin remodeling for transcription occurs in a context specific manner[35,36], with different histone variants, structural proteins and modifications specifying control of different genomic regions[37–39] (Figure 5D).

**Discussion**

Together these studies support a model in which a balance of local and global chromatin remodeling events is necessary to structurally poise all chromosomes for genome-wide changes in transcription. Evidence in support of this conjecture is the stimulus specificity of chromatin structural protein expression, the distinct localization of HMGB target genes across all chromosomes, and the alterations in HMGB protein expression and histone linker to core ratio in the setting of massive gene expression changes in disease. Decreased abundance (or knockdown, in the experimental setting) of chromatin structural proteins including linker histones and HMGs, would tend to shift the equilibrium toward relaxed chromatin (or to the left in Figure 2D). This hypothesis solves the problem of how an active transcription factor "finds" the right regions of the genome to bind: commensurate with transcription factor activation, global chromatin remodeling through altered linker histone/HMG abundance (Figure 5D) shifts the presentation of the interphase genome to a conformation permissive for binding of transcription factors to the

appropriate regions of the genome, while simultaneously masking those regions not targeted for immediate expression. Future work will be required to demonstrate whether these principles are operative in non-cardiac cell types. Likewise, additional studies will be required to test the hypothesis that three dimensional reorganization of the cardiac genome occurs reproducibly in real-time, using methods (likely microscopy-based) that capture the movements of chromosomes in the anatomical context of the cell[40]. It is likely that the function of chromatin structural proteins peripheral to the nucleosome is determined by local DNA sequence and chromatin associated proteins (and their post-translational modifications), rather than being universally activating or inhibiting of transcription across the genome.

The unsupervised clustering of chromatin protein abundance provides novel insights into the global processes that control gene expression. Whether and if so, how, the proteins in the modules identified in Figure 3 are functionally connected in the cell is an exciting area of future work. Our bioinformatic analyses by gene ontology, however, yield interesting leads: modules involved in muscle development, stress response, transcription, differentiation, development and metabolism all suggest that chromatin-bound proteins participate in various aspects of cellular homeostasis, and that many of these processes are targeted in distinct ways during disease.

In addition to controlling some genes previously associated with hypertrophic growth (ANF and SERCA; Figure 4C), our microarray studies following knockdown indicate a role for HMGB2 in the packing and regulation of other interesting cardiac-related genes. Among genes up-regulated when HMGB2 is knocked down (suggesting that the protein normally contributes to their silencing) are Wnt5, a well-studied morphogen with roles in cardiac development and disease[41], and Ras, a GTPase known to induce hypertrophic growth through sarcoplasmic reticulum disruption[42]. Down-regulated genes following knockdown included HDACs, whose

activity is inversely correlated with gene activation in the heart[43], supporting observations in the current study about a shift to more euchromatic phenotype in hypertrophy (Figure 2C) or during HMGB2 knockdown (Figure 5C), and the chromatin remodeling complex member PARP that, along with Brg1, contributes to reprogramming gene expression in hypertrophic cardiomyopathy[44]. How these and many other interesting genomic candidates coordinately regulate the hypertrophic response to pressure overload and/or loss of HMGB2 will be interesting to explore in future work.

Chromatin remodeling is pervasive in cardiovascular development and disease[45,46], although the mechanisms for genome-wide control of genomic structure remain unknown. Previous investigations have revealed key roles for histone modifying enzymes in controlling transcriptional states[47], including in the regulation of cardiac development and disease[12,48]. Other investigators have used mass spectrometry to examine chromatin-associated proteins[14,15,17,18], however the present studies provide the first quantitative blueprint of chromatin binding proteins from an endogenous source combined with unbiased analysis of the changes in chromatin structural proteins during the development of disease. Our data demonstrate global shifts in the hetero- to euchromatin balance during disease progression and implicate specific roles for individual histone and other chromatin structural proteins, including HMGB2. These findings have fundamental implications for the field of chromatin biology as well as for cardiovascular and other diseases, suggesting future therapeutic strategies may be engineered to target entire functional regions of the genome via specific chromatin structural proteins.

**Figure Legends**

**Figure 1: Proteomic measurement of chromatin-bound proteins.** *A,* Chromatin-associated proteins were isolated by acid extraction from mouse cardiac nuclei. Proteins were electrophoretically separated, digested and analyzed by semi-quantitative label-free mass spectrometry. All the procedures are described in greater detail in Experimental Procedures. *B,* Individual peptide peaks were quantified and linked with sequence information by database searching; shown are examples of histone H1.0 (top; peptide YSDM(147.0355)IVAAIQAEK), which was altered with disease compared to histone H1.4 (bottom; peptide TSGPPVSELITK), which was unchanged (see Supplemental Figure 3 for expanded version of this data). In the left panels, the monoisotopic peptide peak from each biological and technical replicate were aligned and quantified as area under the curve. Dotted gold lines indicate the area of integration used for intensity calculations. Intensity data from all 6 replicates was averaged for each condition (right panels, top and bottom). All peaks in the isotopic window were used for intensity calculations, (see Supplemental Figures 3A & 3B). Note the reproducibility in alignment and quantitation between runs. *C*, Pair-wise analyses between basal and hypertrophy reveals remodeling of the chromatin protein backbone during disease. Red lines indicate 2-fold change; similar trends exist at peptide (left panel) and protein (right panel) level. Up- or down-regulated proteins/peptides indicated in green or red had $p<0.01$. *D,* Same as panel C except only peptides for the indicated proteins are shown, demonstrating good agreement between the peptides used for identification and quantification of a given protein. *E,* Principal component analysis of all measured peptide intensities after ANOVA reveals groupings of technical (n=2) and biological replicates (n=3) within basal (blue), hypertrophy (red) or failure (green) chromatin, concomitant with excellent reproducibility of expression variation between these three groups. Axes plot peptide intensities.

**Figure 2: Mechanisms for genome-wide chromatin remodeling: changing the core to linker histone ratio.** *A,* Global changes in gene expression with disease are associated with quantitative changes in histone families H2A, H2B and H1 as measured by mass spectrometry. Histone H3 is absent because of extensive post-translational modification, in agreement with previous reports from other cell types[49]; histone H4 has essentially only one known coding sequence. Findings were determined by spectral counting from manually validated SEQUEST results. *B,* Ratio of linker (histone H1) to core (here, histone H4 is used based on the aforementioned rationale) histone variants exhibits significant global decrease during hypertrophy, returning to baseline during failure (left panel; data confirmed by western blotting, see Supplemental Figure 3G). Ratio among core histones is unchanged (middle and right panels). *C*, The histone modifications H3K4me3 (an established activating mark[50]) and H3K9me3 (an established silencing mark[30]) were increased and decreased, respectively, serving as a readout for a global shift from hetero- to euchromatin. *D,* Model for effects of alterations in core to linker histone ratio on local chromatin structure. Red circle indicates chromatin structural protein, such as linker histone H1. "Loose" and "packed" local features would tend to favor global eu- or heterochromatin, respectively.

**Figure 3: Unsupervised clustering reveals modules of chromatin regulatory proteins.** *A*, Proteins with altered chromatin association during disease are displayed (top left; each line is an individual protein). Trend analysis produced nine clusters (top right and then expanded in lower panels) with similar behavior during disease progression (data is z-score transformed to compare patterns of change across the entire dynamic range). *B,* Heat map display of modules from **A**, indicating the GO terms enriched in the given modules. In addition to these novel features, GO,

Interpro and KEGG analyses revealed enrichment in various protein domains and established pathways in the cardiac chromatin data (Supplemental Figure 4).

**Figure 4: Specific chromatin structural proteins control distinct regions of the genome, enabling global changes in gene expression.** *A,* Knockdown of HMGB2 leads to compensatory up-regulation of HMGB1 although the inverse does not occur. RT-PCR was also used to confirm efficacy of knockdown at mRNA level. Knockdown of HMGB2, but not HMGB1, induces neonatal cardiac myocytes to undergo hypertrophic growth akin to disease (*B*; Bar indicates 50 μM; For each condition at least 50 cells were quantified in three separate experiments, yielding 150 cells per condition.). **C,** This shift in phenotype recapitulates some but not all of the gene expression changes associated with disease *in vivo* (compare panel *C* with Supplemental Figure 1C), suggesting that different chromatin structural proteins package distinct populations of genes. Error bars represent SEM. Asterisk indicates p<0.001 *D,* To examine regions of the genome whose structure is maintained—and therefore transcriptional activity modulated—by HMGB2, expression microarrays were performed in the absence and presence of HMGB2 knockdown. The genomic distribution map shows up- (green) and down- (red) regulated genes following HMGB2 knockdown, demonstrating global effects that include all chromosomes (yellow indicates a bin in which some genes were up and other down regulated).

**Figure 5: Functional impact of HMGB knockdown on gene expression and chromatin structure.** *A,* Pie charts demonstrate the processes enriched in the up- or down-regulated genes detected by microarray following HMGB2 knockdown. Additional bioinformatic analyses (Supplemental Figure 5) demonstrate enrichment of nucleotide binding domains and many

pathways implicated in cell growth and cardiac function/dysfunction in the genes under control of HMGB2. ***B,*** Changes in HMGB1/2 expression at distinct times following different hypertrophic agonists in neonatal rat ventricular myocytes (these stimuli are sufficient to induce hypertrophic cell growth). ***C***, HMGB2 knockdown induces a shift from hetero- to euchromatin, as documented by increased histone H3 K4 methylation, decreased K9 methylation, and increased abundance of heterochromatin binding protein 1 alpha. ***D***, Representation of the local chromatin remodeling effects that can orchestrate global shifts in DNA accessibility. In this model, linker histones or non-histone chromatin structural proteins (like HMGs; indicated by the red and purple circles) induce specific genomic regions (indicated with the green or blue nucleosomes) to assume a more compact structure (Figure 2D); regions not under structural regulation by the given proteins remain unaffected (grey nucleosomes). For simplicity, only nucleosomes and chromatin structural proteins are shown; *in vivo*, nucleosome positioning (through DNA sequence, proteins interactions and post-translational modifications) and other position-specific chromatin modifiers, likely are contributory to specifying these hypothetical populations of chromatin. This model suggests these distinct local principles are applied globally based on different expression profiles and genomic localization of linker histones and non-histone chromatin structural proteins.
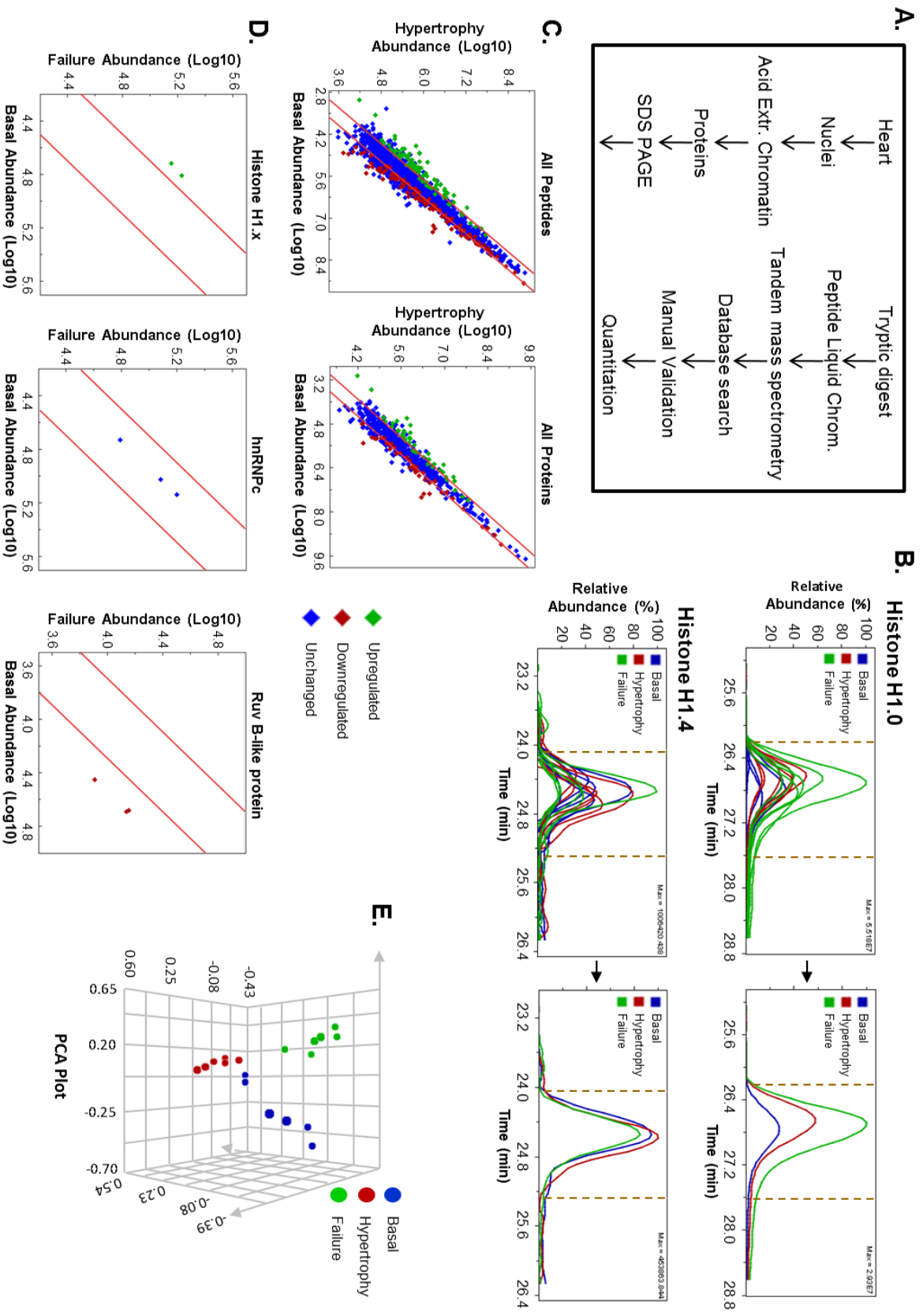
**Figure 1**
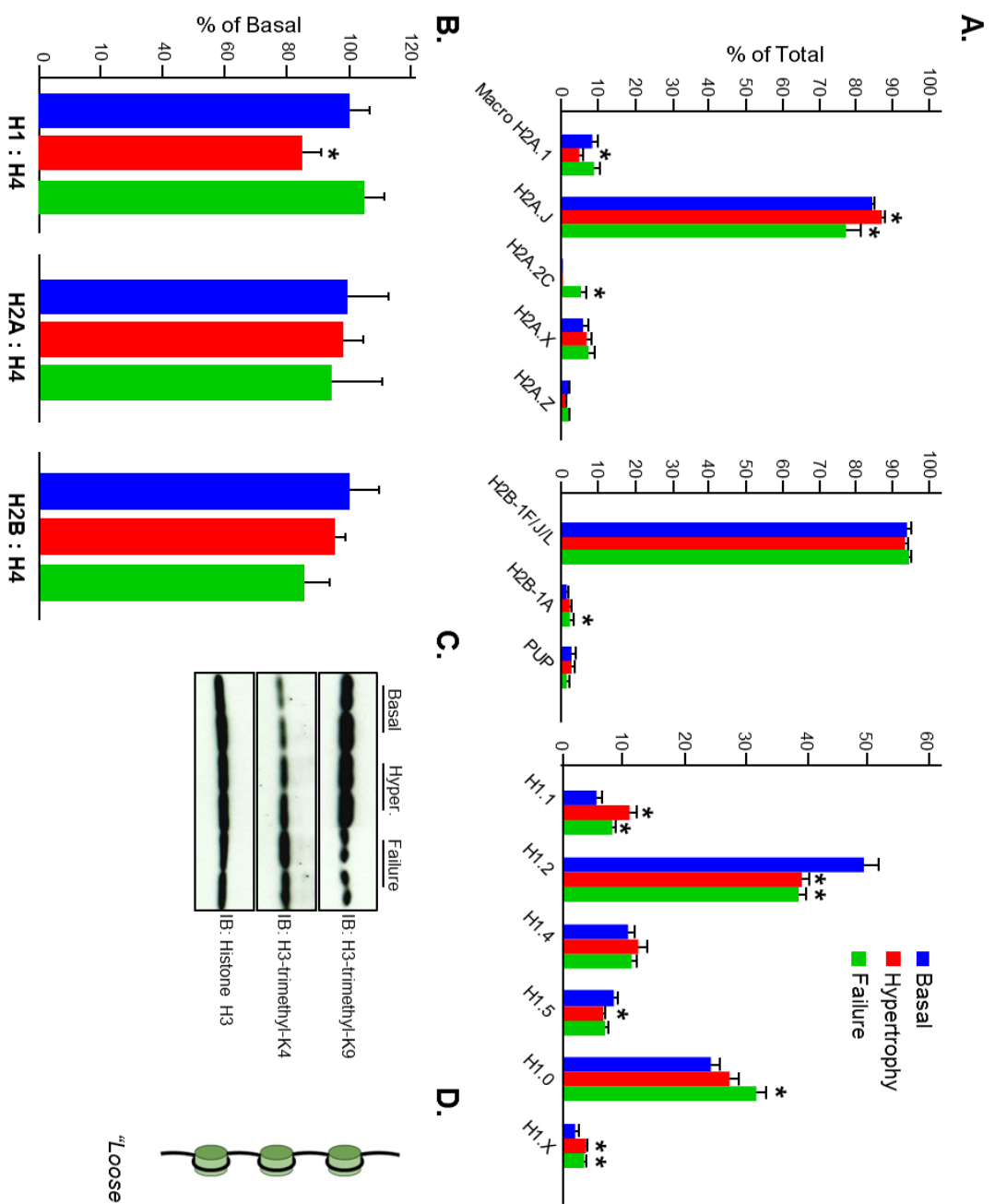
**Figure 2**
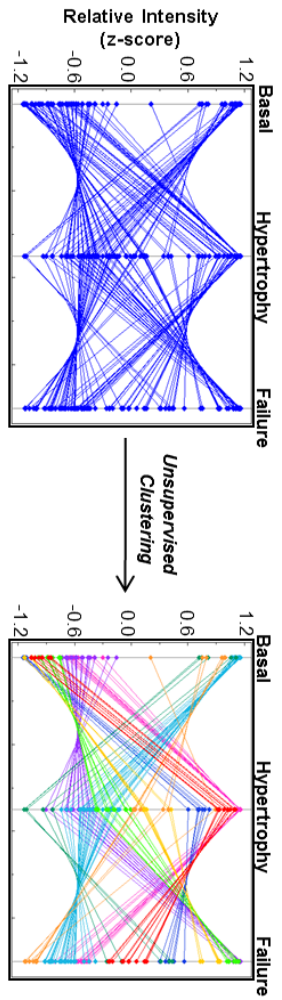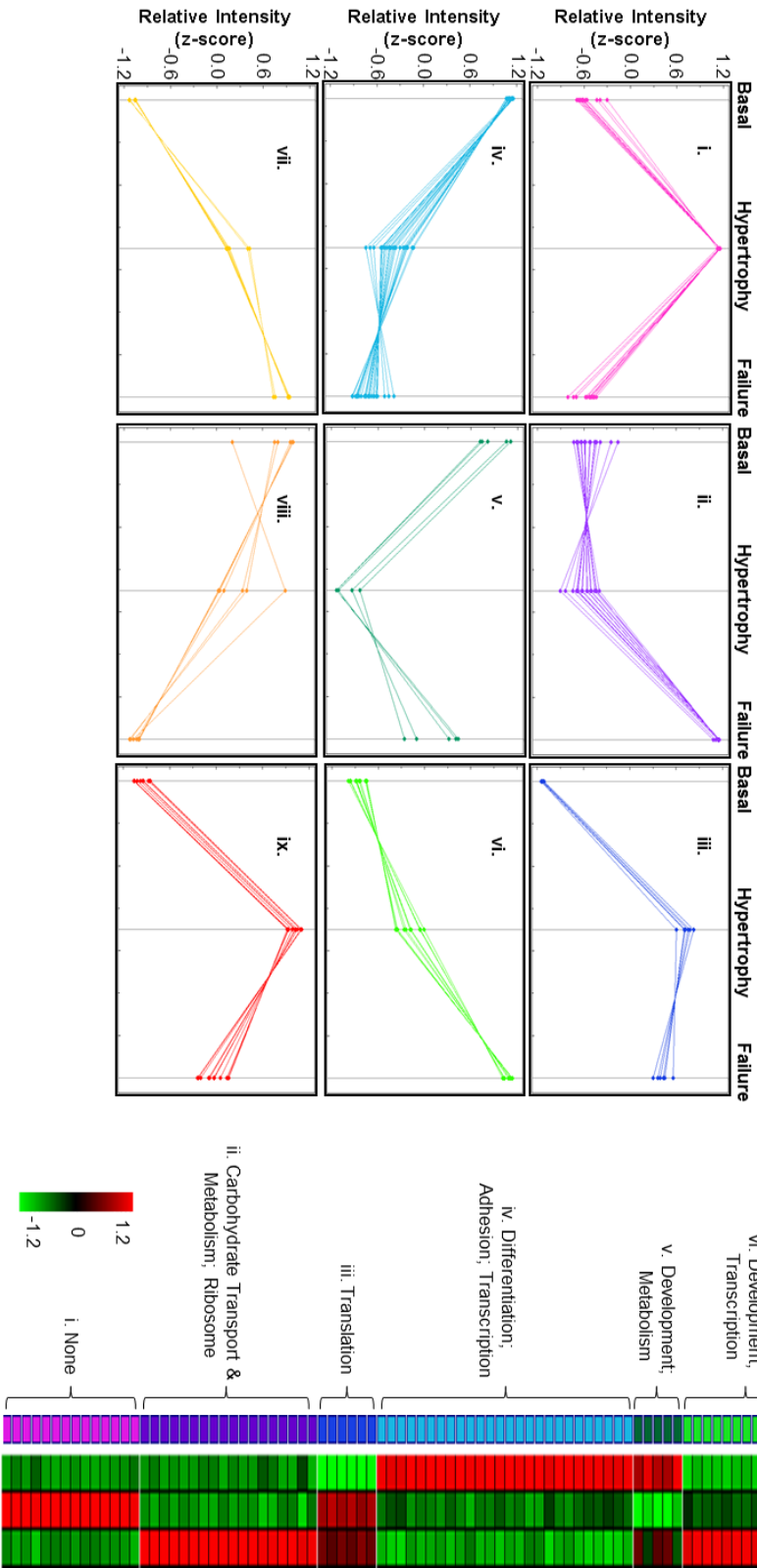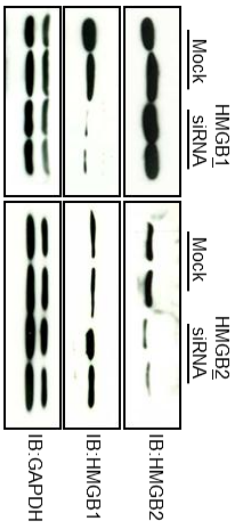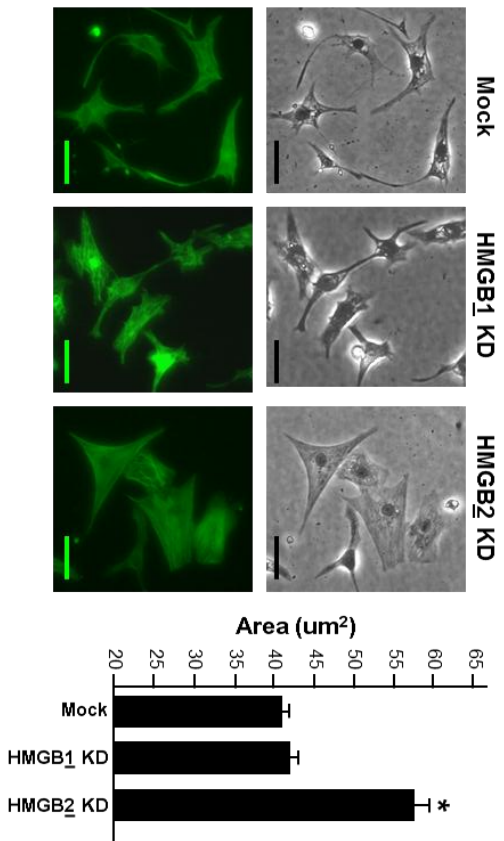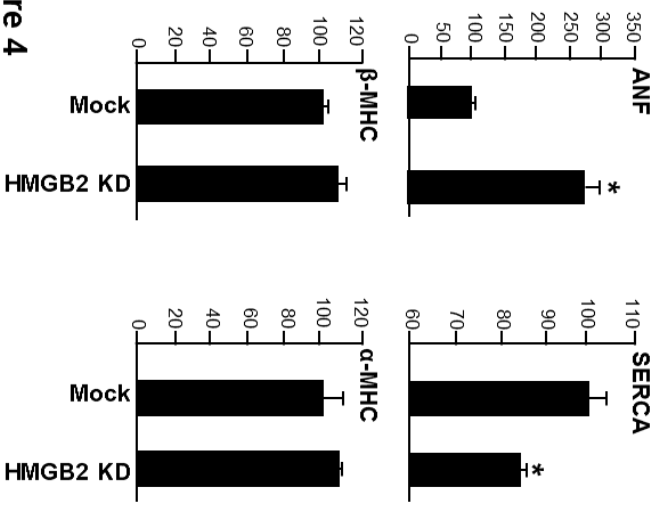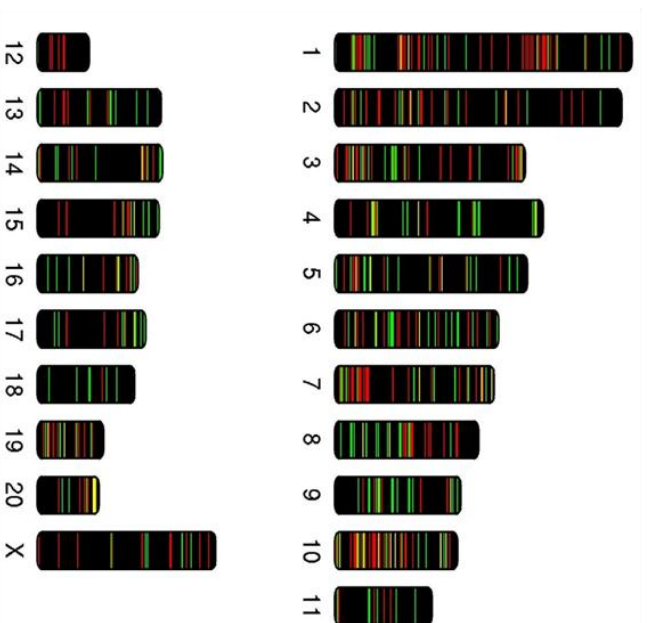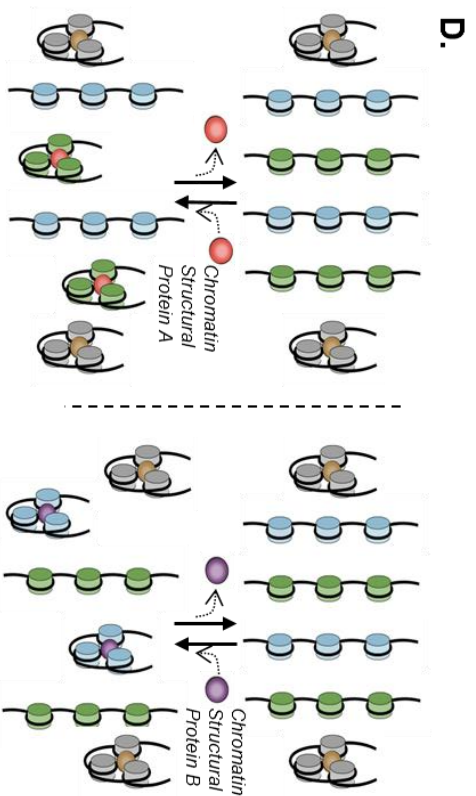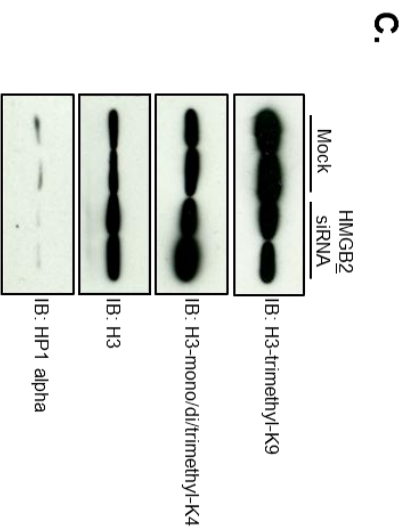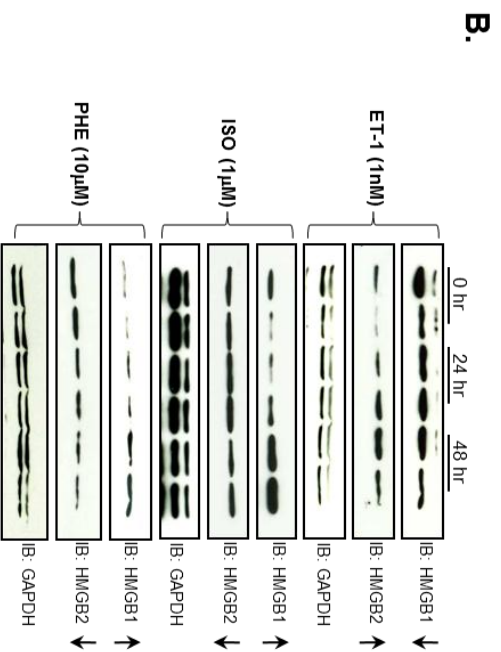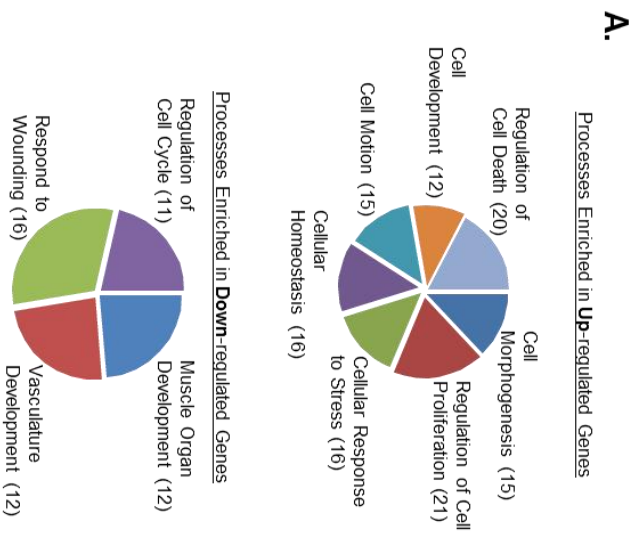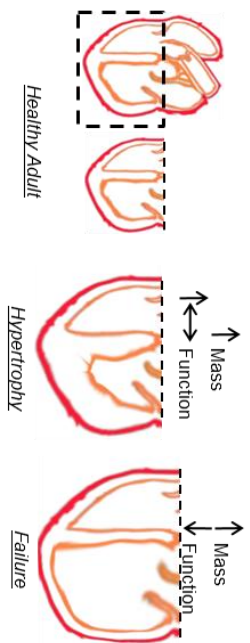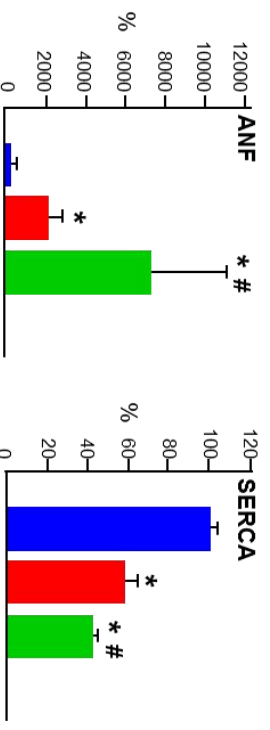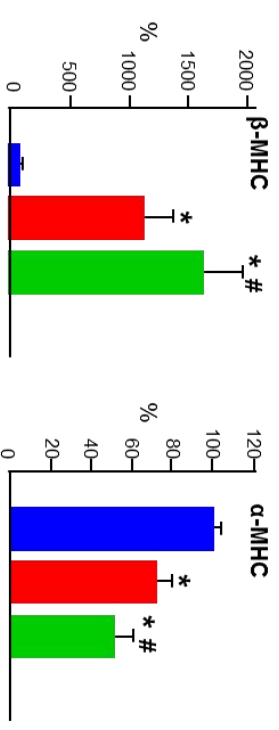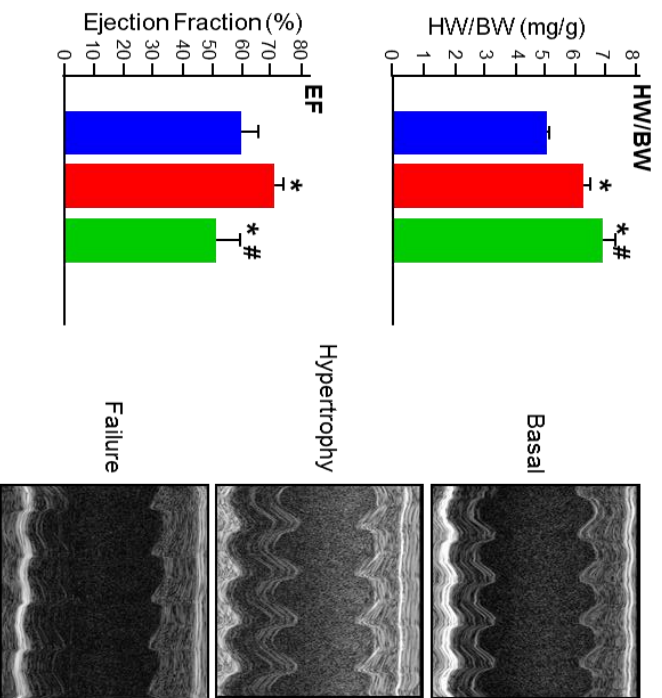
Figure 3

Figure 4

Figure 5

**Supplemental Figure 1**

**Supplemental Figure 2**
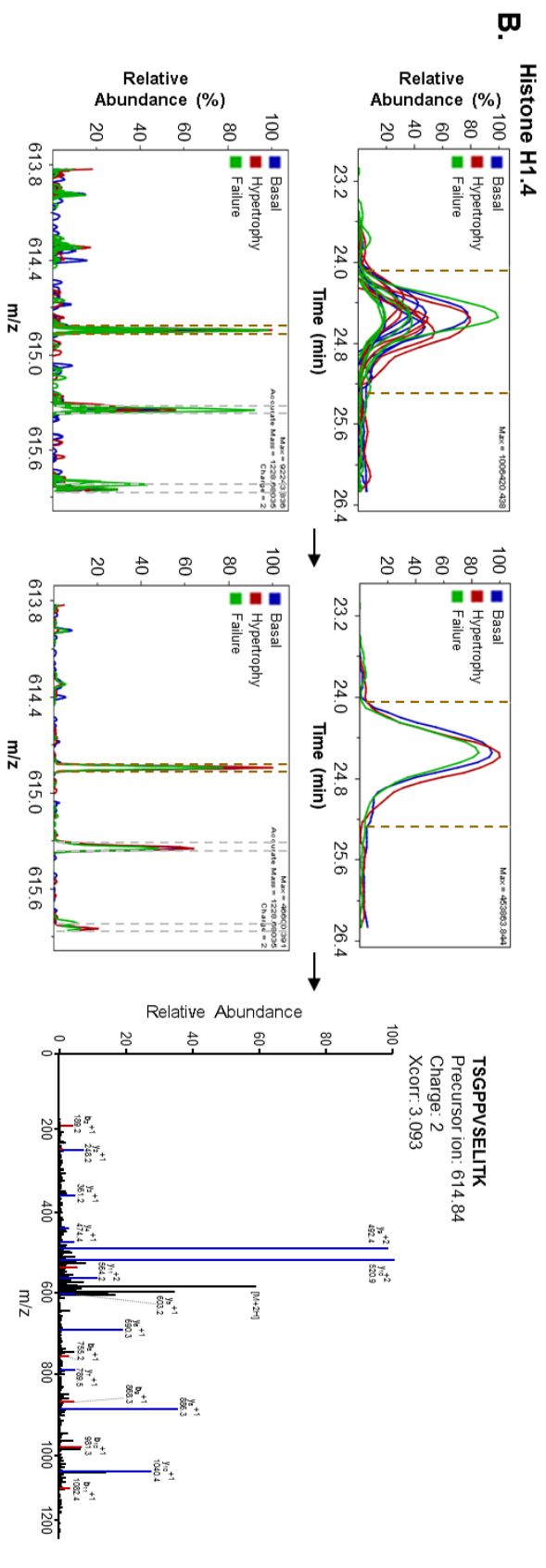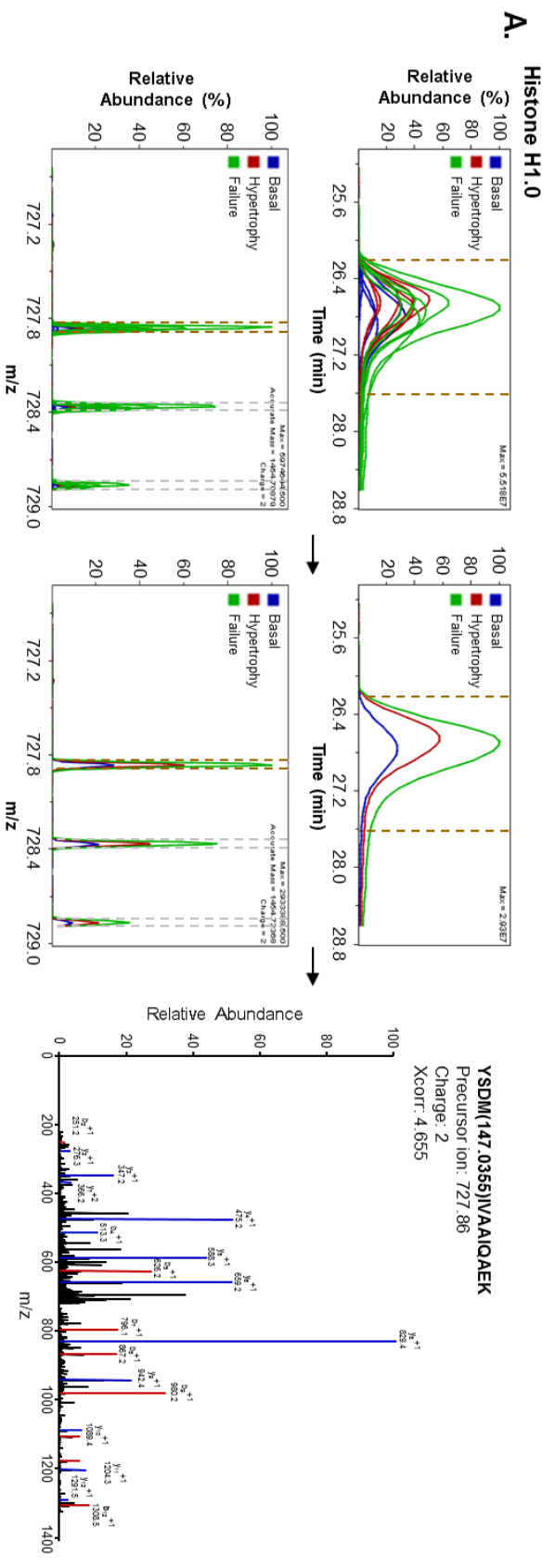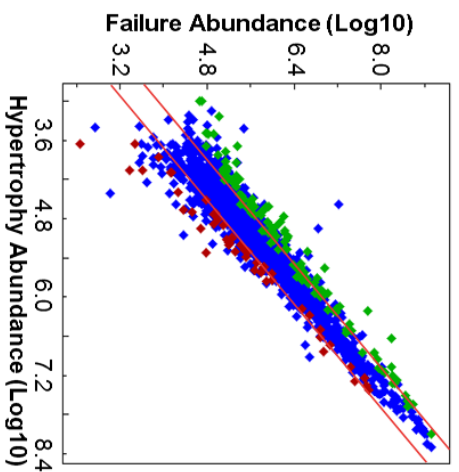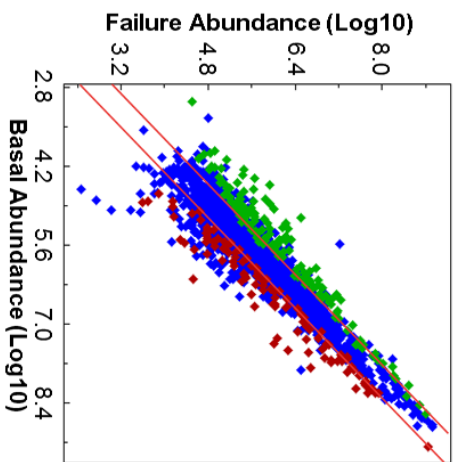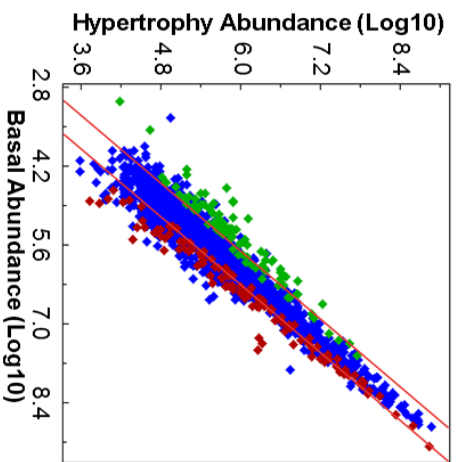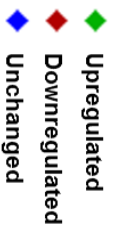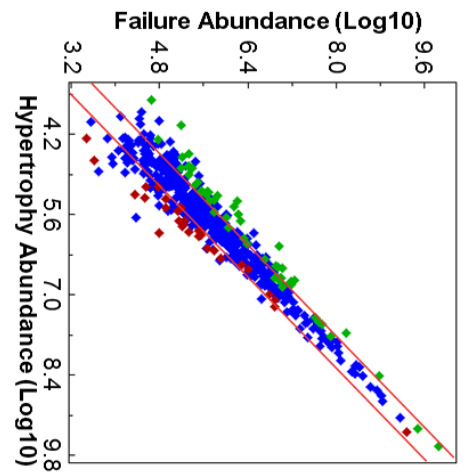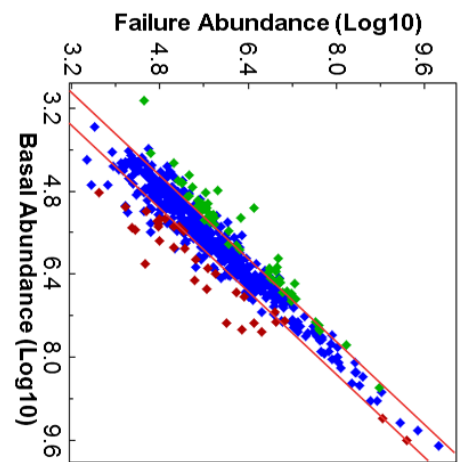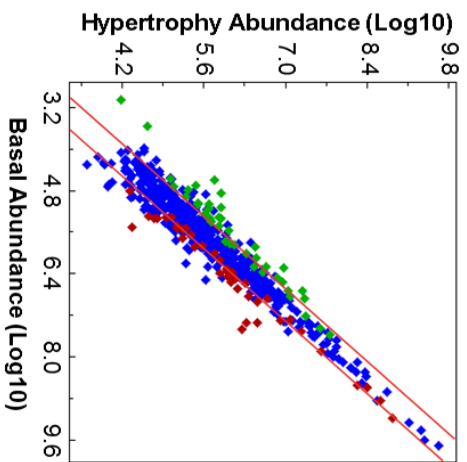
**A.**

IB:Histone H3

IB:GAPDH

WHL — Nuc Lys

IB: Histone H3

IB: U170K

NuP, CH, AE, Nuc Lys, Hela Lys

**B.**

Basal — Hypertrophy

Basal — Failure

Supplemental Figure 3

Supplemental Figure 3

Supplemental Figure 3

E.

F.

**Molecular Weight**

**Isoelectric Point**

**Hydropathy**

Legend:
- IPI Mouse v3.76
- Basal
- Hypertrophy
- Failure

G.

IB: Histone H3

IB: Ruvb-like 1

Basal — Failure

IB: Histone H4

IB: Histone H1.0

Basal — Hypertrophy

Supplemental Figure 4

A.

B. Enriched Protein Domains

C. Enriched KEGG Pathways

34

**D.**



(v.)
(iv.)
(iii.)
(ii.)
(i.)

(v.) Failure
(iv.) Hypertrophy
(iii.) Basal Chromatin Proteome
(ii.) Chromosomes
(i.) Genes
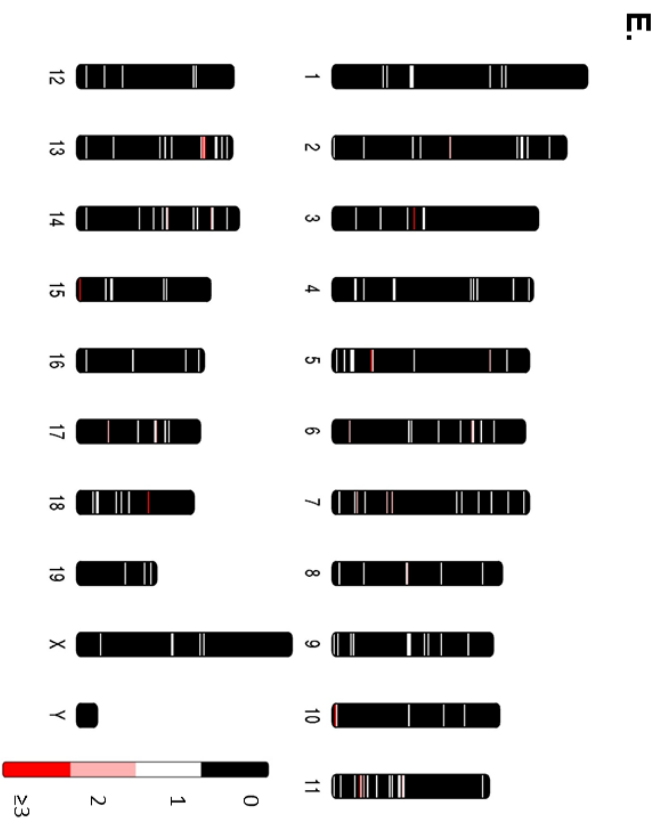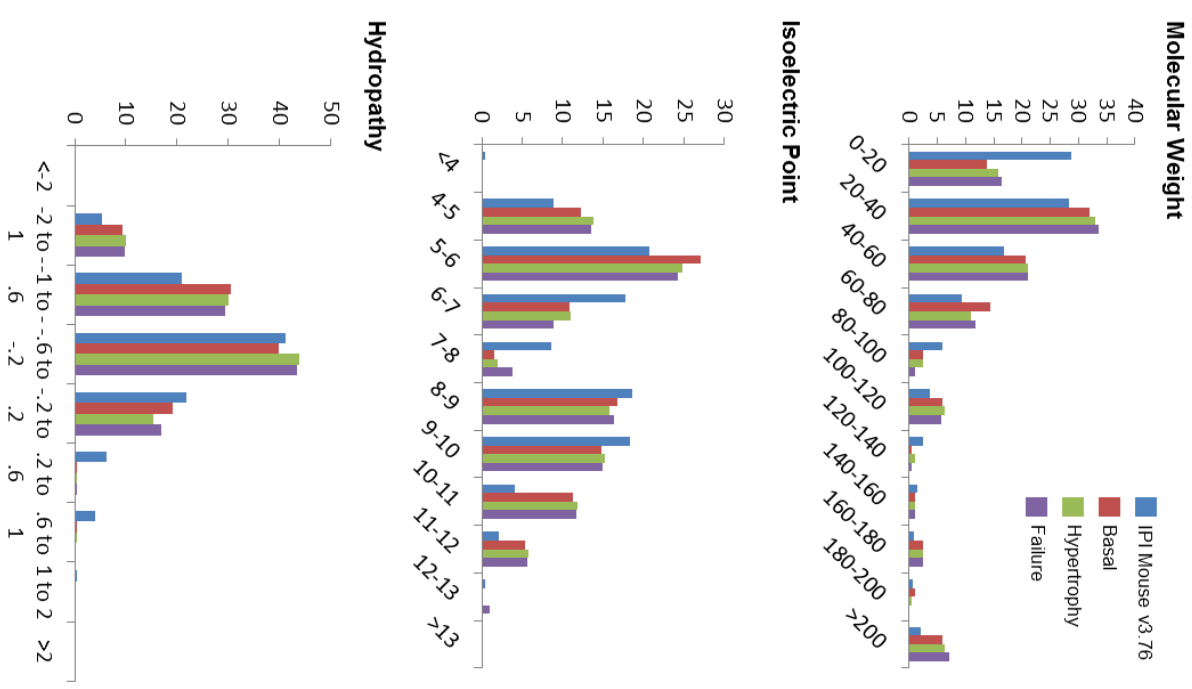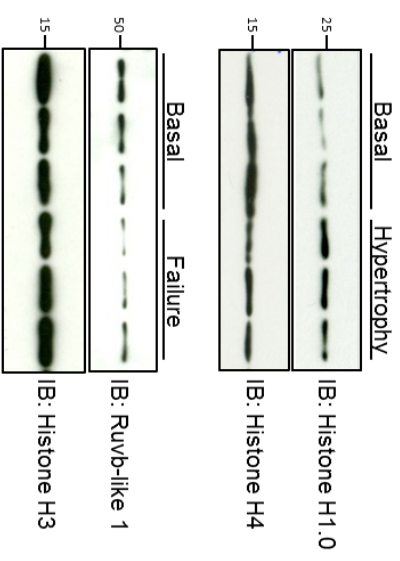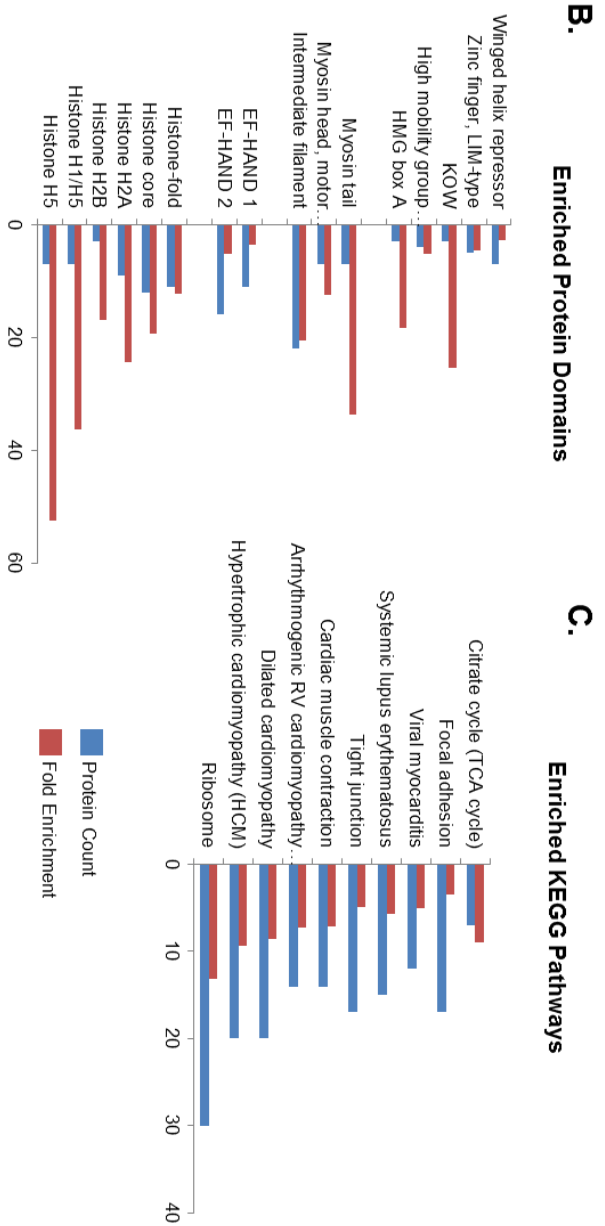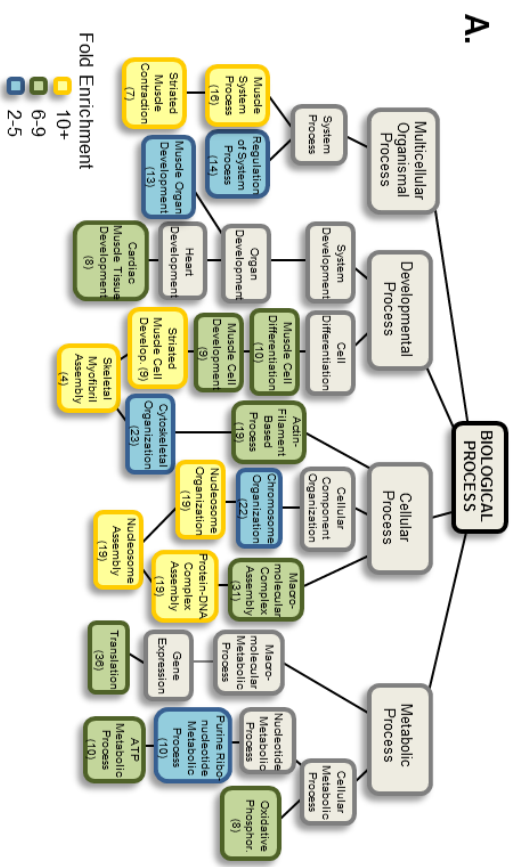
**Supplemental Figure 4**
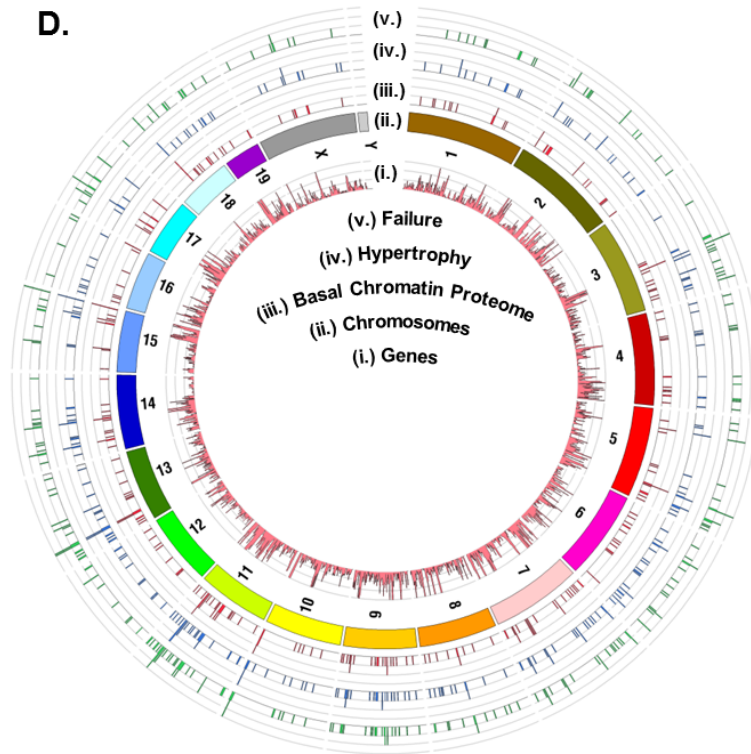
Supplemental Figure 5

A. Domains Enriched
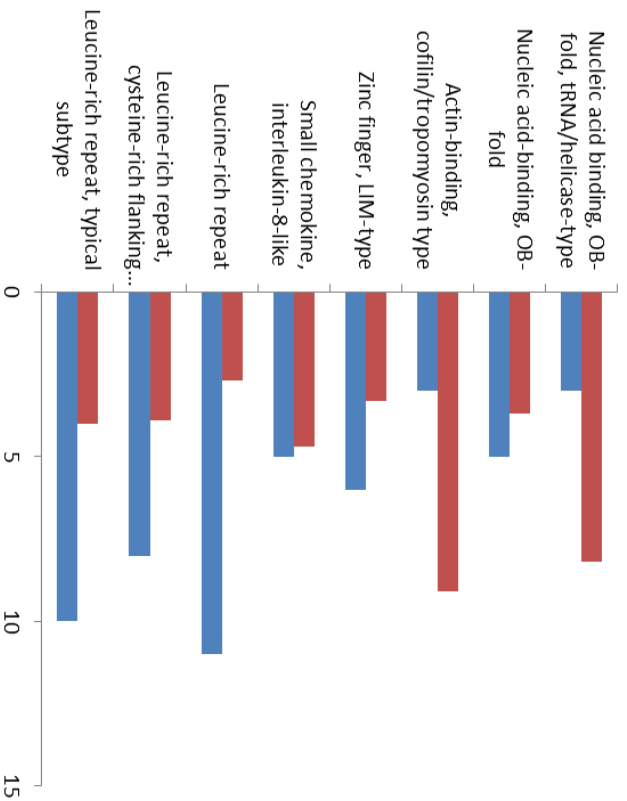
B. Pathways Enriched
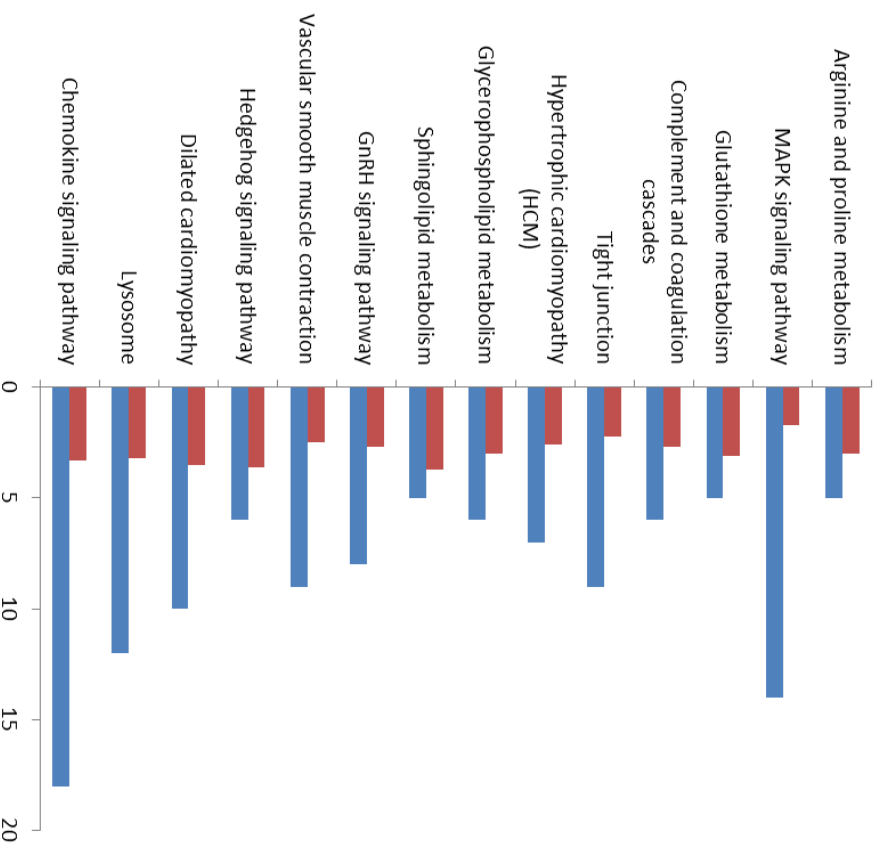
**Acid Extraction of Chromatin Proteins**
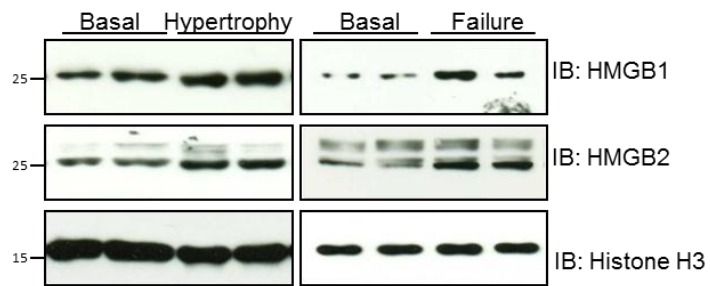


**Detergent Extraction (no Acid) of Chromatin**
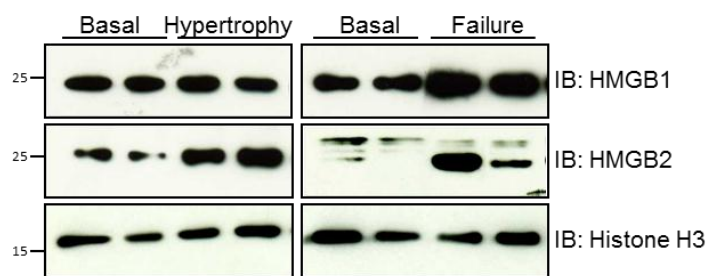


**Supplemental Figure 6**

*Reference*

1.      Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature reviews. Genetics* **2,** 292–301 (2001).

2.      Van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nature biotechnology* **28,** 1089–1095 (2010).

3.      Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–60 (1997).

4.      Tremethick, D. J. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell* **128,** 651–4 (2007).

5.      Pruss, D. *et al.* An asymmetric model for the nucleosome: a binding site for linker histones inside the DNA gyres. *Science* **274,** 614–7 (1996).

6.      Graziano, V., Gerchman, S. E., Schneider, D. K. & Ramakrishnan, V. Histone H1 is located in the interior of the chromatin 30-nm filament. *Nature* **368,** 351–4 (1994).

7.      Weintraub, H. Histone-H1-dependent chromatin superstructures and the suppression of gene activity. *Cell* **38,** 17–27 (1984).

8.      Bustin, M. Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Molecular and cellular biology* **19,** 5237–46 (1999).

9.      Goodwin, G. H., Sanders, C. & Johns, E. W. A new group of chromatin-associated proteins with a high content of acidic and basic amino acids. *European journal of biochemistry* **38,** 14–9 (1973).

10.     Talbert, P. B. & Henikoff, S. Histone variants--ancient wrap artists of the epigenome. *Nature reviews. Molecular cell biology* **11,** 264–75 (2010).

11.     Cairns, B. R. The logic of chromatin architecture and remodelling at promoters. *Nature* **461,** 193–8 (2009).

12.     Haberland, M., Montgomery, R. L. & Olson, E. N. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nature reviews. Genetics* **10,** 32–42 (2009).

13.     Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463,** 474–84 (2010).

14.    Ohta, S. *et al.* The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics. *Cell* **142,** 810–21 (2010).

15.    Zougman, A. & Wiśniewski, J. R. Beyond linker histones and high mobility group proteins: global profiling of perchloric acid soluble proteins. *Journal of proteome research* **5,** 925–34 (2006).

16.    Gerstein, M. B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330,** 1775–87 (2010).

17.    Bartke, T. *et al.* Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143,** 470–84 (2010).

18.    Chu, D. S. *et al.* Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. *Nature* **443,** 101–5 (2006).

19.    Rockman, H. A. *et al.* Segregation of atrial-specific and inducible expression of an atrial natriuretic factor transgene in an in vivo murine model of cardiac hypertrophy. *Proceedings of the National Academy of Sciences of the United States of America* **88,** 8277–81 (1991).

20.    Takimoto, E. *et al.* Chronic inhibition of cyclic GMP phosphodiesterase 5A prevents and reverses cardiac hypertrophy. *Nature medicine* **11,** 214–22 (2005).

21.    Mitchell-Jordan, S. A. *et al.* Loss of Bmx nonreceptor tyrosine kinase prevents pressure overload-induced cardiac hypertrophy. *Circulation research* **103,** 1359–62 (2008).

22.    Franklin, S. *et al.* Specialized compartments of cardiac nuclei exhibit distinct proteomic anatomy. *Molecular & cellular proteomics* **10,** M110.000703 (2011).

23.    Sadygov, R. G., Cociorva, D. & Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature methods* **1,** 195–202 (2004).

24.    Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19,** 1639–45 (2009).

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

26.    Kash, J. C. *et al.* Genomic analysis of increased host immune and cell death responses induced by 1918 influenza virus. *Nature* **443,** 578–81 (2006).

27.    Neubert, H. *et al.* Label-free detection of differential protein expression by LC/MALDI mass spectrometry. *Journal of proteome research* **7,** 2270–9 (2008).

28.  Lloyd-Jones, D. *et al.* Heart disease and stroke statistics--2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation* **119,** 480–6 (2009).

29.  Rajabi, M., Kassiotis, C., Razeghi, P. & Taegtmeyer, H. Return to the fetal gene program protects the stressed heart: a strong hypothesis. *Heart failure reviews* **12,** 331–43 (2007).

30.  Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410,** 120–4 (2001).

31.  Strahl, B. D., Ohba, R., Cook, R. G. & Allis, C. D. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 14967–72 (1999).

32.  Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *The Journal of cell biology* **83,** 403–27 (1979).

33.  Kwon, J.-H. *et al.* Overexpression of high-mobility group box 2 is associated with tumor aggressiveness and prognosis of hepatocellular carcinoma. *Clinical cancer research* **16,** 5511–21 (2010).

34.  Lee, D., Kwon, J.-H., Kim, E. H., Kim, E.-S. & Choi, K. Y. HMGB2 stabilizes p53 by interfering with E6/E6AP-mediated p53 degradation in human papillomavirus-positive HeLa cells. *Cancer letters* **292,** 125–32 (2010).

35.  Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461,** 186–92 (2009).

36.  Segal, E. & Widom, J. What controls nucleosome positions? *Trends in genetics* **25,** 335–43 (2009).

37.  Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–40 (2010).

38.  Goldberg, A. D. *et al.* Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell* **140,** 678–91 (2010).

39.  Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–83 (2011).

40.  Mewborn, S. K. *et al.* Altered chromosomal positioning, compaction, and gene expression with a lamin A/C gene mutation. *PloS one* **5,** e14342 (2010).

41.    Rao, T. P. & Kühl, M. An updated overview on Wnt signaling pathways: a prelude for more. *Circulation research* **106,** 1798–806 (2010).

42.    Zheng, M. *et al.* Sarcoplasmic reticulum calcium defect in Ras-induced hypertrophic cardiomyopathy heart. *American journal of physiology. Heart and circulatory physiology* **286,** H424–33 (2004).

43.    McKinsey, T. A., Zhang, C. L. & Olson, E. N. Control of muscle development by dueling HATs and HDACs. *Current opinion in genetics & development* **11,** 497–504 (2001).

44.    Hang, C. T. *et al.* Chromatin regulation by Brg1 underlies heart muscle development and disease. *Nature* **466,** 62–7 (2010).

45.    Bruneau, B. G. Chromatin remodeling in heart development. *Current opinion in genetics & development* **20,** 505–11 (2010).

46.    Han, P., Hang, C. T., Yang, J. & Chang, C.-P. Chromatin remodeling in cardiovascular development and physiology. *Circulation research* **108,** 378–96 (2011).

47.    Narlikar, G. J., Fan, H.-Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108,** 475–87 (2002).

48.    Ago, T. *et al.* A redox-dependent pathway for regulating class II HDACs and cardiac hypertrophy. *Cell* **133,** 978–93 (2008).

49.    Garcia, B. A., Pesavento, J. J., Mizzen, C. A. & Kelleher, N. L. Pervasive combinatorial modification of histone H3 in human cells. *Nature methods* **4,** 487–9 (2007).

50.    Nishioka, K. *et al.* Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes & development* **16,** 479–89 (2002).

**Part II.    Hypothesis: Structural Considerations for Chromatin State Models with Transcription as a Functional Readout**

[This hypothesis was originally published in FEBS Letters by Chen, H. et al. Structural considerations for chromatin state models with transcription as a functional readout. FEBS letters 586, 3548–54 (2012). The manuscript is re-used under the License Agreement between Haodong Chen and Elsevier.]

Two established principles of chromatin biology present an obvious paradox for gene regulation: eukaryotic chromosomes are long, contiguous molecules that must be packaged in an ordered manner to fit into the nucleus *and* global changes in gene expression occur with specificity, reproducibility and speed, across the genome. The first principle has been known for decades; the second has emerged as a result of rigorous molecular dissection of individual transcriptional events in isolation and, in the past decade or so, through the application of global measures of gene expression (microarrays and RNA sequencing). The structure of the nucleosome, the octameric protein complex which binds ~146-147 base pairs of DNA and is the building block of chromatin, is known with atomic level resolution[1]. Likewise, the characteristic appearance of mitotic chromosomes has been known for nearly a century[2]. How the genome is organized in three dimensions in other phases of the cell cycle, or in non-dividing cells, is only beginning to be understood. Moreover, much less is known about the principles of chromatin packing that convert the beads on a string nucleosomal DNA into an ordered three dimensional unit, and how this structure is subsequently reorganized in a non-random way to enable gene expression. Lastly, the intermediate structural states the genome can assume in the non-mitotic somatic nucleus are unknown, although emerging evidence supports discrete intermediate fiber dimensions of 10 or 30 nm (based mostly on reconstitution studies[3] and some microscopic

evidence *in situ*[4]) as well as reproducible inter- and intra-chromosomal interactions between endogenous genetic elements (based largely on chromosomal capture techniques[5–7]).

We propose that the rapidly evolving field of chromatin regulation is inherently limited by a lack of understanding of the structural basis for packaging of the genome. In addition to work addressing this problem experimentally through techniques such as chromosomal conformation capture[5], we believe the field can benefit from a state model of chromatin. Such a model could serve as a framework to bring together experimental data from reductionist transcription assays in cell culture and *in vitro*, along with global measures of protein occupancy across the genome and direct measurements of genomic structure. A chromatin state model should be able to evolve as new experiments reveal additional states (or disprove existing hypothesized ones) and importantly, provides a basis for mathematical representation of local and global transitions in chromatin structure.

### *Considerations for Models of Chromatin Structure*

<u>Transcription Classified in a Binary Manner</u>

For the purpose of this thesis we will define *transcriptionally on* regions of the genome as those areas directly being transcribed to RNA, including genes (exons and introns), validated non-protein-coding RNAs, and characterized proximal promoters. *Transcriptionally off* regions include distal enhancers (with some exceptions[8]), intergenic DNA and portions of the genome currently unrecognized to code for RNA. Off regions also include genes and other coding regions that can be transcribed but that are not actively being transcribed in the given cell or cellular environment. Sequencing studies estimate that at any given time, ~70% of the mRNA encoding genome is transcribed[9], keeping in mind that this amounts to a little greater than one-third of the total genome, which is only ~3% exon and ~35% intron in humans and mice (the rest is intergenic DNA; source: NCBI). It is tempting to immediately classify this group of genes (i.e. those that can be activated/transcribed, but that are not, given the overriding cellular condition)

as a third group of transcriptional activity. This is avoided in the current analysis based on the following rationale: differentiating off permanently (that is, areas that can never be transcribed) from off temporarily (the type described in the previous sentence) requires omniscient knowledge of all transcriptional states, for all RNAs, in all cells. A more conservative analysis that avoids a false negative assignment of one region as permanently inactive due to a lack of evidence leaves two transcriptional states—on and off—with genetic elements moving between these transcriptional states when polymerases use them as a template for RNA synthesis.

Non-coding Regions and Other Structural Elements Contribute to Chromatin States

Again, we define *on* as a region of the genome being read as a template to synthesize RNA in a given population of cells. It is possible to argue, however, that the physically contiguous nature of the genome makes it difficult or impossible to functionally separate two regions of DNA adjacent to each other on the physical strand—regardless of whether these regions are undergoing active transcription. Unlike bacteria which lack introns and large regions of intergenic DNA, it is well established that in eukaryotes, coding areas are widely interspersed with regions of non-protein coding, intergenic DNA. In humans, the size of intergenic DNA varies widely from several bases to several megabases. The implication, therefore, is that it is virtually impossible to separate regions involved in transcription from those that are not solely on the basis of whether RNA is actively produced. Indeed these large regions of untranscribed DNA are clearly essential for the genome to assume the convoluted three dimensional structure[2,5] in the interphase nucleus that enables specific gene expression: without intergenic DNA, long range connections between disparate loci required for establishment of chromosomal territories could not be formed. Furthermore, the large amounts of intergenic DNA also likely play a role in genome evolution, acting as a substrate for gene duplication, modification and functional selection. Lastly, these non-coding regions of the genome have been shown to host non-nucleosomal chromatin structural proteins[10] and extensive DNA methylation[11], both of which are

dynamically regulated over the life of a cell, highlighting the role of DNA-binding proteins and DNA modification of intergenic regions in transcriptional regulation (be it indirectly, from a physical standpoint).

The three dimensional configuration of the genome, which is limited (but not encoded) by the primary sequence and the distribution of coding and non-coding regions, then, determines the phenotype of the cell by limiting the range of possible transcriptional states—gene expression states—at any given time. Converting between these states requires changes in the structure of the genome; since the DNA component is unchanged (over transcriptional time scales), this change must be endowed by the non-DNA component of the chromatin backbone: transcription factors, histones, non-histone chromatin structural proteins, post-translational modifications and non-coding RNAs. This relationship also establishes the means for the three dimensional structure of the genome—like the three dimensional structure of a protein—to be a substrate of selection and thereby evolution. In such a scheme, rearrangements of the linear DNA in three dimensions that enabled favorable transcriptional states to arise in the third dimension (that is, in the context of the *in vivo* packaging of the genome in the interphase nucleus) could undergo positive selection. Negative selection in this same manner is also possible. This logic can explain how transcriptional modules or gene expression profiles that contain genes distributed throughout the genome can be selected for, and thereby conserved, *en toto*.

The two premises we have to this point are: firstly, from the perspective of chromatin, there are (only) transcriptionally on and off states (how many *chromatin* states there are will be addressed shortly); secondly, intergenic DNA along with DNA-bound chromatin structural proteins are involved in and probably necessary for the immense variety of cell types in individual eukaryotic organisms, for the countless transcriptional programs they can exhibit, and for their evolution. These two premises present an obvious problem, in that the second seems to contradict the digital logic of the first. Our requirements for a chromatin state model follow from these premises and attempt to resolve this conflict: the model must have a quantifiable

transcriptional readout and it must have a detectable structural impact on the genome. An ancillary goal is to provide a rhetoric for understanding the time dimension in chromatin biology, but structuring results in a state model whereby transitions between, and dwell times within, individual states can be captured and compared across different cell types. While pioneering work has suggested that chromatin can directly do things other than endow transcriptional states—such as functioning as a lens in the rod photoreceptor of the nocturnal eye[12]—these actions are at present esoteric and not codified (and certainly not known to be present in all nuclei). For this discussion, we restrict our remarks to actions of chromatin as they relate to transcription, although it is certainly possible that more bizarre behaviors of chromatin remain to be discovered and may be proven evolutionarily important.

### *Chromatin Model with Two States*

A two state model is appealing because it is for this model alone that unequivocal evidence exists for a transcriptional phenotype: a segment of DNA is either being transcribed, or it is not. A binary model does not require a more nuanced transcriptional or structural readout. Accordingly, in this model the functional chromatin states are either on or off, open or closed, euchromatin or heterochromatin (Panel A in Figure 1). One should not conclude that a two state model is equivalent to a simple one; the everyday example of the enormous complexity that can be generated with 1's and 0's in computer programming is an intuitive example of how two states, applied across a continuum and with differential reading frame, can create complexity (this has also been shown in more relevant examples of biological networks[13]).

Superimposition of signal transduction preceding transcription, transcript/protein abundance and lifetime in the cell, and the role of interaction networks (physical and functional) onto the two state model can give rise to all the variation in transcriptional activities observed experimentally. In other words, the chromatin model itself need have no more than two states to account for the known behaviors of genomes. A cogent argument can be made, in fact, that

unequivocal evidence exists for no more than two states of chromatin across model systems, genomes and importantly, individual loci (in contrast to detailed analyses of varied transcriptional behavior observed at isolated loci *in vitro* or in heterologous cell systems).

While perhaps appealing from an analytical standpoint, there are several practical problems with a two state chromatin model, perhaps foremost that it is hard to disprove and thus a weak vehicle for hypothesis generation. There is also a wealth of knowledge about global chromatin regulation that is not incorporated into two states, including: distal enhancers/repressors, chromosomal territories, three dimensional genomic structure, nucleosome positioning and altered histone variant deposition, distinct activities of RNA polymerase, different rates of transcription, DNA methylation and histone post-translational modifications, to name the major categories that have been shown to affect the two properties—transcriptional readout and structural rearrangement of chromatin—we propose as  necessary components of a strong model. While we would argue that these factors have not been shown to universally control gene expression in a given manner, research in these areas is certainly not equivocal and an ideal chromatin state model should incorporate as much of our knowledge on transcription and chromatin structure as possible—even if some of it will ultimately be disproven—to stimulate further experimentation. Thus, there is a need for a chromatin model with greater than two states. Note that in such a model, our definition of two states of *transcription* still remain; it is a *chromatin* model that we are positing has a greater number of states.

### Multi-State Chromatin Models

The recent deluge of chromatin immunoprecipitation studies coupled with either microarrays or next generation DNA sequencing has fueled the urge to define transcriptional states on the basis of protein binding profiles[14]. While these studies have led to unprecedented insights into genome-wide protein occupancy, a chromatin state model that can be tested and refined across laboratories is unlikely to result solely from this approach based on the following

pieces of evidence: global studies are rarely all done in the same cell type, making universality difficult to conclude; the chromatin proteome is known to contain hundreds of proteins, as confirmed from multiple cell types[15,16]; and histone proteins, themselves occurring in numerous isoforms and variants, can undergo scores of modifications in the same cell type[17]. Just considering the last two, a state model incorporating only the 4 core histone proteins and only 10 post-translational modifications could theoretically specify 2^40 (or ~10^12) states, although the actual number would likely be less as some modifications are mutually exclusive. In our view, this is not a meaningful synthesis and, as has been recognized in other areas of cell biology, identifying modular features of biological systems can make their representation tractable[18–21]. Therefore, it is not experimentally (or conceptually) helpful to define states in a model of chromatin based solely on combinatorial patterns of protein occupancy—states in a manageable model should exhibit a transcriptional readout and a chromatin structure phenotype, rather than just reproducible patterns of protein occupancy.

There are numerous histone variants[22] and post-translational modifications[23,24] for which transcriptional readouts, and/or chromatin features, have been described. This is true for states ostensibly "on" or "off" in terms of transcription. For some modifications, effects on higher order chromatin structure have been observed. The recent development of genome-wide ChIP-seq studies has in turn provided extensive information about localization of these modifications in distinct cell types[25–27]. In a similar vein, it is apparent that nucleosome positioning is non-random, dynamic and plays a fundamental role in gene expression[28–30]. What is lacking is a universal dogma for how these features operate in concert to specify global gene expression and thereby phenotype. An emerging theme is that no modification acts in isolation and that, to the extent such modifications specify chromatin states, global changes in gene expression are achieved through combinatorial complexity. A classic example of this phenomenon is bivalent chromatin marks[31], that is, genes labeled with histone post-translational modifications that alone

specify opposing chromatin accessibility and/or transcriptional activity, but combined create decision-making nodes in gene regulation networks.

Permanent inactivation of genes, a well-known example being inactivation of the X chromosome in the female, leads to a distinct type of off state. This process is mediated in large part by histone post-translational modifications, long non-coding RNAs and DNA methylation[32]. Genes in this region of DNA are permanently silenced in all cell types in the organism ("off permanently silenced", panel B in Figure 1), although which copy of the X chromosome is inactivated varies between cells.

Another functionally obvious chromatin state is characterized by genes not expressed in a given cell type due to silencing associated with normal differentiation[33]. Such genes clearly exist, but a universal dogma to explain their regulation in the context of chromatin states—why each cell has a different gene/protein expression profile—is lacking. Nevertheless, these genes clearly occupy a different off state, in that they can be activated in other cells but never will be in the normal life of the given cell ("off temporarily silenced", panel B in Figure 1). We propose this chromatin state to be plastic as evinced by dedifferentiation in heart disease[34], cancer[35] and induced pluripotency[36].

The last species of functionally distinct off state we seek to define is that containing genes that can be expressed in the normal life of a differentiated cell but that are off at a given point in time ("off inactive", panel B in Figure 1). This includes genes activated by stress, mitogens, injury, environmental cues and so forth. One can conceptually envision a "poised off" state that is distinct from those discussed so far and which may exist, for example when RNA polymerase disengages from DNA before nucleosomes reassemble, however we know of no functional readout for such a state in terms of transcription and/or chromatin structure.

The most basic "on" state is that in which transcription is actively occurring ("on, active", panel B in Figure 1) and is identical to the "on" state in the two state model. There is ample evidence that an additional on state exists when RNA polymerase "pauses" on the template,

temporarily delaying elongation[37,38]. Another distinct on state is one in which chromatin is "poised" for transcription[39,40], at which sites transcription is imminent but not active. Both of these states imply directionality and whether they are biologically (rather than just semantically) distinct, remains unknown.

It is well established that RNA polymerase exhibits different firing rates—that it, different rates at which the RNA is generated on a given template. In the present definition of a chromatin state model, we avoid delineation of these different rates as different states due to the absence of evidence for coordinate variable chromatin structure; it is very possible that such structural differences exist and have yet to be detected (or, perhaps these are subspecies of the "on, active" state).

Based on currently available experimental data and the preceding structure-function considerations, we favor a six state chromatin model (Panel B in Figure 1). Of course these states are defined not for entire genomes but for regions of the genome within a single nucleus— different regions of the same genome can simultaneously exist in multiple states. Constitutive heterochromatin is thought to be the same between cell types (although there are exceptions to this, for example in the regulation by centromeric satellite DNA[41]) whereas facultative heterochromatin will vary between cell types and within a cell given the state of development. In the present synthesis, these forms of chromatin would be considered "off permanently silenced" and "off temporarily silenced", respectively (Panel B, Figure 1). In Panel C of Figure 1, we consider the types of genes that would reside in individual states in the example of a fully differentiated cell like a cardiac myocyte; while we can predict functional classes of genes for on active, off temporarily silenced, off permanently silenced, it is *a priori* possible to distinguish neither between on active poised and off inactive, nor between on paused, on active and on poised inactive. These distinctions can only be made experimentally. Our goal for this model is to create a framework that incorporates as much of the current data (and ongoing technique development) as possible while at the same time restricting the number of states based on the

field's experimental knowledge of chromatin and transcription. An example of how we envision specific transcriptional phenotypes, along with to-be-determined structural features, contributing to chromatin states, is represented in Figure 2. Undoubtedly the result of this approach is an over-simplification of the endogenous behavior of chromatin and the actions of various proteins. To the latter point, non-nucleosomal chromatin structural proteins, such as CCCTC binding factor (CTCF)[42,43] and high mobility group proteins (HMG)[10,15], have been shown to control gene expression and phenotype, while their roles in chromatin packing appear to include formation of intermediate chromatin domains and global endogenous genomic structure. Incorporating the actions of these types of proteins into a chromatin state model with transcription as a readout requires conceptualizing their actions in the context of the endogenous genome: unlike a conventional transcription factor model, these proteins do not have their effects only in and around the transcription start site.

### *Implications and Future Experiments*

The objective of this thesis is to lay the groundwork for a chromatin model that incorporates three-dimensional structural changes with a transcriptional readout. Such a model is an essential analytical tool, we reason, to resolve the issue of how genome-wide regulation of gene expression is accomplished in eukaryotes *in vivo*. We present two models that fulfill the criteria of having structural changes in chromatin and transcriptional readouts: a two state and greater than two state model. For reasons of completeness with existing data and ability to generate hypotheses, we favor a model with greater than two states and suggest that a six state model is best representative of current knowledge from both isolated genetic elements and genome-wide studies. While other models of chromatin have been proposed[14,27,44,45], there are none to our knowledge with these criteria linking structure to transcription as described herein. Of course a model of chromatin states must be linked inextricably with data—our point herein is that incorporating knowledge of the biological function of chromatin can result in a more

tractable model. The next experimental step is to directly measure structural features that define individual states.

## Use of a Chromatin State Model

Implicit in a transcriptional model of chromatin states are the structural changes at intermediate levels of organization, that is, below the level of the chromosome and above the level of the individual nucleosome. Recent work[30,46] with DNase hypersensitivity, MNase-sequencing analyses and computational approaches has revolutionized the way we think about how nucleosomes associate with different regions of the genome and the logic for the repositioning of these protein complexes commensurate with transcription factor binding. For instance, we now know that upstream of transcription start sites, nucleosomes exhibit very regular spacing[29], likely contributing to one of the "on" states in this or other chromatin state models. On a more global level, nucleosomes are enriched in exons and depleted in introns[47], further supporting the role nucleosome density plays in chromatin structure. One limitation of these nuclease-sequencing studies is that the information, although now acquired with single base resolution, is projected onto the linear representation of the genome; in the process, three-dimensional information is lost. Indeed these intermediate areas of chromatin structure represent, in our view, a key frontier in the study of chromatin and chromatin states: new techniques that can directly measure intermediate chromatin states (e.g. by imaging[12,48] and/or chromosomal capture techniques[5,6])—between the level of the nucleosome and the whole genome—and link these states with transcription, will advance our understanding of genome packing and chromatin biology. These approaches can enable representations of chromatin that include both three-dimensional structure and gene expression networks, such that features of cell type-specific chromatin structure (and thereby, transcriptome generation) can be modeled[7], and ultimately, compared.

The worth of this model is that it provides testable hypotheses. Converting between chromatin states requires something other than the DNA substrate itself; two obvious and extensively studied candidates are RNA and protein. Notwithstanding DNA sequence preferences for remodelers and transcription factors, how inter-state conversion is coordinated on a genome-wide scale is unknown. To test this, one would need to experimentally characterize transitions to establish directionality of movement among states. To the extent possible, this would include recapitulation of large-scale chromatin structures *in vitro* and/or use of mathematical modeling combined with experimental data to define distinct three-dimensional structural features, should they exist, that are hallmarks of different states. Once structural hallmarks can be established, dwell times in each state can be measured (similar to what has been done genome-wide for individual nucleosomes[49]) and this quantitative information added to the model. A major challenge with this type of experimentation is that it at present is conducted on populations of (usually millions) cells or on tissues. The heterogeneity present in such populations presents a major technical hurdle for rigorous delineation of any discrete subcellular event, chromatin remodeling included, and new methods to measure single-cell level changes[50] will revolutionize the way we examine subcellular processes.

Importantly, should a model such as the one proposed herein be tested experimentally, analyses of genome occupancy (ChIP-seq) must be coupled with expression (microarrays/sequencing) and chromosomal conformation capture analysis, to define whether nuanced structural states accompany the various patterns that emerge from ChIP-seq and gene expression studies. *Any of these three global methods alone is insufficient to define chromatin states.* If one accepts the premise of chromatin states defined by structural features and transcriptional states, then any investigation of chromatin measuring these outputs directly or indirectly could be interpreted in the context of this model—that is, assigning individual loci measured in a given system to one chromatin state. In more comprehensive studies, this would allow determination of what percentage of the coding genome resides in any given state, and if

carried out before and after stimuli (or perhaps in healthy and diseased cells), would shed light on how physiological processes result from shifting portions of the genome between different chromatin states. Such comprehensive studies, which are increasingly common, would allow the model to be disproven by showing that additional chromatin states exist based on significant portions of the genome exhibiting structural and transcriptional features not defined in the existing states, or to demonstrate the absence of structural features distinguishing two states posited to exist, implying only transcriptional states, not those of chromatin, determine a physiological response. Regardless of the system, key data inputs for such a model include loci-specific transcriptional data (microarrays and RNA seq—the higher the resolution, the better) and chromatin structural readouts (principally techniques like MNase-seq, DNase hypersensitivity analysis and ChIP-seq, but increasingly super resolution imaging and chromosomal capture techniques like 3C/HiC will play a role). We speculate that regions of the genome near each other in three-dimensions based on 3C/HiC data would occupy the same chromatin state. Insights into cellular phenotype will certainly aid in the use of these models of chromatin states but in principle are not required for their generation. We envision these models being probabilistic to begin with, due to the large amounts of data and distinct number of loci to be considered, although when possible it will be appealing to apply numerical and potentially analytical methods (employing differential equations) to encapsulate the behaviors of chromatin.

Other Contributors to Chromatin States?

While DNA methylation is clearly associated with modulating chromatin function[51], the discrete structural features it endows make challenging its direct incorporation into a state model. Like histone post-translational modification, context specificity clearly plays a role in determining the readout. Future studies into histone and DNA modifications need to address directly whether there is universal employment of combinations of marks to endow different readouts. Whether there is a code in the strict sense will require omniscient knowledge of all

histone modifications (an unrealistic proposition) or a reimagining of what constitutes the digits in such a code. Regardless, if the goal is to understand global remodeling of chromatin, more studies examining histone and DNA modifications need to test readouts of chromatin structure directly, including: DNase sensitivity, MNase digestion, sedimentation analysis in reconstituted chromatin, high resolution microscopy and perhaps most prominently, chromosomal conformation capture techniques. This model posits six states for coding DNA, but only one ("off permanently silenced") for non-transcribed regions of the genome. A major challenge for structural studies of chromosomes *in situ* is to determine structures of non-coding regions in a manner that allows them to be linked to phenotype, regardless of whether these non-coding regions directly or indirectly affect transcription.

How does one distinguish whether a structural or transcriptional action is of paramount importance, when evidence for both exists (whereas often when evidence for which comes first, on an atomic scale, does not)? In the case of a transcription factor binding to the promoter of a gene, it is not much of a jump to assume that if the transcription factor induces a structural change important for transcription, then this change should precede transcription. However, for chromatin structural proteins that decorate various coding and non-coding areas of the genome, the simple linear view of local structural changes preceding transcription of the same region quickly becomes inadequate. Thus, when no data on time-course exist (from a straightforward cause-effect point of view), our preference is to consider them as a single event. Novel approaches that can link structural changes to transcriptional outputs when these two events are not connected by the linear DNA strand (i.e. that rely on the endogenous three-dimensional architecture of the genome) will enhance this understanding.

Genomes—the term here referring to the DNA and all the chromatin structural proteins and RNAs that bind it—are self-organizing systems; there is no master regulator mechanism that assembles the three dimensional structure of the genome *in vivo*. In this property genomes are not unlike proteins themselves[52], in that structural features arise at secondary, tertiary and

quaternary levels to endow functionality not present in the primary protein/DNA sequence. Like protein structures, then, we can examine genome structures in different cells to reveal features endowing cell type specific gene expression profiles. One fundamental difference between these two systems is that proteins are thought to be structurally super-imposable between copies within a cell or between cells; we certainly do not imply this to be the case with the genome, where structural similarities between copies of a genome are more likely global patterns (think: cloud formations). From an evolutionary standpoint, structural features that arose in proteins, untraceable to amino acid sequence, are selected for based on function. So too it may be with genomes, in that the structural features of the genome in three dimensions—and the consequent properties thereby determining how different regions shift their chromatin states—determine phenotype.
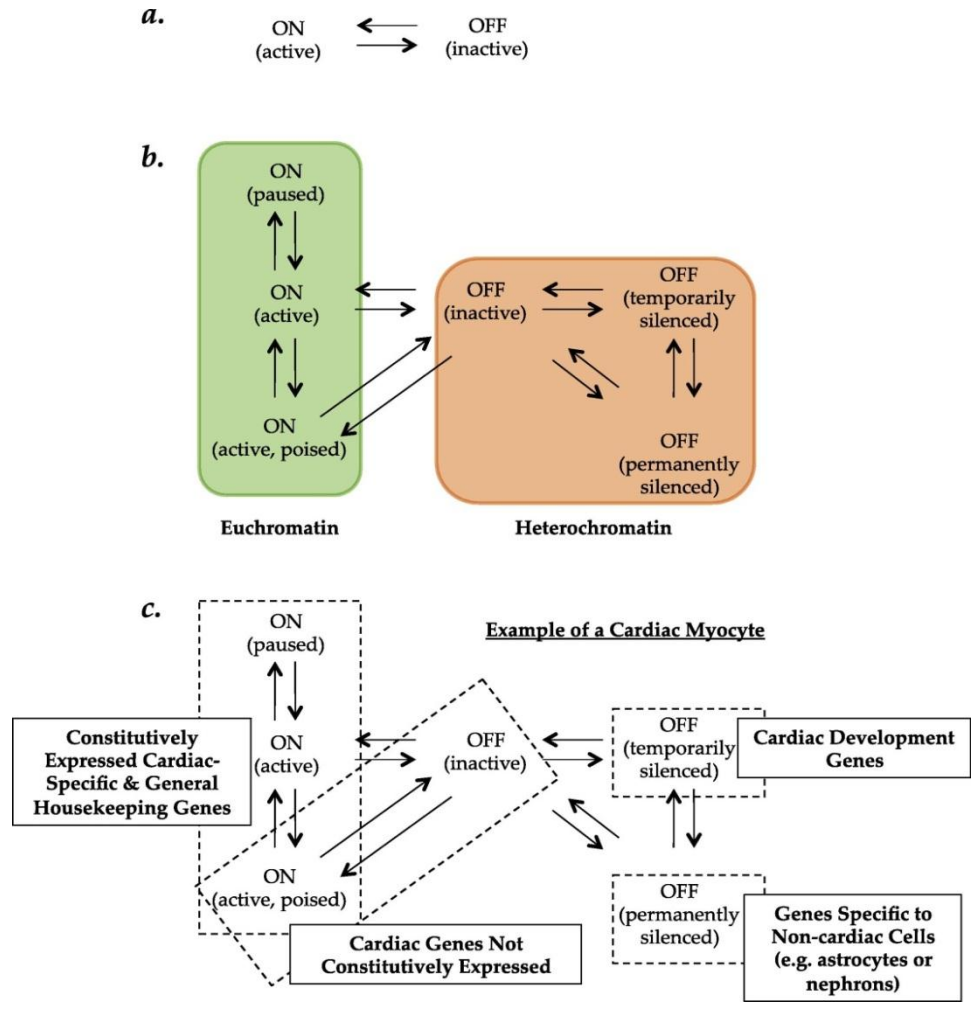
**Figure 1. Chromatin state models.** (A) Although unequivocal data may exist for only a two state model of chromatin with transcription as a readout, with on and off states corresponding to DNA being transcribed or not, this model does not facilitate hypothesis generation for unraveling new properties of chromatin. (B) We propose a model with six states based on the following criteria: each state is defined by chromatin structural features and a transcriptional readout; major chromatin behaviors meeting the previous requirement are incorporated in one of the states; an open format is used in which more states can be added based on the determination of global features meeting the structure/transcription criteria; and the model incorporates high-throughput ChIP-seq and related data and is constrained by the known physiological properties of chromatin. The arrows indicate hypothesized routes of transition between states and the shaded areas indicate how we envision these distinct states fitting into a broader concept of chromatin as either hetero- or euchromatic. (C) Different types of genes in various states are given for the example of a cardiac myocyte to illustrate the hypothesized relationship between chromatin states and cell-type specific gene expression and phenotype.
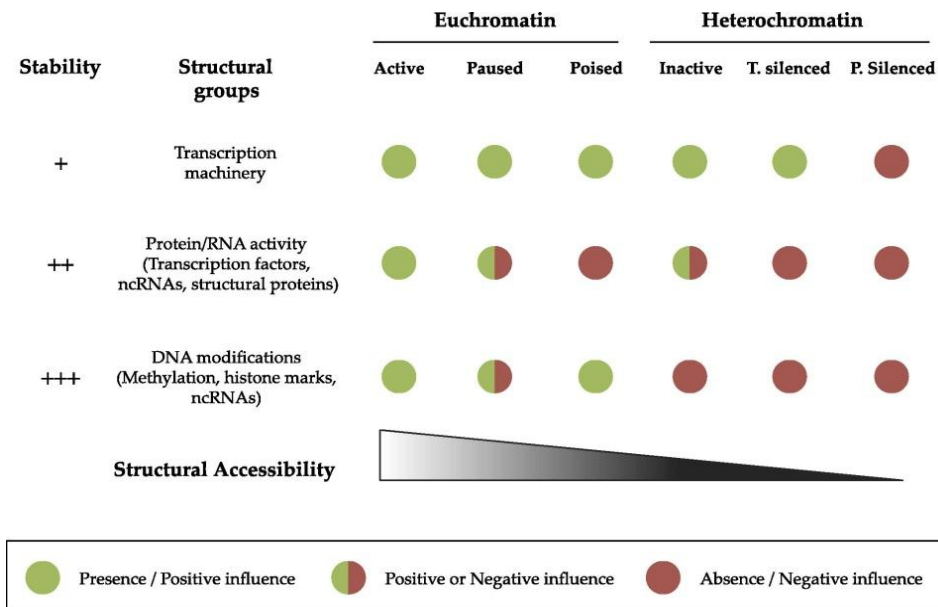
**Figure 2. Hypothesized structural differences between chromatin states.** We hypothesize that three key structural groups can, in conjunction with endogenous 3D genomic structure, distinguish the different proposed states at the level of the gene, namely the accessibility of the gene to RNA polymerase II and transcriptional machinery binding (top row), the presence of the appropriate transcription factor in the nucleus and the localization of specific chromatin structural proteins/non-coding RNAs to the gene (middle row) and the presence of conducive, and absence of inhibitory, DNA and histone modifications at the gene (bottom row). Together these help define the structural accessibility of the gene (bottom triangle, with decreasing accessibility as one moves from left to right through the states as displayed). For each state we propose which class of structural elements are conducive to transcription (green) and which are limiting (red). We further propose that for two states in particular, paused and inactive, more experimental data is needed to determine which class of structural features cause the different transcriptional read-outs (half red/half green indicates possibilities which must be confirmed experimentally). Finally, for each class of structural groups we propose their relative stability and thus the energy required to interconvert between states when these levels of structural regulation are modified.

# Reference

1.  Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–60 (1997).

2.  De Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & development* **26,** 11–24 (2012).

3.  Tremethick, D. J. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell* **128,** 651–4 (2007).

4.  Fussner, E., Ching, R. W. & Bazett-Jones, D. P. Living without 30nm chromatin fibers. *Trends in biochemical sciences* **36,** 1–6 (2011).

5.  Van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nature biotechnology* **28,** 1089–1095 (2010).

6.  Baù, D. *et al.* The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology* **18,** 107–14 (2011).

7.  Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148,** 84–98 (2012).

8.  Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182–7 (2010).

9.  Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology* **5,** e1000598 (2009).

10. Cuddapah, S. *et al.* Genomic profiling of HMGN1 reveals an association with chromatin at regulatory regions. *Molecular and cellular biology* **31,** 700–9 (2011).

11. Guo, J. U. *et al.* Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nature neuroscience* **14,** 1345–51 (2011).

12. Solovei, I. *et al.* Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. *Cell* **137,** 356–68 (2009).

13. Kauffman, S. *At Home in the Universe*. (Oxford University Press, 1996).

14. Baker, M. Making sense of chromatin states. *Nature methods* **8,** 717–22 (2011).

15. Franklin, S. *et al.* Quantitative analysis of the chromatin proteome in disease reveals remodeling principles and identifies high mobility group protein B2 as a regulator of hypertrophic growth. *Molecular & cellular proteomics* **11,** M111.014258 (2012).

16. Vermeulen, M. *et al.* Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell* **142,** 967–80 (2010).

17. Young, N. L. *et al.* High throughput characterization of combinatorial histone codes. *Molecular & cellular proteomics* **8,** 2266–84 (2009).

18. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–52 (1999).

19. Vondriska, T. M., Klein, J. B. & Ping, P. Use of functional proteomics to investigate PKC epsilon-mediated cardioprotection: the signaling module hypothesis. *American journal of physiology. Heart and circulatory physiology* **280,** H1434–41 (2001).

20. Franklin, S. & Vondriska, T. M. Genomes, proteomes, and the central dogma. *Circulation. Cardiovascular genetics* **4,** 576 (2011).

21. Koch, C. Systems biology. Modular biological complexity. *Science* **337,** 531–2 (2012).

22. Talbert, P. B. & Henikoff, S. Histone variants--ancient wrap artists of the epigenome. *Nature reviews. Molecular cell biology* **11,** 264–75 (2010).

23. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–5 (2000).

24. Turner, B. M. Decoding the nucleosome. *Cell* **75,** 5–8 (1993).

25. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–83 (2011).

26. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–9 (2011).

27. Filion, G. J. *et al.* Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143,** 212–24 (2010).

28. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–98 (2008).

29. Zhang, Z. *et al.* A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **332,** 977–80 (2011).

30.     Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–6 (2009).

31.     Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–26 (2006).

32.     Morey, C. & Avner, P. Genetics and epigenetics of the X chromosome. *Annals of the New York Academy of Sciences* **1214,** E18–33 (2010).

33.     Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463,** 474–84 (2010).

34.     Rajabi, M., Kassiotis, C., Razeghi, P. & Taegtmeyer, H. Return to the fetal gene program protects the stressed heart: a strong hypothesis. *Heart failure reviews* **12,** 331–43 (2007).

35.     Levine, A. J. & Puzio-Kuter, A. M. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science* **330,** 1340–4 (2010).

36.     Yamanaka, S. & Blau, H. M. Nuclear reprogramming to a pluripotent state by three approaches. *Nature* **465,** 704–12 (2010).

37.     Gilchrist, D. A. *et al.* Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143,** 540–51 (2010).

38.     Rougvie, A. E. & Lis, J. T. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54,** 795–804 (1988).

39.     Gross   Garrard WT, D. S. Poising chromatin for transcription. *Trends Biochem Sci* **12,** 293–297 (1987).

40.     Barski, A. *et al.* Chromatin poises miRNA- and protein-coding genes for expression. *Genome research* **19,** 1742–51 (2009).

41.     Ugarkovic, D. Functional elements residing within satellite DNAs. *EMBO reports* **6,** 1035–9 (2005).

42.     Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148,** 335–348 (2012).

43.     Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–211 (2009).

44.     Orkin, S. H. & Hochedlinger, K. Chromatin connections to pluripotency and cellular reprogramming. *Cell* **145,** 835–50 (2011).

45.    Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology* **28,** 817–25 (2010).

46.    Segal, E. & Widom, J. What controls nucleosome positions? *Trends in genetics* **25,** 335–43 (2009).

47.    Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466,** 388–92 (2010).

48.    Mitchell-Jordan, S. *et al.* Features of endogenous cardiomyocyte chromatin revealed by super-resolution STED microscopy. *Journal of molecular and cellular cardiology* (2012). doi:10.1016/j.yjmcc.2012.07.009

49.    Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nature structural & molecular biology* **12,** 46–53 (2005).

50.    Bendall, S. C. & Nolan, G. P. From single cells to deep phenotypes in cancer. *Nature biotechnology* **30,** 639–47 (2012).

51.    Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* **11,** 204–20 (2010).

52.    Grosberg, A., Rabin, Y., Havlin, S. & Neer, A. Crumpled Globule Model of the Three-Dimensional Structure of DNA. *Europhysics Letters (EPL)* **23,** 373–378 (1993).

# Chapter 1: Characterization of Nucleosome Positioning in Cardiac Hypertrophy

## A. Abstract

Nucleosome organization in the cardiac myocyte is not well characterized. Local chromatin packing around the nucleosome protein complex is understood at the atomic resolution; however, higher level properties of chromatin packing in the cardiovascular system are unknown. We hypothesized that nucleosome occupancy (i.e. exact positioning of nucleosomes at given loci) is a feature that regulates altered gene expression in hypertrophy. To investigate mechanisms that establish chromatin structure, genome-wide nucleosome location in heart was determined by Micrococcal Nuclease treatment followed by next-generation sequencing (MNase-Seq). Resulting occupancy profiles from normal and diseased cardiac myocytes exhibit global (i.e. chromosome scale) similarities, but local (i.e. gene/promoter scale) differences. In depth analyses of the data showed that nucleosome positioning patterns at many regulatory chromatin elements are regulated by factors other than primary DNA sequence (presumably, other proteins), especially the periodicity of nucleosome positioning at enhancers, insulators and DNase I Hypersensitive sites. Moreover, our data reveal different periods of nucleosome positioning at these sites. Comparing the nucleosome occupancy profiles between normal and hypertrophic cardiac myocytes, we find a decrease of nucleosome occupancy at many functional elements including DNase I Hypersensitive sites and Long Terminal

Repeats. This study is the first to examine nucleosome positioning in heart and reveals a dynamic process of nucleosome remodeling during cardiac hypertrophy.

## B. Introduction

DNA within the eukaryotic nucleus is packed in the form of chromatin. The basic structure unit of chromatin is the nucleosome, containing eight histone core proteins and a segment of DNA wrapped around the histone core. The classical model of chromatin is called "beads-on-a-string", which means multiple nucleosomes bind one DNA molecule non-specifically and form a string. However, the location of nucleosomes throughout the whole genome is not random, and precise positioning at certain regions is important in maintaining normal gene expression[1–3]. For example, the removal of nucleosomes at promoter region, thereby creating a nucleosome free region, is associated with active gene expression[4,5].

Cardiac hypertrophy is a disease that involves multiple molecular events at the cellular level. Several high-throughput studies have shown that cardiac hypertrophy features a genome-wide co-regulation of numerous genes[6–9]. Many kinds of regulatory machineries are required to coordinate such a complex network. Although it is difficult to examine all molecular behaviors within the cardiac nuclei comprehensively, we reason that we can provide some clues as to how global cardiac chromatin structure is modulated under disease status by studying nucleosome positions on a genome-wide scale.

To our knowledge, no studies of global nucleosome positioning in the heart have been reported. With the development of high throughput techniques, the impact of global nucleosome positioning on the chromatin structure has been increasingly appreciated in

lower organisms such as yeast, as well as in mammalian cells[10–12]. Many experimental and bioinformatic tools have been developed to study how nucleosome positioning is regulated. In this chapter, we wanted to characterize chromatin structure within cardiac nuclei from the perspective of nucleosome positioning, and to explore whether nucleosome positions are altered in hypertrophic cardiac myocytes.

## C.  Materials and Methods

*Mice.* All experiments were approved by the UCLA Animal Research Committee and conform to the National Institutes of Health Guide for the Care and Use of Laboratory Animals. BALB/c mice were purchased from Charles River Lab, Inc. All experiments were performed on 8~12-week-old adult male mice.

*Transverse aortic constriction surgery.* The cardiac hypertrophy was induced by transverse aortic banding surgery as described previously[13–15]. Briefly, animals were anesthetized with 1.5–2.0% isoflurane, intubated, and ventilated with 1.5–2.0% isoflurane in 95% $O_2$, 5% $CO_2$. After shaving the hair from the animal, the chest was entered from the left side via the third intercostal space; the aorta was identified at the T8 region; and a venous vascular clamp (Fine Science Tools), outfitted with a band of silastic tubing at the distal edge of one of the clamps, was placed around the vessel. The internal diameter of the resulting modified clamp was that of a 27.5-gauge needle, a common diameter for aortic banding in the mouse; the clamp is of a size that does not impede pulmonary function. Once the aorta was clamped, distal blood flow was measured and quantified using a flow probe (Transonic Instruments). The chest was then

closed using 6–0 proline suture, during which negative pressure in the thorax was reestablished by removing air with a PE-50 chest tube attached to a syringe.

*Echocardiography (ECHO).* ECHO was used to determine cardiac parameters in live mice as described[13,15]. ECHO was performed on mice sedated with isoflurane vaporized (2.5% for induction, 1.0% for maintenance) in oxygen. The animal's chest was shaved and positioned in the left lateral decubitus position for ultrasonic imaging with a Vevo 770 high resolution ECHO system equipped with a 35 MHz transducer. The short axis view (M-mode) measurements were performed, and data were stored for off-line analysis. Left ventricular (LV) chamber dimensions, including LV size (end diastolic, LVID;d, and end systolic, LVID;s), intra-Ventricular septum (IVS; d) and posterior wall thickness (LV PW;d ), were obtained from M-mode images. LV systolic function is assessed from these measurements by calculating fractal shortening (FS), and left ventricular size assessed by calculating LV Mass and relative wall thickness (RWT; d).

$$\text{LV \%FS} = \frac{LVID;d - LVID;s}{LVID;d} \times 100$$

$$\text{RWT;d} = \frac{LV\ PW;d}{LVID;s}$$

$$\text{LV Mass (Corrected)} = 1.05 \times [(LVID;d + LV\ PW;d + IVS;d)^3 - LVID;d^3] \times 0.8$$

*Isolated adult cardiac myocytes.* Adult cardiac myocytes were isolated using the protocol adopted from others[16,17]. Mouse was injected i.p. with 0.5 cc heparin and anesthetized with isoflurane. After the mouse was fully anesthetized, chest was opened, heart excised out and cannulated gently to the Langendorff perfusion apparatus. Heart was perfused with perfusion buffer (113 mM NaCl, 4.7 mM KCl, 0.6 mM $KH_2PO_4$, 0.6

mM $Na_2HPO_4$, 1.2 mM $MgSO_4$, 4.6 mM $NaHCO_3$, 10mM $KHCO_3$, 10 mM HEPES

buffer, 30mM Taurine, 5.5 mM Glucose, pH 7.5) for 4 min and switch to myocyte

digestion buffer (2.4 mg/mL collagenase II dissolved in perfusion buffer) for another 8

min. Then, ventricles were cut and teased into several small pieces with fine forceps.

Cells were filtered through cell strainer (100 μm Nylon, BD Falcon), and sediment by

gravity for 8 to 10 mins. After a brief centrifuge, cell pellet was resuspended in myocyte

digestion stopping buffer (12.5 μM $CaCl_2$, 10% Bovine calf serum in perfusion buffer),

and washed in PBS.

*Purification of cardiac chromatin.* Cardiac chromatin was isolated as previously

described[15,18]. Briefly, isolated cardiac myocytes were resuspended in ice cold cell lysis

buffer containing 10 mM Tris (pH 7.4), 250 mM sucrose, 1 mM EDTA, 0.15% Nonidet

P-40. The crude nuclei were pelleted by centrifuging and resuspended in nuclear lysis

buffer (20 mM HEPES, pH 7.6, 7.5 mM $MgCl_2$, 30 mM NaCl, 1 M urea, 1% Nonidet P-

40). After solubilization, the samples were centrifuged at $13,000 \times g$ for 10 min to pellet

the insoluble chromatin and remove the nucleoplasm fraction.

*Micrococcal nuclease digestion and DNA-sequencing (MNase-Seq).* Chromatin

pellet was washed twice in MNase digestion buffer (10 mM Tris-HCl pH 7.4, 15 mM

NaCl, 60 mM KCl, 0.15 mM spermine, 0.5mM spermidine), and resuspended in 100 μL

MNase digestion buffer supplementing with 1 mM $CaCl_2$. 50 units MNase (Worthington,

NJ) was added into nuclei and incubated at 37°C for 5 min. Digestion was stopped by

adding 20 μL MNase stop buffer (100 mM EDTA, 10 mM EGTA, pH 7.5) and 100 μL

MNase digestion buffer. Proteins were removed by incubating with 3μL of proteinase K

(25 mg/mL) and 10 μL of 20% SDS at 37°C overnight, and DNA was purified by phenol/chloroform extraction and ethanol precipitation. DNA fragment size was checked by DNA agarose gel electrophoresis. Sequencing libraries were prepared according to the manufacturer provided genomic library preparation protocol version 2.3(Illumina, La Jolla, CA), and were sequenced on Illumina GAII sequencer on single-end mode.

*Bioinformatics.* 76-nt reads were aligned to mouse genome (mm9) using Novoalign V2.07.00. Read trimming is turned on to progressively trim read trailing bases until reads can be aligned or length reduces to less 21nt. Reads aligned to locations are reported as no alignment. For normal mouse, we generated a total of 45.3 million reads, and uniquely aligned 24.8 million reads; for TAC mouse, we generated a total of multiple 43.7 million reads, and uniquely aligned 25.9 million reads. After the alignment, each read was extended to 148nt. Nucleosome score at a given position is defined as the number of reads covering that position. Nucleosome positioning prediction was downloaded from Segal's lab[19]. CpG island and Long Terminal Repeat (LTR)[20] coordinates were downloaded from UCSC Genome Browser.

*Cardiac transcription factor binding sites, histone modification peaks and DNaseI hypersensitive sites.* The ENCODE project[21] and many other high-throughput ChIP-Seq studies[22,23] provide a high quality source of protein-DNA interaction datasets. Cardiac transcription factor (TF) binding sites for GATA binding protein 4 (GATA4) (GSM558904)[23], NK2 homeobox 5 (NKX2.5) (GSM558906)[23], serum response factor (SRF) (GSM558906)[23] and T-box 3 (TBX3) (GSM862695)[22] were downloaded from Gene Expression Omnibus[24]. ENCODE ChIP-Seq peak data were download through

USCS Genome Browser. Binding peak coordinates for mouse heart p300, RNA polymerase II (Pol II), zinc finger CCCTC-binding factor (CTCF), histone modification peaks, were from Dr. Bing Ren's laboratory at the Ludwig Institute for Cancer Research (LICR). Mouse heart DNaseI hypersensitive sites (DHS) were from the University of Washington ENCODE group. Because binding peaks for GATA4, NKX2.5 and SRF are too broad, we identified their binding positions from ChIP-Seq peaks using FIMO from the MEME Suite (version 4.7.0)[25].

*Local normalization of nucleosome positioning profile.* In Figures 1-4, 1-5, 1-6 and 1-7, we applied a local normalization method to each nucleosome positioning profile. Briefly, both normal and TAC nucleosome positioning profiles were normalized to the average read counts at both -3,000 bp ~ -2,500 bp and +2,500 bp ~ +3,000 bp regions. This normalization cancels the background difference between normal and TAC conditions.

*Nucleosome periodicity analysis.* To measure the period of nucleosome positions at p300, CTCF and DHS sites, a similar method[26] was adopted with the following modifications. Nucleosome positioning profiles were smoothed using simple arithmetic moving average with a window size of 50 bps. Local minima of the smoothed curve were chosen as boundaries of nucleosomes. The length of nucleosome repeat was then determined from a linear fit of the peak positions, relative to the first nucleosome, versus the nucleosome number.

## D.  Results

### *Pressure overload-induced heart hypertrophic growth*

In this chapter, we used transverse aortic constriction (TAC) as an experimental mouse model to study pressure overload-induced cardiac hypertrophy. Ligation of the transverse aorta (Fig 1-1, A) generates a back force that increases the afterload on the left ventricle, resulting in a compensatory growth of the heart muscle. When measured by M-mode echocardiography (Fig 1-1, B), altered cardiac morphology can be seen, as indicated by the increased chamber wall thickness (Fig 1-1, C). This growth also manifests as increased heart weight as a ratio of body weight (Fig 1-1, D). We collected samples at the stage of compensatory hypertrophy, when heart function is still normal (Fig 1-1, E).

### *Measurement of cardiac global nucleosome positioning by MNase-Seq*

MNase digests linker DNA between two nucleosomes and produces nucleosome associated DNA (Fig 1-2, A)[1]. These ~147 bp DNA fragments (Fig 1-2, B), when sequenced, can indicate the position of nucleosomes. Our data first confirmed other groups' conclusion that regions with high GC content prefer nucleosome occupancy[19,27] (Fig 1-3, A). When examined at the scale of entire chromosomes, we did not observe major differences between normal and TAC animals (Fig 1-3, A). However, chromatin-remodeling events, as indicated by the relocation of nucleosomes, are clearly visible when zooming in to smaller scales (Fig 1-3, B).

### *Nucleosome occupancy at cardiac TF binding sites*

We examined the nucleosome occupancy profiles for the binding sites of four cardiac TFs that are known to be important in regulating cardiac gene expression and development. Previous studies have shown that the co- localization of GATA4, NKX2.5 and SRF can be used to identify cardiac active transcriptional enhancers[23]. Interestingly, we observed that these three TFs share a similar nucleosome occupancy profile (Fig 1-4, A-C, left panels), implying that the positioning of nucleosomes in these TF sites is determined by primary sequence. Moreover, we observed a pattern of single split peak for these three TFs. To our surprise, nucleosome-positioning prediction gives a similar pattern, including the split peak phenomena (Fig 1-4, A-C, right panels). We also examined the nucleosome occupancy profile for TBX3, which is known as a transcription repressor and plays an important role in cardiac development[28,29]. The nucleosome occupancy profile shows a depletion of nucleosome at the TBX3 binding site (Fig 1-4, D, left panel). We speculate that this depletion may be caused by TBX3 binding rather than reduced nucleosome affinity of DNA sequence, because the nucleosome binding prediction showed an increased nucleosome binding signal (Fig 1-4, D, right panel), which is inconsistent with the experimental profile.

### *Nucleosome occupancy and positioning at cardiac RNA Pol II, p300, CTCF sites and CpG islands*

To further explore the role of nucleosome positioning in regulating chromatin structure, we examined the nucleosome positioning profiles at active transcribed regions, enhancers, insulators and CpG islands (Fig 1-5). Surprisingly, we observed that

nucleosomes are depleted at the center of all these four elements (Fig 1-5, A-D, left panels), while nucleosome occupancy algorithms predict high nucleosome affinity (Fig 1-5, A-D, right panels). This discrepancy indicates protein-DNA complexes other than nucleosomes located at these regions play a role in repelling nucleosome formation. Moreover, the nucleosome positioning profile at CpG island displays a ~60% depletion in nucleosome occupancy, which is the largest nucleosomal depletion among the four elements we surveyed.

We also observed periodicity of nucleosome positioning at p300 and CTCF binding sites, but at neither RNA Pol II nor CpG islands. Interestingly, even though nucleosome occupancy profiles at both P300 and CTCF show periodicity, P300 site has a wider nucleosome free region comparing with CTCF site. We speculate that this periodicity is caused by interactions of other non-nucleosomal proteins with DNA, as DNA sequence-based prediction algorithms failed to detect this periodicity.

### *Nucleosome occupancy at histone modification peaks*

Next, we investigated whether nucleosome occupancy varies depending on chromatin states as specified by histone modification. We studied both euchromatic marks, including H3K4me3, H3K36me3, H3K79me2, H3K9ac and H3K27ac, and the heterochromatic mark H3K27me3. Due to the lack of histone modification data in disease, our data cannot examine the nucleosome positioning at disease-specific histone modification peaks.

Based on the shape of nucleosome positioning profiles from normal mouse, it can be inferred that H3K4me1, H3K36me3, H3K79me2, H3K27me3 and H3K27ac

modifications stabilize nucleosome formation, as indicated by the increase of nucleosome scores at the center of histone modification (Fig 1-6). In contrast, H3K4me3 and H3K9ac repel nucleosome formation (Fig 1-6). Comparing the nucleosome occupancy between normal and TAC mice, we found that nucleosome scores decreased at H3K4me1, H3K4me3, H3K27me3 and H3K9ac, indicating chromatin remodeling events at these sites.

***Nucleosome occupancy at DNaseI hypersensitivity sites***

DHS mark open chromatin regions that often serve as cis-regulatory elements including promoters, enhancers, and insulators[30]. The high chromatin accessibility at DHS usually indicates a low level of nucleosome occupancy. To investigate the nucleosome positioning within DHS to test this hypothesis for the heart, we calculated the average nucleosome occupancy at DHS using published heart DHS position data from two replicates (Fig 1-7). We observed a reduced nucleosome occupancy at DHS (Fig 1-7, left panels), which is inconsistent with nucleosome occupancy predictions for these regions (Fig 1-7, right panels), indicating that the major cause of DHS is not due to the low nucleosome affinity of primary DNA sequence. Moreover, nucleosomes are well positioned at DHS and form a pattern similar, although not exactly the same, as CTCF or p300. Again, this periodicity is not predicted by nucleosome positioning algorithms, suggesting that this pattern is caused by protein-DNA interactions, not primary DNA sequence.

*Nucleosome repeat length at enhancers, insulators and DHS*

Our nucleosome positioning profiles showed well-positioned nucleosomes at enhancers (Fig 1-5, B), insulators (Fig 1-5, C) and DHS (Fig 1-7). The periodicity of nucleosome positioning at these elements makes it possible to measure the nucleosome repeat length, i.e. the length of nucleosomal DNA plus linker DNA between nucleosomes. We chose ±1,000 bp region at the center of these three elements because this region gives a strong periodicity signal (Fig 1-8, A-C). This region is large enough to give us 8~10 well resolved nucleosomes, and the periodicity signal starts to decay outside of this region. Using a linear fit model, we obtained the value of 188bp and 186bp, for enhancers and insulators, respectively. Interestingly, this value decreases to 177bp for DHS. A decrease of nucleosome repeat length indicates shorter distance between nucleosomes. Whether there is a difference in the ratio of linker histone H1 to core histone at these regions is not clear. We did not observe dramatic difference in the nucleosome repeat length at these elements between normal and TAC mice (Fig 1-8, D-F).

*Enhanced nucleosome occupancy at long terminal repeats*

The periodicity of nucleosome positioning at enhancers, insulators and DHS sites raised a question of what causes this periodicity. One potential explanation is the nucleosomal barrier model[31]. Based on this model, highly phased nucleosomes, either due to the barrier effect of nucleosome free regions or highky compacted chromatin regions, should be able to form well-positioned arrays of nucleosomes. Long terminal repeats (LTR) are insertions of large DNA fragments caused by retrotransposons. According to

the data from UCSC Genome Browser, there are about $7.7 \times 10^6$ LTRs in the mouse genome, and the length of LTR ranges from 12bp to about 7,000 bp. It is believed that heterochromatin, formed either by non-coding RNA or by histone modification, plays an important role in silencing transposable elements[32]. Therefore, it can be expected that the nucleosomes at LTR and adjacent regions are well positioned due to the formation of heterochromatin at LTR elements.

To test this hypothesis, we generated nucleosome positioning profiles for different length of LTR for both 5' end and 3' end (Fig 1-9). We first observed an enrichment of nucleosome signals at the LTR, but not its upstream or downstream regions. There is also a small but observable dip of nucleosome occupancy signal at the boundary of each LTR. This may indicate the formation of heterochromatin. We observe only very weak periodicity upstream, downstream, or within the LTRs. Strikingly, comparing with the healthy animal, genomic DNA from the TAC mice shows a decrease of nucleosome levels within LTRs, especially the long ones (Fig 1-9, C, F). This may indicate a potential stress-induced transformation at LTR from heterochromatin to euchromatin, or chromatin remodeling at LTRs as part of a global reprogramming of chromatin structure.

## E. Discussion

In this chapter, we asked the question whether genome-wide nucleosome positioning is remodeled during cardiac hypertrophy. We first generated nucleosome occupancy maps for both normal and hypertrophic cardiac myocytes. We then evaluated nucleosome positioning at cardiac TF binding sites, RNA polymerase II binding sites, enhancers, insulators, CpG islands, histone modification peaks and DHS. Finally, we

calculated nucleosome repeat length based on the nucleosome positioning profiles at enhancer, insulators or DHS. Interestingly, we found substantial differences between experimentally-measured nucleosome positioning and that predicted based on primary DNA sequence. We also identified a number of rearrangements of nucleosome positioning in hypertrophic cardiac myocytes at different histone modification peaks. Our study concludes that many factors other than primary DNA sequence determine nucleosome positioning, and genome-wide nucleosome reorganization occurs during cardiac hypertrophy development. More studies will be needed to examine whether nucleosome rearrangement at histone modification peaks facilitates alterations of chromatin states and how nucleosome positioning and histone PTMs are co-regulated at individual loci.

The discovery that Brahma-related gene 1 (Brg1), an ATPase subunit of Brg1/Brm-associated factor (BAF) chromatin remodeling complex, is involved in heart development[33] and heart failure[34] suggests that chromatin remodeling is an important mechanism of cardiac gene regulation. Unfortunately, no direct evidence was available to examine whether nucleosomes are globally repositioned in cardiovascular diseases. In fact, very few studies have looked at the global nucleosome positioning in disease states in any mammalian cell or tissue. Our study provides the first global view of nucleosome positioning in the heart, a fully differentiated organ, under normal and disease state, and reveals some basic principles for how nucleosomes coordinate with other chromatin features to affect phenotype.

Based on the observation that the arrangement of DNA nucleotides affects DNA flexibility and bendability, several groups have proposed various models to predict nucleosome occupancy based on primary DNA sequences[19,35–37]. The success of these models confirmed the idea that DNA sequence is an important factor that regulates nucleosome formation. However, the fact that prediction results matched *in vitro* formed chromatin better than *in vivo* chromatin clearly suggests the existence of more complex mechanisms other than DNA sequence that determine nucleosome positioning. In this study, we found that a few of nucleosome-positioning patterns obtained through MNase-Seq are consistent with prediction, while others are not (Table 1-1). Among the sixteen functional elements we examined in this chapter (Fig 1-4, Fig 1-5, Fig 1-6 & Fig 1-7, right panels), all of them showed a DNA sequence-based preference of nucleosome formation, while MNase-Seq results showed variance on nucleosome occupancy. Unlike transcription start sites which have an intrinsic nucleosome reduction due to DNA sequence[19], our data suggests that the depletion of nucleosomes at several regulatory regions, such as CpG island and enhancers, is mainly due to multiple protein-DNA complexes competition, not primary sequence.

The split peak pattern of nucleosome positioning profiles at GATA4, NKX2.5 and SRF is very interesting (Fig 1-4, A-C). This cannot be explained by two well positioned nucleosomes because the distance between peaks is too close (~70 bps). Remarkably, predicted nucleosome occupancy profiles are similar to, but not exactly like, experimental data. It is difficult to provide an explanation for this pattern solely based on nucleosome positioning data, but at least this shows that the nucleosome positioning at

some TFs is partially controlled by primary DNA sequence. Whether this is a specific pattern for cardiac TFs or a general pattern for TFs in other organs is unknown.

Unlike model organisms with smaller genome, such as *Arabidopsis thaliana*[38] or *Caenorhabditis elegans*[39], mammals generally have a big genome with billions of nucleotides. This makes it difficult to measure the genome-wide nucleosome distribution accurately, due to the need to perform significantly more sequencing to precisely call mammalian nucleosome positions with base resolution. Chodavarapu, et al. reported a study of nucleosome positioning and DNA methylation using *Arabidopsis*[40]. In that study, they obtained a total number of 48.1 million MNase-Seq reads and generated a 68-fold coverage of nucleosome space. Considering the size of mouse genome is about 20-fold bigger than *Arabidopsis* genome, a total number of 960 million reads will be needed to achieve the same coverage in mice, which is very expensive to conduct. Using MNase-Seq in mouse cardiac myocytes, we obtained ~25 million uniquely aligned reads for one mouse, which converts into a 1~2 -fold coverage. This coverage is not deep enough to do single gene/locus analysis. Nevertheless, it is still possible to study mammalian nucleosome positioning using MNase-Seq. Averaging nucleosome signals at one kind of regulatory element will cancel noise and enhance the real signal, resulting in a more smoothed distribution of nucleosome positioning profile. Nonetheless, this will require certain amount of regions to generate a well-smoothed curve. For example, the ChIP-Seq peak data we used to generate CTCF profile contained 33,172 CTCF binding sites. To effectively cancel noise and smooth nucleosome positioning profile, a large number of

protein-DNA binding sites as identified by ChIP-Seq is required, and this number can be lowered if the depth of MNase-Seq is high enough.

The results of nucleosome spacing, or nucleosome repeat length, from different studies reveal heterogeneity in nucleosome organization, depending on cell type, chromatin state, gene activity and measurement method. Valouev, et al. reported 178-187 bp at active promoter-associated regions, 190-195 within gene bodies and 205 in heterochromatin, in human CD4+ T cells[41]. Teif, el al. using a similar method but different cell types, reported 186 bp for mouse embryonic stem cells and 193bp for mouse neural progenitor cells, and observed an ~10 bp smaller value for CTCF binding sites at these cell types[26]. In this chapter, we reported 188 bp for enhancers, 186 bp for insulators, and 178 bp for DHS. It should be noted that short nucleosome spacing does not always indicate a highly compact chromatin. The binding of H1 to linker DNA may increase the nucleosome spacing and at the mean time form the higher order structure of heterochromatin[42–46]. Therefore, a smaller nucleosome repeat length at DHS indicates less H1 binding and more euchromatic structure at DHS.

In our data, we did not observe any difference in nucleosome repeat length between TAC and normal mice at enhancer, insulator or DHS regions. However, this cannot rule out the possibility that some regions other than the above ones feature a difference of nucleosome repeat length. In fact, our previous study has shown that the ratio between linker histone H1 and core histone H4 decreased in hypertrophic heart[15]. If the hypothesis that the binding of H1 increases linker DNA length and forms higher order structure is correct, then in hypertrophic hearts we should be able to observe a decrease of

linker DNA length. Due to the lack of cardiac H1 localization data, we cannot examine

whether there is a difference of linker DNA length at H1 binding sites after hypertrophy.

A combination of H1 ChIP-Seq study with our data will be able to answer this question.
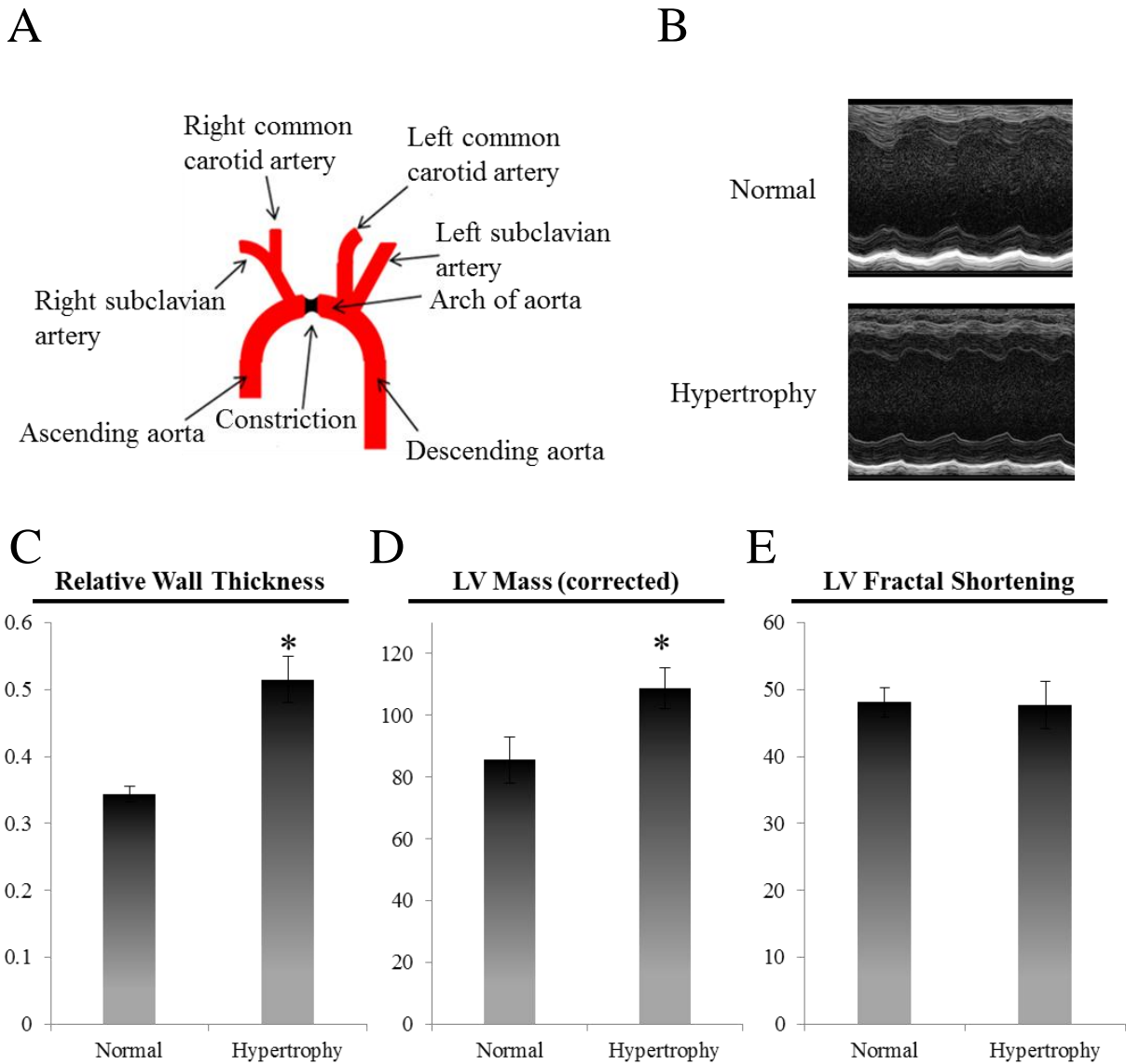
**Figure 1-1. Pressure overload-induced cardiac hypertrophy in mice.** (A) Transverse aortic constriction (TAC) surgery was performed on mice as described in methods. (B) Increased afterload leads to left ventricular growth as measured by echocardiography. Left ventricular wall size, as measured by diastolic relative wall thickness (C), and left ventricular mass (D) significantly increase after one week of TAC, whereas left ventricular fractional shortening remains unchanged (E).
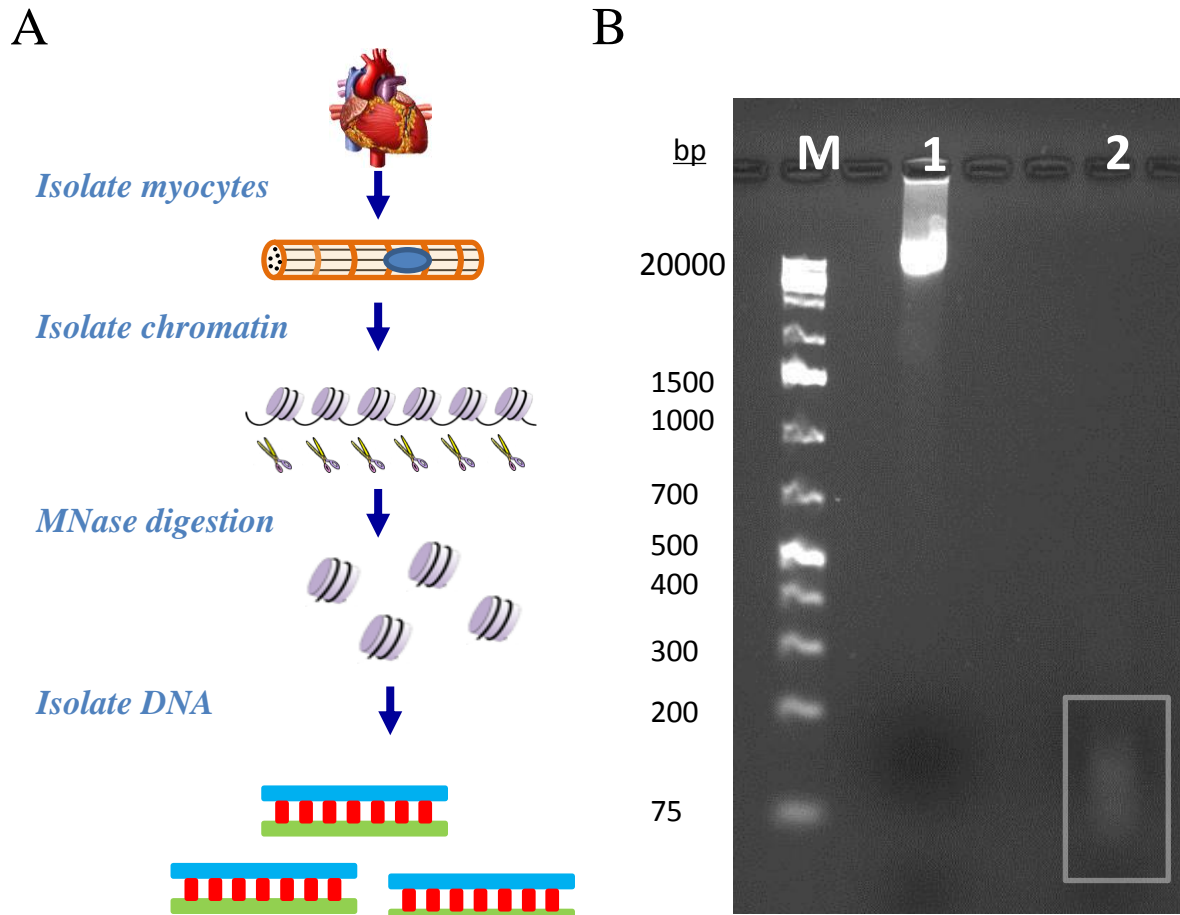
**Figure 1-2. MNase digestion generates mononucleosomal DNA fragments.** (A) Flowchart showing the experimental design of this study. Adult cardiac myocytes are isolated from whole heart, followed by Mnase digestion to remove linkder DNA between nucleosomes. DNA from mononucleosomes are then isolated. (B) DNA argarose gel showing the length of MNase digestion products. Lane M, DNA molecular weight ladder. Lane 1, genomic DNA. Lane 2, mononucleosomal DNA after MNase digestion.
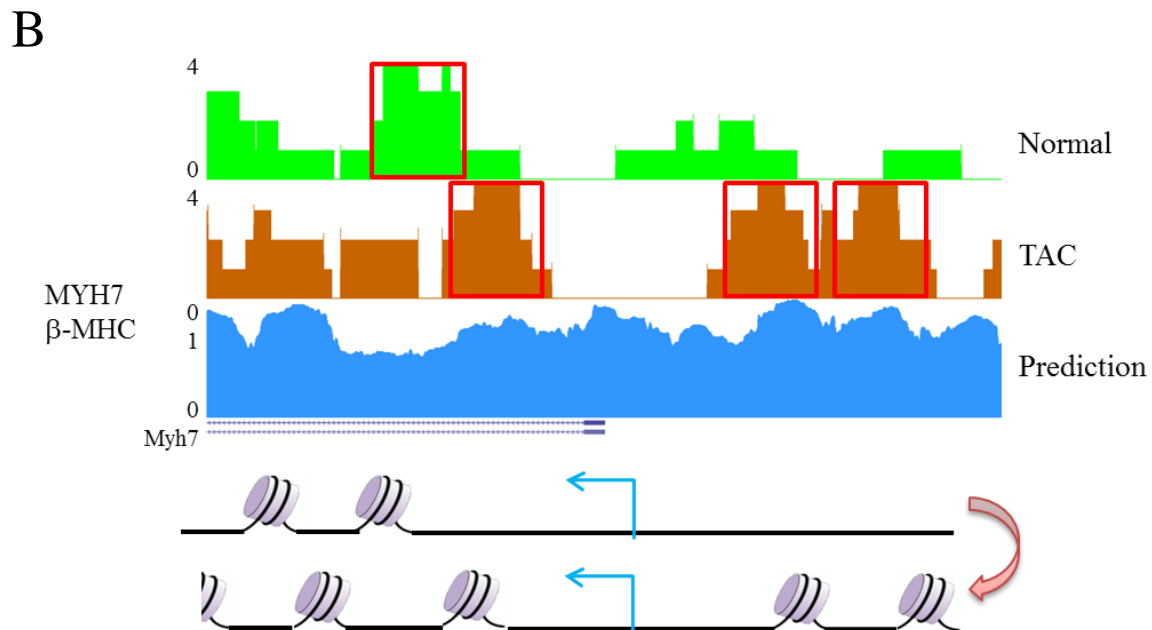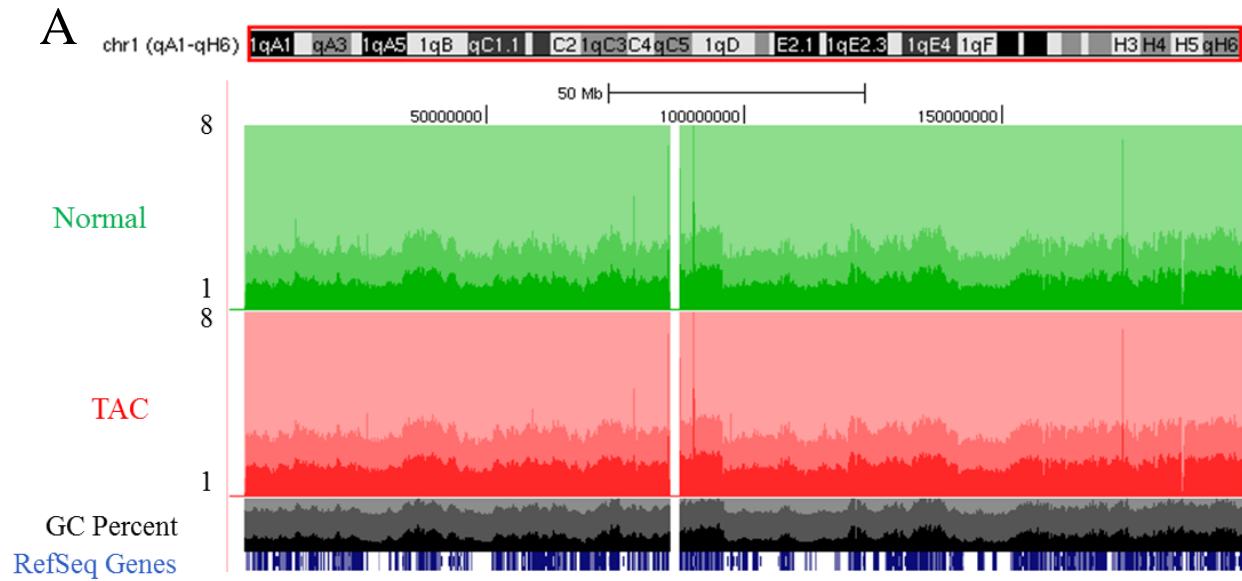
**Figure 1-3. Pressure overload induced alteration of nucleosome occupancy profiles in promoter regions.** (A) Chromosome 1 is shown as an example to display the overall nucleosome positioning at the whole chromosome level. Normal and TAC mice are shown as green and red colors, respectively. GC percent and RefSeq gene tracks are shown at the bottom. Correlation between nucleosome positioning and GC percent, at the chromosome scale level, can be observed. (B) The promoter of β-MHC is shown as an example. Nucleosome scores at the promoter region are shown, along with the nucleosome position prediction based on primary DNA sequence. A diagram of nucleosome position is shown at the bottom panel.

**A**

GATA4

GATA4 (prediction)

Control
TAC
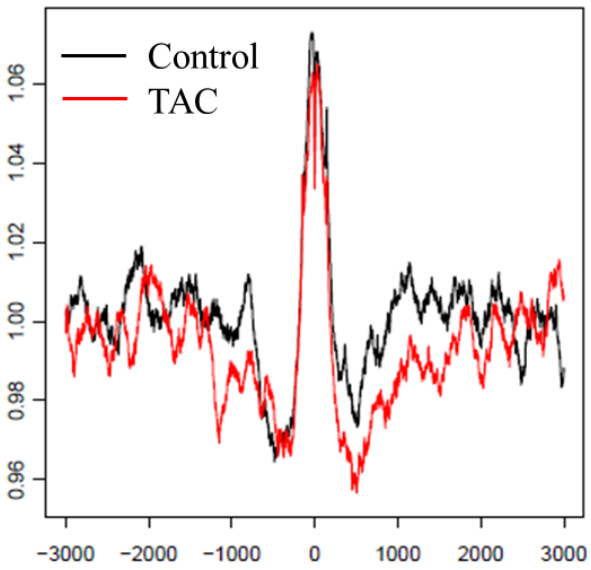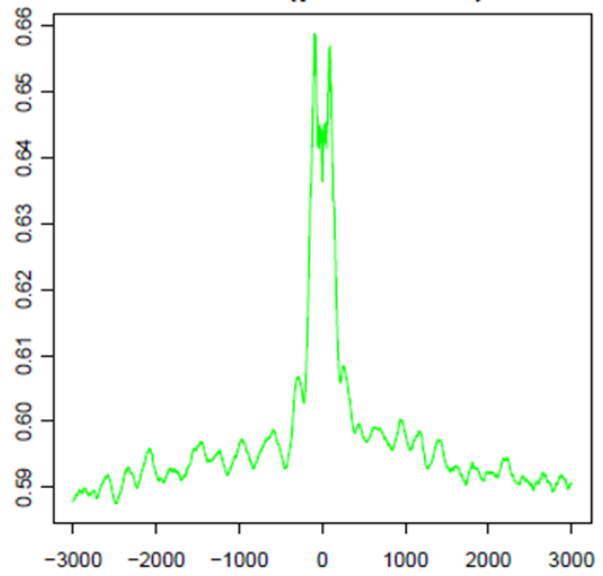
**B**
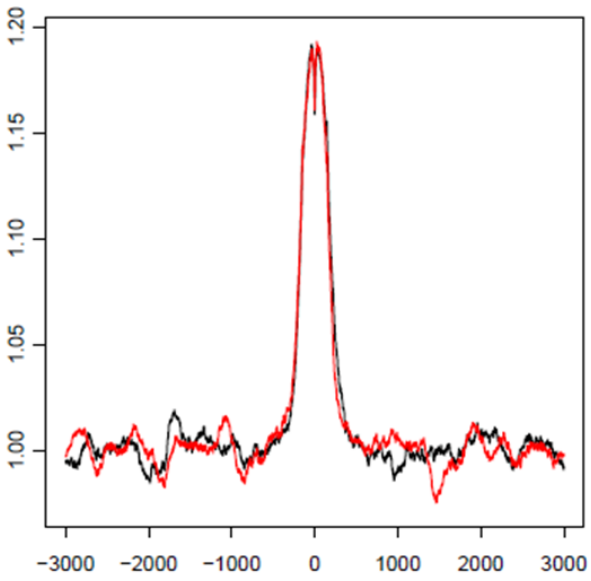
NKX2.5

NKX2.5 (prediction)
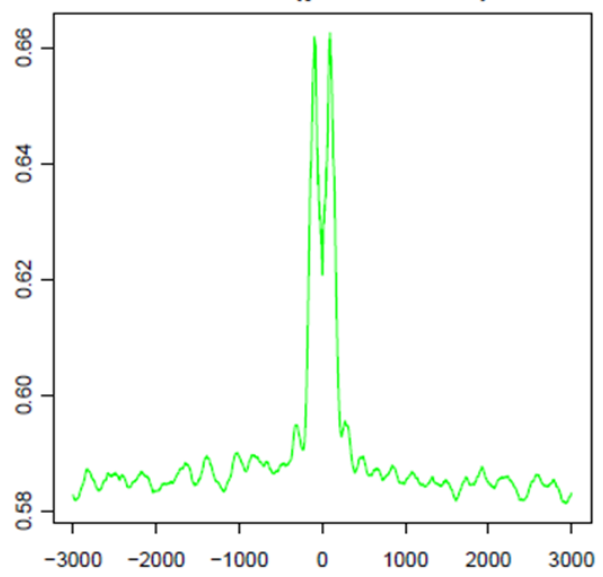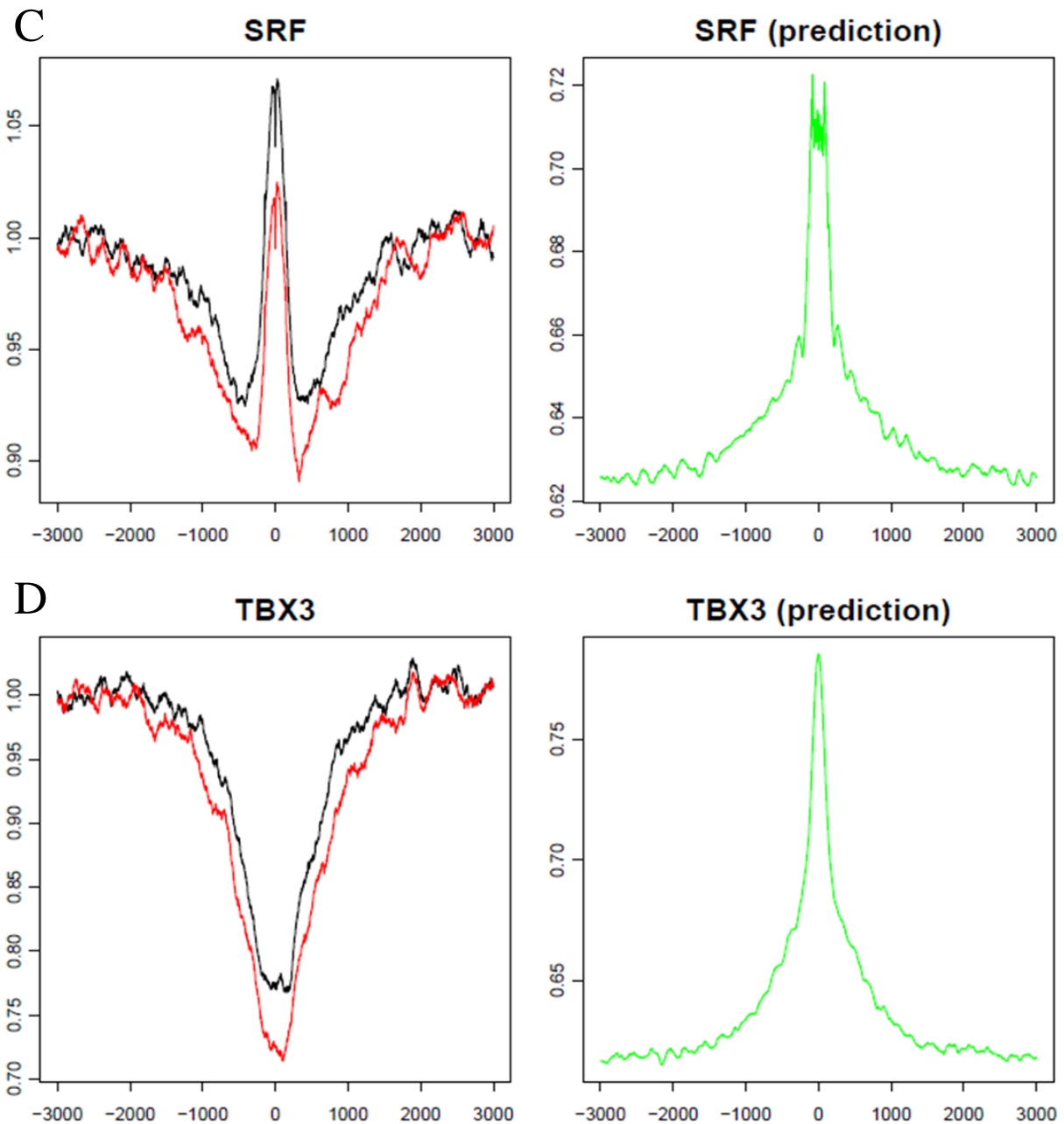
**Figure 1-4. Nucleosome positioning profiles at transcription factor binding sites.** Averaged nucleosome positioning profiles at four cardiac transcription factors, including GATA4 (A), NKX2.5 (B), SRF(C) and TBX3 (D), are shown. Black, control. Red, TAC. X-axis, distance to transcription factor binding site ($\pm$3,000bp). Y-axis, averaged nucleosome score. Left panel, data experimentally obtained. Right panel, nucleosome occupancy prediction based on primary DNA sequences.
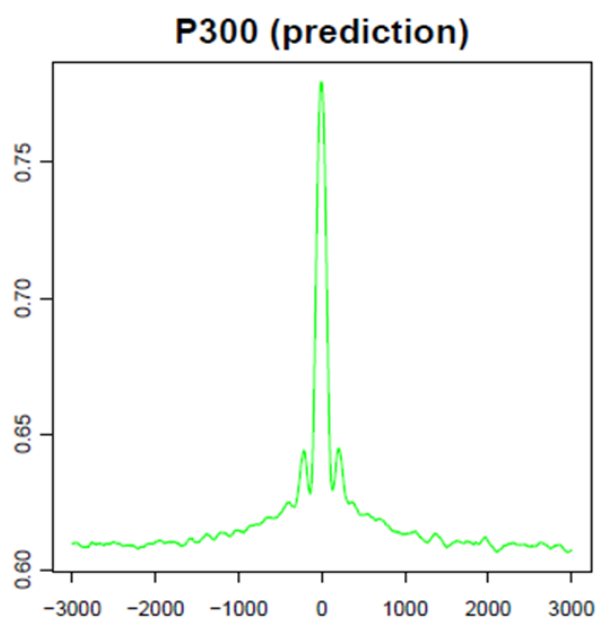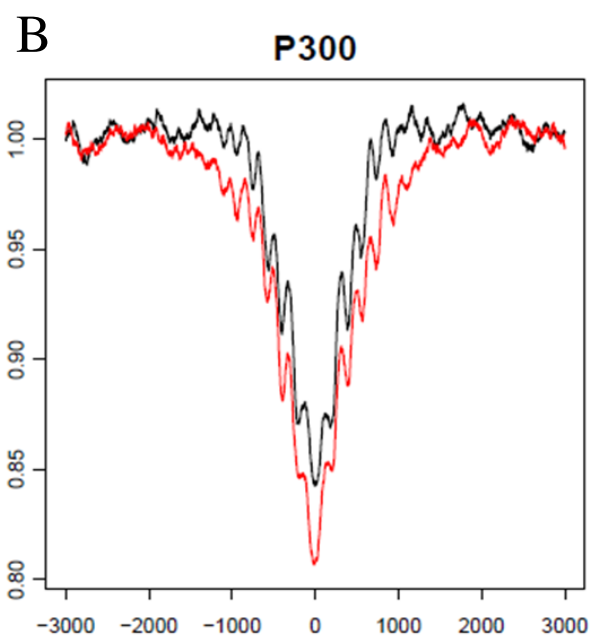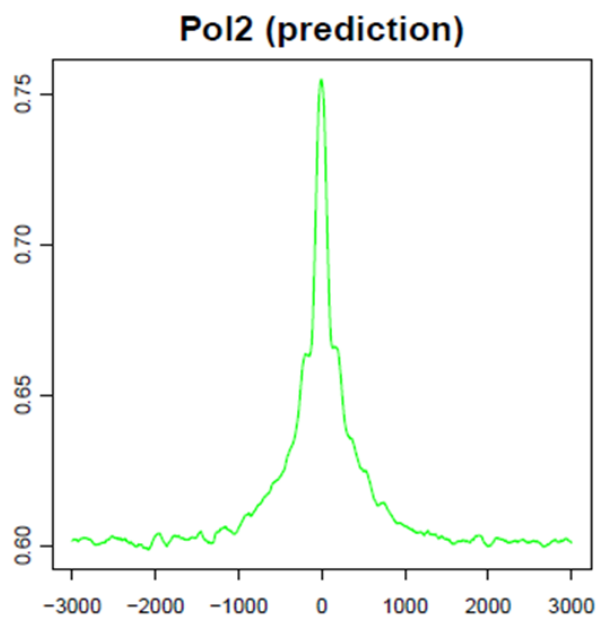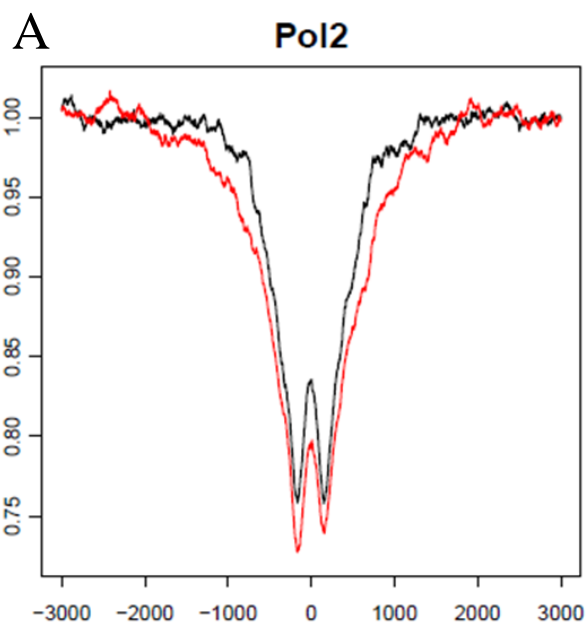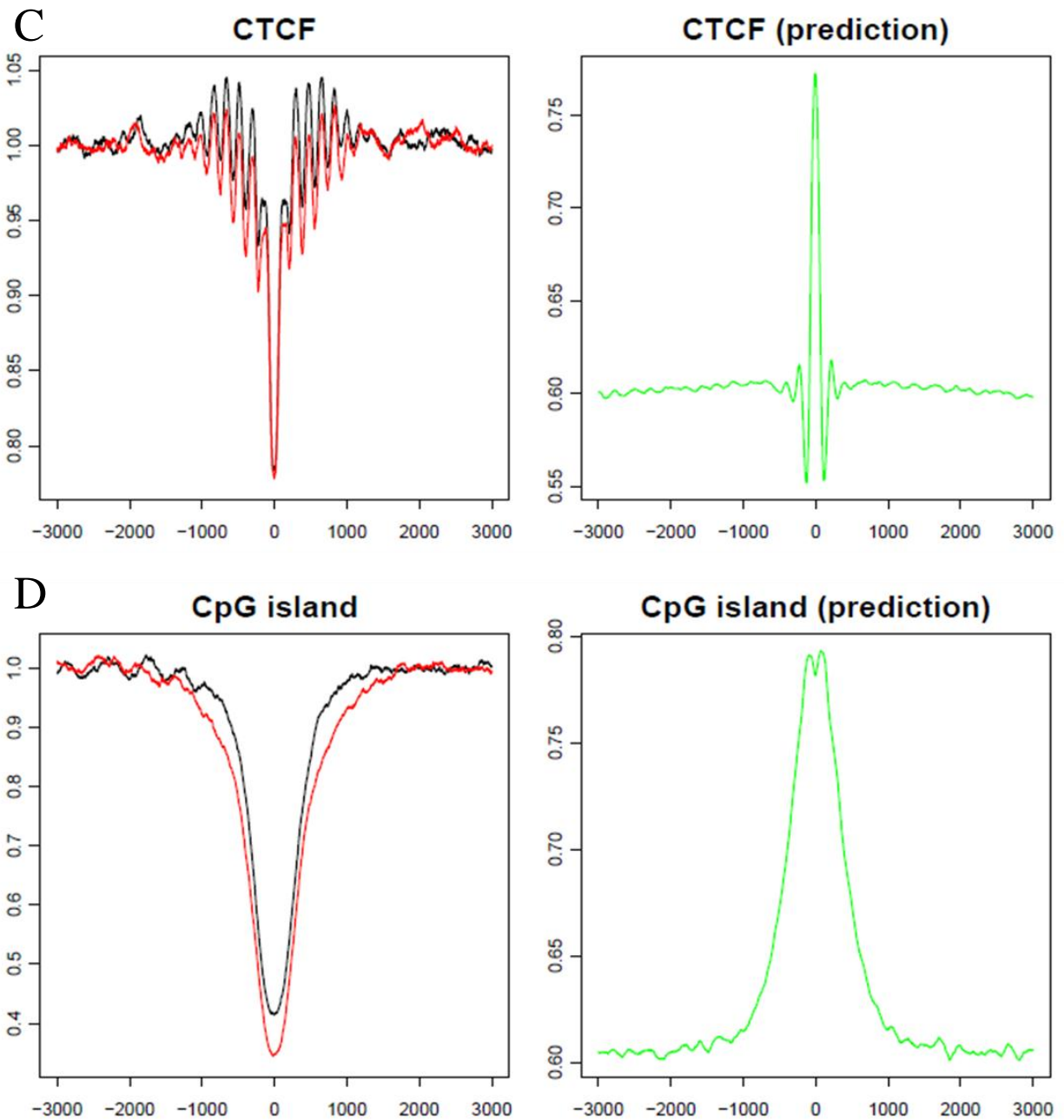
**Figure 1-5. Nucleosome positioning profiles at chromosomal functional elements.** Averaged nucleosome positioning profiles at active transcription regions, (RNA Pol II binding site, A), enhancers (p300 binding site, B), insulators (CTCF binding site, C) and CpG island(D), are shown. Black, control. Red, TAC. X-axis, distance to the center of the functional element ($\pm$3,000bp). Y-axis, averaged nucleosome score. Left panel, data experimentally obtained. Right panel, nucleosome occupancy prediction based on primary DNA sequences.

D

| H3k79me2 | H3k79me2 (prediction) |

E

| H3k09ac | H3k09ac (prediction) |

F

| H3k27ac | H3k27ac (prediction) |
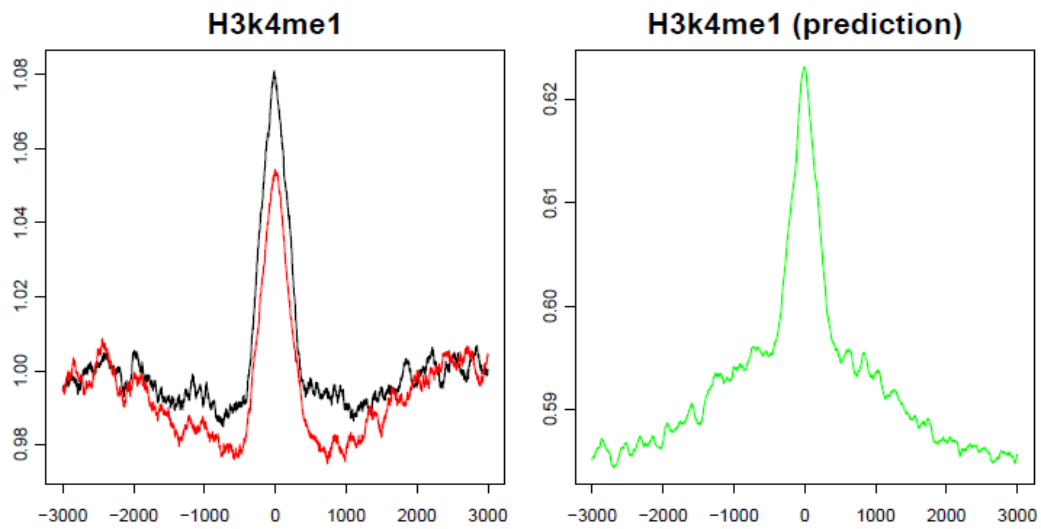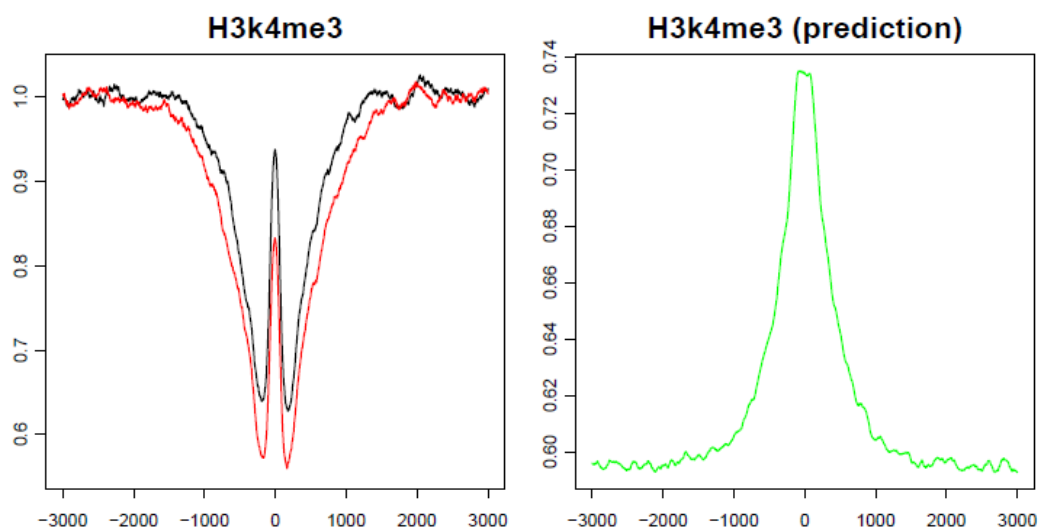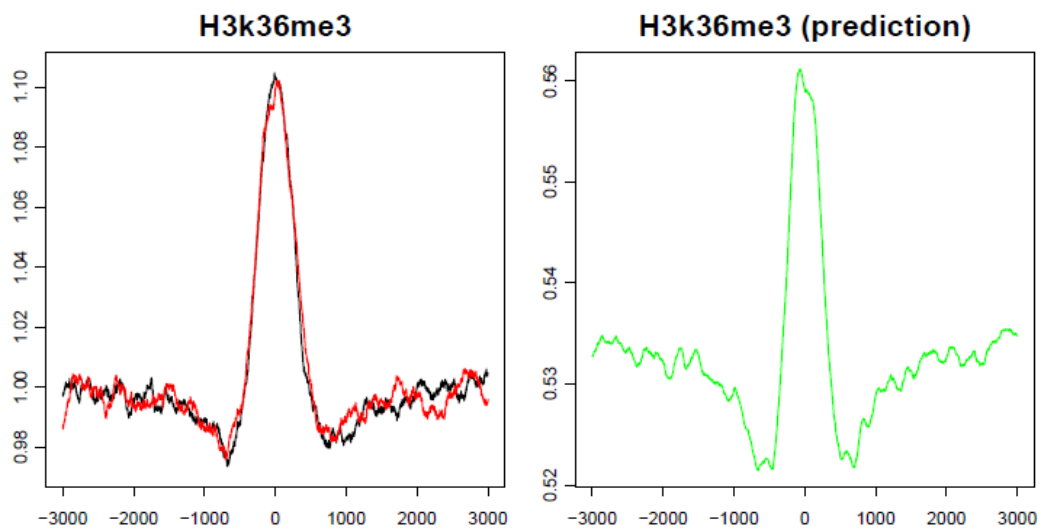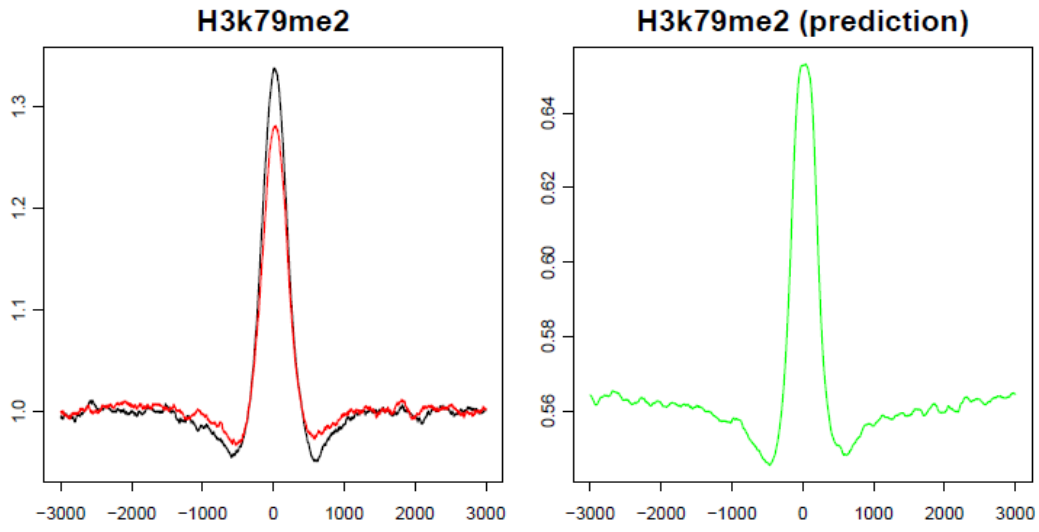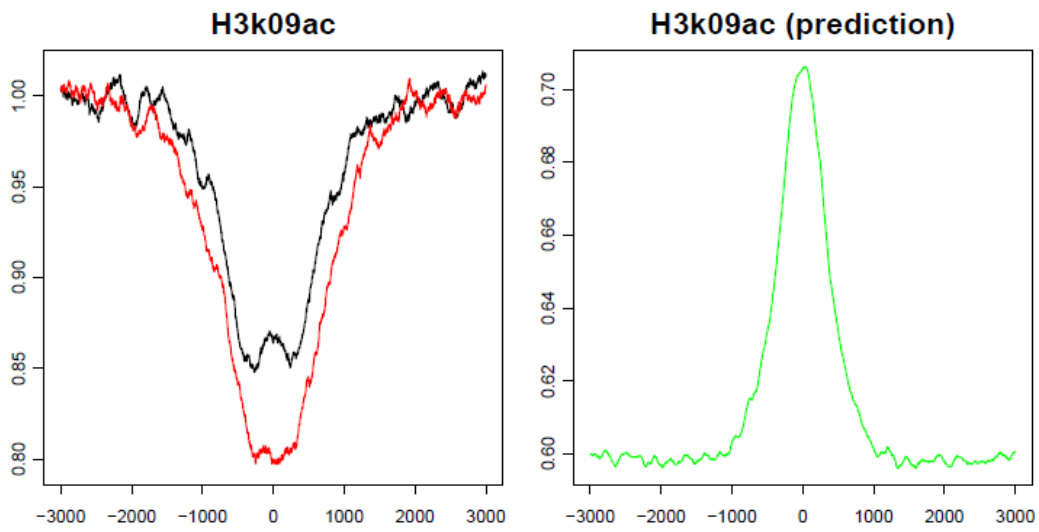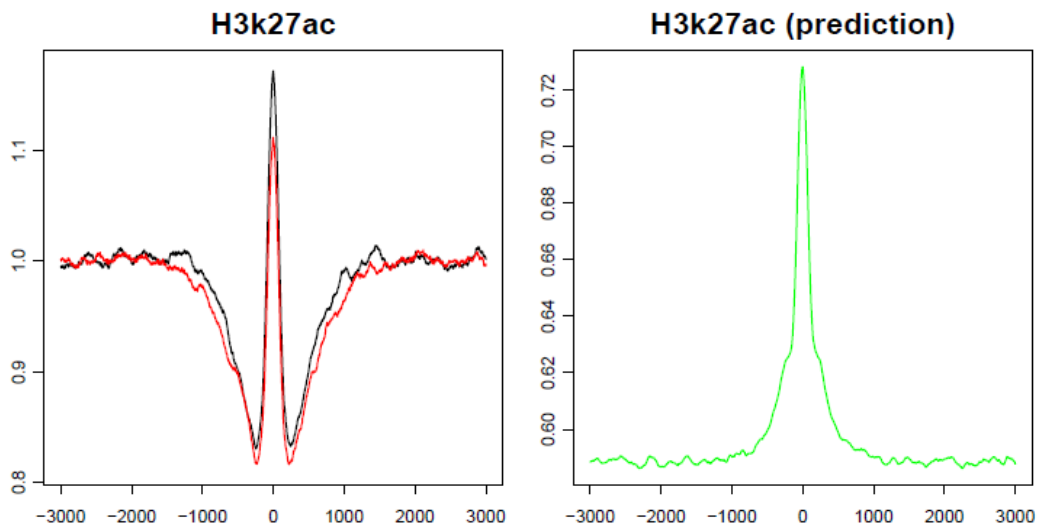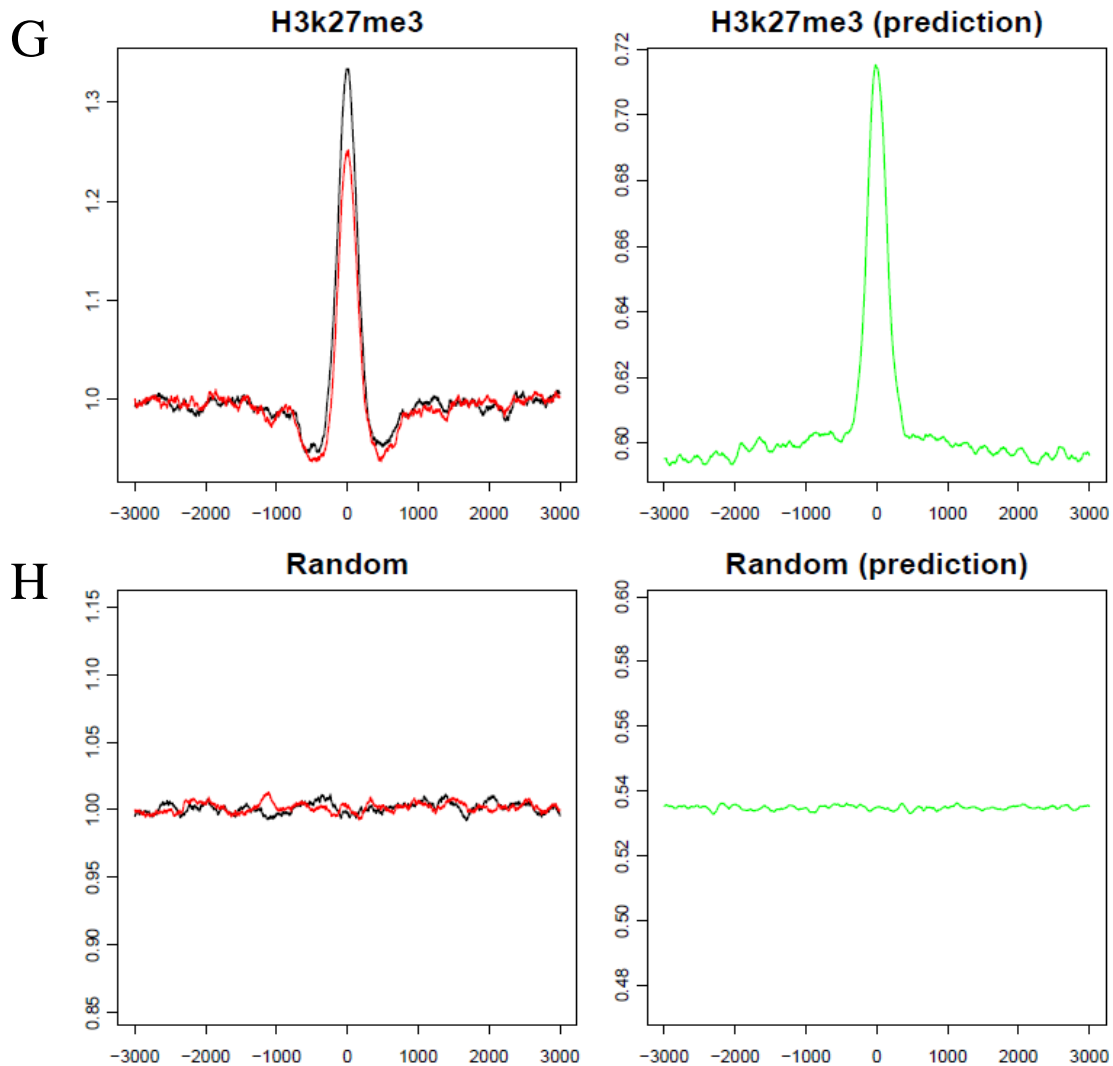
**Figure 1-6. Nucleosome positioning profiles at histone modification peaks.** Averaged nucleosome positioning profiles at histone modification peaks are shown. Black, control. Red, TAC. X-axis, distance to the center of histone modification peaks(±3,000bp). Y-axis, averaged nucleosome score. Left panel, data experimentally obtained. Right panel, nucleosome occupancy prediction based on primary DNA sequences. Nucleosome profiles at randomly chosen peaks are shown in H.
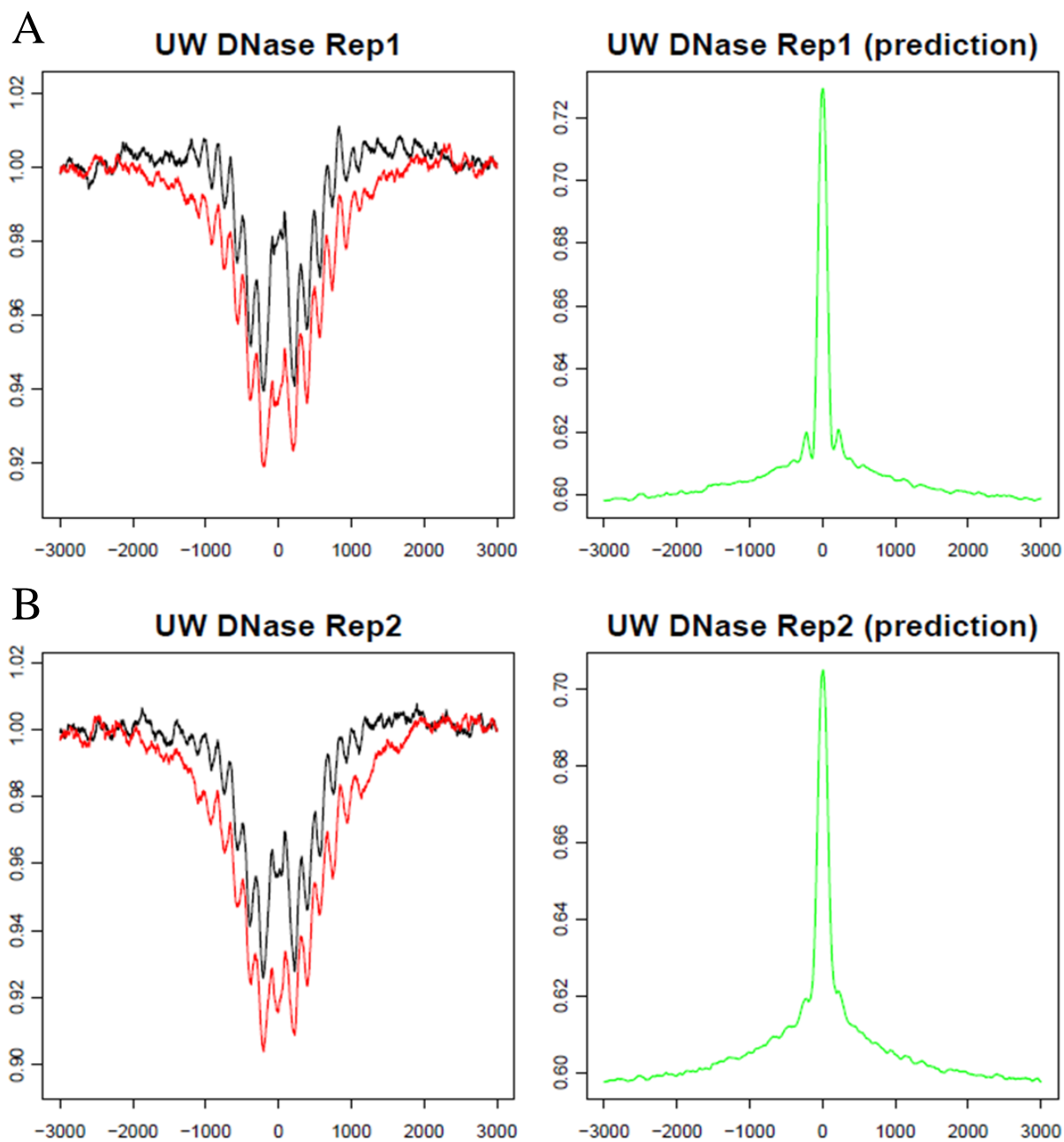
**Figure 1-7. Nucleosome positioning profiles at DNaseI HS sites.** Averaged nucleosome positioning profiles at DNaseI HS are shown. Profiles are plotted for the two replicates of DNaseI HS datasets. Black, control. Red, TAC. X-axis, distance to DNaseI HS site (±3,000bp). Y-axis, averaged nucleosome score. Left panel, data experimentally obtained. Right panel, nucleosome occupancy prediction based on primary DNA sequences
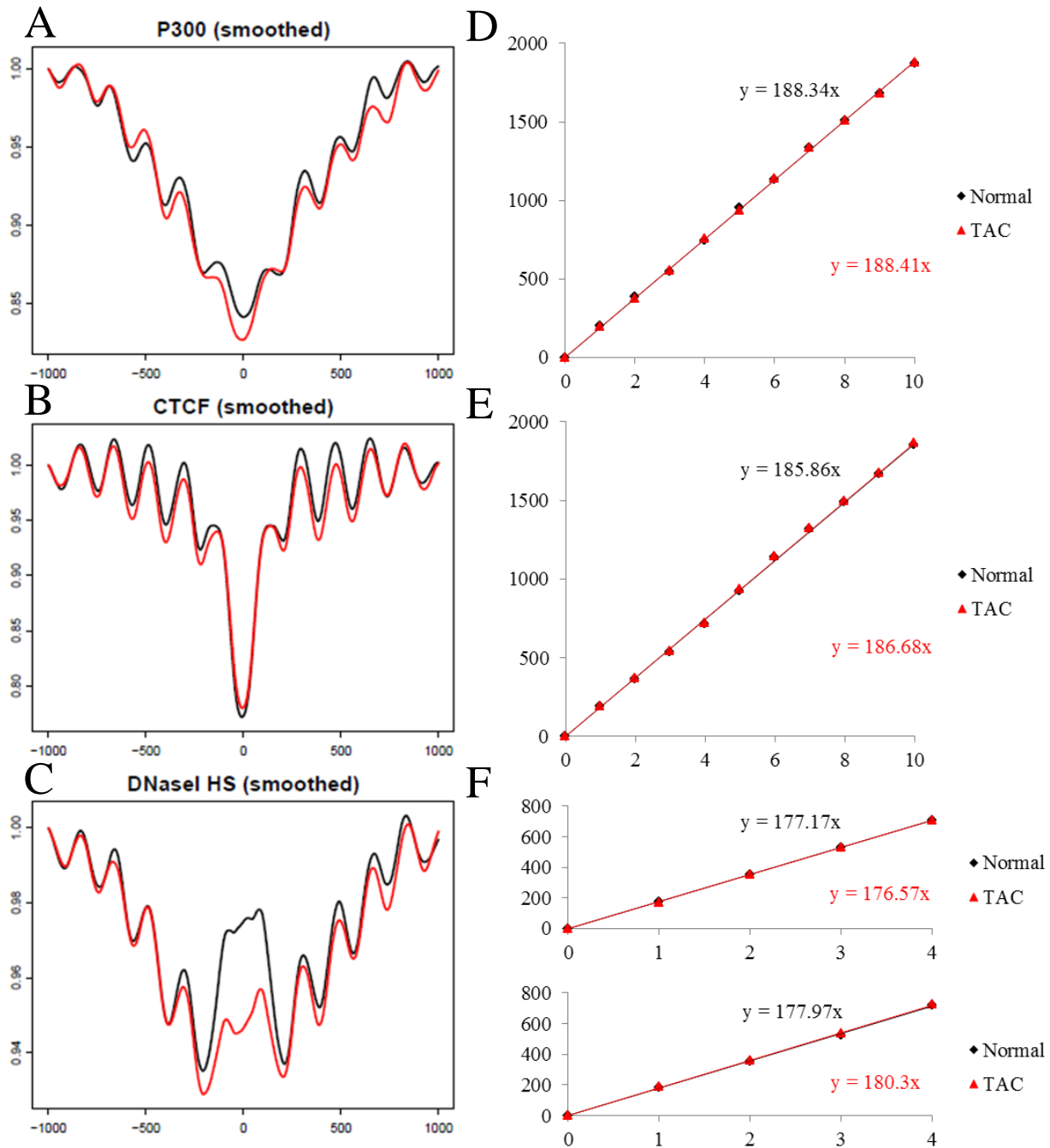
**Figure 1-8. Nucleosome repeat length at enhancer (A, D), insulator (B, E) and DNaseI HS (C, F) regions.** (A-C) Smoothed nucleosome positioning profiles at these three regions are shown. Black, control. Red, TAC. X-axis, distance to the center of these elements ($\pm$1,000 bp). Y-axis, averaged nucleosome score. (D-F) Linear regression fit of peak positions (y-axis) versus the nucleosome number (x-axis). The slope of trendline equals to the nucleosome repeat length. Due to the broad peak at the center of DNaseI HS, nucleosomes from the left or right side are calculated separately.
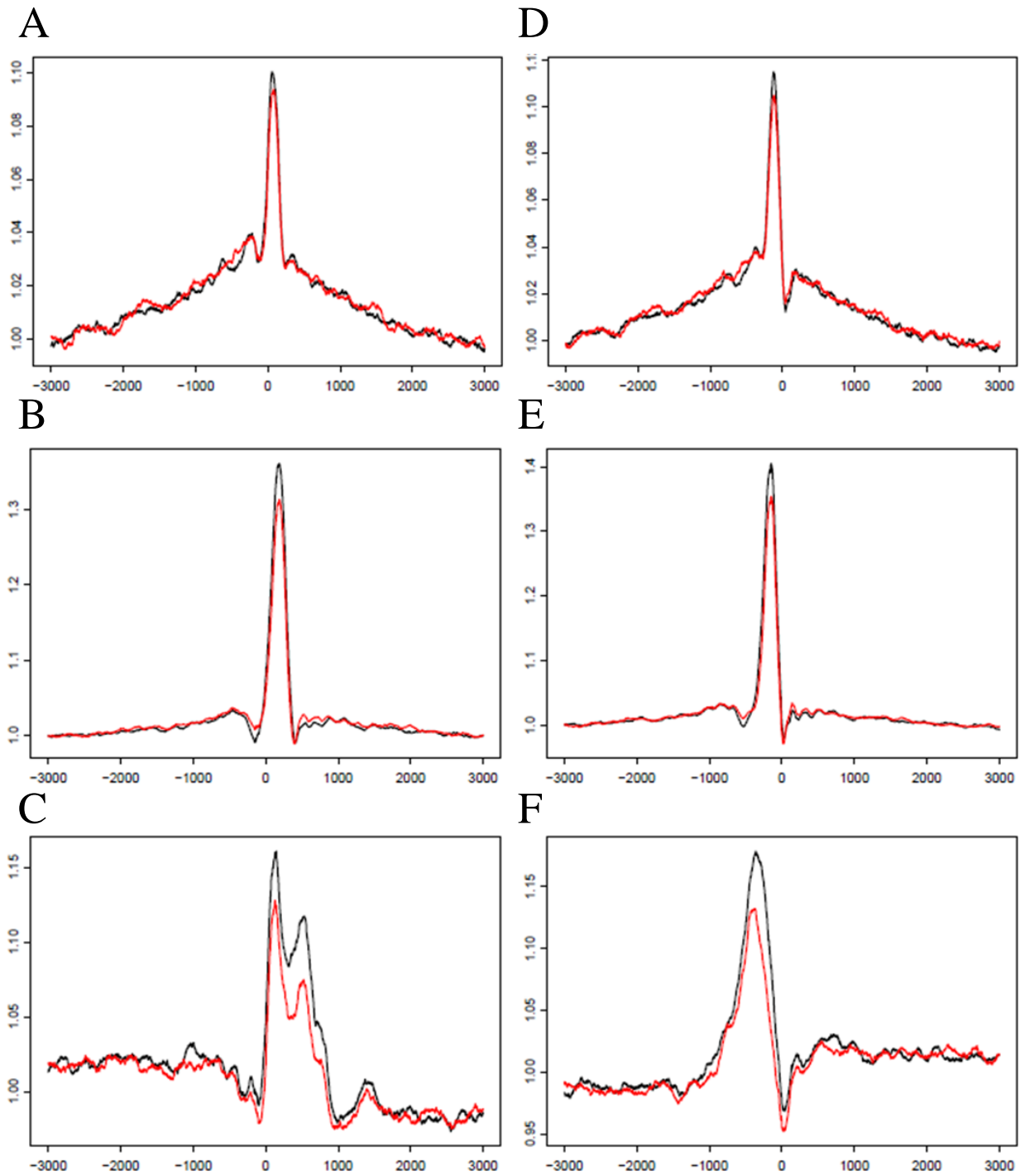
**Figure 1-9. Nucleosome positioning profiles at the boundaries of LTR.** Both 5' ends (A-C) and 3' ends (D-F) are shown. LTRs are classified based on the length. (A, D) 50~250bp. (B, E) 250~450bp. (C, F) > 450bp. Black, control. Red, TAC. X-axis, distance to the boundaries of LTR ($\pm$3,000 bp). Y-axis, averaged nucleosome score.

**Table 1-1: Summary of nucleosome positioning comparison between MNase-Seq and predicted results**

| Marks | MNase-Seq | Prediction | Comparison |
|---|---|---|---|
| GATA4 | Inverted 'V' shape, split peak | Inverted 'V' shape, split peak | Consistent |
| NKX2.5 | Inverted 'V' shape, split peak | Inverted 'V' shape, split peak | Consistent |
| SRF | 'W' shape, split peak | Inverted 'V' shape, split peak | Partial consistent |
| TBX3 | 'V' shape, single peak | Inverted 'V' shape, single peak | Inconsistent |
| RNA Pol II | 'V' shape, split peak | Inverted 'V' shape, split peak | Inconsistent |
| P300 | 'V' shape, periodic peaks | Inverted 'V' shape, single peak | Inconsistent |
| CTCF | Sharp 'V' shape, periodic peaks | Inverted 'V' shape | Inconsistent |
| CpG island | 'V' shape | Inverted 'V' shape | Inconsistent |
| H3K4me1 | Inverted 'V' shape | Inverted 'V' shape | Consistent |
| H3K4me3 | 'W' shape | Inverted 'V' shape | Inconsistent |
| H3K36me3 | Inverted 'V' shape | Inverted 'V' shape | Consistent |
| H3K79me2 | Inverted 'V' shape | Inverted 'V' shape | Consistent |
| H3K9ac | 'V' shape | Inverted 'V' shape | Inconsistent |
| H3K27ac | 'W' shape | Inverted 'V' shape | Partial consistent |
| H3K27me3 | Inverted 'V' shape | Inverted 'V' shape | Consistent |
| DNaseI HS | 'W' shape, central split peak, peripheral periodic peaks | Inverted 'V' shape, single peak | Inconsistent |

# Reference

1.  Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–98 (2008).

2.  Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS biology* **6,** e65 (2008).

3.  Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nature reviews. Genetics* **10,** 161–72 (2009).

4.  Mavrich, T. N. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453,** 358–62 (2008).

5.  Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics* **39,** 1235–44 (2007).

6.  Chen, H. B., Wang, L. & Jiang, J. F. Re-analysis of expression profiles for revealing new potential candidate genes of heart failure. *European review for medical and pharmacological sciences* **17,** 903–11 (2013).

7.  Hwang, J.-J. *et al.* Microarray gene expression profiles in dilated and hypertrophic cardiomyopathic end-stage heart failure. *Physiological genomics* **10,** 31–44 (2002).

8.  Lee, J.-H. *et al.* Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circulation research* **109,** 1332–41 (2011).

9.  Moreno-Moral, A., Mancini, M., D'Amati, G., Camici, P. & Petretto, E. Transcriptional Network Analysis for the Regulation of Left Ventricular Hypertrophy and Microvascular Remodeling. *Journal of cardiovascular translational research* (2013). doi:10.1007/s12265-013-9504-x

10. Tolstorukov, M. Y. *et al.* Swi/Snf chromatin remodeling/tumor suppressor complex establishes nucleosome occupancy at target promoters. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 10165–70 (2013).

11. Portela, A. *et al.* DNA methylation determines nucleosome occupancy in the 5'-CpG islands of tumor suppressor genes. *Oncogene* (2013). doi:10.1038/onc.2013.162

12. Van Bakel, H. *et al.* A compendium of nucleosome and transcript profiles reveals determinants of chromatin architecture and transcription. *PLoS genetics* **9,** e1003479 (2013).

13. Mitchell-Jordan, S. A. *et al.* Loss of Bmx nonreceptor tyrosine kinase prevents pressure overload-induced cardiac hypertrophy. *Circulation research* **103,** 1359–62 (2008).

14. Takimoto, E. *et al.* Chronic inhibition of cyclic GMP phosphodiesterase 5A prevents and reverses cardiac hypertrophy. *Nature medicine* **11,** 214–22 (2005).

15. Franklin, S. *et al.* Quantitative analysis of the chromatin proteome in disease reveals remodeling principles and identifies high mobility group protein B2 as a regulator of hypertrophic growth. *Molecular & cellular proteomics* **11,** M111.014258 (2012).

16. O'Connell, T. D., Ni, Y. G., Lin, K.-M., Han, H. & Yan, Z. Isolation and culture of adult mouse cardiac myocytes for signaling studies. *AfCS Research Reports* **1,** (2003).

17. O'Connell, T. D., Rodrigo, M. C. & Simpson, P. C. Isolation and culture of adult mouse cardiac myocytes. *Methods in molecular biology* **357,** 271–96 (2007).

18. Franklin, S. *et al.* Specialized compartments of cardiac nuclei exhibit distinct proteomic anatomy. *Molecular & cellular proteomics* **10,** M110.000703 (2011).

19. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–6 (2009).

20. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development* **9,** 657–63 (1999).

21. ENCODE Project. at <http://encodeproject.org/ENCODE/>

22. Van den Boogaard, M. *et al.* Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer. *The Journal of clinical investigation* **122,** 2519–30 (2012).

23. He, A., Kong, S. W., Ma, Q. & Pu, W. T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 5632–7 (2011).

24.   Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic acids research* **39,** D1005–10 (2011).

25.   Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–8 (2011).

26.   Teif, V. B. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology* **19,** 1185–92 (2012).

27.   Tillo, D. & Hughes, T. R. G+C content dominates intrinsic nucleosome occupancy. *BMC bioinformatics* **10,** 442 (2009).

28.   Bakker, M. L. *et al.* T-box transcription factor TBX3 reprogrammes mature cardiac myocytes into pacemaker-like cells. *Cardiovascular research* **94,** 439–49 (2012).

29.   Hoogaars, W. M. H. *et al.* The transcriptional repressor Tbx3 delineates the developing central conduction system of the heart. *Cardiovascular research* **62,** 489–99 (2004).

30.   Sabo, P. J. *et al.* Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 16837–42 (2004).

31.   Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research* **18,** 1073–83 (2008).

32.   Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics* **8,** 272–85 (2007).

33.   Takeuchi, J. K. *et al.* Chromatin remodelling complex dosage modulates transcription factor function in heart development. *Nature communications* **2,** 187 (2011).

34.   Hang, C. T. *et al.* Chromatin regulation by Brg1 underlies heart muscle development and disease. *Nature* **466,** 62–7 (2010).

35.   Tanaka, Y. & Nakai, K. An assessment of prediction algorithms for nucleosome positioning. *Genome informatics. International Conference on Genome Informatics* **23,** 169–78 (2009).

36.   Mitra, R. & Gupta, M. A continuous-index Bayesian hidden Markov model for prediction of nucleosome positioning in genomic DNA. *Biostatistics* **12,** 462–77 (2011).

37. Van der Heijden, T., van Vugt, J. J. F. A., Logie, C. & van Noort, J. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proceedings of the National Academy of Sciences of the United States of America* **109,** E2514–22 (2012).

38. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408,** 796–815 (2000).

39. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282,** 2012–8 (1998).

40. Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466,** 388–92 (2010).

41. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–20 (2011).

42. Fan, Y. *et al.* H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. *Molecular and cellular biology* **23,** 4559–72 (2003).

43. Bustin, M., Catez, F. & Lim, J.-H. The dynamics of histone H1 function in chromatin. *Molecular cell* **17,** 617–20 (2005).

44. Woodcock, C. L., Skoultchi, A. I. & Fan, Y. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome research* **14,** 17–25 (2006).

45. Hashimoto, H. *et al.* Histone H1 null vertebrate cells exhibit altered nucleosome architecture. *Nucleic acids research* **38,** 3533–45 (2010).

46. Perišić, O., Collepardo-Guevara, R. & Schlick, T. Modeling studies of chromatin fiber structure as a function of DNA linker length. *Journal of molecular biology* **403,** 777–802 (2010).

# Chapter 2: High-Mobility Group Protein B2 is Essential in Regulating Normal Cardiac Gene Expression and Participates in Cardiac Hypertrophy

## A. Abstract

Heart hypertrophy is a complex disease that involves differential expression of hundreds of genes. Unknown, however, are the highly coordinated chromatin remodeling events that must facilitate such genome-wide changes in gene expression. Chromatin architectural proteins such as the linker histone H1 and high-mobility group proteins have been shown to occupy specific regions in chromatin, but whether their location is static or dynamic is ucclear. To address this question, we examined the genome-wide distribution of HMGB2 in isolated cardiac myocytes with or without adrenergic receptor agonist stimulation and compare it with other cardiac related chromatin elements. We report a comprehensive map of HMGB2 binding in both normal and hypertrophic cardiac myocytes, and find that HMGB2 preferentially localizes to promoters, CpG islands, enhancers and transcription factor binding sites. Moreover, we find that upon hypertrophic stimulation, HMGB2 peaks move from regulatory elements to intergenic regions. These studies provide the first evidence for the fundamental mechanisms by which the chromatin structural protein HMGB2 facilitates global chromatin remodeling in heart hypertrophy.

## B. Introduction

One significant difference between eukaryotes and prokaryotes is that in eukaryotes, DNA is organized into chromosomes and is contained within a nucleus. The formation of chromosomes provides an important gene regulation mechanism that is also known as epigenetic regulation. A chromosome is composed of DNA, protein and RNA, the latter two being the

major components of epigenetic regulatory machinery. To maintain the structure of chromosome, both histone proteins and nonhistone proteins work in a coordinative way to form euchromatin or heterochromatin, which consists of actively transcribed genes or inactive genes, respectively.

High-Mobility Group Protein B2 (HMGB2), also known as high mobility group box 2, or HMG2[1], is a highly abundant non-histone chromosomal structural protein expressed in many eukaryotes[2]. Despite the long history of study since its discovery in 1970s[3,4], its function is still far from clear. Before the development of high-throughput sequencing technique, studies on HMGB2 mainly focused on its interaction with a few genomic loci. These studies demonstrated the roles of HMGB2 in nucleosome remodeling[5], DNA double-strand break repair[6], and enhancing transcription factor binding to DNA[7]. Based on the crystal structure studies of HMG box-DNA complex[8], it is believed that HMGB proteins bind to the minor groove of DNA and bend DNA towards the major groove, stabilizing distorted DNA structure[9–12]. To further reveal the molecular role of HMGB2 under physiological conditions, a genome-wide binding study is necessary. Genome-wide distribution of another high-mobility group protein, named HMGN1, showed that it is not randomly distributed throughout the genome, but preferentially localizes to functional elements[13]. It remains unknown whether HMGB2 has a similar behavior.

Using quantitative mass spectrometry, we previously found that the protein abundance of chromatin-associated HMGB2 is up-regulated in heart failure, and demonstrated that in isolated cardiac myocytes, interfering RNA-mediated HMGB2 knock down, but not HMGB1, can induce hypertrophic growth[14]. It has been known that chromatin remodeling is involved in cardiovascular development and disease[15,16]. Several histone modification enzymes and chromatin remodelers have been identified to explain the abnormal gene expression during heart

failure[17,18]. These molecules, however, are insufficient to explain the global changes of chromatin structure, which lead to pathological gene expression in disease. Using high-throughput techniques, hundreds to thousands of genes have been shown to be dynamically regulated in different stages of heart failure development[19–23]. The disease-associated gene expression pattern cannot be explained by single transcription factor or signaling pathway, and genes with altered expression in disease are located across the whole genome without a clear chromosomal pattern. This indicates that genome-wide regulatory machinery is needed to regulate pathological gene expression. To test the hypothesis that chromosomal structural proteins are dynamically altered across the genome to coordinate gene expression during heart failure, and to address the general principle for how the entire genome is remodeled as a result of chromosomal architectural protein modulation, we used ChIP-Seq to identify the genome wide localization of HMGB2 in normal and hypertrophic cardiac myocytes. Our data reveal another layer of information on how chromatin is globally remodeled and provide important new insights into the molecular mechanism by which HMGB2 coordinates the overall protein-DNA interactions.

## C. Materials and Methods

### *Neonatal rat ventricular myocytes*

Neonatal rat ventricular myocytes (NRVMs)[24] were obtained by enzymatic dissociation from 1-day-old litters and plated in Dulbecco modified Eagle medium (Invitrogen, #11965) containing 1% penicillin, 1% streptomycin, 1% insulin-transferrin-sodium selenite supplement and 10% fetal bovine serum for the first 24 h after which the cells are cultured in serum-free

media. To induce hypertrophy, cells were treated with phenylephrine (PHE) (10µM dissolved in phosphate-buffered saline) for 48 hours.

### *Chromatin immunoprecipitation and sequencing*

Cells were cross-linked by adding formaldehyde to the culture media to a final concentration of 1% and were incubated at room temperature for 10 minutes, followed by adding glycine to a final concentration of 125 mM to the media to quench the cross-linking. Cells were scraped and pelleted in cold PBS, and resuspended in lysis buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM EDTA pH 8, 1% Triton X-100, 0.1% Sodium Deoxycholate, 0.1% SDS, supplied with protease inhibitor cocktail tablet from Roche). Chromatin was sonicated into fragment size of 500bp, and diluted in RIPA buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM EDTA, 1% NP-40, 0.5% Sodium Deoxycholate, 0.1% SDS, supplied with protease inhibitor cocktail tablet). DNA-HMGB2 complexes from chromatin preparation were purified by incubating anti-HMGB2 antibody (ChIP Grade, ab67282, Abcam) at 4°C overnight, and ligating to protein A conjugated magnetic beads (Millipore, LSKMAGA10). After washing beads two times in Wash Buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA pH8, 150 mM NaCl, 20 mM Tris-HCl pH 8) and one time in high salt Wash Buffer (500 mM NaCl in Wash Buffer), DNA was eluted from beads by incubating beads in Elution Buffer (1% SDS, 100 mM NaHCO$_3$) and rotating for 15 min, 30°C. DNA fragments were purified using phenol:chloroform extraction and ethanol precipitation. Sequencing adapters were ligated using the Illumina Paired-End sample prep kit, and sequenced on Illumina Genome Analyzer IIx in the paired-end sequencing mode.

To confirm HMGB2 binding peaks identified from ChIP-Seq data, the immunoprecipitated DNA was amplified on the MyiQ$^{TM}$ Real-Time PCR Detection System

(Bio-Rad). Primer sets used for ChIP-PCR are shown in Table 2-1. Results are reported as fold change normalized to IgG IP control.

### *Bioinformatics*

Paired-end 100-bp Illumina reads were aligned to rat reference genome (rn4) using Bowtie[25] (0.12.7), allowing up to two mismatches in the seed region (the first 28nt of one read) of alignment. HMGB2 binding peaks were called using MACS 1.4.1[26], with a p-value cutoff at $10^{-5}$ . Genome annotation tracks, including rat gene structure and CpG island positions, were downloaded from the UCSC Genome Browser. Promoter was defined as the region within 2kb upstream of transcription start site.

### *DNA bendability scores*

DNA bendability score was calculated using the trinucleotide parameters as proposed by Brukner et al[27]. Rat genome was divided into 10bp bins. For each bin with the DNA sequence of $N_1 N_2 \cdots N_9 N_{10}$, the bending propensity was calculated as

$$S_{bin} = \exp\left(\sum_{i=1}^{8} p_{i:i+2}\right)$$

where $p_{i:i+2}$ represents the trinucleotide DNA bendability parameter for $N_i N_{i+1} N_{i+2}$. According to Brukner et al, a high score indicates bending towards the major groove. Bins with unknown nucleic acid residue were disregarded.

### *Transcription factor binding sites*

The ENCODE project[28] and many other high-throughput ChIP-Seq studies[29] provide a high quality source of protein-DNA interaction datasets. However, most of these studies were

conducted in either human or mouse tissue samples; to date, these datasets also do not include HMGB2. To compare our rat data with other protein-DNA interaction data, we utilized the following method to overcome this cross-species problem. Binding peak coordinates for mouse heart p300[28], CTCF (from human CD4+ T cells[30], mouse ES[31] and mouse heart tissue[28]), and other cardiac transcription factors[29] were converted into rat genome (version rn4) using LiftOver from UCSC Genome Bioinformatics Site, with the parameter of –minMatch=0.85. Binding sites were then identified from converted DNA sequences using FIMO from the MEME Suite(version 4.7.0)[32].

### *Microarray datasets*

Microarray data were obtained from two previously published studies[14,33]. Expression data before and after HMGB2 knock down were obtained from our previously published report[14]. Total RNA from either HMGB2 knock down or control NRVMs was analyzed using Illumina RatRef-12 Expression BeadChip. Expression data before and after PHE treatment were downloaded from Gene Expression Omnibus (Series GSE925)[33], which was conducted on Affymetrix Rat Expression 230A Array. Probe annotation information was extracted from Ensembl database (version 69) using Biomart[34]. Probes mapped to multiple genes were removed.

## D. Results

### *Genome wide distribution of HMGB2 identified by ChIP-Seq and validated by ChIP-PCR.*

Our HMGB2 ChIP-Seq data showed a genome-wide distribution of HMGB2 in isolated myocytes, without preferential localization to any given chromosome or chromosome-scale feagure (e.g. centromeres, telomeres) (Fig 2-1). Although it is believed that HMGB proteins bind

to DNA in a sequence-independent manner, this distribution in cardiac myocytes is neither uniform nor random. Peaks and valleys of HMGB2 ChIP-Seq intensity are clearly visible when plotting for each chromosome (Fig 2-1), indicating the preference of HMGB2 binding to certain regions. To confirm the ChIP-Seq signals, we chose 19 HMGB2 binding peaks that are located in promoters of genes whose expression is altered after HMGB2 knock down, and validated the binding events using anti-HMGB2 ChIP-qPCR for 18 of them.

***Genomic annotation of HMGB2 binding peak reveals its enrichment in promoters and CpG islands.***

Next, we examined the binding of HMGB2 to promoter, exon, intron and intergenic regions (Fig 2-3, A-C). HMGB2 regulates gene expression by directly interacting with nucleosomes or with other transcription factors[35,36]. This activity requires the binding of HMGB2 to functional DNA elements. We found that promoter and exonic regions have a higher preference for HMGB2 binding compared with intergenic regions. In particular, using the whole genome annotation as the background (Fig 2-3, A), HMGB2 binding peaks have an 8.8-fold enrichment in promoters and a 4.6-fold enrichment in exons (Fig 2-3, B). These data decrease to 5.2-fold and 3.6-fold, respectively (Fig 2-3, C), in PHE-induced hypertrophic cardiac myocytes.

Given that many promoters are associated with CpG islands, the enrichment of HMGB2 at promoters indicates that CpG islands may be another HMGB2-enriched regulatory regions. After comparing HMGB2 binding peaks with CpG island annotation, we found a 16-fold enrichment at CpG islands (Fig 2-3, E), confirming our prediction. Consistent with the observation of PHE-induced decrease at promoters, we observed a decrease of HMGB2 binding peaks in CpG islands, to 10.9-fold after PHE treatment (Fig 2-3, F).

***CpG island associated promoters are targets of HMGB2.***

Given our previous study suggesting an enrichment of HMGB2 at CpG island and promoter regions, we asked the question what percent of CpG island associated promoters are targets of HMGB2. To answer this question, we first counted the number of genes in rat genome as well as genes with CpG island-associated promoters, and found that 49% of rat genes have CpG islands in their promoters (Fig 2-4, A). This ratio, however, increased dramatically to 89%, when only the genes bound by HMGB2 at their promoters were considered (Fig 2-4, A.25%/28% ≈ 89%). In another word, this means that 89% HMGB2 targeted genes have CpG island promoters. Meanwhile, we did a similar calculation for other TFs and found that some TFs prefer CpG island-associated promoters while others do not (Supplementary Fig 2-1). We also show that almost half of the genes with CpG island associated promoters are targets of HMGB2 (Fig 2-4 A, right, 25%/49% ≈ 50%).

Previous studies have indicated that CpG island associated promoters behave differently from non-CpG island associated promoters[37,38]. We then asked the question whether HMGB2 targeted promoters have a different behavior. To assess the relationship between gene expression level and HMGB2 binding, we plotted the distribution of Illumina microarray probe signals measured in unstimulated NRVMs (Fig 2-4, B). Interestingly, we found that genes with HMGB2 bound promoters have a high expression level comparing with average expression level for all genes examined in that microarray experiment. These genes also exhibit wider expression variance. Since HMGB2 most likely targets CpG island promoters, this observation may be the effect of CpG islands located at promoters. To our surprise, we found that genes with CpG island promoters, but not HMGB2 binding, show no difference with the average of the whole genome. It seems that genes whose promoters are both HMGB2 targeted and CpG island associated, but

not those promoters with either feature alone, have a higher expression level than the overall genome.

### *HMGB2 knock down or PHE treatment induces hypertrophic growth through different molecular mechanisms.*

Our previous study has shown that loss of HMGB2 in isolated cardiac myocytes can induce hypertrophic growth[14]. To examine the role of HMGB2 in gene regulation, we conducted siRNA knock down to remove HMGB2 in isolated cardiac myocytes, and measured the changes of transcriptome using expression microarray. Meanwhile, we compared our HMGB2 knock down data with the published PHE-treated microarray data. Surprisingly, we found that these two types of hypertrophy remodel the myocyte transcriptome in distinct patterns (Fig 2-5, A,D&F). To tease out the effect of HMGB2 on gene regulation, we plotted the genes that are not only regulated by HMGB2 knock down or PHE treatment, but also bound by HMGB2 at their promoters (Fig 2-5, B,C&E). It is clear that the remove of HMGB2 from promoters, either through PHE treatment (Fig 2-5, B) or siRNA interference (Fig 2-5, E), results in both up-regulation and down-regulation of gene expression. Therefore, unlike transcription factors, which mainly facilitate gene expression, the effect of HMGB2 at gene promoters depends on the local chromatin environment.

### *HMGB2 binding profiles at chromosomal structural proteins and cardiac transcription factor binding sites.*

HMGB family proteins have been known to interact with other chromatin-associated proteins, including linker histone H1[39], estrogen receptor[40], Sterol regulatory element-binding protein[41] and transcription factor Oct4[42] to facilitate their binding to DNA. To study the

prospective effects of HMGB2 in facilitating chromatin remodeling and cardiac transcription factor binding, we profiled the HMGB2 binding intensity at different regulatory protein binding sites.

The first regulatory element that we wanted to study is enhancer. In this study, we chose p300 as the marker for enhancers, which was also used by other groups to identify heart enhancers in both human[43] and mouse[44]. Because there is no published genome-wide p300 binding data from rat tissue, we converted the mouse binding data into rat genome based on the sequence homology. After normalizing to the average genome-wide read count, we found that there is a 3-fold enrichment of HMGB2 binding intensity at the p300 regions, and this enrichment is not altered by PHE treatment (Fig 2-6, A). To rule out the possibility that this enrichment is due to the enhanced binding of HMGB2 at promoters, we plotted HMGB2 binding profile at p300 binding peaks far away from any transcription start sites, and still found a 2.5-fold enrichment (Fig 2-6, B).

Next, we examined the pattern of HMGB2 at the binding site of the zinc finger protein CCCTC-binding factor (CTCF). CTCF is a conserved zinc finger protein known to be involved in transcriptional insulation, X chromosome inactivation and gene activation/repression[45]. Again, we converted mouse and human CTCF binding data to rat genome before comparison. To rule out the possible tissue-specific or species-specific bias, we studied three datasets from human CD4+ T cells (Fig 2-6, C), mouse heart tissue (Fig 2-6, D) and mouse embryonic stem cells (Fig 2-6, E). Interestingly, we observed a similar colocalization pattern of HMGB2 among these three datasets, and this pattern remained the same after PHE treatment.

To further test the role of HMGB2 in regulating cardiac gene expression, we examined four cardiac transcription factors, including Myocyte-specific enhancer factor 2A (MEF2A) (Fig

2-7, A), NK2 transcription factor-related, locus 5 (NKX2-5) (Fig 2-7, B), GATA-binding protein 4 (GATA4) (Fig 2-7, C) and serum response factor (SRF) (Fig 2-7, D). It was reported that the co-occupancy of these transcription factors can identify a distinct set of enhancers and consequently direct cardiac gene expression in a collaborative way[29]. We observed 2.5-fold, 1.6-fold, and 2-fold enrichment of HMGB2 at MEF2A, NKX2.5 and GATA4 binding sites, respectively, and a moderate decrease of HMGB2 binding at the SRF binding regions. Because our transcription factor binding data were converted from the mouse genome, it is possible that this observation was caused by conservation between mouse and rat. To rule out this possibility, we randomly chose similar amount of peaks from mouse genome and converted them into rat, and observed no enrichment of HMGB2. This demonstrated that the binding patterns in Figure 2-6 and 2-7 are not artifacts induced by cross-species conversion.

***DNA bending propensity facilitates HMGB2 binding.***

If the binding position of HMGB2 affects global gene expression, then its location should be strictly regulated. However, HMGB2 has no known DNA binding consensus motif. Therefore, one question remaining unanswered is how HMGB2 finds the right place to bind. HMGB2 has been known to bend DNA by interacting with the minor grooves of DNA backbone[11,12]. To examine whether HMGB2 has a high affinity to DNA regions thermodynamically favoring bending, we calculated the bendability of DNA as measured by DNase I digestion[27], and compared the binding affinity of HMGB2 with DNA bendability within each 500Kbp bin (Fig 2-8, A&B). We found a positive correlation between HMGB2 binding and DNA bendability. This observation provides some clues as to how HMGB2 finds its place to bind.

To further assess the relationship between HMGB2 binding and DNA bending propensity, we generate the averaged DNA bendability score profiles for HMGB2 binding peaks. Comparing with the randomly chosen peaks, an increase of DNA bendability at HMGB2 binding peaks is clearly observed (Fig 2-8 C, blue line). In comparison, after PHE treatment, HMGB2 molecules move to regions with less flexible regions, as indicated by the lower DNA bendability score (Fig 2-8 C, red line). Meanwhile, we calculated the DNA bendability score for mouse cardiac TFs and CTCF and showed that CTCF and GATA4 favor more bendable regions while NKX2.5 and MEF2A favor less bendable regions (Supplementary Fig 2-2).

## E. Discussion

In this chapter, we asked the question what is the role of HMGB2 to maintain normal cardiac gene expression and how is this changed during cardiac hypertrophy. To answer this question, we started with the HMGB2 ChIP-Seq experiment to measure genome-wide HMGB2 binding in normal and hypertrophic cardiac myocytes, and found that HMGB2 binding is enriched in functional elements including promoters and CpG islands. Because both HMGB2 knock-down and PHE treatment can induce hypertrophic growth, we decided to compare the alterations of gene expression between HMGB2 knock-down and PHE treatment using expression microarrays, and found that the hypertrophy-related gene expression pattern upon HMGB2 knock-down is different from the one after PHE treatment (Fig 2-5). This also indicates the existence of multiple pathways that can lead to hypertrophic growth of cardiac myocytes. Further study revealed that HMGB2 co-localizes with many other chromosomal protein binding sites including p300 and CTCF. Interestingly, we found that HMGB2 binding profiles are different at the binding sites of four cardiac transcription factors, MEF2A, NKX2.5, GATA4 and

SRF. More specifically, HMGB2 localization is higher at MEF2A binding sites, indicating that HMGB2 may be involved in facilitating the functioning of MEF2A. Finally, we calculated the DNA bendability score based on the primary sequence of the rat genome and correlated it with HMGB2 binding intensity. Consistent with the understanding that HMGB2 bends DNA toward DNA major groove, we found a positive correlation between HMGB2 binding and DNA bending propensity. Surprisingly, PHE treatment in NRVMs induced the relocation of HMGB2 to less bendable intergenic regions. Taken together, our data conclude that HMGB2 is required for maintaining the normal cardiac gene expression pattern, and its binding to cardiac chromatin is dynamically altered during hypertrophic growth.

**HMGB2 preferentially binds to CpG islands and CpG island promoters**. The promoter of a protein-coding gene is the regulatory region located upstream of transcription start site. Based on the frequency of CpG dinucleotide, promoters can be classified into CpG island associated promoter and non-CpG island promoter. Strikingly, we found that half of CpG island promoters are bound by HMGB2, while the overall genome has only 28% of promoters associated with HMGB2 binding peaks (Fig 2-4, A). In mammalian cells, CpG island associated promoters usually have a low level of DNA methylation. However, under pathological conditions CpG islands can be hypermethylated[46], which will disrupt the binding of chromatin binding proteins such as CTCF[47]. It is unclear whether DNA methylation prevents HMGB2 binding, and whether binding of HMGB2 to CpG islands affects DNA methylation, which can be tested by anti-HMGB2 ChIP followed by bisulfite sequencing, ChIP-chop, or confocal immunofluorescent co-staining microscopy.

*Biophysical characterization of HMGB2 reveals its role in chromatin remodeling.* Early X-ray crystallographic studies on DNA/HMG-domain complex displayed the interaction

between HMG box and widened DNA minor groove[8]. Later on, many groups studied the kinetics of HMGB2-DNA interaction using optical tweezers[48,49], showing that the binding of HMGB2 to dsDNA is able to generate sharp DNA bends with an angle of 70~100°. The binding of HMGB2 to the minor groove of DNA -- which contributes to its apparent lack of a strict binding consensus motif -- makes it an ideal molecule for global chromatin remodeling. We found that HMGB2 prefers binding DNA regions thermodynamically favoring bending toward major groove (Fig 2-8), confirming that the behavior of HMGB2 predicted from biophysical properties is manifest under physiological conditions. This observation also provides another theory on how HMGB2 localizes itself to the right chromatin position. Fluorescent microscopy examination has shown the nucleus location of HMGB2-EGFP in interphase cells[50], forming an ocean of HMGB2 molecules submerging DNA molecules. Based on the theory of dynamic chemical equilibrium, DNA regions that are either intrinsically favoring bending, or bended by protein-DNA interactions, will have a higher chance to become HMGB2 targets. HMGB2, together with other chromosomal structural proteins, maintains chromatin confirmation and regulates gene expression.

*HMGB2 is a versatile regulator of multiple DNA-protein interactions*. HMGB proteins orchestrate protein-DNA interaction through numerous ways including direct interacting with both DNA and transcription factors[51], binding to protein but not DNA[7], and/or facilitating nucleosome sliding[52]. HMGB2 binding profiles at different transcription factor binding sites (Fig 2-6 & 2-7) suggests that HMGB2 mainly facilitates the binding of chromatin structural proteins (p300 & CTCF) as well as some, but not all transcription factors (in the case of the myocyte, MEF2A but not SRF). This observation, however, should not be interpreted as demonstrating co-localization of HMGB2 with other proteins. In fact, ternary complexes with HMGB-protein-

DNA were not detected in most cases, with the exception of TBP[12]. Instead, it is likely that HMGB2 assists the binding of other proteins secondarily through its primary role in sculpting higher-order chromatin structure and enhancing the DNA structural flexibility, which opens the binding window for other proteins, without simultaneous direct interaction with both protein and DNA at a given locus (Fig 2-9, model 2). This is different from many transcription factors or co-factors which, when initiating gene transcription, form a protein complex and co-occupy the same site. Moreover, HMGB2 may compete with other proteins and repress gene expression. The bending toward major groove induced by HMGB2 may impede the binding of proteins that recognize DNA major groove. The high-order chromatin structure form by HMGB2 binding may also suppress DNA-protein binding. An *in vitro* study using different concentrations of HMGB2 found that the protein binds to DNA in a flexible hinge model at low concentration, but forms an irreversible filament at high concentration[48]. Whether both modes exist under physiological conditions is unknown, but this information indicates that HMGB2 may have distinct functions depending on its concentration within nucleus. This may also explain why increased amount of HMGB2 in pressure overload-induced cardiac hypertrophy[14] failed to protect cells from the hypertrophic growth. More than enough HMGB2 in nucleus may trigger malicious events that lead to abnormal cell growth, resulting in a different gene expression pattern as HMGB2 knock down (Fig 2-5).

**The pathophysiological role of HMGB2 in different diseases indicates its involvement in cell growth.** Many clinical oncology studies have shown that the level of HMGB2 is up-regulated in cancer cells, including hepatocellular carcinoma[53], squamous cell carcinoma[54], bladder carcinoma[55] and glioblastoma[56]. Tumor cells lacking HMGB2 are more sensitive to chemotherapy drugs[55–57], while overexpression of HMGB2 activates antiapoptotic pathways[53,54].

Moreover, studies using isolated cells from gene knockout mice showed that the lack of HMGB2 can result in impaired development of neural stem cells[58] and mesenchymal stem cell differentiation[59]. These studies indicate that HMGB2 promotes cell growth by up-regulating cell cycle-related genes and down-regulating apoptosis-related genes. A detailed study in a cell line showed that HMGB2 expression increases markedly after S phase and reaches maximum at G2 phase, and the progression of G1 to S phase is repressed by anti-HMGB2 antisense RNA[60]. It is very likely that HMGB2 serves as a checkpoint for cell proliferation, and its misregulation could result in a variety of diseases including cancer. Cardiovascular diseases are different from cancer. Nevertheless, up-regulation of some proto-oncogenes such as c-fox and c-myc was reported as part of the fetal gene reprogramming in adult heart failure[61]. It has also been proposed that inhibition of Myc is a potential therapeutic for hypertrophic cardiomyopathy[62]. Therefore, it will be interesting to investigate in preclinical models whether HMGB2 has a similar effect in hypertrophic cardiac myocytes as in tumor cells.
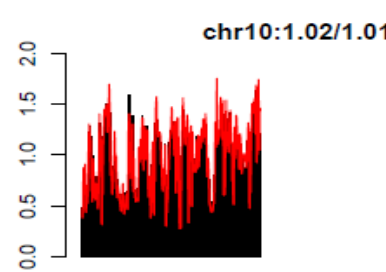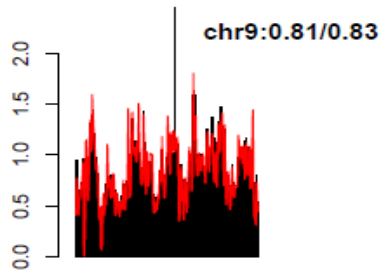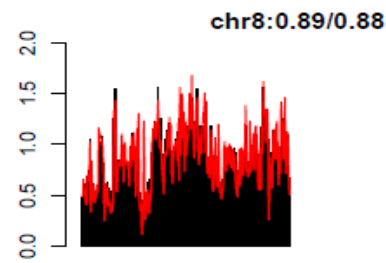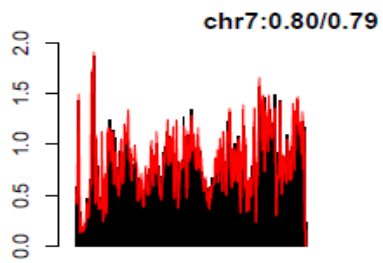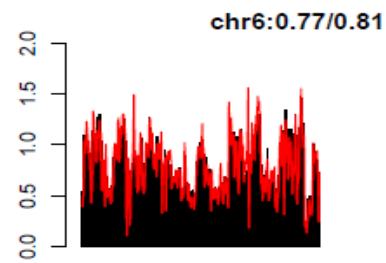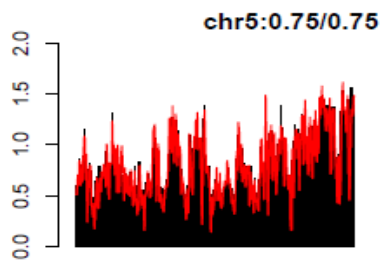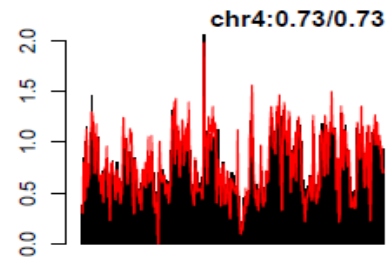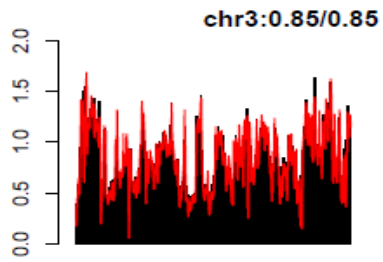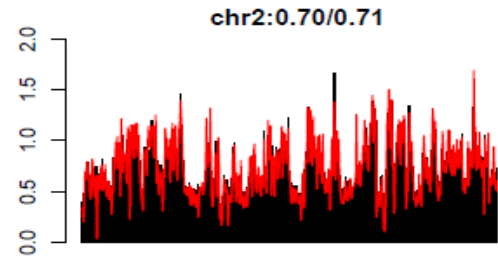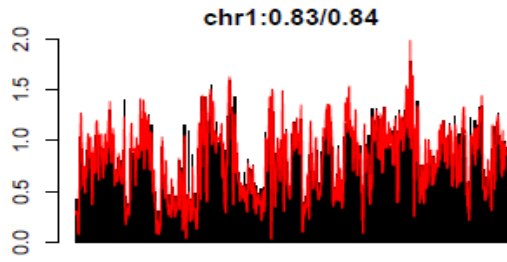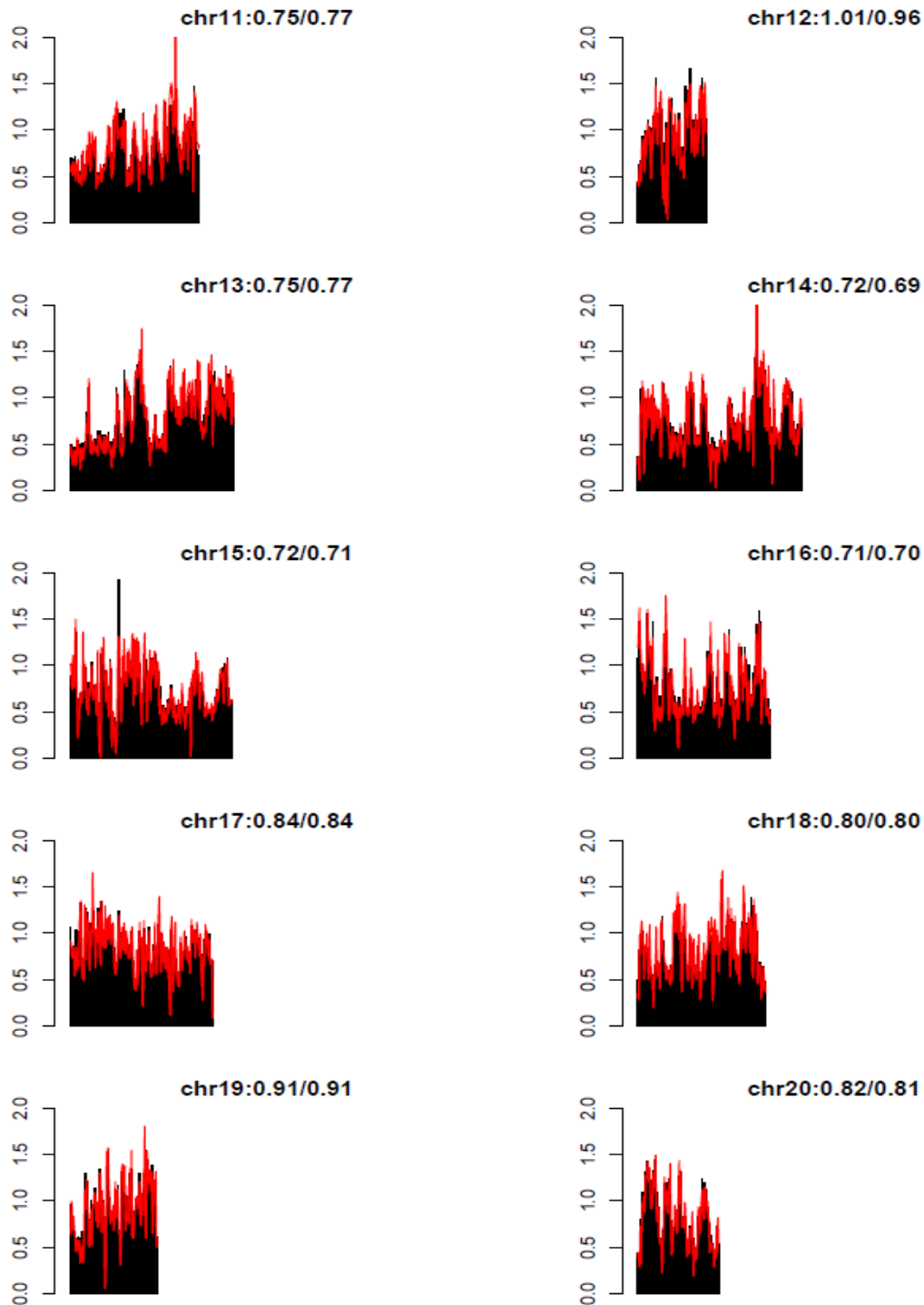
**Figure 2-1. Genome-wide distribution of HMGB2.** Rat genome is divided into 500Kbp bins. HMGB2 binding intensity, i.e. the mean number of HMGB2 ChIP-Seq reads, is plotted for both control NRVMs (black) and PHE treated NRVMs (red). Chromosome names are labeled on top of each panel, followed by the median intensity of control and PHE treated myocytes, respectively, for each chromosome. Chromosome X is removed.
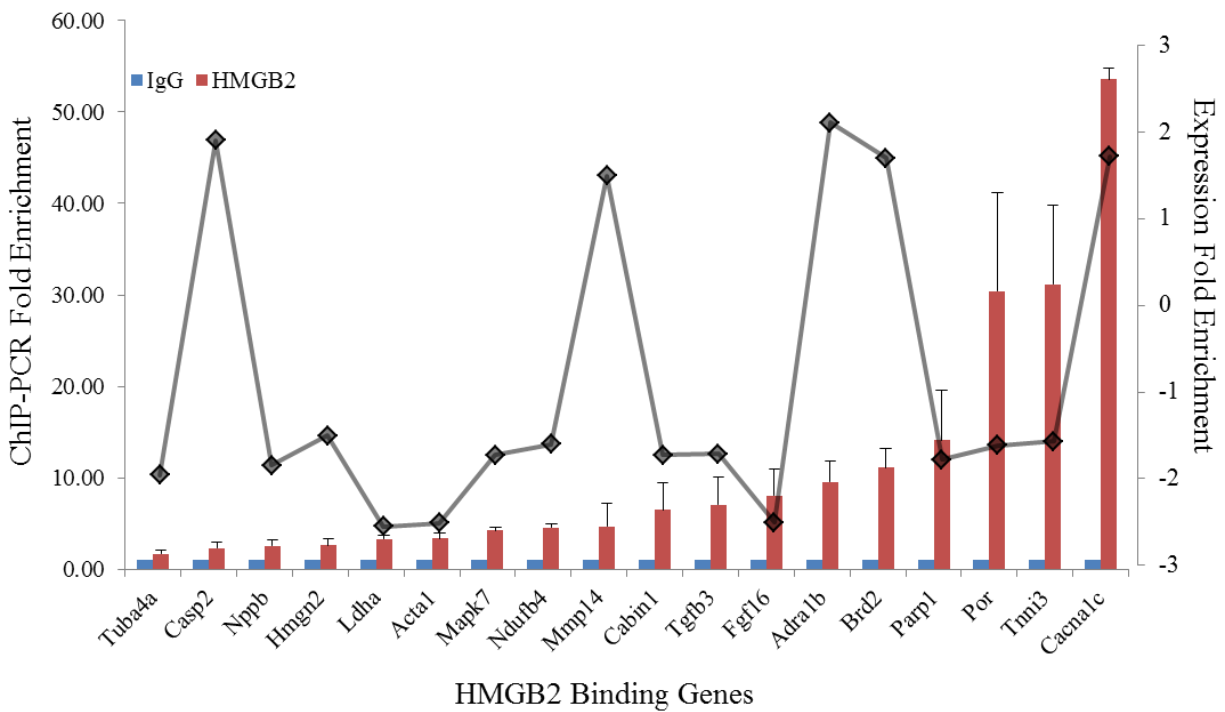
**Figure 2-2. HMGB2 binding distribution along the genome.** ChIP analysis showing binding to 18 HMGB2 positive sites identified by ChIP-seq. Relative enrichment is quantified by real-time PCR with the indicated primer sets. Data are normalized against the IgG negative enrichment. Errors bars represent s.d. of at least 3 independent experiments. (Dr. Manuel Rosa Garrido & Ricardo Gray) Gene expression fold changes, as measured by microarray after HMGB2 knock-down, are shown as the line-plot.
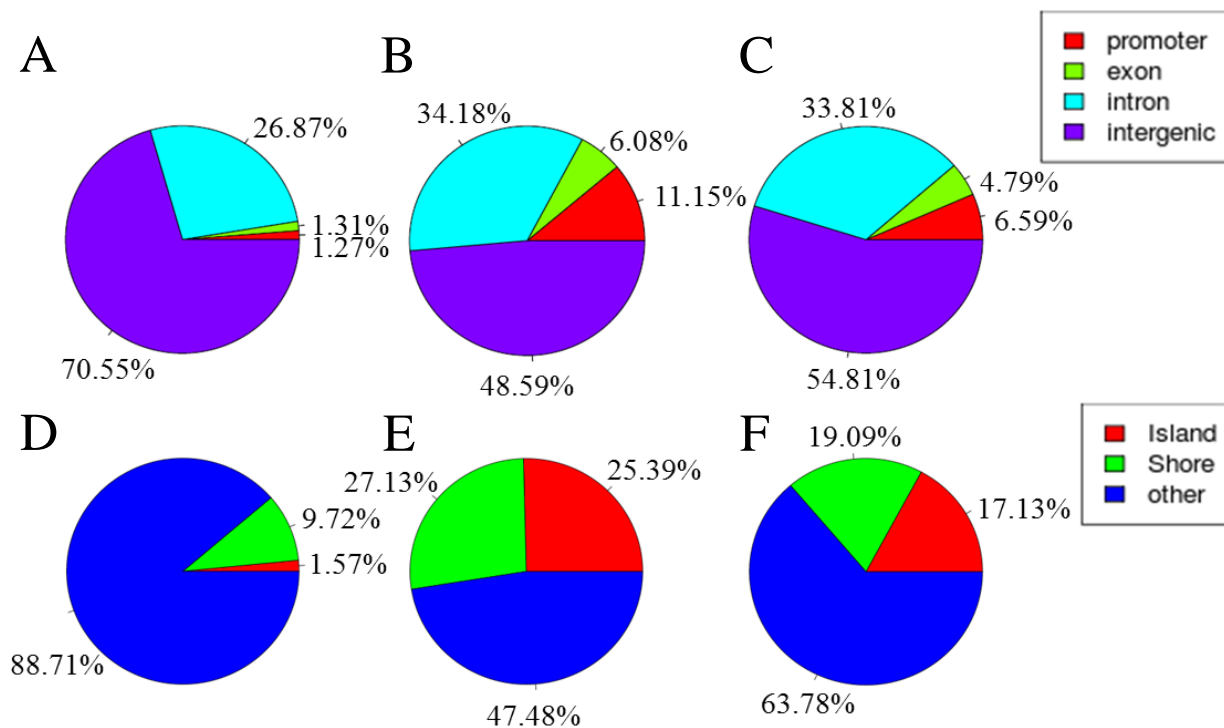
**Figure 2-3. Annotation of HMGB2 binding peaks.** HMGB2 binding peaks as identified by MACS are annotated based on its location on promoter, exon, intron or intergenic regions (A-C), or CpG island, island shore or far from CpG island (D-F). A & D, annotation of rat reference genome (rn4). B & E, annotation of HMGB2 binding peaks from control cardiac myocytes. C & F, annotation of HMGB2 binding peaks from ISO-treated cardiac myocytes. Promoter, regions within 2kb upstream of transcription start site. CpG island shore, regions within 2kb flanking CpG island.
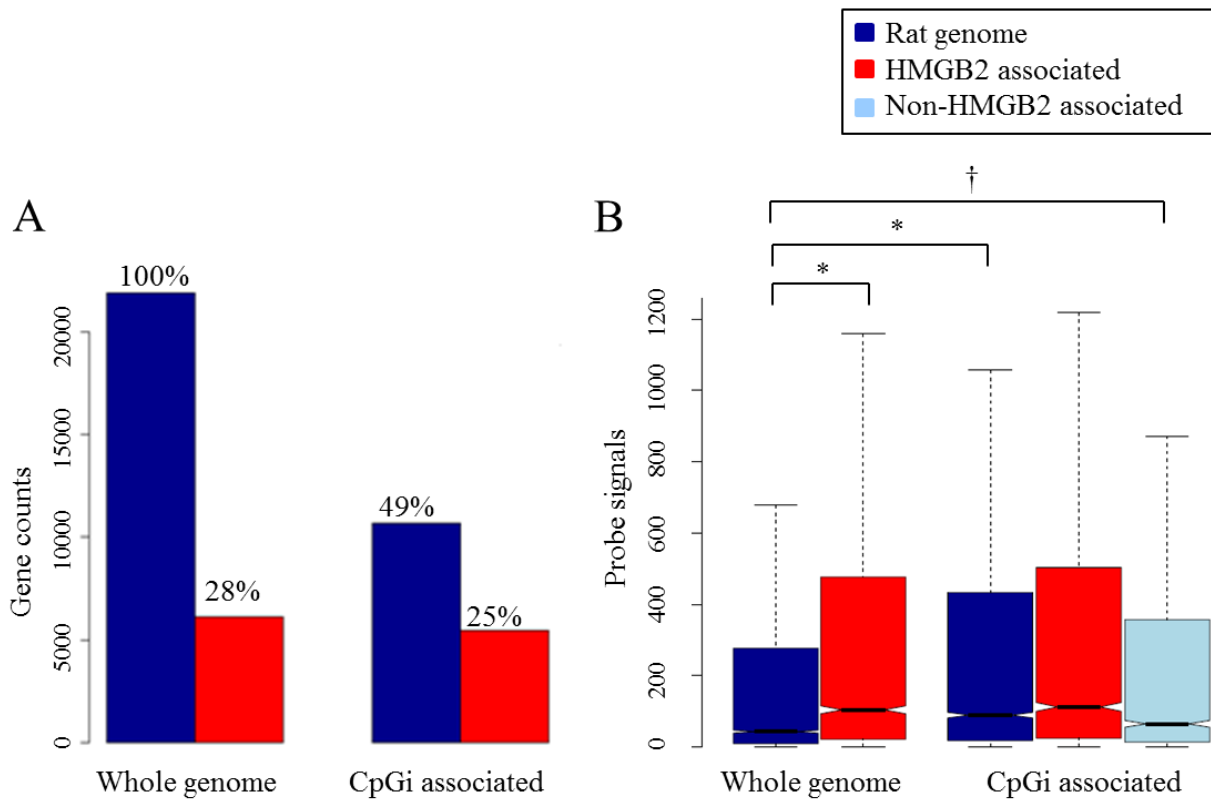
**Figure 2-4. Binding of HMGB2 to CpG island associated promoters.** (A) Number of genes bound by HMGB2 and/or overlapped with CpG island. Genes are separated based on whether their promoters overlap with CpG island. Left, the whole rat genome. Right, genes with CpG island associated promoters. Red, genes with HMGB2 binding in their promoters. (B) Box plot showing microarray probe signal distribution for each group of genes. Expression data are from NRVMs without PHE stimulation. Genes are clustered in the same way as panel A.*, $p < 10^{-5}$. †, $p = 0.23$.
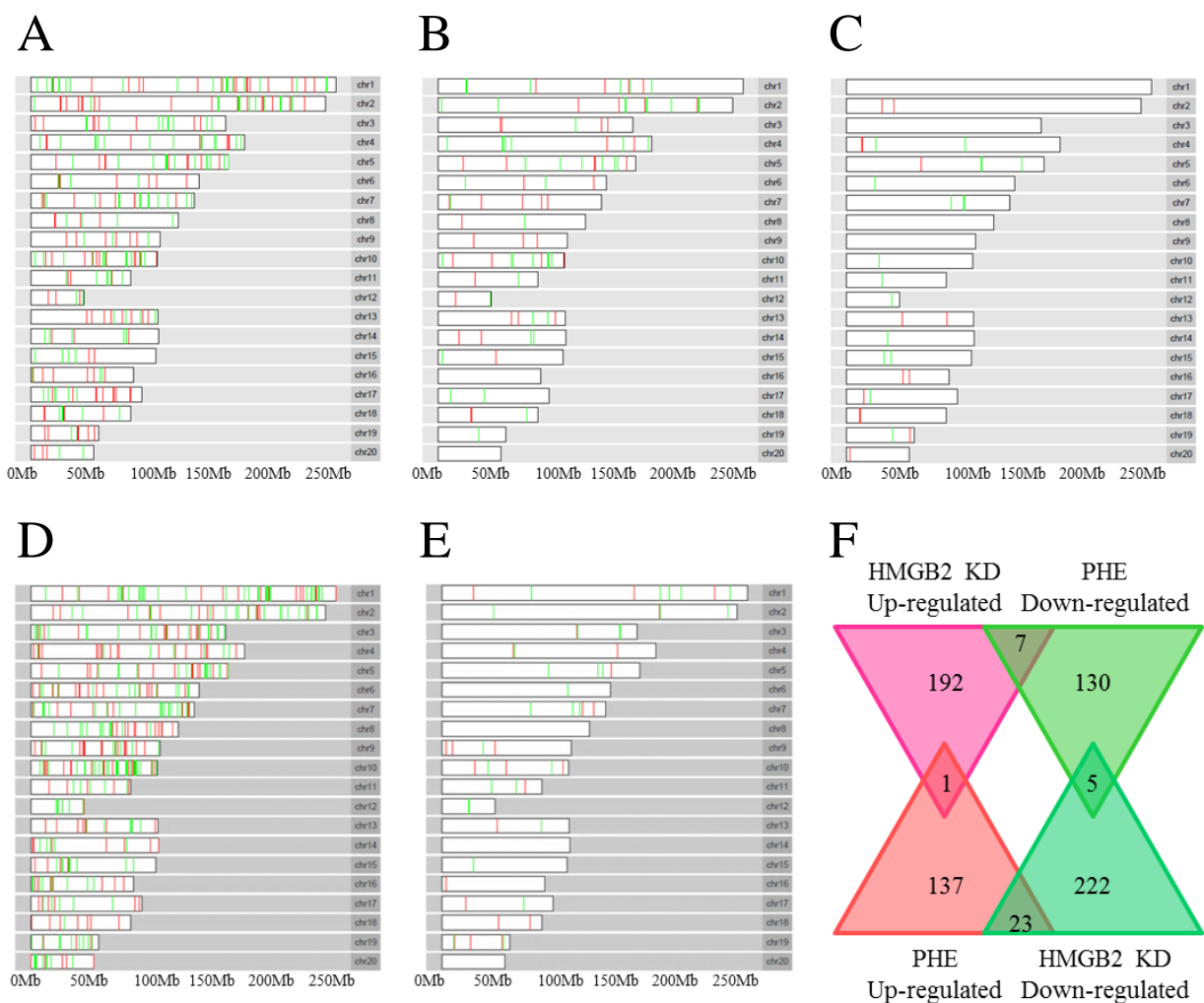
**Figure 2-5. Distribution of genes affected by PHE treatment or HMGB2 knock down.** (A) Genomic location of genes affected by PHE treatment in NRVMs. Gene expression microarray data are downloaded from Gene Expression Omnibus using the following accession IDs: unstimulated, GSM13869; PHE treated, GSM13870. Subset of these genes whose promoters featuring loss of HMGB2 binding (B), or gain of HMGB2 (C) after PHE treatment, are shown. (D) Genomic location pattern of genes affected by HMGB2 knock down in NRVMs. (E) Genes affected by HMGB2 knock down and occupied by HMGB2 at promoter before stimulation. Red, up-regulated genes after either PHE or HMGB2 siRNA treatment. Green, down-regulated genes. (F) Venn diagram showing the overlap of genes affected by PHE treatment or HMGB2 siRNA treatment.
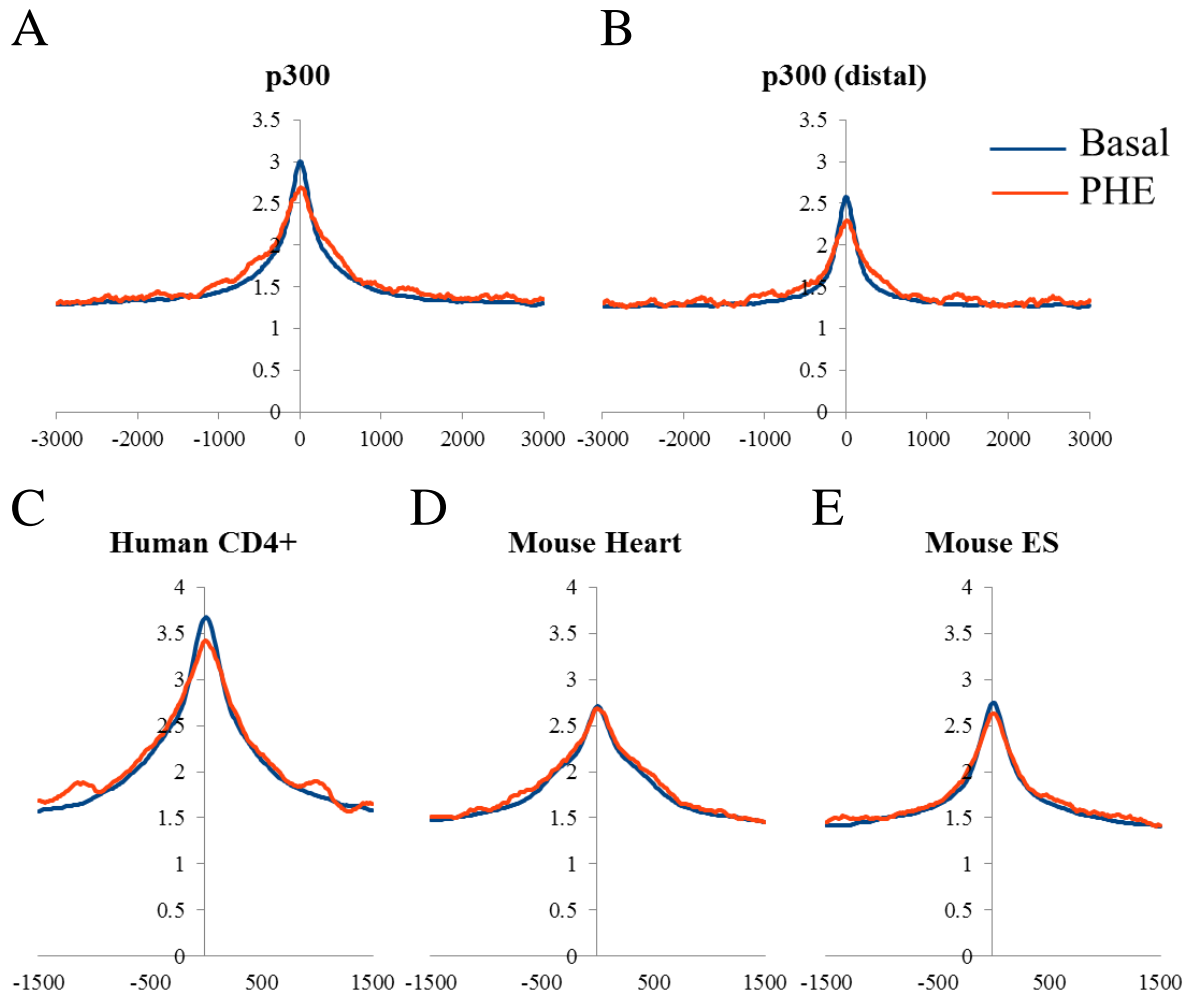
**Figure 2-6. Illustration of HMGB2 binding enrichment at p300 binding enhancers and CTCF binding sites.** Averaged HMGB2 read counts, after normalization, were plotted around the binding sites of p300 (A,B) and CTCF (C-E). Genome wide p300 binding sites (A), as well as distal binding sites (B), i.e. more than 2.5kb far away from any transcription start sites, are shown. CTCF binding sites as identified from human CD4+ T cells (C), mouse heart tissue (D), and mouse embryonic stem cells (E) are also shown. x-axis, relative distance to binding site. y-axis, scaled density of HMGB2 ChIP-Seq reads. Dark blue, reads from unstimulated NRVMs. Red, reads from PHE treated NRVMs.
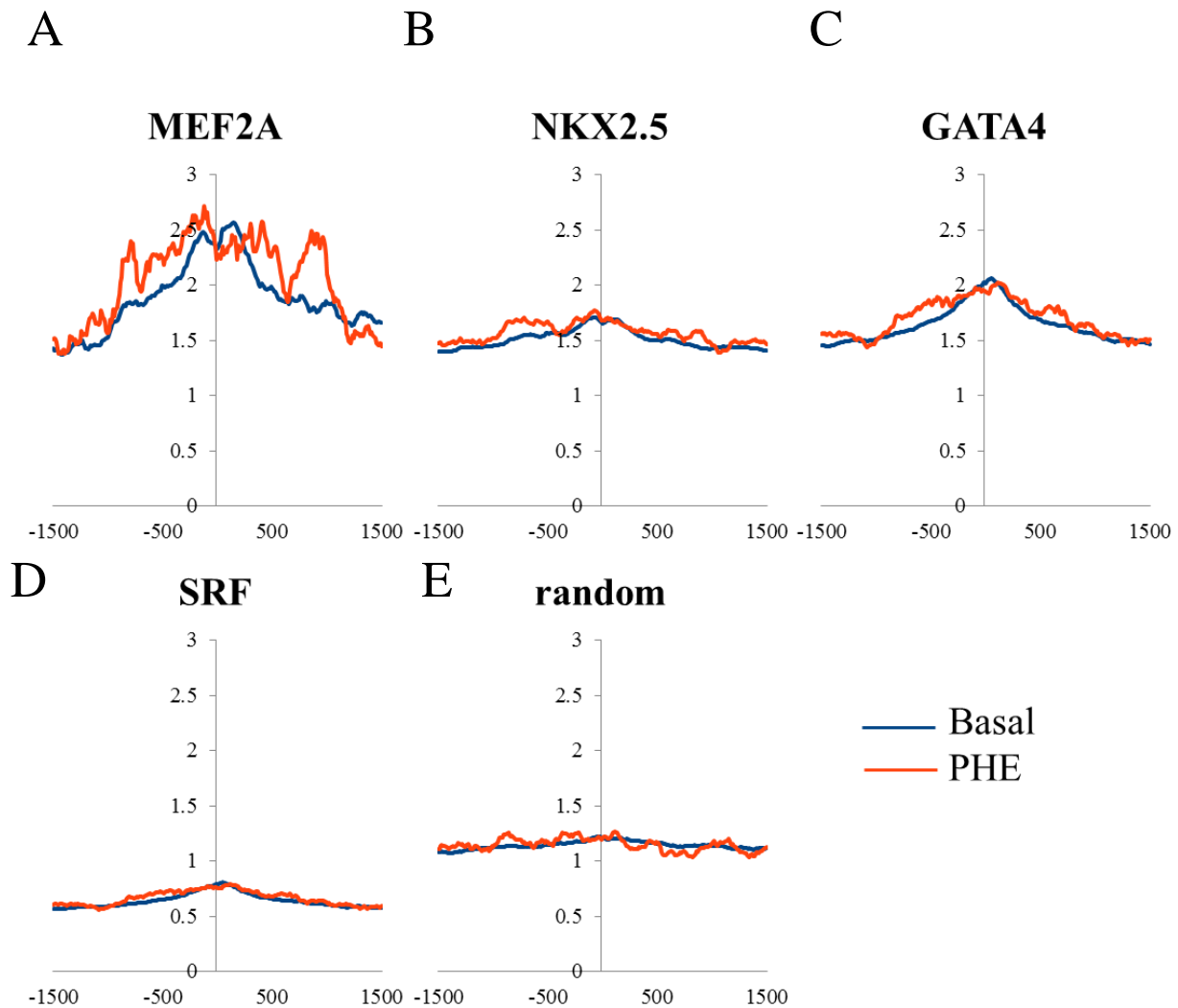
**Figure 2-7. HMGB2 binding profiles in cardiac transcription factor binding sites.** Averaged HMGB2 read counts, after normalization, were plotted around the binding sites of MEF2A (A), NKX2.5 (B), GATA4 (C) and SRF (D). To rule out the effects of cross-species conversion, regions from mouse genome were randomly chosen and converted into rat genome. This conversion generates a flat HMGB2 binding profile (E). x-axis, relative distance to binding site. y-axis, scaled density of HMGB2 ChIP-Seq reads. Dark blue, reads from unstimulated NRVMs. Red, reads from PHE treated NRVMs.
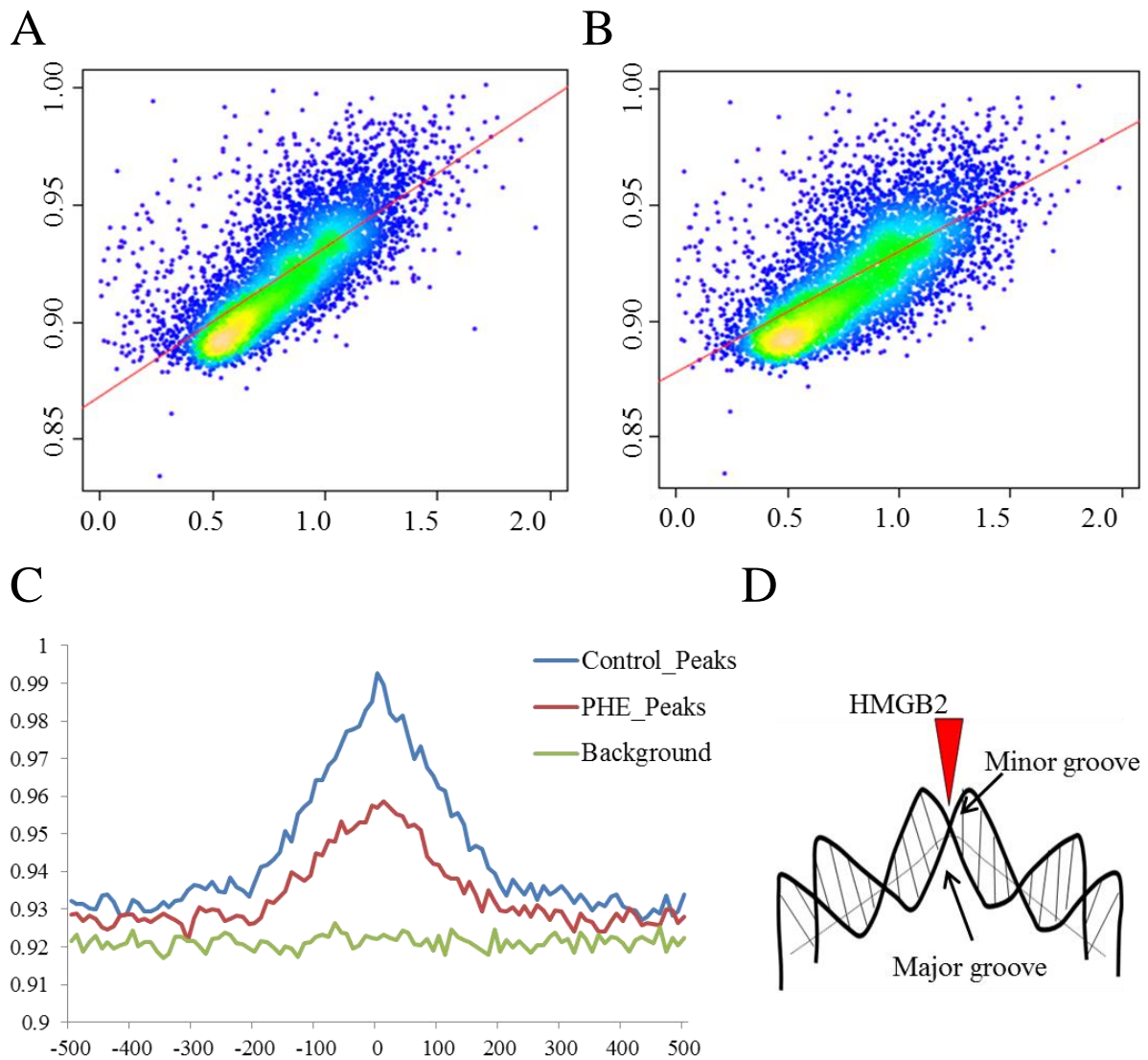
**Figure 2-8. HMGB2 binding intensity versus DNA bending propensity.** (A,B) Each point represents one 500Kbp bin of rat genome. HMGB2 signals from both control (A) and PHE-treated (B) NRVMs are shown. x-axis, normalized HMGB2 ChIP-Seq read counts. y-axis, averaged DNA bendability scores. Red line represents the linear trend. The positive slope of the trend line indicates the positive correlation between HMGB2 binding intensity and the DNA bending propensity toward major groove. (C) DNA bendability is profiled near HMGB2 bind peaks. x-axis, distance from HMGB2 binding peak represented as the zero point. y-axis, averaged DNA bendability score. (D) Diagram showing that HMGB2 interacts with DNA through the minor groove, and bends DNA toward the major groove. HMGB2 may search locally for the place with a higher DNA bending propensity, i.e. thermodynamically favoring bending toward major groove, to find the place to bind.
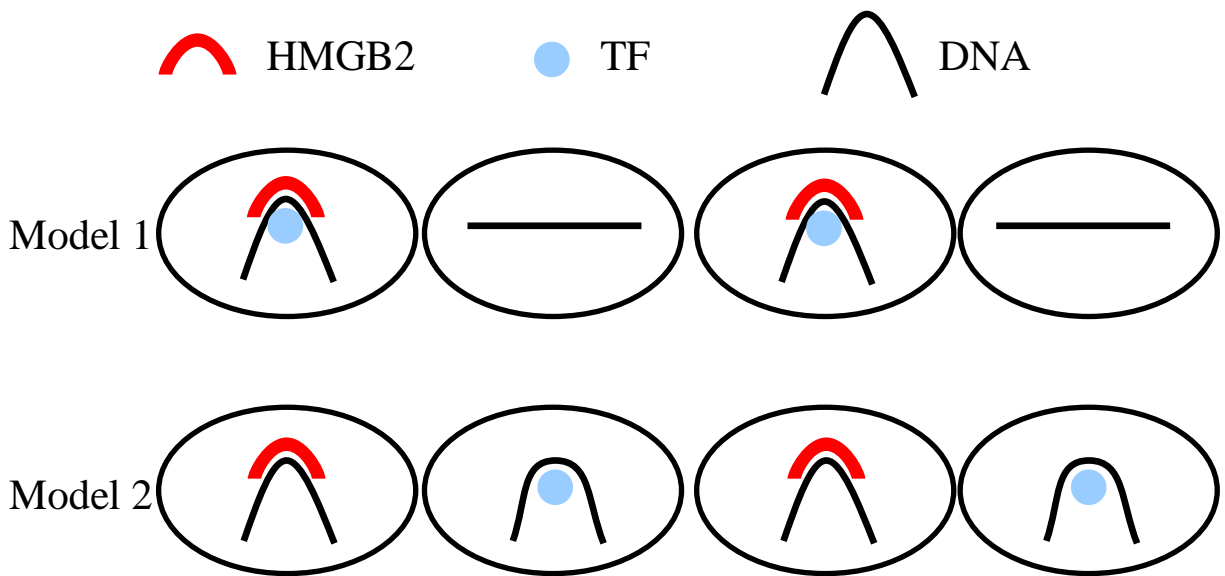
**Figure 2-9. Diagrammatic figure showing models of how HMGB2 facilitates transcription factor binding.** Two models are proposed to explain how HMGB2 facilitates transcription factor binding by altering DNA conformation. Each circle represents one cell within a population. In model one, HMGB2 and TF bind to the same position simultaneously. In model two, HMGB2 stabilizes the structure of bended DNA to enable TF binding, but does not co-localize with TF.
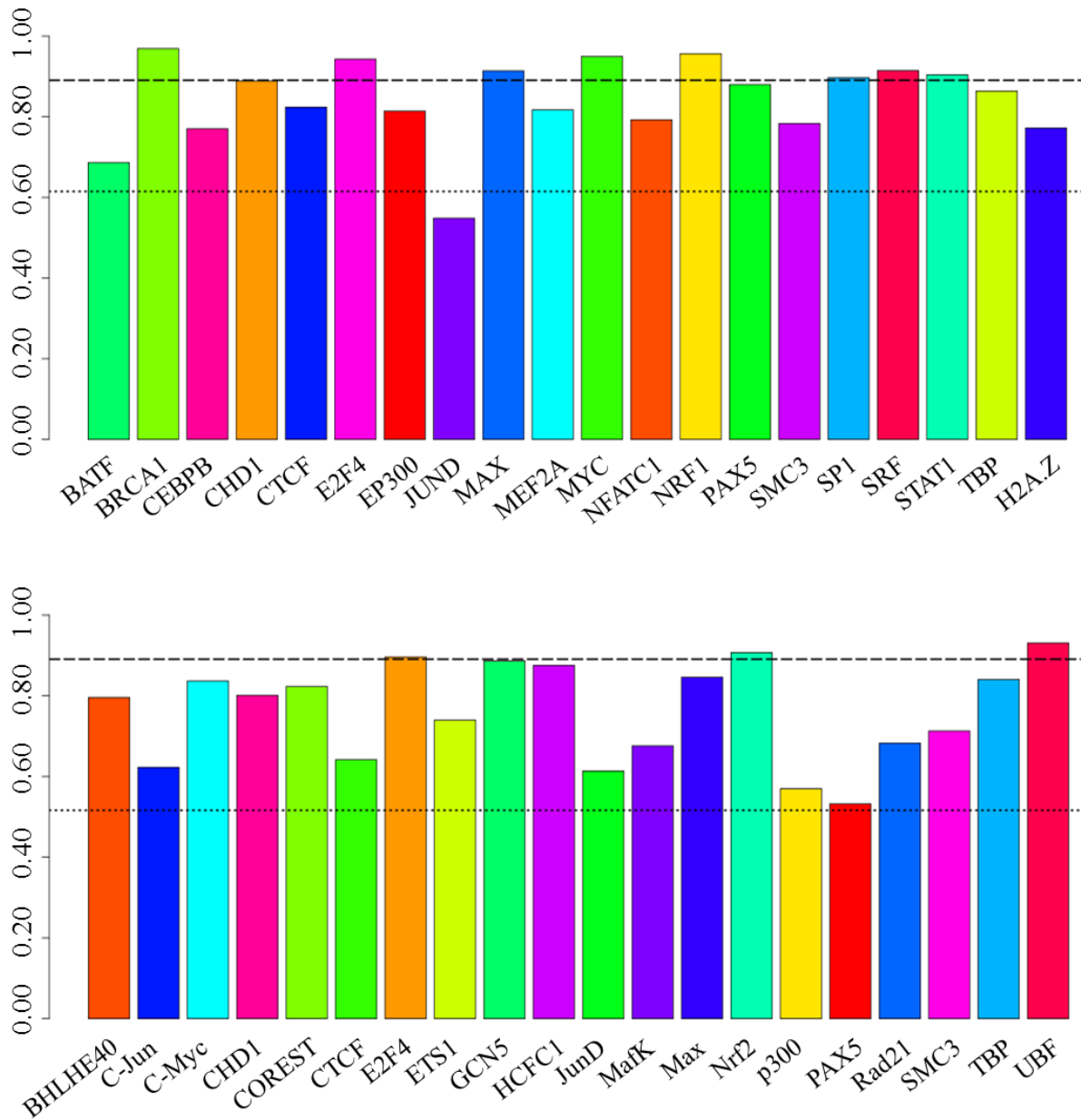
**Supplementary Figure 2-1. Preference of transcription factors for CpG island promoters.** Percentage is calculated as the ratio between the counts of TF-bound CpG island-associated promoters to the counts of TF-bound promters. 100% means all TF-bound promoters are associated with CpG islands, while 0% means none of the TF-bound promoters are CpG island-associated. Dotted line represents the percentage of CpG island-associated promoters as a fraction of the whole genome. Dashed line represents that percentage (89%) of HMGB2 associated promoters as calculated from figure 2-4. (A) Data from human B-lymphocyte cell line GM12878. (B) Data from mouse B-lymphoma cell line CH12. ENCODE ChIP-Seq Peak files used in this analysis are included in Supplementary Table 2-1.

# A

## Cardiac TFs



GATA4      SRF      NKX2_5      MEF2A

# B
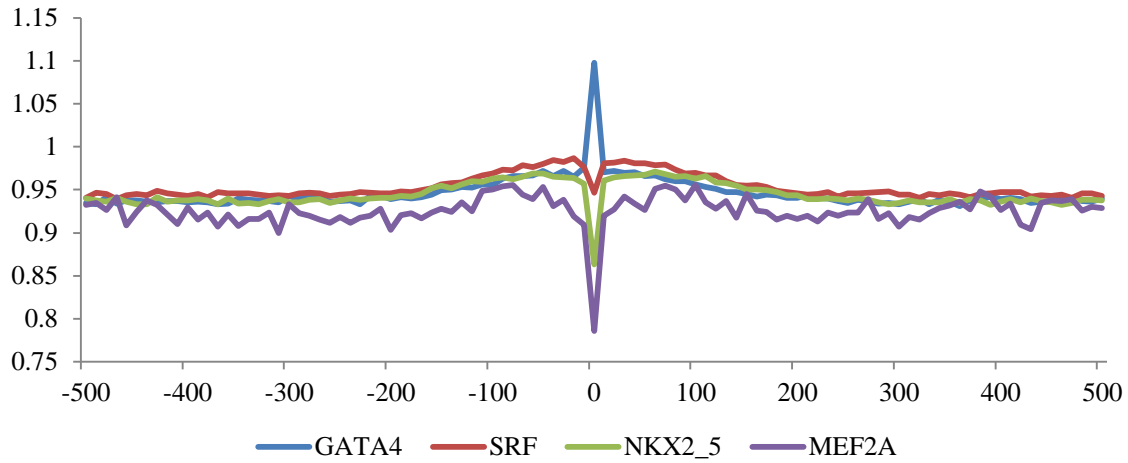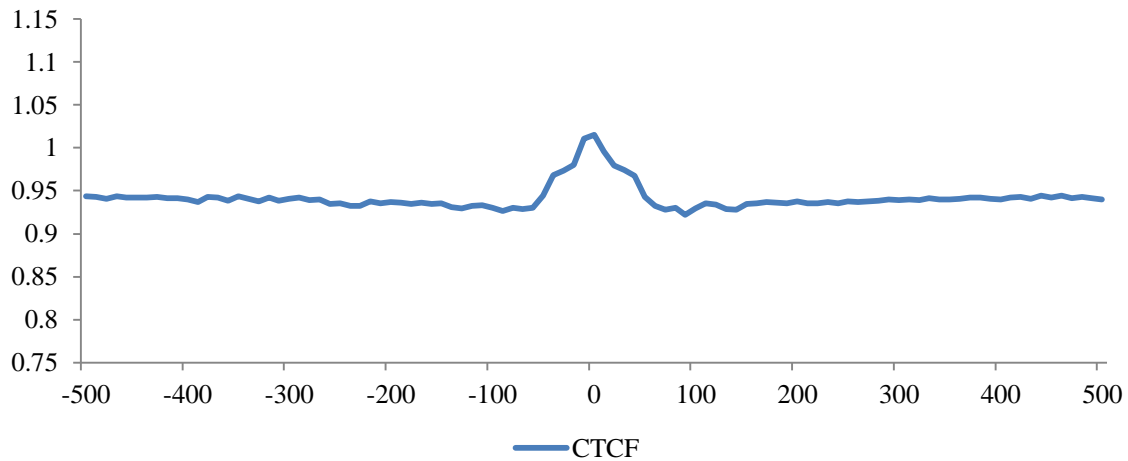
## Cardiac CTCF



CTCF

**Supplementary Figure 2-2. Bendability scores for cardiac transcription factors and CTCF.** DNA bendability scores are profiled near cardiac TFs (A) and CTCF (B) bind peaks. Same as figure 2-8 (C). x-axis, distance (bp) from binding peak represented as the zero point. y-axis, averaged DNA bendability score.

**Table 2-1: Primers used for anti-HMGB2 ChIP-PCR**

| Gene Name | Chromosome | TSS | Strand | Peak Position | Relative Position (TSS +/-) | Forward Sequence | Reverse Sequence |
|---|---|---|---|---|---|---|---|
| Cacna1c | 4 | 155517389 | -1 | 155515177 | 2212 | ACTGACGCCAGTCAGCCACCA | ACCTGAGGGCTTTGTCCTGGCTC |
| Tnni3 | 1 | 68028198 | 1 | 68028711 | 513 | AACCCGTGGCCCAGAGAGGG | AGCGACGTCGGACAGGAGCA |
| Por | 12 | 22154339 | -1 | 22154343 | -4 | CCCGCGGGTCCTGTAGGTCTCTG | CCGCAGCCTTCTGGGTCGGTG |
| Parp1 | 13 | 96309670 | 1 | 96309758 | 88 | CTGCGCCACGAGAGGGAGGA | TGCGGAGGCGAGTCCTTGGGG |
| Brd2 | 20 | 4864606 | 1 | 4862151 | -2455 | GCGCGTCCCTGAGCTCCTT | CCGAGGCAGAGCCTCCAGCA |
| Adra1b | 10 | 28946889 | -1 | 28946231 | 658 | GAAGATGGCGCCCAGCACCC | CTGCGGACGCGGCCACCAACTAC |
| Fgf16 | X | 93870484 | 1 | 93869552 | -932 | CCCCTTAAGCGGCTCCCACCCT | TCCCCTAGTCCCACTCCCCACC |
| Tgfb3 | 6 | 110195215 | -1 | 110196463 | -1248 | CGCGATCCTGGCAGCGGTT | CAGAGGGCACCCTCGGCCTT |
| Cabin1 | 20 | 13569453 | -1 | 13569563 | -110 | CCTGACGCGGACGGACCAAC | CGGGCGAACTGAGTTGGAAGCC |
| Mmp14 | 15 | 32493852 | 1 | 32493760 | -92 | AAGGAGGGCATTGGGCGGG | TGCGCCAGACACACACAG |
| Ndufb4 | 11 | 64852589 | 1 | 64852755 | 166 | CGGGAGACCCGAAAAGGGCAG | CACGATGTGCGTTTGGTTTGG |
| Mapk7 | 10 | 47654316 | -1 | 47654329 | -13 | CGGAGTGGAACGGGGTTCTGCCT | TCCCTCTGCCTGCTGCTC |
| Acta1 | 19 | 54084508 | -1 | 54083225 | 1283 | CGCTTGCTCTGGGCCTCGTC | GTTCTCCCTGCCCCCTGGCT |
| Ldha | 1 | 97403077 | 1 | 97403263 | 186 | CTGGGGTGGAGGTGCAGGGT | CAGGCCCGCCATCCCCTAA |
| Hmgn2 | 5 | 152715527 | -1 | 152715853 | -326 | TGCGCGACACTGGGCACATC | GCCAGGCCTCGCAAACCCT |
| Nppb | 5 | 165062348 | 1 | 165061827 | -521 | ACCAGAGTGCCCGGAAGTGGTG | AGGCCCTGCCCGGCTACCAA |
| Casp2 | 4 | 70029728 | 1 | 70029840 | 112 | AAGGGGCTGATGGCGGCTGA | CGCGGGGACCAGGCCAAGAAG |
| Tuba4a | 9 | 74500192 | -1 | 74499751 | 441 | TGGCTCAGGAGGGGGTGCTG | GCGCGGGGTTGGGTGTTAGGGG |

# Supplementary Table 2-1: ENCODE files used to generate supplementary figure 2-1

| Human | | Mouse | |
|---|---|---|---|
| BATF | wgEncodeAwgTfbsHaibGm12878BatfPcr1xUniPk.narrowPeak.gz | BHLHE40 | wgEncodeSydhTfbsCh12Bhlhe40nb100IggrabPk.narrowPeak.gz |
| BRCA1 | wgEncodeAwgTfbsSydhGm12878Brca1a300IggmusUniPk.narrowPeak.gz | c-Jun | wgEncodeSydhTfbsCh12CjunIggrabPk.narrowPeak.gz |
| CEBPB | wgEncodeAwgTfbsHaibGm12878Cebpbsc150V0422111UniPk.narrowPeak.gz | c-Myc | wgEncodeSydhTfbsCh12CmycIggrabPk.narrowPeak.gz |
| CHD1 | wgEncodeAwgTfbsSydhGm12878Chd1a301213alggmusUniPk.narrowPeak.gz | CHD1 | wgEncodeSydhTfbsCh12Chd1nb1006041111ggrabPk.narrowPeak.gz |
| CTCF | wgEncodeAwgTfbsSydhGm12878Ctctfsc15914c20UniPk.narrowPeak.gz | COREST | wgEncodeSydhTfbsCh12Corestsc30189IggrabPk.narrowPeak.gz |
| E2F4 | wgEncodeAwgTfbsSydhGm12878E2f4IggmusUniPk.narrowPeak.gz | CTCF | wgEncodePsuTfbsCh12CtcfImmortal2a4bInputPk.broadPeak.gz |
| EP300 | wgEncodeAwgTfbsSydhGm12878P300IggmusUniPk.narrowPeak.gz | E2F4 | wgEncodeSydhTfbsCh12E2f4IggrabPk.narrowPeak.gz |
| JUND | wgEncodeAwgTfbsSydhGm12878JundUniPk.narrowPeak.gz | ETS1 | wgEncodeSydhTfbsCh12Ets1IggrabPk.narrowPeak.gz |
| MAX | wgEncodeAwgTfbsSydhGm12878MaxIggmusUniPk.narrowPeak.gz | GCN5 | wgEncodeSydhTfbsCh12Gcn5IggrabPk.narrowPeak.gz |
| MEF2A | wgEncodeAwgTfbsHaibGm12878Mef2aPcr1xUniPk.narrowPeak.gz | HCFC1 | wgEncodeSydhTfbsCh12Hcfc1nb100068209IggrabPk.narrowPeak.gz |
| MYC | wgEncodeAwgTfbsUtaGm12878CmycUniPk.narrowPeak.gz | JunD | wgEncodeSydhTfbsCh12JundIggrabPk.narrowPeak.gz |
| NFATC1 | wgEncodeAwgTfbsHaibGm12878Nfatc1sc17834V0422111UniPk.narrowPeak.gz | MafK | wgEncodeSydhTfbsCh12Mafkab50322IggrabPk.narrowPeak.gz |
| NRF1 | wgEncodeAwgTfbsSydhGm12878Nrf1IggmusUniPk.narrowPeak.gz | Max | wgEncodeSydhTfbsCh12MaxIggrabPk.narrowPeak.gz |
| PAX5 | wgEncodeAwgTfbsHaibGm12878Pax5c20Pcr1xUniPk.narrowPeak.gz | Nrf2 | wgEncodeSydhTfbsCh12Nrf2IggrabPk.narrowPeak.gz |
| SMC3 | wgEncodeAwgTfbsSydhGm12878Smc3ab9263IggmusUniPk.narrowPeak.gz | p300 | wgEncodeSydhTfbsCh12P300sc584IggrabPk.narrowPeak.gz |
| SP1 | wgEncodeAwgTfbsHaibGm12878Sp1Pcr1xUniPk.narrowPeak.gz | PAX5 | wgEncodePsuTfbsCh12Pax5cFImmortal2a4bInputPk.broadPeak.gz |
| SRF | wgEncodeAwgTfbsHaibGm12878SrfPcr2xUniPk.narrowPeak.gz | Rad21 | wgEncodeSydhTfbsCh12Rad21IggrabPk.narrowPeak.gz |
| STAT1 | wgEncodeAwgTfbsSydhGm12878Stat1UniPk.narrowPeak.gz | SMC3 | wgEncodeSydhTfbsCh12Smc3ab9263IggrabPk.narrowPeak.gz |
| TBP | wgEncodeAwgTfbsSydhGm12878TbpIggmusUniPk.narrowPeak.gz | TBP | wgEncodeSydhTfbsCh12TbpIggmusPk.narrowPeak.gz |
| H2A.Z | wgEncodeBroadHistoneGm12878H2azStdPk.broadPeak.gz | UBF | wgEncodeSydhTfbsCh12Ubfsc13125IggrabPk.narrowPeak.gz |

# Reference

1.  Bustin, M. Revised nomenclature for high mobility group (HMG) chromosomal proteins. *Trends in biochemical sciences* **26,** 152–3 (2001).

2.  Bustin, M. & Reeves, R. High-mobility-group chromosomal proteins: architectural components that facilitate chromatin function. *Progress in nucleic acid research and molecular biology* **54,** 35–100 (1996).

3.  Smerdon, M. J. & Isenberg, I. Interactions between the subfractons of calf thymus H1 and nonhistone chromosomal proteins HMG1 and HMG2. *Biochemistry* **15,** 4242–7 (1976).

4.  Baker, C., Isenberg, I., Goodwin, G. H. & Johns, E. W. Physical studies of the nonhistone chromosomal proteins HMG-U and HMG-2. *Biochemistry* **15,** 1645–9 (1976).

5.  Travers, A. A. Priming the nucleosome: a role for HMGB proteins? *EMBO reports* **4,** 131–6 (2003).

6.  Nagaki, S. *et al.* Non-histone chromosomal proteins HMG1 and 2 enhance ligation reaction of DNA double-strand breaks. *Biochemical and biophysical research communications* **246,** 137–41 (1998).

7.  Boonyaratanakornkit, V. *et al.* High-mobility group chromatin proteins 1 and 2 functionally interact with steroid hormone receptors to enhance their DNA binding in vitro and transcriptional activity in mammalian cells. *Molecular and cellular biology* **18,** 4471–87 (1998).

8.  Ohndorf, U. M., Rould, M. A., He, Q., Pabo, C. O. & Lippard, S. J. Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins. *Nature* **399,** 708–12 (1999).

9.  Malarkey, C. S. & Churchill, M. E. A. The high mobility group box: the ultimate utility player of a cell. *Trends in biochemical sciences* **37,** 553–62 (2012).

10. Churchill, M. E. A., Klass, J. & Zoetewey, D. L. Structural analysis of HMGD-DNA complexes reveals influence of intercalation on sequence selectivity and DNA bending. *Journal of molecular biology* **403,** 88–102 (2010).

11. McCauley, M. J., Rueter, E. M., Rouzina, I., Maher, L. J. & Williams, M. C. Single-molecule kinetics reveal microscopic mechanism by which High-Mobility Group B proteins alter DNA flexibility. *Nucleic acids research* **41,** 167–81 (2013).

12. Stros, M. HMGB proteins: interactions with DNA and chromatin. *Biochimica et biophysica acta* **1799,** 101–13 (2010).

13. Cuddapah, S. *et al.* Genomic profiling of HMGN1 reveals an association with chromatin at regulatory regions. *Molecular and cellular biology* **31,** 700–9 (2011).

14. Franklin, S. *et al.* Quantitative analysis of the chromatin proteome in disease reveals remodeling principles and identifies high mobility group protein B2 as a regulator of hypertrophic growth. *Molecular & cellular proteomics* **11,** M111.014258 (2012).

15. Han, P., Hang, C. T., Yang, J. & Chang, C.-P. Chromatin remodeling in cardiovascular development and physiology. *Circulation research* **108,** 378–96 (2011).

16. Bruneau, B. G. Chromatin remodeling in heart development. *Current opinion in genetics & development* **20,** 505–11 (2010).

17. Montgomery, R. L. *et al.* Histone deacetylases 1 and 2 redundantly regulate cardiac morphogenesis, growth, and contractility. *Genes & development* **21,** 1790–802 (2007).

18. Hang, C. T. *et al.* Chromatin regulation by Brg1 underlies heart muscle development and disease. *Nature* **466,** 62–7 (2010).

19. Lee, J.-H. *et al.* Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circulation research* **109,** 1332–41 (2011).

20. Hwang, J.-J. *et al.* Microarray gene expression profiles in dilated and hypertrophic cardiomyopathic end-stage heart failure. *Physiological genomics* **10,** 31–44 (2002).

21. Zhao, M., Chow, A., Powers, J., Fajardo, G. & Bernstein, D. Microarray analysis of gene expression after transverse aortic constriction in mice. *Physiological genomics* **19,** 93–105 (2004).

22. Dewey, F. E. *et al.* Gene coexpression network topology of cardiac development, hypertrophy, and failure. *Circulation. Cardiovascular genetics* **4,** 26–35 (2011).

23. Van den Bosch, B. J. C. *et al.* Early and transient gene expression changes in pressure overload-induced cardiac hypertrophy in mice. *Genomics* **88,** 480–8 (2006).

24. Franklin, S. *et al.* Specialized compartments of cardiac nuclei exhibit distinct proteomic anatomy. *Molecular & cellular proteomics* **10,** M110.000703 (2011).

25. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10,** R25 (2009).

26. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9,** R137 (2008).

27. Brukner, I., Sánchez, R., Suck, D. & Pongor, S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *The EMBO journal* **14,** 1812–8 (1995).

28. ENCODE Project. at <http://encodeproject.org/ENCODE/>

29. He, A., Kong, S. W., Ma, Q. & Pu, W. T. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 5632–7 (2011).

30. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research* **19,** 24–32 (2009).

31. Handoko, L. *et al.* CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics* **43,** 630–8 (2011).

32. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–8 (2011).

33. Bush, E. *et al.* A small molecular activator of cardiac hypertrophy uncovered in a chemical screen for modifiers of the calcineurin signaling pathway. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 2870–5 (2004).

34. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011,** bar030 (2011).

35. Melvin, V. S. *et al.* The role of the C-terminal extension (CTE) of the estrogen receptor alpha and beta DNA binding domain in DNA binding and interaction with HMGB. *The Journal of biological chemistry* **279,** 14763–71 (2004).

36. Ueda, T. & Yoshida, M. HMGB proteins and transcriptional regulation. *Biochimica et biophysica acta* **1799,** 114–8 (2010).

37. Vavouri, T. & Lehner, B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome biology* **13,** R110 (2012).

38. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & development* **25,** 1010–22 (2011).

39. Cato, L., Stott, K., Watson, M. & Thomas, J. O. The interaction of HMGB1 and linker histones occurs through their acidic and basic tails. *Journal of molecular biology* **384,** 1262–72 (2008).

40. Das, D., Peterson, R. C. & Scovell, W. M. High mobility group B proteins facilitate strong estrogen receptor binding to classical and half-site estrogen response elements and relax binding selectivity. *Molecular endocrinology* **18,** 2616–32 (2004).

41. Najima, Y. *et al.* High mobility group protein-B1 interacts with sterol regulatory element-binding proteins to enhance their DNA binding. *The Journal of biological chemistry* **280,** 27523–32 (2005).

42. Campbell, P. A. & Rudnicki, M. A. Oct4 interaction with Hmgb2 regulates Akt signaling and pluripotency. *Stem cells* **31,** 1107–20 (2013).

43. May, D. *et al.* Large-scale discovery of enhancers from human heart tissue. *Nature genetics* **44,** 89–93 (2012).

44. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nature genetics* **42,** 806–10 (2010).

45. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–211 (2009).

46. Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21,** 5427–40 (2002).

47. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome research* **22,** 1680–8 (2012).

48. McCauley, M., Hardwidge, P. R., Maher, L. J. & Williams, M. C. Dual binding modes for an HMG domain from human HMGB2 on DNA. *Biophysical journal* **89,** 353–64 (2005).

49. McCauley, M. J., Zimmerman, J., Maher, L. J. & Williams, M. C. HMGB binding to DNA: single and double box motifs. *Journal of molecular biology* **374,** 993–1004 (2007).

50. Pallier, C. *et al.* Association of chromatin proteins high mobility group box (HMGB) 1 and HMGB2 with mitotic chromosomes. *Molecular biology of the cell* **14,** 3414–26 (2003).

51. Das, D. & Scovell, W. M. The binding interaction of HMG-1 with the TATA-binding protein/TATA complex. *The Journal of biological chemistry* **276,** 32597–605 (2001).

52. Bonaldi, T., Längst, G., Strohner, R., Becker, P. B. & Bianchi, M. E. The DNA chaperone HMGB1 facilitates ACF/CHRAC-dependent nucleosome sliding. *The EMBO journal* **21,** 6865–73 (2002).

53. Kwon, J.-H. *et al.* Overexpression of high-mobility group box 2 is associated with tumor aggressiveness and prognosis of hepatocellular carcinoma. *Clinical cancer research* **16,** 5511–21 (2010).

54. Sharma, A., Ray, R. & Rajeswari, M. R. Overexpression of high mobility group (HMG) B1 and B2 proteins directly correlates with the progression of squamous cell carcinoma in skin. *Cancer investigation* **26,** 843–51 (2008).

55. Wang, W. *et al.* Overexpression of high mobility group box 1 and 2 is associated with the progression and angiogenesis of human bladder carcinoma. *Oncology letters* **5,** 884–888 (2013).

56. Wu, Z. B. *et al.* High-mobility group box 2 is associated with prognosis of glioblastoma by promoting cell viability, invasion, and chemotherapeutic resistance. *Neuro-oncology* (2013). doi:10.1093/neuonc/not078

57. Shin, Y.-J. *et al.* High-mobility group box 2 (HMGB2) modulates radioresponse and is downregulated by p53 in colorectal cancer cell. *Cancer biology & therapy* **14,** 213–21 (2013).

58. Abraham, A. B. *et al.* Members of the high mobility group B protein family are dynamically expressed in embryonic neural stem cells. *Proteome science* **11,** 18 (2013).

59. Taniguchi, N. *et al.* Expression patterns and function of chromatin protein HMGB2 during mesenchymal stem cell differentiation. *The Journal of biological chemistry* **286,** 41489–98 (2011).

60. Yamazaki, F., Nagatsuka, Y., Shirakawa, H. & Yoshida, M. Repression of cell cycle progression by antisense HMG2 RNA. *Biochemical and biophysical research communications* **210,** 1045–51 (1995).

61. Zimmer, H. G. Catecholamine-induced cardiac hypertrophy: significance of proto-oncogene expression. *Journal of molecular medicine* **75,** 849–59

62. Wolfram, J. A., Lesnefsky, E. J., Hoit, B. D., Smith, M. A. & Lee, H.-G. Therapeutic potential of c-Myc inhibition in the treatment of hypertrophic cardiomyopathy. *Therapeutic advances in chronic disease* **2,** 133–44 (2011).

# Chapter 3: The Role of DNA Methylation in Cardiac Chromatin Structure and in the Susceptibility to Heart Failure

## A. Abstract

DNA methylation is an epigenetic mechanism that controls gene expression. To determine how DNA methylation participates in the development of heart failure, we measured the DNA methylome in healthy and hypertrophic mouse hearts using reduced representational bisulfite sequencing (RRBS). To mimic heart failure, Isoproterenol (ISO) minipumps were implanted into two mouse strains with opposite phenotypes: BUB/BnJ, which is resistant to ISO-induced hypertrophy/heart failure, and BALB/cJ, which is susceptible. Contrary to the dogma of DNA methylation being a stable modification, we observed noteworthy ISO-induced methylation differences in genes related to heart diseases. Interestingly, the DNA methylome of the two mouse strains respond differently to ISO treatment. We selected the top 15,000 DNA fragments with high variations and carried out an unbiased weighted gene correlation network analysis (WGCNA). We found that CpGs can be clustered into modules based on their behavior after ISO treatment and these modules in turn correlate with the localization of different epigenetic marks. Our data demonstrate global changes in DNA methylation in the diseased adult heart and reveal networks of modified CpG loci that may be associated with heart failure development.

## B.  Introduction

DNA methylation is a covalent modification that adds a methyl group to the 5' carbon of cytosine nucleotides[1]. Extensive studies have shown that DNA methylation is an important epigenetic modification that regulates chromatin structure and gene expression. The deregulation of DNA methylation has been linked to many diseases including different types of cancers[2–6]. Traditionally, DNA methylation is thought to be a marker of heterochromatin because condensed chromatin has a higher level of DNA methylation, as observed during X-chromosome inactivation, for example. Conversely, in cancer cells, chromatin is often globally hypomethylated, which is believed to be a cause of oncogenesis[2]. Unlike other epigenetic marks, DNA methylation has long been considered a stable modification of DNA and expected to remain unchanged throughout the life of the fully differentiated somatic cell, but this idea has received many challenges recently.

Recent studies using human samples have indicated the pathological changes of DNA methylation in heart diseases[7–10]. These studies, however, cannot provide a comprehensive analysis of DNA methylation alteration in a case-control fashion, and do not fully establish the causal relationship between DNA methylation and disease development. Meanwhile, most of these studies do not give single base resolution for the whole genome. To solve this problem and provide a clear map on how DNA methylation is altered in heart disease, we used animal models to better understand this feature of chromatin. With the recent development of bisulfite high-throughput sequencing, we are able to study the genome-wide DNA methylation in any tissues or cell types of interest.

Attempts to study alterations in cardiac DNA methylation in human dilated cardiomyopathy did not identify global DNA methylation pattern that is specific to the heart and explains the actions of DNA methylation in this organ[9,10]. Most of these studies, although conducted using genome-wide techniques such as microarray or sequencing methods, only identified a few loci with altered DNA methylation. Importantly, this contradicts the fact that heart diseases involve dozens of genes and molecular pathways. Moreover, these studies did not examine the relationship between DNA methylation and other chromatin modifications due to limited availability of epigenetic data from human samples

In this study, we aimed to investigate DNA methylation in heart failure by: i) measuring the genome-wide DNA methylation pattern in both healthy and failing hearts using high resolution (single base specificity) and quantitatively rigorous next generation DNA sequencing, ii) comparing alterations in DNA methylation between two mouse strains with different sensitivity to heart failure, iii) understanding the role of differential methylation in phenotype by analyzing the annotations of the nearby or associated genes, and iv) studying the crosstalk between DNA methylation and other epigenetic marks using systems approach.

## C.  Materials and Methods

### *Mice*

Adult female BALB/cJ or BUB/BnJ mice aged 8-10 weeks were obtained from Jackson Laboratories (Bar Harbor, ME). All protocols involving animals were approved by the UCLA Animal Research Committee. Mice were housed in pathogen-free conditions according to NIH guidelines.

## Isoproterenol treatment

The treatment protocol was adapted from the Hybrid Mouse Diversity Panel[11]. Briefly, mice were implanted with an Alzet (Cupertino, CA) Micro-Osmotic Pump (Model 1004) filled with 420 μg/g body weight of Isoproterenol to provide a constant supply of 20 μg/g (body weight)/day of ISO. Mice were monitored by echocardiography before surgery and at days 7, 14 and 21 after surgery. Mice were sacrificed at day 21 after surgery and left ventricles were harvested for further analysis.

## Genomic DNA extraction

Tissues were cut into small pieces, put in a tube with 0.7mL Extraction Buffer (50 mM Tris-Cl pH 8.0, 50 mM EDTA, 100 mM NaCl, 0.5% SDS, 0.5 mg/mL Proteinase K) and incubated at 55°C overnight. DNA was extracted using Phenol:Chloroform, followed by ethanol precipitation, and dissolved in 150 μL TE buffer (10 mM Tris-Cl pH 8.0, 0.5 mM EDTA).

## Bisulfite conversion for reduced representational sequencing

The RRBS method was adapted from other studies[12,13]. Genomic DNA was digested with MspI overnight at 37°C and purified using Phenol:Chloroform as described before. DNA fragments were end repaired using Klenow DNA polymerase. An excess amount of dATP was added to provide sufficient 5' adenosine overhangs. Illumina Genome Analyzer adapters containing methylated cytosines instead of unmethylated cytosines were annealed and ligated to Klenow treated DNA fragments, and purified by Phenol:Chloroform extraction. Size selection of adapter-ligated fragments was conducted through agarose gel purification. To select the fragments with the size of 40-280 bp, adapter-ligated fragments with the size of 160-400 bp were

excised and purified. The EpiTect system from Qiagen was used for bisulfite treatment to convert unmethylated C to U, according to the provided protocol. The final library was built by large-scale PCR amplification using Pfu Turbo Cx Hotstart DNA Polymerase and 1.1/2.1 Illumina Primers, and cleaned up by AMPure magnetic beads.

### *Illumina sequencing and alignment*

The RRBS library sequencing and alignment protocol was adapted from our collaborators[12]. Briefly, 100-base Illumina single-end sequencing reads were generated from previously prepared RRBS libraries using Solexa sequencing technology (Illumina HiSeq 2000 sequencers). The reads were aligned to the reference mouse genome (mm9) using BS-Seeker2, allowing up to ten mismatches, and methylation call was conducted by the script provided with BS-Seeker2. The methylation level for each covered cytosine is calculated as the number of methylated reads divided by the total number of reads. To achieve an accurate estimate of methylation level, we only included cytosines that are covered by at least ten reads. Cytosines overlapping with known mouse SNPs were removed. Because the methylation occurs almost exclusively in the CpG context, we focused only on those cytosines in CpG dinucleotides.

### *Identification of differentially methylated regions*

To identify differentially methylated fragments (DMF), we applied the two-sided unpaired t-test to each fragment to calculate the t-score and p-value between two groups of comparison. Each fragment was defined as the *in silico* MspI digestion product of mouse reference genome with length ranging from 50 bp to 300 bp, which is wider than the size range of RRBS gel selection to compensate errors from gel running and cutting. The unpaired two-

sided t-test was conducted using the function t.test in R 3.0.1. Each CpG within one fragment was considered as an independent measurement and the methylation level for an entire fragment is defined as the mean methylation level for all CpG sites within that fragment. Fragments with a difference of methylation larger than 5% and p value of less than $10^{-3}$ were considered as DMFs. This threshold gave us a reasonable amount of DMFs.

### Co-methylation network analysis

The weighted Gene Co-Expression Network Analysis (WGCNA) package was used for co-methylation network analysis[14], following the standard protocol provided online[15]. To preserve the information of both correlation and anti-correlation of co-methylation, the similarity of methylation profiles between two fragments is defined as $S_{ij} = \frac{1+cor(i,j)}{2}$, where $i$ and $j$ represent the methylation levels of two DNA fragments. The top 15,000 DNA fragments with the highest methylation variation among samples were chosen for co-methylation network construction. To achieve a scale-free topology, a soft-thresholding power of 10 was chosen, resulting in a $R^2$ of 0.8 as measured by the scale-free topology model linear regression[14]. This $R^2$ value meets the recommendation from the authors of WGCNA package[14]. In total, we generated 82 clusters with the smallest cluster having 30 DNA fragments.

### Epigenetic mark enrichment in DNA methylation modules

To call an overlap between one DNA fragment and an epigenetic mark, we required at least 50bp overlap between one DNA fragment and the peak of one epigenetic mark. A binomial test P-value was calculated to assess the significance of overlap between one track of epigenetic mark and one DNA methylation. The background was defined as the fraction of epigenetic mark

associated DNA fragments within all 15K DNA fragments used for the DNA co-methylation network analysis. To show the gradient of P values on a heatmap, a ceiling of six was set for the −log10 of P value to minimize the effects of extreme P values on heatmap color code, but this does not introduce dramatic difference to the clustering pattern.

### *Bioinformatics*

The mouse reference genome (mm9) was downloaded from UCSC Genome Bioinformatics (ftp://hgdownload.cse.ucsc.edu/). The UCSC Genes track from UCSC Genome Browser was used as the genome annotation. The CpG islands (cpgIslandExt) track from UCSC Genome Browser was used as CpG island annotation. Functional analysis of cellular pathways was performed using DAVID Bioinformatics Resources 6.7[16]. ENCODE data of ChIP-Seq peaks for histone modifications and other epigenetic marks were downloaded from the ENCODE Project at UCSC[17].

## D. Results

### *Genomic annotation of CpG sites covered by RRBS libraries*

A brief report on sequencing depth and alignment is provided in table 3-1. Given that the size of the mouse genome is ~2.65 billion bp, ~265 million uniquely aligned 100-bp reads would be needed to achieve a ten-fold coverage of the mouse genome using traditional shotgun sequencing. To reduce the data size while maintaining data quality, we use RRBS instead of shotgun sequencing. Although RRBS cannot cover every CpG within the genome, this method provides us a very high coverage on the CpGs within RRBS libraries (Table 3-1).

RRBS uses MspI to digest genomic DNA into fragments during library preparation. MspI is a restriction endonuclease that recognizes the CCGG motif and generates a -CGG sticky end. This introduces a fragment selection bias towards CpG rich regions. To quantify this bias, we compared the annotation of CpG sites covered by our RRBS libraries with the total CpGs within the reference genome. Figure 3-1 shows the location of CpGs covered by our RRBS libraries (Fig 3-1, C & F), comparing with the genome background (Fig 3-1, A & D) and all CpG dinucleotides (Fig 3-1, B & E). This data indicate that RRBS provides good coverage for promoters and exons, as well as CpG islands and island shores, and that this method is useful for studying the methylation of gene-related and/or CpG island-related CpG dinucleotides.

The RRBS profile of the mouse heart shows a bimodal distribution of DNA methylation (Fig 3-2A), with 42% CpGs poorly methylated (methylation level < 5%) and 29% CpGs highly methylated (methylation level > 80%). This distribution may be different from other cells/tissues. For example, Jiang, et al. reported a global hypermethylation in oocytes and sperms from Zebrafish[18]. Further analysis shows that almost all CpGs located within CpG islands (Fig 3-2B, upper panel) and most CpGs close to CpG islands (Fig 3-2B, middle panel) are hypomethylated, while most CpGs outside of any CpG islands ("lone" CpGs) are hypermethylated (Fig 3-2B, lower panel). On the other hand, more than 90% of hypomethylated CpGs are located in CpG islands or island shores (Fig 3-2C, panel d), while most hypermethylated CpGs are located far away from CpG islands (Fig 3-2C, panel f).

### Differential methylation induced by ISO treatment

To study how cardiac DNA methylation is altered in heart failure, we used ISO micro-osmotic pumps to induce heart failure in mice. After three-week treatment, BALB/cJ mice showed a severe phenotype of dilated cardiomyopathy, while BUB/BnJ mice showed only a mild compensation (Fig 3-3).

Using a t-test as a statistical tool, we found that while the majority of MspI digested fragments had no substantial methylation difference, some regions showed a significant alteration after ISO treatment (Fig 3-4 A&B). To further show that our statistical tool gave us reliable results, we plotted the distribution of P values for each DNA fragment (Fig 3-4 C-E). The shape of the overall P value distribution for biological replicate comparison, i.e. under null hypothesis (Fig 3-4 E), is similar with the ones for case-control comparison (Fig 3-4 C&D), with the exception that there are more DNA fragments with low P value in the latter two. This indicates that the surplus of small P values in the case-control comparison reflects the overdispersion of test statistics[19].

Table 3-2 shows the number of DMFs induced by ISO treatment in two different mouse strains. Among the 112,616 DNA fragments measured in both strains, 411 of them show differential methylation after ISO treatment in at least one strain. Surprisingly, despite of the similar volcano plot shape between BALB/cJ and BUB/BnJ strains (Fig 3-4 A&B), the overlap of DMFs between two strains is very small. Only 17 fragments show uniform behavior in both strains under our statistical threshold (table 3-2).

To test whether certain chromosomes harbor more DMFs than others, we plot the location of all identified DMFs for two strains in the genome (Fig 3-5). The first observation was

that the genomic locations of DMFs between the two strains show little overlap (Fig 3-5).

Moreover, we showed that DMFs are located across the whole genome without an obvious

preference for a certain chromosome. This is in accordance with our previous report that cardiac

hypertrophy related genes are distributed in all chromosomes[20].

### *DMFs are associated with genes involved in cardiomyopathy*

To understand the potential function of DMFs, we studied genes containing DMFs or

harboring one within 5Kbp upstream of their transcription start sites. In total, we identified 312

genes that are associated with at least one differentially methylated fragment between control

and ISO-treated mice. More details on genes associated with DMFs are listed in table 3-3.

### *Network analysis of DNA methylome shows different patterns of co-methylation*

When looking at the raw DNA methylation data, we found that some CpGs are highly

correlated with each other. To study the co-methylation network of DNA methylation, we

calculated the correlation among the top 15,000 variable DNA fragments and built a weighted

network based on the correlation matrix (Fig 3-6). DNA fragments within one module have

similar behaviors and one EigenCpG is assigned to each module to represent the group behavior

(Fig 3-6 B). To show the higher-order topology, we correlated the EigenCpG network with the

traits of ISO treatment. Interestingly, we found that most modules have some level of correlation

with ISO treatment (Fig 3-6 A). Some modules show a uniform response to ISO treatment in

both BALB/cJ and BUB/BnJ strains (Fig 3-6 B, upper and lower panels), while others show a

different response between two strains (Fig 3-6 B, middle panel). We also found a few modules

that, despite no significant responses to ISO treatment, show distinct methylation levels in two

strains at the basal level (data not shown), indicating methylation polymorphisms between the two strains.

*Overlap between co-methylation modules and epigenetic marks*

Because previous reports have indicated potential interactions between DNA methylation and histone post-translational modification as well as chromatin structure[21,22], we asked the question of whether different modules have diverse preference for epigenetic marks. To answer this question, we determined which regions contained in various DNA methylation modules are also occupied by histone post-translational modifications and other chromatin-binding proteins, based on published data from the heart and other organs (Fig 3-7). To rule out the possibility that some epigenetic marks favor CpG rich regions, we chose the set of 15K DNA fragments used for network construction, rather than the whole genome, as the background to calculate the P value for the overlap amongst these subsets of loci. Surprisingly, we observed modularity of within the heatmap according to both the epigenome track direction (column-wise) and the DNA methylation module direction (row-wise). Moreover, we found that most DNA methylation modules featuring a decrease of methylation after ISO treatment overlap with H3K4me1 and K3K4me3 (as seen in the bottom half of the heatmap). The enrichment pattern of transcription factor binding in different DNA methylation modules is not obvious, probably because of the heterogeneity of transcription factors.

## E. Discussion

In this chapter, we asked the question whether DNA methylation, a major epigenetic regulator of gene expression and chromatin structure, is involved in heart failure. Using the

RRBS method, we surveyed genome-wide methylation in the normal and diseased heart, with an emphasis on promoters, exons, and CpG islands. We observed a bimodal distribution of DNA methylation on a genome-wide level; when only CpG islands or lone CpGs are considered, this distribution takes on a unimodal character. Using the β-adrenoreceptor agonist Isoproterenol, we induced heart failure in two mouse strains that have different responses to treatment and demonstrated that certain DNA regions showed methylation alterations in a strain-specific manner during disease. Further, we clustered DNA methylation fragments based on their correlation in terms of alteration after ISO and revealed that the resulting modules exhibit a range of behaviors after stimulation depending on whether the genome predisposes the animal to be susceptible or resistant to hypertrophy. Finally, we compared our DNA co-methylation modules with published epigenome ENCODE data, and found that different modules are co-occupied by different epigenetic markers. Taken together, we demonstrated that in heart failure, the cardiac DNA methylome is globally remodeled to coordinate chromatin remodeling (Fig 3-8, black solid boxes and arrows) in combination with other protein-based chromatin features.

Despite recent studies using patient samples[7–10,23], the molecular mechanisms for how altered DNA methylation contributed to the progression of cardiomyopathy remains unclear[24]. These studies are important in that they showed the involvement of DNA methylation in cardiovascular disease. However, the main problem with these studies is the genetic heterogeneity of human samples, including different dietary and environmental exposures, etc. In addition, the heterogeneity of the types of heart disease between patients contributes more diversity to human samples, further decreasing the statistical power of recent studies. This is why most of these studies cannot provide a comprehensive picture of the disease-specific methylome

and they also lack the ability to resolve mechanisms of regulation. Our study utilized a strict case-control experimental design, including consistent housing conditions and diet. Furthermore, the genome-wide RRBS method we used here samples randomly from the entire mouse genome, providing more information than a microarray-based method that only studies promoter regions.

Our study found distinct DNA methylation alteration patterns induced by ISO treatment in two different mouse strains. It is well known that different mouse strains exhibit different gene expression patterns[25] and have variance in their susceptibility to diseases[26]. Previous studies of mouse genetics focused on how to explain strain-specificity through genetic variations[11,27], and such studies are similar to Genome Wide Association Studies (GWAS) in humans. However, it is very likely that the strain-specific DNA methylation acts as an amplifier to magnify the effects of primary DNA sequence differences[28]. DNA methylation polymorphism highly depends on genetic polymorphism. One piece of evidence for this is that the F1 hybrids of two different strains maintain the strain-specific methylation pattern from their parental allele[29]. In our study, we found that the functional annotation of regions associated with altered DNA methylation in disease is different between the two strains. For example, when studying the DNA regions with up-regulated methylation in disease, we found genes of the dilated cardiomyopathy pathway enriched in the BLAB/cJ mice, while the Wnt signaling pathway is enriched in the BUB/BnJ mice. This further explains the strain-specific susceptibility to ISO-induced heart failure.

The relationship between DNA methylation and gene expression is not linear. Although some believe that promoter hypermethylation is associated with gene repression, and gene body hypermethylation is associated with gene expression[30], an increasing number of studies have demonstrated that the relationship is not so simple[9,12]. In reviewing reports on the role of DNA

methylation on gene expression, we found that this theory reflects a group behavior, but that it is not a strict rule for all genes. For one single gene, whether and how DNA methylation regulates its expression depends on the combination of many factors including histone modifications, binding of transcription factors, and even the higher order structure of chromatin. van Eijk and de Jone et al[31] published a systematic study of DNA methylation and gene expression levels in human blood cells, reporting that while some modules are highly correlated (in a positive or negative manner) with gene expression, most modules had no correlation between gene expression and DNA methylation. Based on this information, we decided to combine our DNA methylation data with other available epigenetic data to explain how the alteration of DNA methylation affects disease development by regulating chromatin structure.

The traditional method of identifying DMFs using a statistical hypothesis test is based on the assumption that all methylation measurements are independent and identically distributed (i.i.d.). However, this strong assumption does not reflect the reality of biology. In fact, many biological actions are accomplished through network behaviors[32], such as gene expression and protein-protein interactions. The information on co-regulation can be used to infer the topology of a biological network[14]. Instead of using genes as nodes in the network, we used DNA fragments to build the co-methylation network, thus circumventing the difficulty that some DNA fragments are located in enhancers or other functional element-related regions far away from known genes. Using this method, we built an epigenetic interaction network in the heart and showed that DNA fragments can be clustered into modules when they are highly correlated. This clustering also implies that these fragments, their associated genomic loci, are potentially regulated by the same chromatin remodeling and gene transcription mechanisms. Interestingly,

members within the same module can be located far away from each other in the linear position of one chromosome, or even located on different chromosomes. One explanation is that these regions are positioned proximately in 3-D space, or they share a similar DNA motif that can be recognized by the same epigenetic regulatory machinery.

Our study also revealed a noteworthy phenomenon that chromosomal regions with similar DNA methylation changes after stimulation share a common pattern of epigenetic marks. In figure 3-7, clustering, or modularity, of the enrichment heatmap is visible in both directions. The modularity of epigenome tracks (column-wise in figure 3-7) may be the result of correlation between the locations of different epigenetic marks. For example, H3K4me1 and H3K27ac are sometimes co-localized in enhancer regions to activate gene expression[33]. The modularity in the DNA methylation module direction (row-wise), however, cannot be explained by the co-localization of DNA fragments between DNA methylation modules, because each DNA fragment was assigned to only one module and there cannot be an overlap between any DNA methylation modules. Since DNA modules are ordered based on their correlation with ISO treatment, this phenomenon indicates that if the DNA methylation patterns of two DNA regions respond to the environment stimulation in a similar way, these two regions may also have a similar histone modification pattern. In other words, chromosomal regions deposited with the same epigenetic patterns will have similar DNA methylation dynamics.

To understand how our findings from mice may have clinical relevance, we need to consider DNA methylation in the context of endogenous regulation (Fig 3-8). Epigenetic modifications are known to be regulated by cellular metabolites. Previous studies have found that cardiac metabolism is aberrant in disease. This includes S-adenosylhomocysteine (SAH) / S-

adenosylmethionine (SAM) ratio imbalance, as indicated by Hyperhomocysteinemia in cardiovascular disease[34], and accumulation of α-ketoglutarate or 2-oxoglutarate, due to inactivation of ketoglutarate dehydrogenase caused by abnormal mitochondrial function and oxidative stress[35]. Previous studies of heart failure metabolism mainly focused on how metabolites impair normal cellular functions, such as mitochondrial function or cell signaling pathways. In view of those studies showing the involvement of SAH/SAM in DNA methylation[36,37] as well as a-ketoglutarate in active DNA demethylation[38], we believe that the metabolic dysfunction in heart failure may cause abnormal DNA methylation, which will then change the normal cardiac gene expression. This study provides a map of how DNA methylation is remodeled in heart failure. Further studies can reveal whether and how manipulation of cardiac metabolites can modulate DNA methylation patterns to relieve disease symptoms.
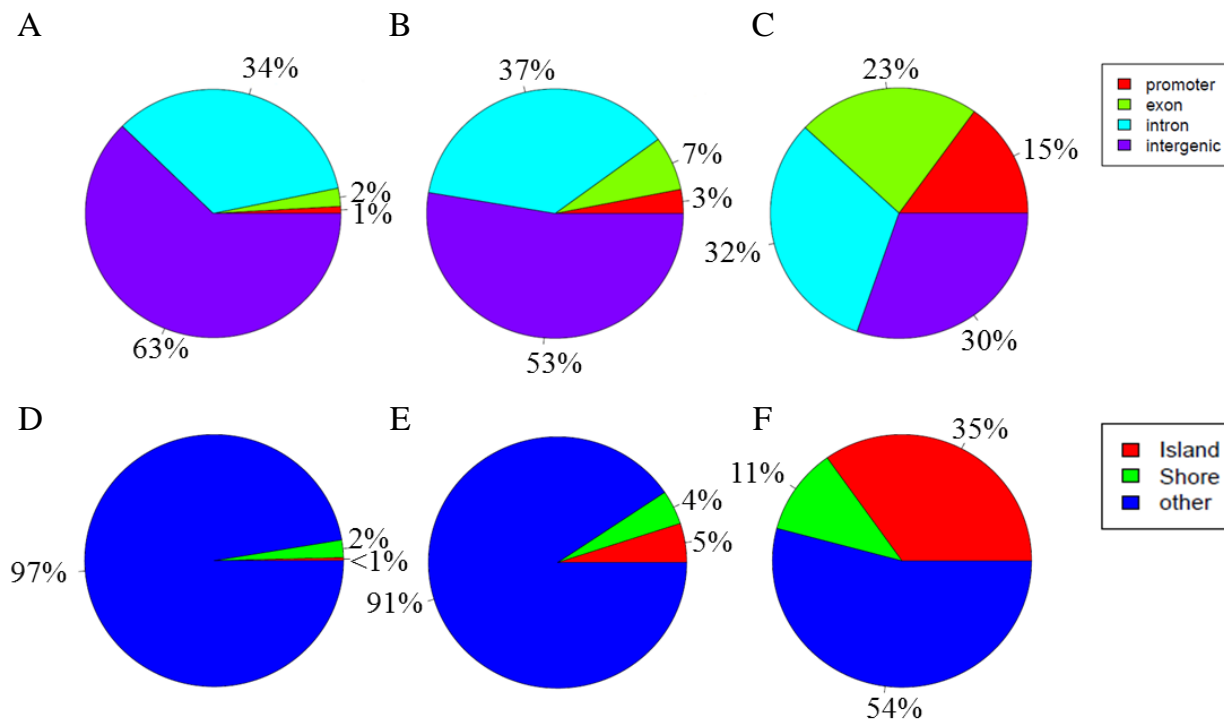
**Figure 3-1. Annotation of CpGs covered by RRBS libraries.** CpGs are annotated based on its location on promoter, exon, intron or intergenic regions (A-C), or CpG island, island shore or far from CpG island (D-F). A & D, Annotation of mouse reference genome. B & E, Annotation of CpGs within the whole genome. C & F, Annotation of CpGs covered by our RRBS libraries. Promoter, regions within 2kb upstream of transcription start site. CpG island shore, regions within 2kb flanking CpG island.
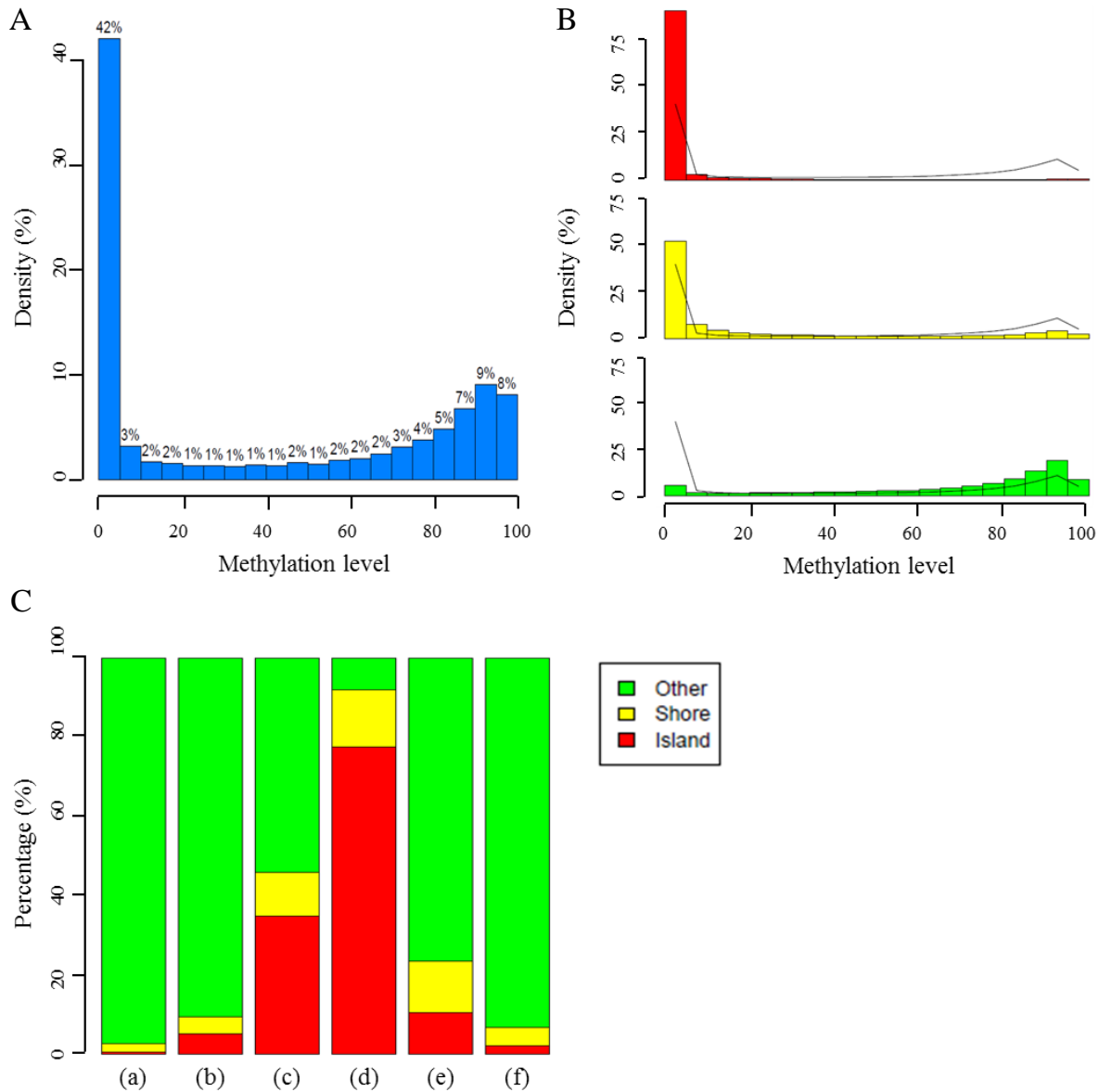
**Figure 3-2. Distributions of DNA methylation within CpG islands, CpG island shores and non-CpG island regions.** (A,B) Histogram of DNA methylation level distribution for the whole RRBS library (A), CpGs located within CpG islands (B, upper panel), CpG island shores (B, middle panel) and non-CpG island regions (B, lower panel). x-axis, DNA methylation level. y-axis, the proportion of total RRBS CpGs. The black line in B represents the whole RRBS library. (C) Stacked histogram of DNA methylation distribution showing the percentage of CpGs located in CpG islands, shores or non-CpG island regions. CpGs are grouped based on their methylation level. (a), the length of each element in the whole genome. (b), all CpGs within the whole genome. (c), RRBS CpGs. (d) RRBS CpGs with lower than 5% methylation. (e) RRBS CpGs with > 5% and < 80% methylation. (f) RRBS CpGs with more than 80% methylation.
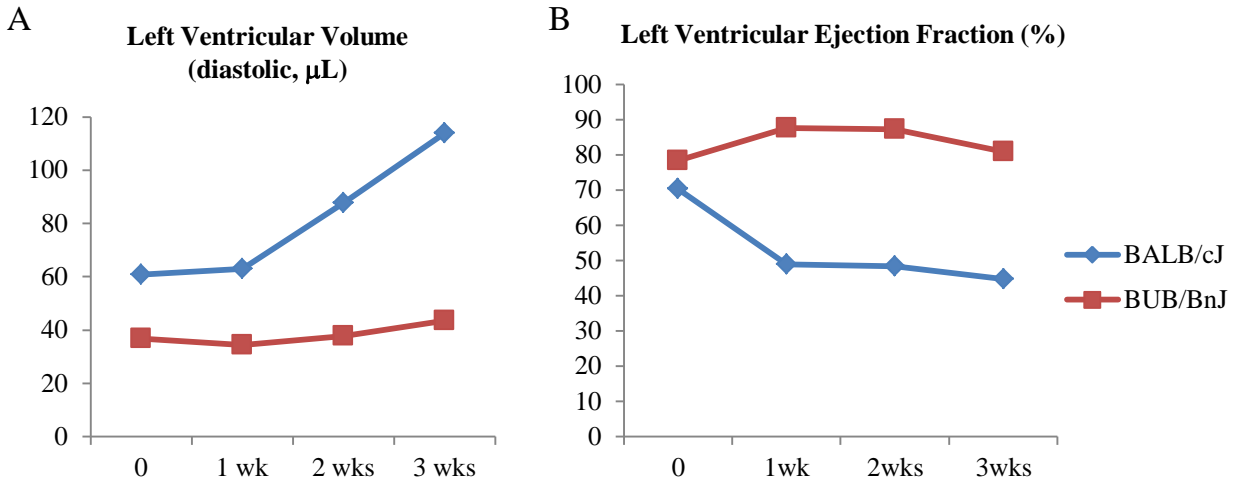
A  **Left Ventricular Volume (diastolic, μL)**

B  **Left Ventricular Ejection Fraction (%)**

BALB/cJ
BUB/BnJ

**Figure 3-3. Isoproterenol induced heart failure in mice of BALB/cJ and BUB/BnJ strains indicating strain-specific responses to β-adrenergic receptor agonist treatment.** The size of heart as indicated by diastolic left ventricular volume (A), and heart function as indicated by ejection fraction (B), were measured using M-mode mouse echocardiography. BALB/cJ mice showed a dramatic increase of LV volume and decrease of ejection fraction after three-week ISO treatment, while BUB/BnJ mice showed only a slight increase of ejection fraction.
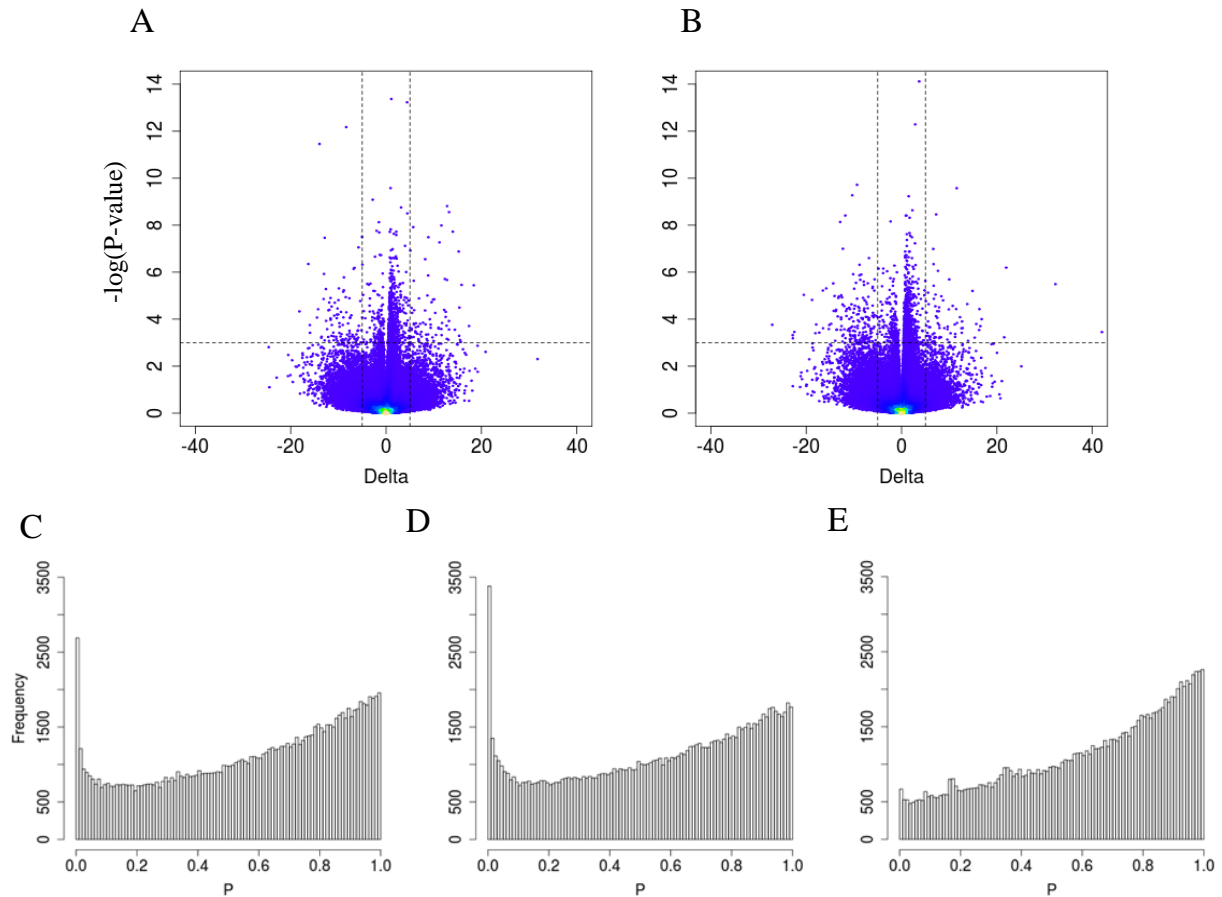
**Figure 3-4. ISO-treatment induces differential methylation on several DNA regions.** (A,B) Volcano plot showing ISO-induced differentially methylated regions in BALB/cJ (A) and BUB/BnJ (B) strains. Each point represents one DNA fragment, i.e. MspI digestion product with length of from 50bp to 300bp. x-axis, difference of DNA methylation level between control and ISO treated mice. y-axis, $-\log_{10}$ of p-value. To qualify as a differentially methylated fragment (DMF), one fragment must have a difference of DNA methylation level bigger than 5%, and a p-value smaller than $1\times10^{-3}$. (C, D, E) Distribution of p-values for all DNA fragments. (C,D) P values of control v.s. ISO treated BALB/cJ (C) and BUB/BnJ (D) strains. (E) P values of two biological replicates that are sequenced at the same time. The enrichment of low P values can be seen as the high bar on the left of the distribution in panels C and D.
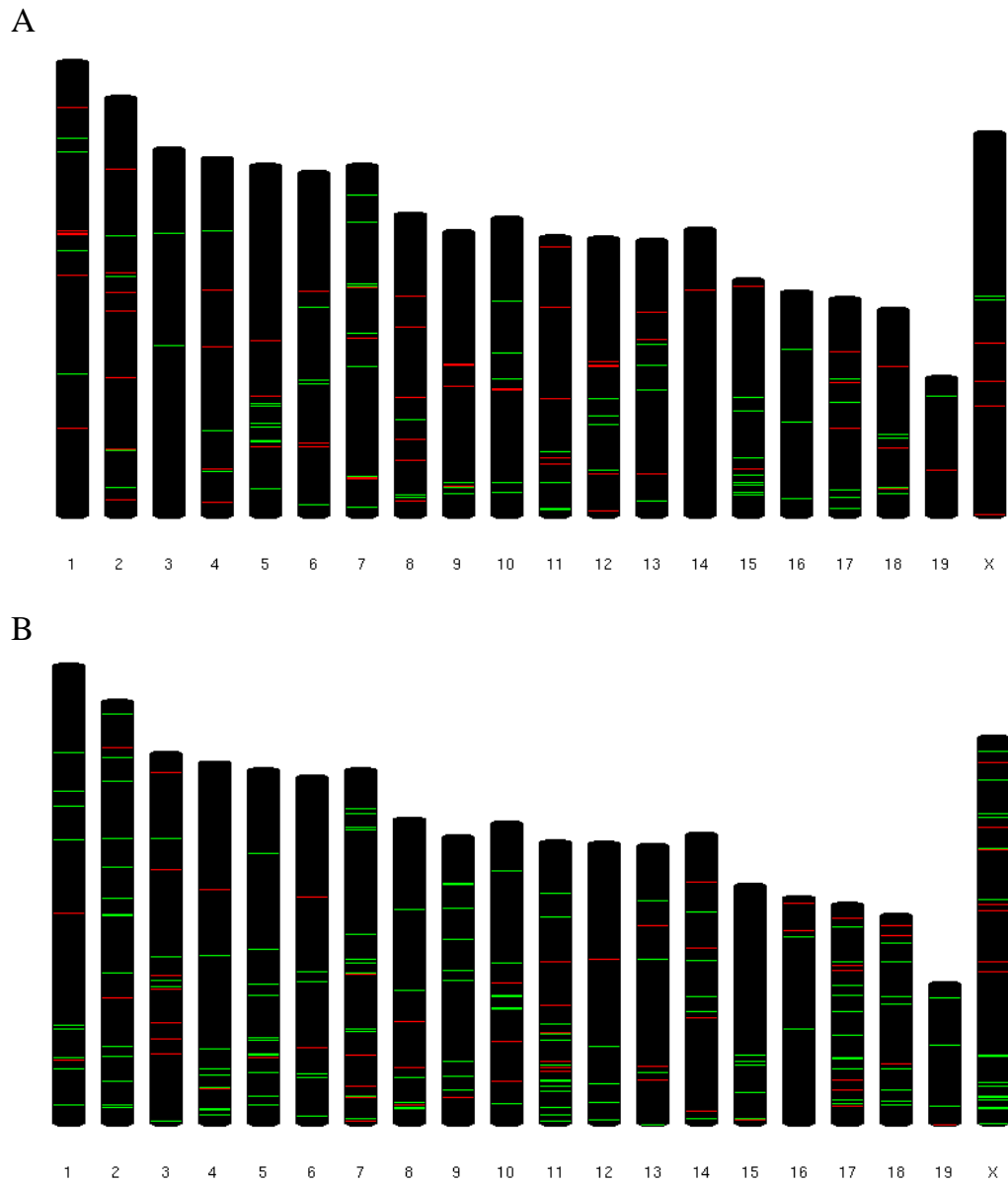
**Figure 3-5. Genomic location of DMFs induced by ISO treatment in BALB/cJ (A) and BUB/BnJ (B) strains.** Red: up-regulated DMFs. Green: down-regulated DMFs.
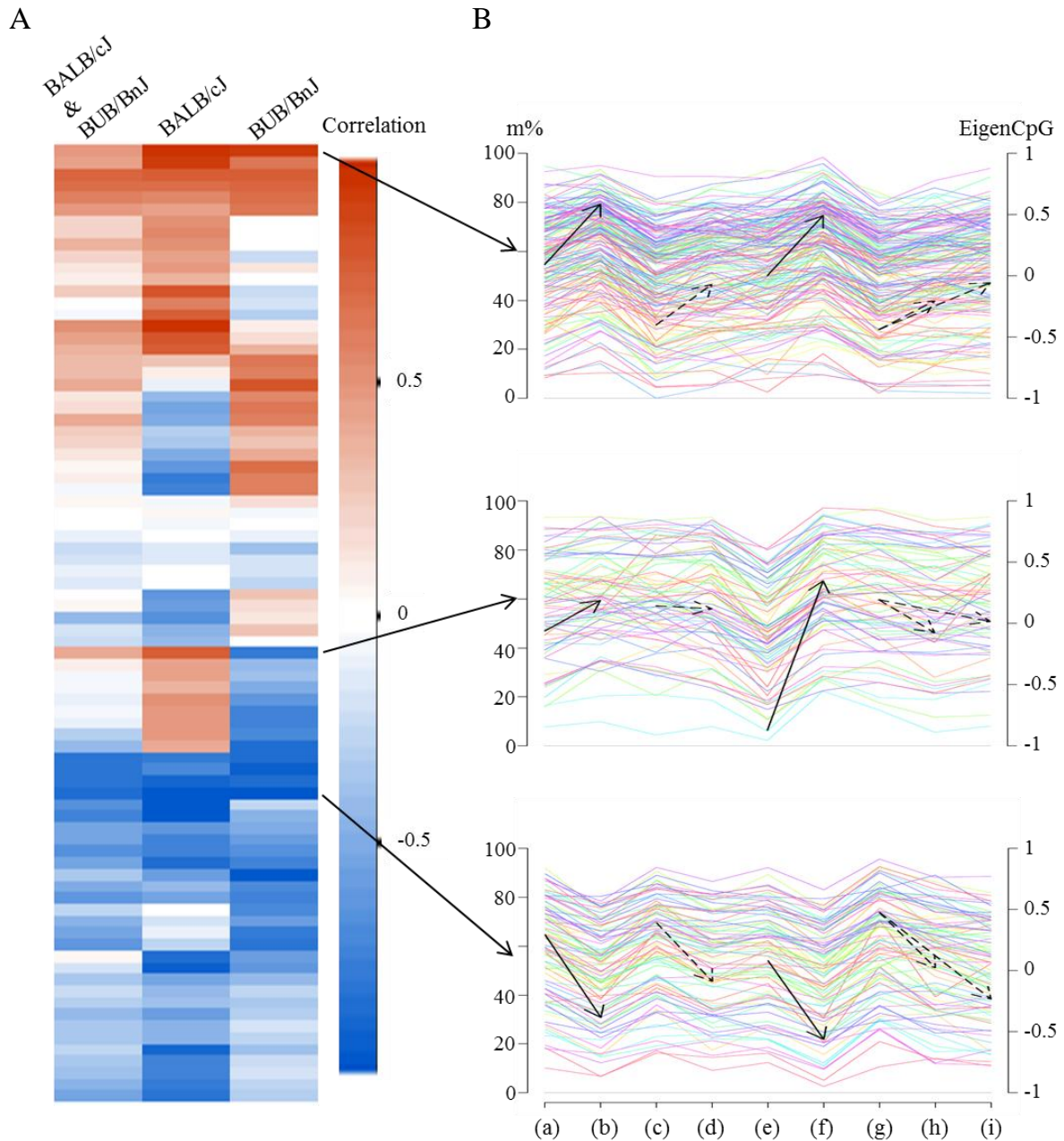
**Figure 3-6. Co-methylation modules of CpGs identified by WGCNA.** (A) Heatmap showing the correlation between EigenCpG methylation and ISO treatment for both strains (left), BALB/cJ (middle) or BUB/BnJ (right) strain. Each row represents one module of CpGs. (B) DNA methylation levels (m%) for individual CpG (colored lines) and eigenvalue for EigenCpG (black solid arrows, BALB/cJ, dashed arrows, BUB/BnJ) in nine samples. Three representing modules are shown. a, BALB/cJ control replicate 1. b, BALB/cJ ISO replicate 1. c, BUB/BnJ control replicate 1. d, BUB/BnJ ISO replicate 1. e, BALB/cJ control replicate 2. f, BALB/cJ ISO replicate 2. g, BUB/BnJ control replicate 2. h, BUB/BnJ ISO replicate 2.1. i, BUB/BnJ ISO replicate 2.2.

155
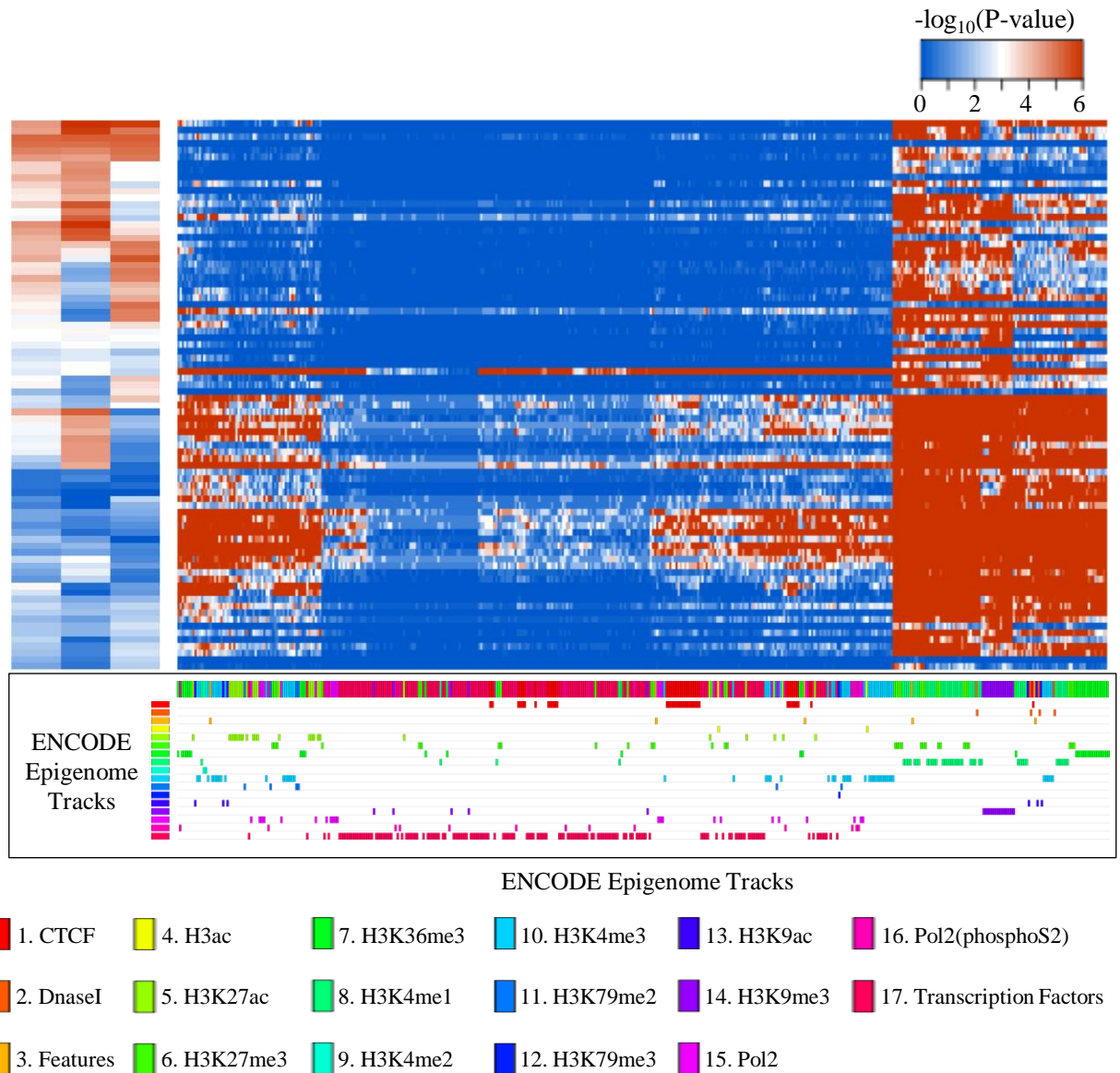
**Figure 3-7. Heatmap of –log$_{10}$ of P values showing the significance of overlapping between CpG modules and epigenetic markers.** CpG modules from Fig 3-6 are shown on the left. Rows are CpG modules and columns are ENCODE ChIP-Seq Peak tracks of epigenetic markers. Columns, but not rows, are re-organized using hierarchical clustering. Colors below heatmap correspond to the epigenetic marks. A decomposed color code is provided below the color bar under heatmap. CTCF, CCCTC-binding factor. DnaseI, DNaseI Hypersensitivity sites. Features, genomic features including promoters, exons, introns, intergenic regions, CpG islands and shores.
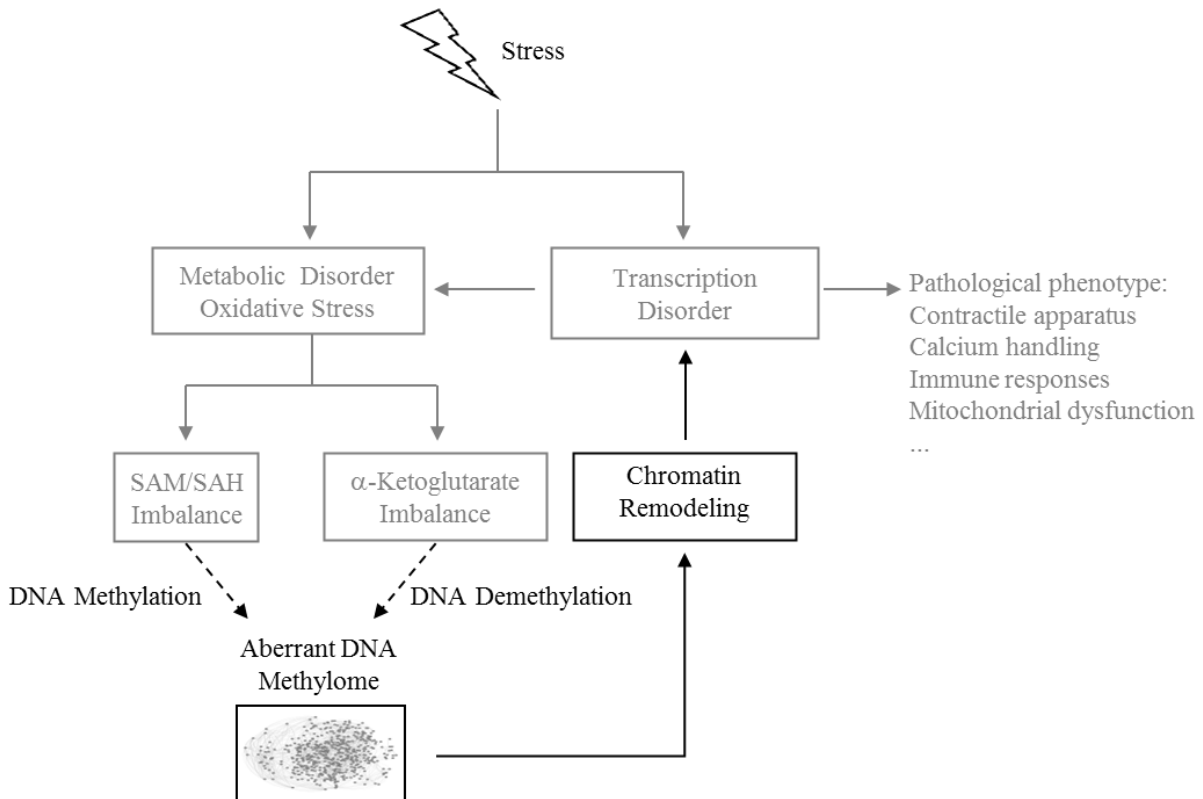
**Figure 3-8. Working hypothesis based on this and other studies on DNA methylation in heart disease pathogenesis.** In this chapter, we demonstrated a disease-specific DNA methylome in heart failure, and showed that alteration of the DNA methylome may affect chromatin remodeling. Gray color represents known information from other publications. How cardiac metabolism affects DNA methylation specifically, as indicated by dashed arrows, remains unclear and will be answered in the future studies.

**Table 3-1: Sequencing depth and CpG coverage for each sample**

| Mouse strain | Treatment | # of reads 100bp/read | # of mapped reads | Mapability | Covered CpGs | Median coverage |
|---|---|---|---|---|---|---|
| BALB/cJ | Control | 200,156,685 | 71,690,017 | 35.82% | 1,283,325 | 100 |
| BALB/cJ | Control | 82,021,527 | 32,706,080 | 39.99% | 1,096,414 | 78 |
| BALB/cJ | ISO | 190,915,791 | 93,904,992 | 49.19% | 1,275,863 | 109 |
| BALB/cJ | ISO | 87,681,466 | 34,310,208 | 39.13% | 1,093,479 | 79 |
| BUB/BnJ | Control | 86,248,467 | 33,503,903 | 38.85% | 1,092,989 | 73 |
| BUB/BnJ | Control | 209,747,249 | 58,269,857 | 27.78% | 1,217,147 | 87 |
| BUB/BnJ | ISO | 84,834,698 | 33,985,555 | 40.06% | 1,088,683 | 80 |
| BUB/BnJ | ISO | 88,872,451 | 42,826,920 | 48.18% | 1,107,335 | 80 |
| BUB/BnJ | ISO | 186,157,332 | 88,058,202 | 47.30% | 1,279,830 | 103 |
| Average | | | | 40.70% | 1,170,562 | 87 |

158

**Table 3-2: Number of DMFs induced by ISO treatment**

| Number of DNA Fragments | | BALB/cJ | | | |
|---|---|---|---|---|---|
| | | → | ↓ | ↔ | Total |
| BUB/BnJ | → | 9 | 0 | 69 | 78 |
| | ↓ | 3 | 8 | 171 | 182 |
| | ↔ | 66 | 85 | 112,205 | 112,356 |
| | Total | 78 | 93 | 112,445 | 112,616 |

↑: DNA Fragments with methylation more than 5% increase.

↓: DNA Fragments with methylation more than 5% decrease.

↔: DNA Fragments with no significant difference.

**Table 3-3: Number of DMFs and genes associated with DMFs**

| | | DMFs | Genes | Enriched Pathways | P-Value |
|---|---|---|---|---|---|
| BALB/cJ | Up-regulated methylation | 78 | 67 | Dilated cardiomyopathy | 0.036 |
| | | | | Vascular smooth muscle contraction | 0.059 |
| | | | | Pathways in cancer | 0.088 |
| BALB/cJ | Down-regulated methylation | 93 | 72 | Type II diabetes mellitus | 0.007 |
| | | | | Regulation of actin cytoskeleton | 0.017 |
| | | | | Focal adhesion | 0.093 |
| | | | | Integrin signaling pathway | 0.055 |
| | | | | TGF-beta signaling pathway | 0.071 |
| BUB/BnJ | Up-regulated methylation | 78 | 63 | Cell cycle | 0.079 |
| | | | | Wnt signaling pathway | 0.029 |
| BUB/BnJ | Down-regulated methylation | 182 | 143 | Regulation of actin cytoskeleton | 0.011 |
| | | | | Chronic myeloid leukemia | 0.012 |

DMFs: Differentially methylated fragments.
Genes: Genes with DMFs in gene bodies or within 5Kbp upstream of transcription start sites.
Pathway databases: KEGG_PATHWAY and PANTHER_PATHWAY

160

# References

1.    Singal, R. & Ginder, G. D. DNA methylation. *Blood* **93,** 4059–70 (1999).

2.    Das, P. M. & Singal, R. DNA methylation and cancer. *Journal of clinical oncology* **22,** 4632–42 (2004).

3.    Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21,** 5427–40 (2002).

4.    Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 8681–6 (1999).

5.    Yamashita, K., Dai, T., Dai, Y., Yamamoto, F. & Perucho, M. Genetics supersedes epigenetics in colon cancer phenotype. *Cancer cell* **4,** 121–31 (2003).

6.    Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell* **17,** 510–22 (2010).

7.    Koczor, C. A. *et al.* Detection of differentially methylated gene promoters in failing and nonfailing human left ventricle myocardium using computation analysis. *Physiological genomics* **45,** 597–605 (2013).

8.    Movassagh, M. *et al.* Distinct epigenomic features in end-stage failing human hearts. *Circulation* **124,** 2411–2422 (2011).

9.    Haas, J. *et al.* Alterations in cardiac DNA methylation in human dilated cardiomyopathy. *EMBO molecular medicine* **5,** 413–29 (2013).

10.   Movassagh, M. *et al.* Differential DNA methylation correlates with differential expression of angiogenic factors in human heart failure. *PloS one* **5,** e8564 (2010).

11.   Van Nas, A. *et al.* The systems genetics resource: a web application to mine global data for complex disease traits. *Frontiers in genetics* **4,** 84 (2013).

12.   Chen, P.-Y. *et al.* Intrauterine calorie restriction affects placental DNA methylation and gene expression. *Physiological genomics* **45,** 565–76 (2013).

13.   Smith, Z. D., Gu, H., Bock, C., Gnirke, A. & Meissner, A. High-throughput bisulfite sequencing in mammalian genomes. *Methods* **48,** 226–32 (2009).

14.  Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4,** Article17 (2005).

15.  Langfelder, P. & Horvath, S. Tutorials for the WGCNA package. at <http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html>

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

17.  Myers, R. M. *et al.* A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9,** e1001046 (2011).

18.  Jiang, L. *et al.* Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. *Cell* **153,** 773–84 (2013).

19.  Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42,** 348–54 (2010).

20.  Franklin, S. *et al.* Quantitative analysis of the chromatin proteome in disease reveals remodeling principles and identifies high mobility group protein B2 as a regulator of hypertrophic growth. *Molecular & cellular proteomics* **11,** M111.014258 (2012).

21.  Wade, P. A. *et al.* Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature genetics* **23,** 62–6 (1999).

22.  Qian, W. *et al.* A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science* **336,** 1445–8 (2012).

23.  Movassagh, M., Vujic, A. & Foo, R. Genome-wide DNA methylation in human heart failure. *Epigenomics* **3,** 103–9 (2011).

24.  Mahmoud, S. A. & Poizat, C. Epigenetics and Chromatin Remodeling in Adult Cardiomyopathy. *The Journal of pathology* n/a–n/a (2013). doi:10.1002/path.4234

25.  Tabibiazar, R. *et al.* Mouse strain-specific differences in vascular wall gene expression and their relationship to vascular disease. *Arteriosclerosis, thrombosis, and vascular biology* **25,** 302–8 (2005).

26.  Paigen, B., Morrow, A., Brandon, C., Mitchell, D. & Holmes, P. Variation in susceptibility to atherosclerosis among inbred strains of mice. *Atherosclerosis* **57,** 65–73 (1985).

27.    Orozco, L. D. *et al.* Unraveling inflammatory responses using systems genetics and gene-environment interactions in macrophages. *Cell* **151,** 658–70 (2012).

28.    Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome biology* **12,** R10 (2011).

29.    Schilling, E., El Chartouni, C. & Rehli, M. Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. *Genome research* **19,** 2028–35 (2009).

30.    Suzuki, M. M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics* **9,** 465–76 (2008).

31.    Van Eijk, K. R. *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC genomics* **13,** 636 (2012).

32.    Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* **5,** 101–13 (2004).

33.    Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research* **21,** 1273–83 (2011).

34.    Chen, N. C. *et al.* Regulation of homocysteine metabolism and methylation in human and mouse tissues. *FASEB journal* **24,** 2804–17 (2010).

35.    Dunn, W. B. *et al.* Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate. *Metabolomics* **3,** 413–426 (2007).

36.    Ulrey, C. L., Liu, L., Andrews, L. G. & Tollefsbol, T. O. The impact of metabolism on DNA methylation. *Human molecular genetics* **14 Spec No,** R139–47 (2005).

37.    Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annual review of biochemistry* **74,** 481–514 (2005).

38.    Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324,** 930–5 (2009).