IDENTIFICATION OF TRANSCRIPTION FACTOR TARGET GENES BY INTEGRATIVE OMICS DATA ANALYSIS

by

Yuxuan Liu

APPROVED BY SUPERVISORY COMMITTEE:

Michael Q. Zhang, Chair

Stephen Spiro

Zhenyu Xuan

Min Chen

This dissertation is dedicated to my parents,

who offered unconditional support and encouragement

IDENTIFICATION OF TRANSCRIPTION FACTOR TARGET GENES BY INTEGRATIVE OMICS DATA ANALYSIS

by

YUXUAN LIU, MS

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

MOLECULAR AND CELL BIOLOGY

THE UNIVERSITY OF TEXAS AT DALLAS

December 2016

ACKNOWLEDGMENTS

I could not be where I am without the support of many people. First of all, I would like to express my thanks to my advisor, Dr. Michael Zhang, for his continuous support. Not only did he offer great help in my research but he also encouraged me to think independently and rigorously, from which I will benefit for the rest of my career. I would also like to take this opportunity to express my sincere gratitude to my committee members, Dr. Stephen Spiro, Dr. Zhenyu Xuan and Dr. Min Chen. I thank Dr. Spiro for believing in me and giving me a chance to join the project in his lab. He provided the best help and training when I was working in his lab. I thank Dr. Xuan for all the valuable discussions and insightful comments. Whenever I had questions and went to him, he was always there with great patience. I thank Dr. Min Chen for his expert suggestions in mathematics. His questions and ideas inspired me to widen my research from a different perspective.

I will forever be thankful to my dad and mom for their unconditional love and support. I thank them for being by my side to cheer me up all the time. Without their encouragement and inspiration, I couldn't go through all those hard times.

Last but not least I would also like to express my thanks to my fellow labmates and friends for their support and all the happy times that we have had for the last six years.

September 2016

PREFACE

This dissertation was produced in accordance with guidelines which permit the inclusion as part of the dissertation the text of an original paper or papers submitted for publication. The dissertation must still conform to all other requirements explained in the "Guide for the Preparation of Master's Theses and Doctoral Dissertations at The University of Texas at Dallas." It must include a comprehensive abstract, a full introduction and literature review, and a final overall conclusion. Additional material (procedural and design data as well as descriptions of equipment) must be provided in sufficient detail to allow a clear and precise judgment to be made of the importance and originality of the research reported.

It is acceptable for this dissertation to include as chapters authentic copies of papers already published, provided these meet type size, margin, and legibility requirements. In such cases, connecting texts which provide logical bridges between different manuscripts are mandatory. Where the student is not the sole author of a manuscript, the student is required to make an explicit statement in the introductory material to that manuscript describing the student's contribution to the work and acknowledging the contribution of the other author(s). The signatures of the Supervising Committee which precede all other material in the dissertation attest to the accuracy of this statement.

Portions of the text of the second chapter are reprinted from our work (Mehta et al., 2015).

IDENTIFICATION OF TRANSCRIPTION FACTOR TARGET GENES BY INTEGRATIVE OMICS DATA ANALYSIS

Publication No. _____

Yuxuan Liu, PhD
The University of Texas at Dallas, 2016

Supervising Professor: Michael Q. Zhang

Transcription factors (TFs) are proteins that control the rate of transcription. They are main regulators of gene transcription. Knowing their targets is very important for understanding developmental processes, cellular stress response and genetic causes of disease. Most of prokaryotic genome is coding and TF binding sites are usually close to genes. However, for the mammalian system, most of its genome is non-coding and TFs usually bind to gene distal regions and they regulate gene transcription via chromosome looping. In our study, we were trying to identify TF targets in both the simple prokaryotic system and the complex mammalian system by integrative omics data analysis. Considering the differences between prokaryotic and mammalian systems, we integrated different omics data in each system to identify TF targets. In prokaryotes, DNA is organized in operon which contains a cluster of genes under the control of a single promoter. There is stronger correlation between TF binding and gene expression in prokaryotes than in the mammalian system. And TF motif in prokaryotes is usually longer and more specific than that in eukaryotes. Therefore, in prokaryotes, we integrated TF genome-wide binding data, expression data and motif infor-

mation to identify TF targets. We conducted our study using TF NsrR and tried to identify its genome-wide binding targets in Uropathogenic *Escherchia coli* (UPEC) CFT073 to understand UPEC's response to nitric oxide. In the mammalian system, DNA is wrapped on histone to form nucleosome. Histone modification and chromatin accessibility are important for transcription factor binding. DNA can form looping interactions to regulate gene expression. Therefore for TF targets identification in the mammalian system, we integrated TF genome-wide binding data, epigenetic data and chromatin looping interaction data. We built a classifier to predict TP53-associated looping interactions and genome-wide long-distance targets of TP53.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Transcription factors (TFs) are proteins that typically bind to specific DNA sequence to regulate gene transcription. They are main regulators of gene expression. Transcriptional activators can stimulate the transcription of target genes by binding to DNA elements of enhancer or promoter. Transcriptional repressors, as opposed to activator, bind to specific DNA elements to prevent transcription of target genes. Knowing TF targets is important for us to understand complex developmental processes and cellular environmental responses. It also allows us to better understand genetic cause of diseases by identifying TF binding sites. Mutations in regulatory regions like TF binding sites were found to be correlated with complex diseases like cancer. For example mutations in binding sites of CEBP factors were highly enriched in cancer and mutations of CEBP sites likely alter transcriptional regulation (Melton et al., 2015).

In our study, we were motivated to identify TF targets in both the simple prokaryotic system and the complex mammalian system by integrative omics data analysis. Chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq) allows for identifying genome-wide binding sites of DNA-associated proteins such as TFs. The prokaryotic genome is efficiently compacted by protein coding sequences. In *E. coli*, for example, only 12% of the genomes is occupied by non-coding DNA (Rogozin et al., 2002). TF binding sites in prokaryotes are usually close to genes, whereas the mammalian genome is composed mostly of non-coding elements. For instance, over 98% human genome is non protein coding (The International Human Genome Sequencing Consortium, 2004). Also, many TF binding sites in the mammalian system are in gene distal regions. TFs regulate gene expression through

chromosome looping interactions. Therefore, in the mammalian system, it is important to accurately assign long-distance targets to TFs.

To identify TF targets in prokaryotic and mammalian systems, we integrated different omics data in each system according to the differences of two systems. In prokaryotes, DNA is organized in operon which contains a cluster of genes under the control of a single promoter and the correlation between TF binding and gene expression is stronger than that in the mammalian system. Besides, binding motif of TFs in prokaryotes is often longer and more specific. Therefore, we integrated TF genome-wide binding data, expression data and motif information to identify TF targets in prokaryotes. We used TF NsrR in Uropathogenic *Escherchia coli* (UPEC) CFT073 to conduct our study and tried to identify its genome-wide targets. NsrR is nitric oxide (NO) sensitive, so knowing its targets can help us to understand how UPEC respond to NO stress. In the mammalian system, DNA is wrapped on histone to form nucleosome. Histone modification and chromatin accessibility is important for transcription factor binding. And DNA can form looping interactions to regulate gene expression. Therefore for long-distance TF targets identification in the mammalian system, we integrated ChIP-seq data, epigenetic data and chromatin looping interaction data. We built a classifier to predict TP53-associated looping interactions and its genome-wide long-distance targets. TP53 is a famous tumor suppressor and knowing its targets can help us to understand complex disease like cancer. The biological background of NsrR and TP53 will be introduced separately in the following parts.

1.1 Identification of genome-wide targets of NsrR in UPEC CFT073

1.1.1 NO detoxification system in *E. coli*

UPEC CFT073 is a pathogenic strain that accounts for 80% of all symptomatic and asymptomatic urinary tract infections (UTIs) (Roos and Klemm, 2006). UTI can cause the migration of neutrophils from the blood to the urine (Godaly et al., 2001), which exposes UPEC

to the defense mechanisms of the innate immune system. NO, produced by the inducible nitric oxide synthase (iNOS) in phagocytic cells through the oxidation of arginine, is part of an effective host immune response to infection (Fang, 1997; Fang and Vazquez-Torres, 2002; Mowat et al., 2010). The antimicrobial effect of NO is due to its ability to target proteins containing iron−sulfur clusters, haem and thiols (Fang, 1997; Kim et al., 1995; Ren et al., 2008). Bacteria that inhabit the host environment utilize NO detoxification strategies to convert NO to less toxic compounds. There are three known NO detoxification enzymes in *E. coli*. Flavohaemoglobin (Hmp) dioxygenates NO to nitrate and reduces NO to N_2O under anaerobic condition (Gardner and Gardner, 2002; Hausladen et al., 2001). Flavorubredoxin (FlRd or NorV) coupled with NorW, a NADH-linked reductase, catalyzes the reduction of NO to N_2O (Gardner et al., 2002). The periplasmic nitrite reductase (NrfA) can detoxify NO by reducing it to ammonia in the absence of oxygen (Poock et al., 2002). Figure 1.1 shows a summary of NO consumption pathways and the regulation of enzymes involved in NO consumption.

1.1.2 NO sensitive regulator NsrR

One important regulatory proteins in the NO response in *E.coli* is NsrR, shown in Figure 1.1. NsrR is a nitric oxide-sensitive repressor of transcription (Partridge et al., 2009; Bodenmiller and Spiro, 2006a) that belongs to the Rrf2 family of transcriptional repressors (Bodenmiller and Spiro, 2006a). It contains Fe-S clusters that can react with NO (Yukl et al., 2008; Tucker et al., 2008). Binding of NO to the Fe-S clusters leads to the loss of DNA-binding activity of NsrR (Tucker et al., 2008; Rankin et al., 2008; Crack et al., 2016). NsrR plays a very important role in the response to NO by regulating the expression of different genes. Flavohaemoglobin (encoded by the *hmp* gene), one of the key NO detoxifying enzymes, is known to be regulated by NsrR. Apart from *hmp*, the NsrR regulon contains various genes implicated in the NO stress response, such as ytfE, hcp and the *nrf* operon (Tucker et al.,

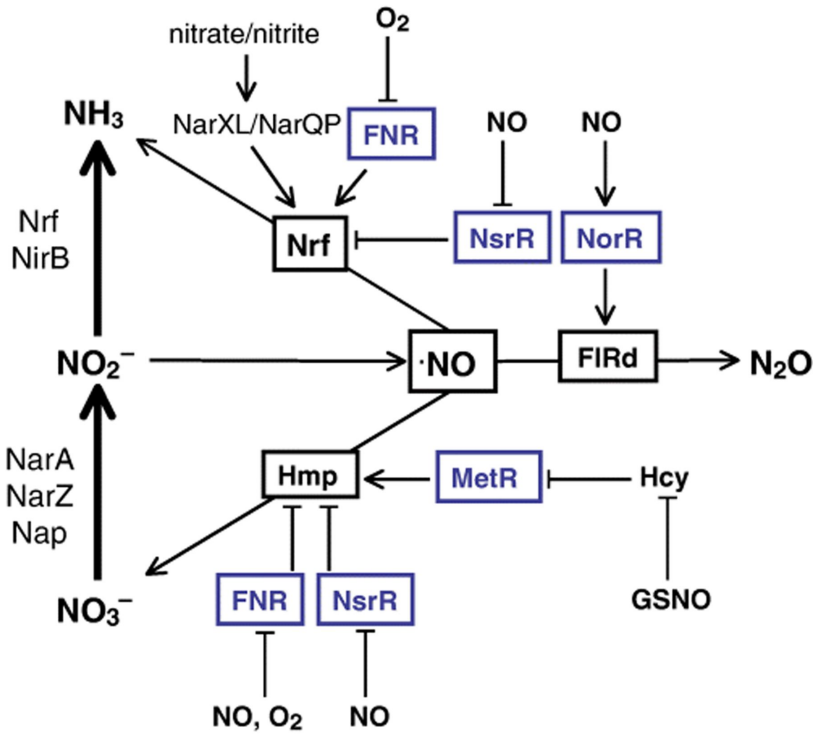

Figure 1.1. NO consumption pathway and the regulation of enzymes involved in NO consumption adapted from (Spiro, 2007). Hmp dioxygenates NO to nitrate. Nrf reduces NO to ammonia and flavorubredoxin (FlRd) reduces NO to nitrous oxide. Blue boxes represent regulators that regulate the expression of NO consumption enzymes. Positive regulation is denoted by arrows, negative regulation by perpendicular lines.

2010). Therefore, identification of members of the NsrR regulon would help us to understand how *E. coli* responds to NO. Chromatin immunoprecipitation and microarray analysis (ChIP-chip) has been performed to identify members of the NsrR regulon in non pathogenic *E.coli* K12. UPEC is more resistant to the stress imposed by acidified nitrite than K-12 strains of *E. coli* (Bower and Mulvey, 2006) and may also be more resistant to a prolonged exposure to NO (Svensson et al., 2006), in which case toxicity might be due to N radicals derived from NO. The genome of strain UPEC CFT073 is larger than that of *E.coli* K-12 strain by ∼0.6Mb. Some genes are only present in K12 and some are unique to CFT073. Compared to K12, five unique inserted prophage genomes which include large proportion of virulence or virulence-associated genes account for the difference in the CFT073 genome (Hacker et al.,

1997; Luo et al., 2009). CFT073-specific islands insert into conserved backbone gene regions in an extensive mosaic manner (Welch et al., 2002; Luo et al., 2009). The annotation related to virulence and virulence-associated genes includes 12 types of fimbriae, 7 autotransporters, and toxin operons such as *hlyCABD* and *upxBDA* (Welch et al., 2002; Luo et al., 2009). Thus we were motivated to understand UPEC biology and how it responds to NO. Here in this study, to understand the NO response in UPEC, we identified the NsrR regulon in CFT073 using ChIP-seq to map NsrR binding sites in the CFT073 genome.

1.2 Identification of genome-wide long-distance targets of TP53

1.2.1 Tumor suppressor TP53

TP53 is a transcription factor that is well known for its function as tumor suppressor. It can protect cells from uncontrolled proliferation and genotoxic stress like DNA damage (Vousden and Lane, 2007). Loss or mutation of TP53 is very common in human cancers and it correlates strongly with the increase of susceptibility to cancer (Vousden and Lane, 2007). TP53 can respond to varied stresses, such as hypoxia, oxidative stress, uncontrolled cell proliferation, and genotoxic stresses (Giaccia and Kastan, 1998; Hu et al., 2012) through the regulation of genes involved in cell cycle arrest, senescence, apoptosis and DNA repair. Figure 1.2 shows the responses of TP53 to diverse stress signals. The cellular responses of TP53 largely rely on its action as transcription activator to induce the expression of different genes. Figure 1.3 shows target genes activated by TP53 in response to different stress signals.

1.2.2 TP53-dependent gene regulation

As mentioned above, TP53 plays very important roles in the response to diverse stresses through acting as a transcription factor to activate gene expressions. Traditionally, it has

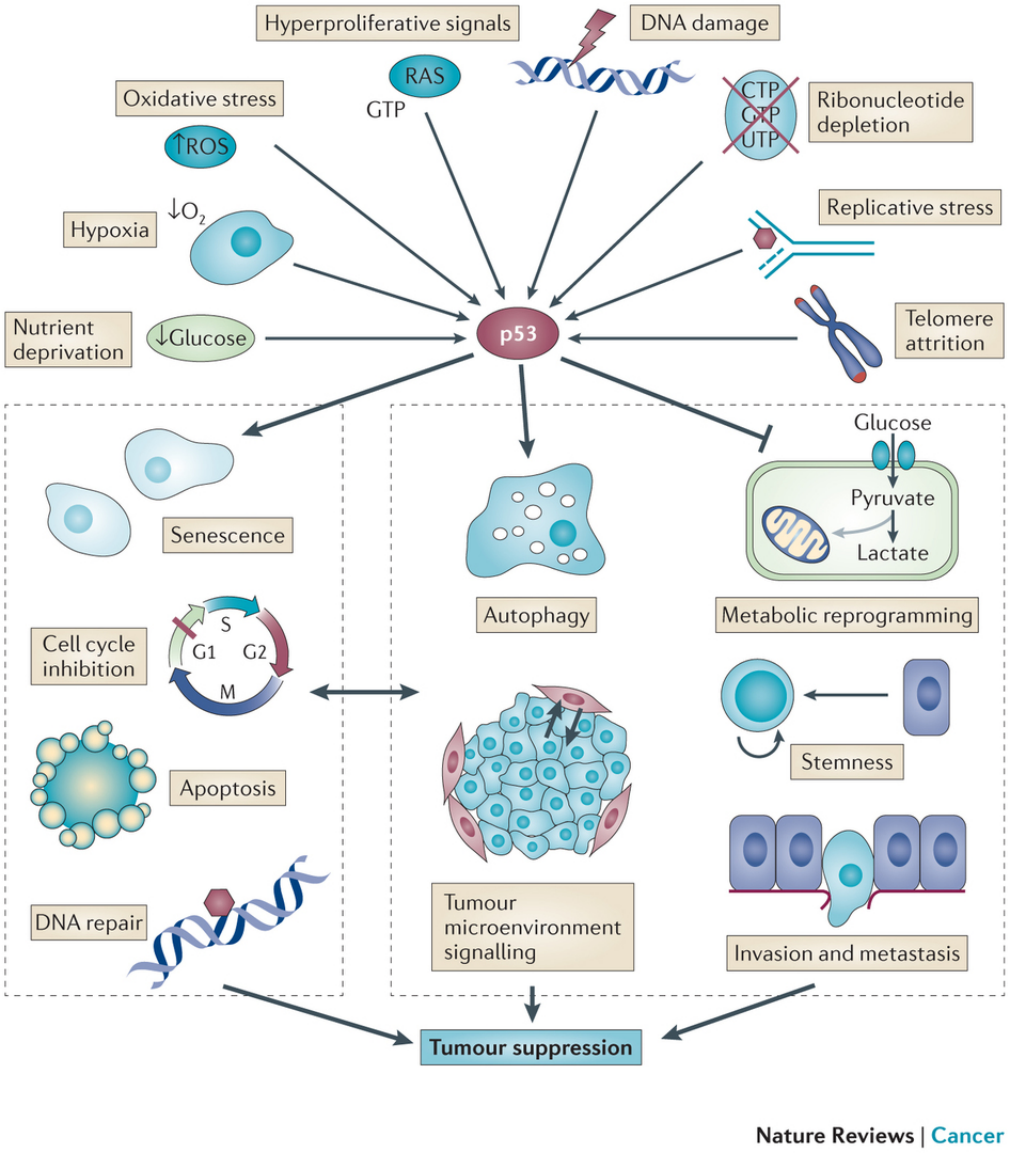

Figure 1.2. Response of TP53 to different stress signals adapted from (Bieging et al., 2014)

been thought that TP53 regulates target gene expressions by binding to the promoter region. However, TP53 genome-wide binding profile revealed that the majority of TP53 binding sites fall to enhancer regions with enriched histone markers of H3K4me1 and distal regions lacking of histone markers of either H3K4me1 or H3K4me3 (Sammons et al., 2015). The recognition of TP53 to non-promoter regions especially to regulatory enhancers might be a general function of TP53. Several studies have been performed to study the regulation of TP53

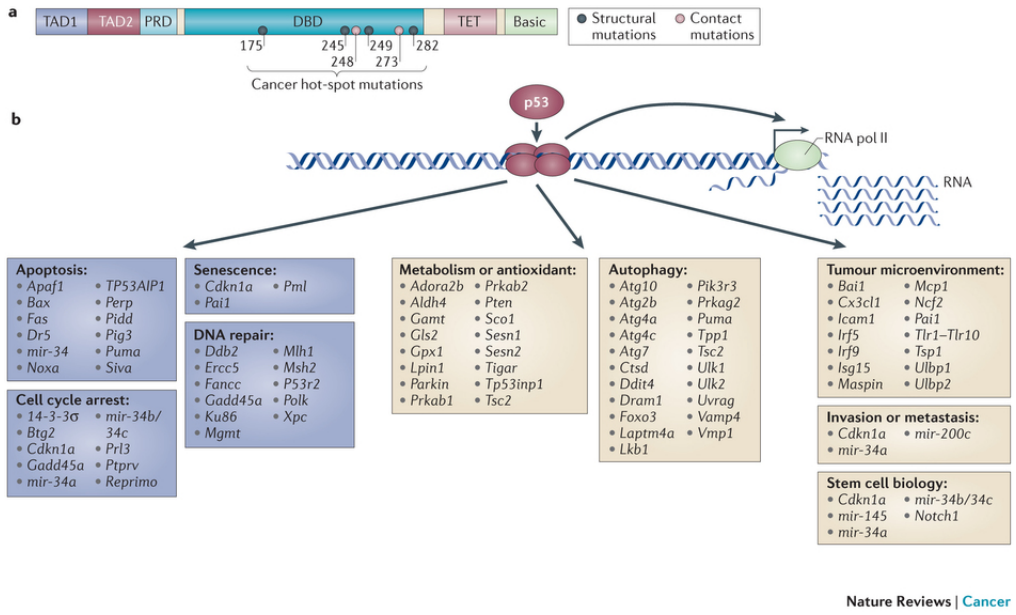

Figure 1.3. Different target genes activated by TP53 in response to different stress signals adapted from (Bieging et al., 2014). a, TP53 protein domains. b, List of key TP53-induced targets involved in processes that are important for tumor suppression.

bound enhancers in both *Drosophila* and human fibroblasts. It was found that TP53-bound enhancer can regulate target gene expression through chromosome looping. A single TP53 enhancer examined in *Drosophila* was discovered to regulate multiple gene expression (Link et al., 2013). Physical looping interactions between that TP53 bound enhancer and its target genes were confirmed by digital chromosome conformation capture (d3C) in combined with fluorescent in situ hybridization(FISH) (Link et al., 2013). Similarly in human fibroblasts, several TP53 bound enhancers were found to convey the regulation of multiple distant genes through intrachromosomal interactions (Melo et al., 2013). And chromosome conformation capture (4C) technology together with next-generation sequencing was applied to confirm the existence of physical interactions (Melo et al., 2013). From the studies of individual TP53 bound enhancers in distant organisms above, it might be a common feature for TP53 bound enhancers to regulate long-distance target genes via chromosome looping. However, its long-distance targets have not been identified genome-widely yet. The development of

high throughput chromatin interaction detection technics can help us to tackle that problem so that we can better understand the function of TP53.

1.2.3 Genome-wide long-range chromatin interaction detection techniques

Capturing chromosome conformation (3C) was first developed to detect the contact frequency between two genomic loci (Dekker et al., 2002). Then related technology 4C (chromosome conformation capture(3C)-on-chip) (Simonis et al., 2006) and 5C (chromosome conformation capture carbon copy) (Dostie et al., 2006) were developed to detect interactions of a given genome locus with all the other genomic loci and interactions between multiple loci. With the development of next generation sequencing, two high-throughput technics Hi-C (Lieberman-Aiden et al., 2009)and ChIA-PET(Fullwood et al., 2009) were developed to allow us to detect genome wide chromatin interactions. Hi-C was adjusted from 3C and biotin-labeled restriction ends were ligated and sequenced. It can detect all the concurring interactions regardless of the proteins that link the interactions. The resolution of Hi-C is about 1Mb based on 10 million paired-end reads (de Wit and de Laat, 2012) which is too low to study interactions between some specific elements like promoter-enhancer interactions. ChIA-PET combines chromatin immunoprecipitation (ChIP) and 3C to analyze the ligation junctions that are enriched by the antibody against the protein of interest. The resolution of ChIA-PET is higher than that of Hi-C, but instead of collecting all the possible interactions, it can only collect interactions associated with some specific protein.

1.2.4 TP53 associated long-range interaction prediction

TP53 ChIA-PET data which directly provides us genome-wide chromatin interaction mediated by TP53 is not available yet. So in order to identify long-range targets of TP53 bound enhancers, we integrated TP53 ChIP-seq data and genome-wide chromatin interaction data. In previous study of TP53 bound enhancer in *Drosophila*, d3C was performed in both wide

type and TP53$^-$ animals. It was found that d3C interaction patterns were similar between wide type and TP53$^-$ animals suggesting that TP53 was not required for the formation of chromatin looping (Link et al., 2013). Similarly, in human fibroblast, it was also found that the long-range interactions existed in the wide-type cells are also present in cells with TP53 stably knocked down (Melo et al., 2013). Similar observations in distant organisms suggest that instead of initiating new loops, TP53 may mainly act on the pre-existed chromatin loops to regulate target gene expressions. Therefore, integrating TP53 ChIP-seq data and pre-existing chromatin interaction data from Hi-C or ChIA-PET can help us to identify some putative functional TP53 bound enhancers and their long-range targets.

However, genome-wide chromatin interaction data is not available for every cell line. We therefore were motivated to predict functional TP53 binding sites that have the potential to regulate gene expression via chromatin looping interactions in cell lines of which genome-wide chromatin interaction data is not available. We were inspired by a study of predicting ERα associated looping interactions (He et al., 2014). We learned pre-configured looping interactions through pol2 ChIA-PET data by integrating multiple transcription factors binding and histone modification profiles. We tried to predict TP53 associated long-range interactions solely from TP53 ChIP-seq and epigenomic data when high throughput chromatin interaction data such as pol2 ChIA-PET is absent. We are able to learn chromatin interaction from epigenetic features for the following reasons. Active enhancers that can form looping interactions are often occupied by multiple transcription factors and form a large protein complex (Borggrefe and Yue, 2011; Young, 2011) which may cause a different epigenetic pattern compared to binding sites with no looping interactions. Pol2 ChIA-PET data showed that basal promoter with pol2 binding but no chromatin interaction displayed different histone modification pattern with pol2 binding site that can form chromatin interactions (Young, 2011). It implies that looking at epigenetic features of TP53 binding sites will be informative for us to predict TP53 binding sites that are associated with looping interactions mediated by pol2.

We performed our analysis in MCF7 because it is a well-characterized human cancer cell model with complementary datasets including TP53 ChIP-seq, ChIA-PET, RNA-Seq and GRO-Seq datasets. We integrated TP53 ChIP-seq and pol2 ChIA-PET data in MCF7 and uncovered hidden features buried beneath TP53 binding sites associated with pol2 mediated long-range interactions through the combination of TF binding, histone modification profiles and open chromatin conformation data . We determined epgenetic features that could discriminate loop associated and non-loop associated TP53 binding sites. Then we used these features to build classifier to predict loop associated TP53 binding sites and to develop DNA looping prediction algorithm to predict TP53 interacting clusters. This allowed us to discover TP53 associated long-range interactions even when genome-wide chromatin interaction data is not available and serve as a complement to the complicated and costly TP53 ChIA-PET experiments.

CHAPTER 2

THE NSRR REGULON OF *E. COLI* CFT073

2.1 Materials and Methods

Chromatin immunoprecipitation (ChIP) was performed as described previously on cultures grown aerobically in L broth to mid-exponential phase (Efromovich et al., 2008). Chromatin samples were sheared by sonication to within a size range of 200-600 bp. DNA fragments were treated using an Epicentre End-It DNA End Repair kit and 3' A overhangs were added with DNA polymerase I (Klenow fragment). Adapters from the IlluminaTruSeq DNA sample preparation kit were ligated using LigaFast (Promega) and DNAs were amplified by PCR using primers provided in the IlluminaTruSeq DNA sample preparation kit and Phusion DNA polymerase (NEB). Products of the ligation reaction and PCR amplification in the range 300-400 bp were purified by 2% agarose gel electrophoresis. DNA concentrations were measured using Qubit dsDNA HS Assay kits (Invitrogen). DNA sequencing was done on the Miseq (Illumina) platform following the manufacturer's instructions. For one replicate, a single-end reads, 60 bp run was performed. For the other two replicates, a paired-end reads, 100 bp run was performed. Sequence reads were aligned with the published *E. coli* CFT073 genome (AE014075.1) using the software package Bowtie with the parameters bowtie -k 1 -X 500 -m 1 (Langmead et al., 2009). Peaks were identified using the peak finding algorithm of MACS2 (Zhang et al., 2008), with default parameters.

For motif analysis, multiple Em for motif elicitation (MEME) was used to identify over-represented sequences (Bailey and Elkan, 1994). PatSer was used to search the genome for the presence of the NsrR position-specific weight matrix (PSWM) (Hertz and Stormo, 1999). A precision-recall curve was constructed to determine the optimal threshold for predicting

high-quality NsrR binding sites. Precision was defined as the ratio of true positives (locations with an NsrR ChIP-seq peak and a predicted NsrR binding site) to true positives plus false positives (locations with a predicted NsrR binding site but no NsrR ChIP-seq peak). Recall was defined as the ratio of true positives divided by true positives plus false negatives (locations with an NsrR ChIP-seq peak but no NsrR predicted binding site)(Myers et al., 2013).

ChIP-seq data have been deposited in the GEO database, accession number GSE69829.

2.2 Results

2.2.1 The NsrR regulon of *E. coli* CFT073

As the *E. coli* CFT073 genome is ∼0.6 Mb larger than that of *E. coli* K-12, it is of interest to determine the extent to which regulatory networks of the two organisms differ. Thus, we used ChIP-seq to identify NsrR binding sites in the *E. coli* CFT073 genome. Cultures expressing 3X flag-tagged NsrR were grown aerobically. After ChIP, libraries constructed from precipitated DNAs were sequenced using the Illumina Miseq platform. The peak finding algorithm MACS2 was used to identify putative NsrR binding sites, with a false discovery rate (FDR) of 0.01. Ninety-four significant peaks (-log $_{10}$(P-value)>10 with fold enrichment greater than 2) were identified in at least two of the three biological replicates as shown in Figure 2.1. In total, 52 % of the binding sites (49 of 94) in *E. coli* CFT073 were located in putative promoter regions (within 350 bp of the start codon) and the remaining 48 % were found either within coding regions or between the coding regions of convergent genes. These potentially functional 49 NsrR binding sites are shown in Table 2.1.

Table 2.1. NsrR binding sites in *E. coli* CFT073 genome

Coordinate[a]	**-log10(P-value)**[b]	**Fold enrichment**[c]	**Flanking genes**[d]	**Distance from summit to start codon**[e]	**Possible NsrR site**[f]	**Sequence**	**PatSer score**	**PatSer ln(P-value)**[g]
1890380	503.52	19.16	*grxD* (<) *mepH* (>)	-138 (*grxD*)	*grxD* (74)	AAATGTTATTT	7	-8.98
					grxD (121)	TTGTTGCATTT	7.19	-9.15
					grxD (111)	AAAATACGTTT	5.44	-7.45
3130600	523.77	16.18	*hycB* (<) hycA (<)	-59 (*hycB*)	*hycB* (54)	AAATGTCATTT	7.6	-9.67
3645158	287.94	11.46	*folB* (<) *plsY* (>)	28 (*folB*)				
967971	248.24	12.00	*hcp* (<) *ybjE* (<)	-22 (*hcp*)	*hcp* (15)	AAGTTATATTT	9.19	-11.59
					hcp (27)	AACATGTATAT	8.78	-11.14
					hcp (11)	AAGTTGCATTA	8.92	-11.34
5049151	247.18	9.65	*ytfE* (<) *ytfF* (<)	-50 (*ytfE*)	*ytfE* (33)	AAGATGCATTT	10.92	-15.25
					ytfE (45)	AAGATGCATTT	10.92	-15.25
					ytfE (128)	CAGATTCAGTT	4.07	-6.32
115339	218.24	10.60	mutT (>) *c0118* (>)	18 (*c0118*)	*c0118* (77)	CATTTGCATAT	3.86	-6.16
					c0118 (6)	AAGGTGCAGTT	7.70	-9.79
227553	163.73	9.01	*c0233* (<) *yaeF* (<)	2 (*c0233*)	*c0233* (29)	AAGTTTTACTT	7.38	-9.39
					c0233 (17)	AACATTCATTT	9.43	-12.21
					c0233 (1)	AAGGTGCAGTT	7.70	-9.79
4375989	117.47	5.31	*dgoK* (<) *dgoR* (<)	222 (*dgoK*)				
3967122	116.05	5.92	*yhgF* (>) feoA*B*(>)	95 (*feoA*)	*feoA* (38)			
2314458	88.47	6.45	*c2470* (<) *c2471* (>)	-66 (*c2471*)	*c2471* (122)	ATGTGATATTT	5.63	-7.62

Table 2.1. NsrR binding sites in *E. coli* CFT073 genome

Coordinate[a]	**-log10(P-value)**[b]	**Fold enrichment**[c]	**Flanking genes**[d]	**Distance from summit to start codon**[e]	**Possible NsrR site**[f]	**Sequence**	**PatSer score**	**PatSer ln(P-value)**[g]
					c2471 (84)	AAGTTTCATGT	7.45	-9.49
					c2471 (72)	TTGATGTTTTT	3.92	-6.21
					c2471 (56)	AGCTTGTATTT	4.55	-6.69
4399249	72.9	4.67	*yieH* (>) *cbrB* (>)	4 (*cbrB*)	*cbrB* (18)	TACTTACCTTT	3.94	-6.22
4224075	72.8	4.56	*waaH* (<) *tdh* (<)	191 (*waaH*)				
698961	54.72	4.83	*ccrB* (<) *ybeM* (>)	345 (*ybeM*)				
4883923	49.17	3.65	*phnC* (<) *phnB* (<)	274 (*phnC*)				
3639613	48.43	3.91	*ygiF* (<) *c3803* (<)	232 (*ygiF*)				
2654737	48.2	4.66	*arnC* (>) *arnA* (>)	316 (*arnA*)				
4037958	46.57	3.97	*livJ* (<) *rpoH* (<)	250 (*livJ*)				
4840905	45.39	3.67	*acs* (<) c5065-*nrfA* (>)	-25 (*c5065*)	*c5065* (94)	AACATGCAGTT	8.12	-10.23
					c5065 (42)	AAGTGGTATTT	8.71	-11.03
					c5065 (31)	TACATGCACTT	6.82	-8.79
					c5065 (4)	ACATTCATAGT	5.41	-7.42
796704	44.53	4.08	*cydB* (>) *c0813* (<) *ybgE* (>)	177 (*c0813*)				
3760376	34.92	3.61	*hflB* (<) *c3935* (<)	152 (*hflB*)				

Table 2.1. NsrR binding sites in *E. coli* CFT073 genome

Coordinate[a]	**-log10(P-value)**[b]	**Fold enrichment**[c]	**Flanking genes**[d]	**Distance from summit to start codon**[e]	**Possible NsrR site**[f]	**Sequence**	**PatSer score**	**PatSer ln(P-value)**[g]
4476183	34.04	2.95	*wecE* (>)*c4712* (<) *wzxE* (>)	-259 (*wzxE*)				
2350649	32.09	3.82	*c2513* (<) *c2514* (>)	200 (*c2514*)				
2948718	32.08	3.79	*glyA* (<) *hmp* (>)	34 (*hmp*)	*hmp* (44)	AAGATGCATTT	10.92	-15.25
					hmp (19)	AAGATGCAAAA	5.00	-7.08
4708187	28.99	3.45	*thrT* (>) *tufB* (>)	-5 (*tufB*)				
2694874	27.98	3.58	*yfbT* (<) *yfbU* (<)	-281 (*yfbT*)				
1088139	27.33	2.24	*yccM* (<) *torS* (<)	-36 (*yccM*)	*yccM* (24)	AAGTTGCATAC	6.86	-8.83
					yccM (36)	TAGTGGCATTT	7.63	-9.69
					yccM (50)	TAGTTGTTCTT	3.97	-6.25
4569962	26.65	3.45	*c4805* (>) *yihF* (>)	314 (*yihF*)				
628066	24.07	3.35	*ydfM* (>) *c0650* (>)	-52 (*c0650*)	*c0650* (50)	AAGATGTATCG	3.95	-6.23
4468281	23.37	2.71	*c4703* (>) *rfe* (>)	201 (*rfe*)				
4342139	23.29	2.93	*c4579* (<) *c4580* (>)	123 (*c4580*)				
2639159	23.00	3.28	*glpT* (<) *glpA* (>)	-41 (*glpA*)	*glpA* (93)	ACGTTTCACTT	4.95	-7.05
					glpA (50)	AACATGAATTG	4.74	-6.86
4792020	22.78	2.78	*zur* (<) *yjbN* (>)	-169 (*yjbN*)				
3999621	21.56	3.05	*c4214* (<) *glgP* (<)	3 (*c4214*)	*c4214* (5)	AAGGTATAAAT	4.35	-6.54

Table 2.1. NsrR binding sites in *E. coli* CFT073 genome

Coordinate[a]	**-log10(P-value)**[b]	**Fold enrichment**[c]	**Flanking genes**[d]	**Distance from summit to start codon**[e]	**Possible NsrR site**[f]	**Sequence**	**PatSer score**	**PatSer ln(P-value)**[g]
					c4214 (69)	AAGTTATATCT	5.80	-7.79
5208290	20.47	2.75	*deoA* (>) *deoB* (>)	18 (*deoB*)				
3131567	17.96	2.67	*hycA* (<) *hypAB* (>)	-126 (*hypB*)	*hypB* (167)	CGAGGTGCAGT	4.13	-6.37
4704549	17.63	2.75	*rrfB* (>) *murB* (>)	244 (*murB*)				
986391	16.91	2.87	*trxB* (<) *lrp* (>)	-92 (*trxB*)	*trxB* (32)	TACTTAAATTT	4.92	-7.01
					trxB (100)	ATGTTGTACTA	4.41	-6.58
					trxB (112)	AACATCGATTT	4.90	-7.00
4961829	15.97	2.80	*c5205* (<) *c5206* (<)	4 (*c5205*)	*c5205* (7)	AACAGGTATTA	6.29	-8.24
3335185	15.31	2.44	*recJ* (<) *dsbC* (<)	-106 (*recJ*)				
240582	15.04	2.73	*aspU* (>) *dkgB* (>)	14 (*dkgB*)	*dkgB* (18)	AAATAGCATTA	4.63	-6.76
					dkgB (6)	AAGAGGCATAT	8.32	-10.55
3016197	14.59	2.70	*recN* (>) *bamE* (>)	6 (*bamE*)	*bamE* (65)	AAGGTCTATTA	5.21	-7.24
					bamE (46)	ATATTACAGAT	3.74	-6.06
4616876	14.49	2.17	*rhaB* (<) *c4854* (>)	229 (*rhaB*)				
2522542	14.14	2.67	*yohJ* (>) *yohK* (>)	-172 (*yohK*)				
3786570	14.05	2.36	*arcB* (<) *yhcC* (<)	-53 (*arcB*)				
3391606	13.75	2.47	*yggT* (>) yggT (>)	-159 (yggT)	yggT (172)	TACAGCCATTT	4.94	-7.03
2961332	13.69	2.64	*pgpC* (<) *c3084* (<)	64 (*pgpC*)				
3795223	13.68	2.25	*nanK* (<) *nanE* (<)	322 (*nanK*)				

Table 2.1. NsrR binding sites in *E. coli* CFT073 genome

Coordinate[a]	**-log10(P-value)**[b]	**Fold enrichment**[c]	**Flanking genes**[d]	**Distance from summit to start codon**[e]	**Possible NsrR site**[f]	**Sequence**	**PatSer score**	**PatSer ln(P-value)**[g]
5139372	12.82	2.57	*fimI* (>) *fimC* (>)	337 (*fimC*)				
4779126	11.28	2.07	*malF* (<) *malE* (<)	-142 (*malF*)	*malF* (82)	ACATACGTTTC	3.98	-6.26
					malF (163)	AAGATGCACAG	5.00	-7.08

a Genomic location of the summit of ChIP-seq peak.

b -log10(P-value) of each peak called by MACS2.

c Fold enrichment of each peak calculated by MACS2.

d The genes flanking the ChIP-seq summit.

e The distance between the peak summit and start codon of the nearest downstream gene.

f Possible NsrR binding motifs identified by PatSer, and the distance from the motif to the start codon of the gene. Only sites upstream of start codons are shown.

g ln(P-value) associated with the PatSer score for each predicted NsrR site.

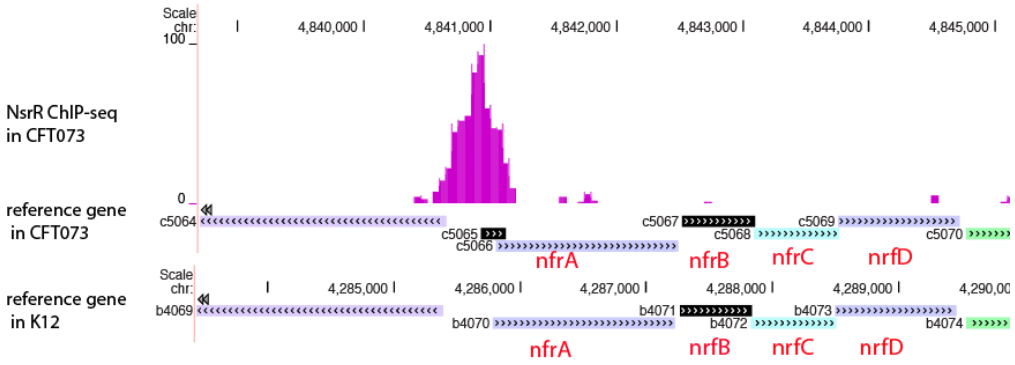

Figure 2.1. UCSC Genome Browser track view of nrfA locus.

The presence of promoter-associated NsrR binding sites identifies target genes that potentially belong to the NsrR regulon. Of these promoters bound by NsrR in vivo (Table 2.1), 19 (*grxD*, *hypA*, *ytfE*, *ygiG/folB*, *hmp*, *ybjW/hcp*, feoA, *ybeM*, *yihF*, *yccM*, *yibD/waaH*, *yieI*, *yohK*, *ygiF*, *trxB*, *yggS/ yggT*, *dgoK*, *rfe*, *yfhB/pgpC*) were identified in a previous ChIP-chip analysis of NsrR binding sites in *E. coli* K-12 (Partridge et al., 2009). Twenty of the remaining sites are associated with genes (*hycB*, *phnC*, *arnA*, *livJ*, *wzxE*, *tufB*, *yfbT*, *glpA*, *yjbN*, *deoB*, *murB*, *recJ*, *dkgB*, *c3139*, *rhaB*, *arcB*, *c3976/ nanK*, *fimC*, *c3934/ hflB* and *malF*) that have homologues in *E. coli* K-12, and 10 (*c0118*, *c0233*, *c2471*,*c5065*, *c0813*, *c2514*, *c0650*, *c4580*, *c4214* and *c5205*) are specific to *E. coli* CFT073.

In *E. coli* K-12, the *nrfA* promoter is bound by NsrR (Partridge et al., 2009) and is repressed by NsrR according to microarray and reporter fusion data (Filenko et al., 2007). In our ChIP-seq data, NsrR binding was also detected upstream of the transcription unit that includes *nrfA* as shown in Figure 2.1. In strain CFT073, an additional gene upstream of *nrfA* (*c5065*) is predicted to be co-expressed with *nrfABCD*. The *c5065* gene encodes a small protein of 65 aa residues. We have confirmed the sequence of this reading frame in the CFT073 genome. The genome location and expression pattern of the *c5065* gene suggest that its product may have a role in the response to NO stress in CFT073.

Nineteen of the 49 potential NsrR targets show differential expression with a fold change greater than 1.5 for the CFT073 3X mutant strain in the presence of a physiological source of NO (Table 2.2). The *hycB*, *c0118*, *feoA*, *ybeM*, *c5065*, *c0813*, *c2514*, *glpA*, *deoB*, *hypAB*, *yohJ*K and *yfhB/pgpC* genes were downregulated, among which *hycB*, *c0118*, *c5065*, *c0813*, *c2514*, *glpA* and *deoB* are newly detected potential NsrR targets in CFT073. The *grxD*, *folB*, *ybjW/hcp*, *ytfE*, *yjbN*, *trxB* and *c5205* genes were upregulated and *c5205* and *yjbN* are potential NsrR targets newly detected in CFT073. The *glpA* gene, which encodes anaerobic glycerol-3-phosphate dehydrogenase subunit A, is downregulated in UPEC strain UTI89 exposed to acidified nitrite (Bower et al., 2009). The *livJ* gene, which encodes a periplasmic Leu/Ile/Val-binding protein, is upregulated during in vitro growth in human urine (Snyder et al., 2004). NsrR binding signals were found close to fimbriae related genes *fimC* and *c4214*. The *fimC* gene whose product is required for the biogenesis of type 1 fimbriae is upregulated in vivo during UTI (Snyder et al., 2004). The product of *c4214* is putative major fimbrial subunit precursor (Figure 2.2). Fimbriae is major determinants of bacterial virulence. The *rfe* gene was upregulated in vivo compared with growth in human urine in vitro (Hagan et al., 2010). The E. coli K-12 homologue of *yfbT* is upregulated in the presence of a source of NO (Hyduke et al., 2007). In *E. coli* K-12, the expression of *arcB* and *malF* was increased and decreased, respectively, after treatment with NO (Hyduke et al., 2007), and *phnC* was upregulated by treatment with 1 mM S-nitrosoglutathione or acidified nitrite (Mukhopadhyay et al., 2004).

2.2.2 Computational analysis of NsrR binding sites in CFT073

The 49 peaks located in putative regulatory regions were used to construct a PSWM for NsrR binding sites in the CFT073 genome. Two hundred base pairs centred on the nucleotide with the largest tag density within each of the peaks was analysed (Myers et al., 2013). The sequence of NsrR in CFT073 is identical to that in *E. coli* K-12, and evidence from

Table 2.2. NsrR binding sites associated with genes that are nitrate-responsive in RNA-seq data

Coordinate[a]	**-log10(P-value)**[b]	**Fold enrichment**[c]	**Flanking genes**[d]	**Distance from summit to start codon**[e]	**Fold change after nitrate treatment**[f]	**PPEE**[g]
1890380	503.52	19.16	*grxD* (<) *mepH* (>)	-138 (*grxD*)	1.59	0.005
3130600	523.77	16.18	*hycB* (<) hycA (<)	-59 (*hycB*)	0.01	0
3645158	287.94	11.46	*folB* (<) *plsY* (>)	28 (*folB*)	1.59	0.01
967971	248.24	12.00	*hcp* (<) *ybjE* (<)	-22 (*hcp*)	43.72	0
5049151	247.18	9.65	*ytfE* (<) *ytfF* (<)	-50 (*ytfE*)	69.63	0
115339	218.24	10.60	*mutT* (>) *c0118* (>)	18 (*c0118*)	0.36	2.8×10^{-7}
3967122	116.05	5.92	*yhgF* (>) feoA*B* (>)	95 (*feoA*)	0.18	0
698961	54.72	4.83	*ccrB* (<) ybeM (>)	345 (ybeM)	0.30	0
4840905	45.39	3.67	*acs* (<) *c5065-nrfA* (>)	-25 (*c5065*)	0.57	0.002
796704	44.53	4.08	*c0813* (<) *ybgE* (>)	177 (*c0813*)	0.62	1.84×10^{-11}
2350649	32.09	3.82	*c2513* (<) *c2514* (>)	200 (*c2514*)	0.41	0
2639159	23	3.28	*glpT* (<) *glpA* (>)	-41 (*glpA*)	0.23	0
4792020	22.78	2.78	*zur* (<) *yjbN* (>)	-169 (*yjbN*)	2.91	0.01
5208290	20.47	2.75	*deoA* (>) *deoB* (>)	18 (*deoB*)	0.28	0
3131567	17.96	2.67	hycA (<) *hypAB* (>)	-126 (*hypB*)	0.05	0.05
986391	16.91	2.87	*trxB* (<) *lrp* (>)	-92 (*trxB*)	1.71	0
4961829	15.97	2.80	*c5205* (<) *c5206* (<)	4 (*c5205*)	1.90	0.001
2522542	14.14	2.67	*yohJ* (>) *yohK* (>)	-172 (*yohK*)	0.06	0
2961332	13.69	2.64	*pgpC* (<) *c3084* (<)	64 (*pgpC*)	0.34	0

a Genomic location of the summit of ChIP-seq peak.
b -log10(P-value) of each peak called by M*acs*2.
c Fold enrichment of each peak calculated by M*acs*2.
d The genes flanking the ChIP-seq summit.
e The distance between the peak summit and start codon of the nearest downstream gene.
f Posterior fold change (the fold change computed from normalized data) calculated by EBSeq, shown for the predicted NsrR target.
g Posterior probability that a gene/transcript is not equally expressed under two conditions, as estimated by EBSeq.

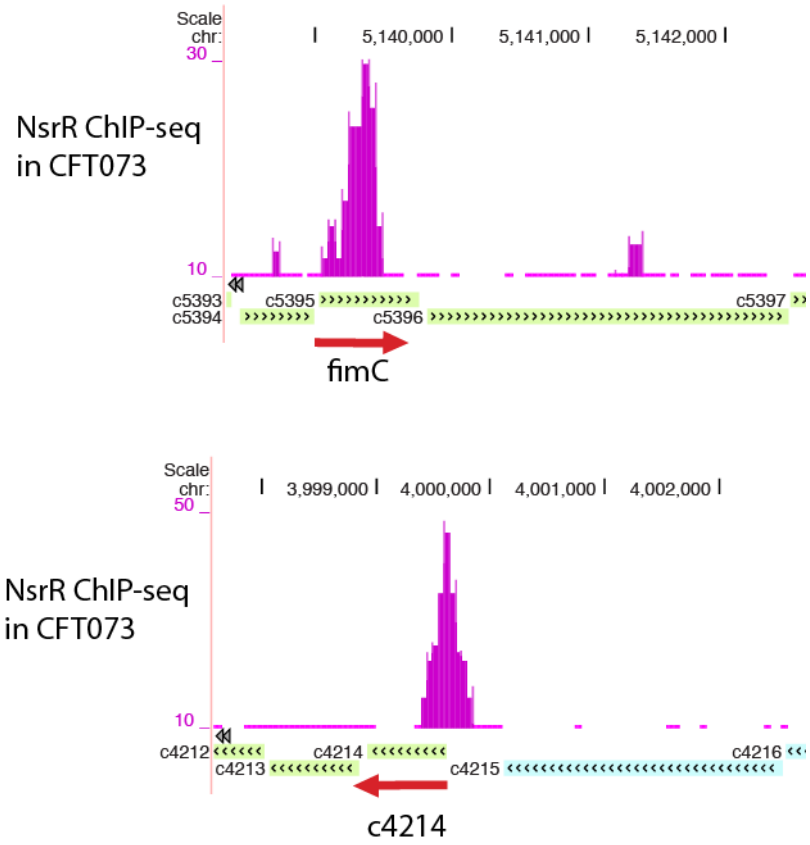

Figure 2.2. UCSC Genome Browser track view of *fimC* and *c4214* loci.

previous studies suggests that NsrR binding sites have two copies of an 11bp motif arranged as an inverted repeat with 1 bp spacing (Partridge et al., 2009). So, we first used MEME to identify over-represented palindromic sequences with the parameters -mod zoops -nmotifs 1 -minw 23 -maxw 23 -revcomp -pal to see if the same motif could be retrieved. Motifs matching the search criteria could be found in 20 of the 49 peak regions. As expected, the predicted NsrR binding site in CFT073 is similar to that for *E. coli* K-12 (Figure 2.3(a)). A precision-recall curve (see Methods) was constructed using the NsrR PSWM with two inverted repeats and searching throughout the genome of CFT073 to determine the optimal threshold for predicting high-quality NsrR binding sites. Using an ln(P-value) of -14.28 as the cut-off, where we had both relatively high precision and recall, there were 27 predicted NsrR

binding sites with the 11-1-11 inverted repeat (palindrome) motif in the CFT073 genome (Table 2.3). Four of these predicted targets were not detected by the ChIP-seq data (*tehA*, *yeaR*, *yhiX* and *ygbA*). Among them, *yeaR* and *ygbA* are known to be regulated by NsrR (Bodenmiller and Spiro, 2006b; Lin et al., 2007), and the *ygbA* promoter was reported to be bound by NsrR in *E. coli* K-12 according to previous ChIP-chip data (Partridge et al., 2009). Likewise, *tehA* was implicated as an NsrR target in *E. coli* K-12 by the same ChIP-chip data and by repressor titration (Bodenmiller and Spiro, 2006b), and it was shown to be upregulated in the urinary tract in an asymptomatic bacteriuria strain of *E. coli* (Roos and Klemm, 2006). By contrast, reporter fusion data suggest that *tehA* is not regulated by NsrR (Bodenmiller and Spiro, 2006b); conflicting reports may reflect differences in growth conditions or genetic background. Minimally, we can conclude that *yeaR* and *ygbA* are probably false negatives in our ChIP-seq data. The *gadX* (*yhiX*) gene was reported to be induced by NO through an indirect NsrR-dependent mechanism in *E. coli* O157:H7 (Branchu et al., 2014), but the presence of an NsrR binding site upstream of *gadX* may indicate a direct regulatory mechanism.

There is evidence that a single 11 bp motif can function as an NsrR-binding site in *E. coli* K-12 (Partridge et al., 2009). So we combined the two halves of the 11-1-11 palindromic motif, and reconstructed a PSWM of 11 bp. The new 11-bp PSWM was used to scan the 49 200 bp sequences flanking all the peak regions using the P-value cut-off of 10^{-6}. In this analysis, 38 of 49 peaks had at least one single motif, and the updated sequence logo for the 11bp motif is shown in Figure 2.3(b) .

2.3 Discussion

By ChIP-seq we identified NsrR binding sites in the CFT073 genome. Of 49 NsrR binding sites in promoter regions, 19 are associated with genes that were nitrate-responsive in the RNA-seq data. This discrepancy may reflect differences in the strains used, or the growth

Table 2.3. NsrR binding sites with 11-1-11 inverted repeats in the *E. coli* CFT073 genome

Peak centre[a]	**Downstream gene**[b]	**ChIP-seq**[c]	**PatSer ln(P-value)**[d]	**Motif sequence**[e]
5049128	*ytfE*	+	-25.13	AAGATGCATTTAAAATGCATCTT
967958	*hcp*	+	-23.87	AAGTTATATTTAATATACATGTT
115376	*c0118*	+	-22.88	AAGTTTTACTTCAAATGAATGTT
227478	*c0233*	+	-22.88	AAGTTTTACTTCAAATGAATGTT
4569952	*yihF*	+	-21.71	AAATTGTATTTGATGTGGATGTT
2948634	*hmp*	+	-18.22	TTGATGTATCTCAAATGCATCTT
2654713	*arnA*	+	-17.73	GAGGTGCATTTAATCTGCATGGT
1088121	*yccM*	+	-17.63	TAGTGGCATTTGGTATGCAACTT
3999611	*c4214*	+	-17.62	AAATTGAATTTCATTTATACCTT
4104796	*yhiX*	-	-17.62	AAGATATATGTTATATGAATGTT
2037249	*yeaR*	-	-17.05	AAATGGTATTTAAAATGCAAATT
1890357	*grxD*	+	-16.81	TTGTTGCATTTCAAATATTCGTT
3130577	*hycB*	+	-16.79	AAATGACATTTCATCGGCATGTT
1693010	*tehA*	-	-16.74	AAAGTATATTTGAAATGCATTTT
3137982	*ygbA*	-	-16.65	AAGGTGCATTTATATTACAACTT
3994013	*c4208*	-	-16.45	AAAGTTTATTTATACTGAATGTT
4840882	*c5065*	+	-16.32	TAAGTGCATGTAAAATACCACTT
4342117	*c4580*	+	-16.16	AAGTTGCATTTTATCTGCACCGG
986393	*trxB*	+	-16.15	ATGTTGTACTAAAAATCGATGTT
1920754	*ydiC*	-	-15.97	AAGTTGCATTGAAAATGACTATT
3391575	*yggS*	+	-15.94	AAGTTGCACGCCAAATGGCTGTA
1651193	*c1819*	-	-14.83	ATATTACATTGGATATGAATGTA
460167	*c0470*	-	-14.57	TAATTGCATATTAAAAATATGTT
4399231	*yieI*	+	-14.5	AAAGGGAGTTTGATATGTCTGTT
3645157	*ygiG*	+	-14.42	ATATTGTATTTATAGAGCAACTT
371521	*c0392*	-	-14.31	TAGTTTCATTATATATGTCTGAT
1830291	*ynfL*	-	-14.28	AAGATGTTTTAAATATGAATCTT

a Sites were identified using PatSer and a precision-recall curve was determined based on an ln(P-value) threshold of -14.28. The coordinate of the centre of the predicted site is shown.
b The gene downstream of the predicted NsrR binding site.
c Presence (+) or absence (-) of an NsrR ChIP-seq peak at the location of each predicted NsrR binding site.
d PatSer ln(P-value) of each predicted NsrR binding site (5'-3').
e Sequence of each predicted NsrR binding site (5'-3').

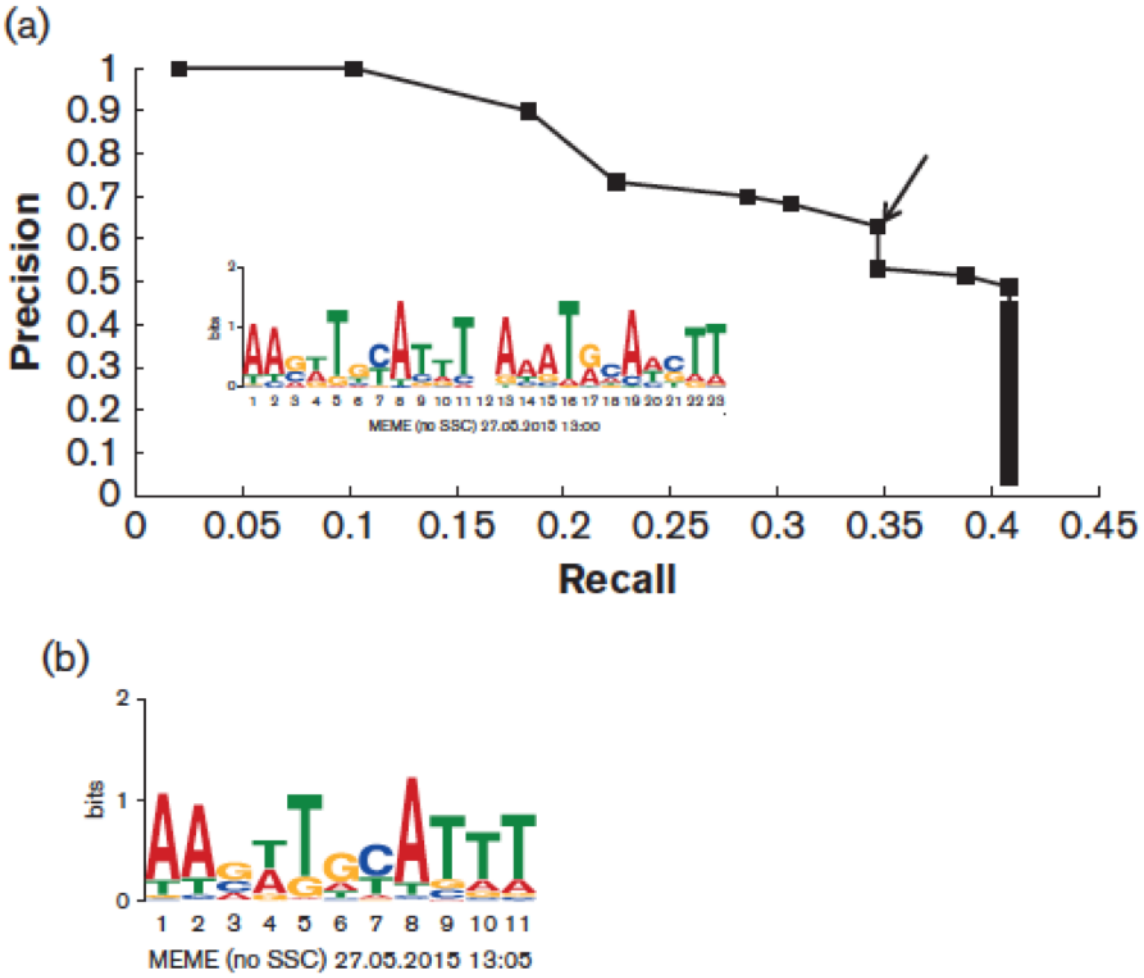

Figure 2.3. Computational prediction of NsrR binding sites. (a) Precision-recall curve used to determine the prediction threshold of NsrR binding sites. The precision and recall values were determined for many ln(P-value) thresholds using the PatSer algorithm and the optimal value (-14.28) is identified by the arrow. The inset shows the NsrR position weight matrix with inverted repeats constructed from the NsrR ChIP-seq sequences. (b) NsrR position weight matrix from NsrR ChIP-seq peak sequences. The height (y-axis) of the letters represents the degree of conservation at that position within the aligned sequences set (in bits), with perfect conservation being 2 bits. The x-axis shows the position of each base (1-11) starting at the 5' end of the motif.

conditions used for the two experiments (aerobic growth for ChIP-seq, anaerobic growth for RNA-seq), although there is no published evidence to suggest that NsrR binding to DNA is sensitive to oxygen in vivo. Another possible explanation is that at some binding sites

NsrR exerts weak or no regulation, as we have observed previously for *E. coli* K-12. As was the case for *E. coli* K-12 (Partridge et al., 2009) around half of mapped sites were within coding regions or between convergently transcribed genes. Similar results have been obtained with other regulatory proteins, for example Fur (Seo et al., 2014), and this is not surprising behaviour for a DNA-binding protein with a relaxed sequence specificity. We assume that most sites in this category have no biological function, although some may regulate the activity of promoters driving expression of small or anti-sense RNAs.

We found strong NsrR binding signals upstream of some hypothetical proteins of unknown function, some of them specific to CFT073 (meaning not present in *E. coli* K-12). Examples are *c0118* and *c0233* (Figure 2.4), which are homologues of each other. Both *c0118* and *c0233* have two copies of a conserved helix-turn-helix domain that is often found in transposases and is likely to bind DNA. Both proteins are implicated as transposases or derivatives in the clusters of orthologous groups of proteins (COGs) database. Transposase genes are frequently associated with pathogenicity islands, and NsrR has been implicated in regulating pathogenicity island genes in *E. coli* O157:H7 (Branchu et al., 2014). Therefore, it would be interesting to study the function of *c0118* and *c0233* to see if they are related to the pathogenicity of CFT073, and to determine if NsrR is involved in the regulation of pathogenicity island genes.

Of the genes implicated as possible NsrR targets by ChIP-seq that were also differentially regulated in response to NO, two-thirds were downregulated in the presence of a source of NO. This behaviour is consistent with positive regulation by NsrR, as has been reported previously (Branchu et al., 2014), or with indirect effects of NsrR. Some genes associated with NsrR binding sites were not differentially regulated in the RNA-seq experiment, which may indicate that these genes are subject to multiple regulatory mechanisms, such that regulation by NsrR is revealed only under specific growth conditions. An additional possibility is that there is a category of promoter that is bound by, but not regulated by, NsrR.

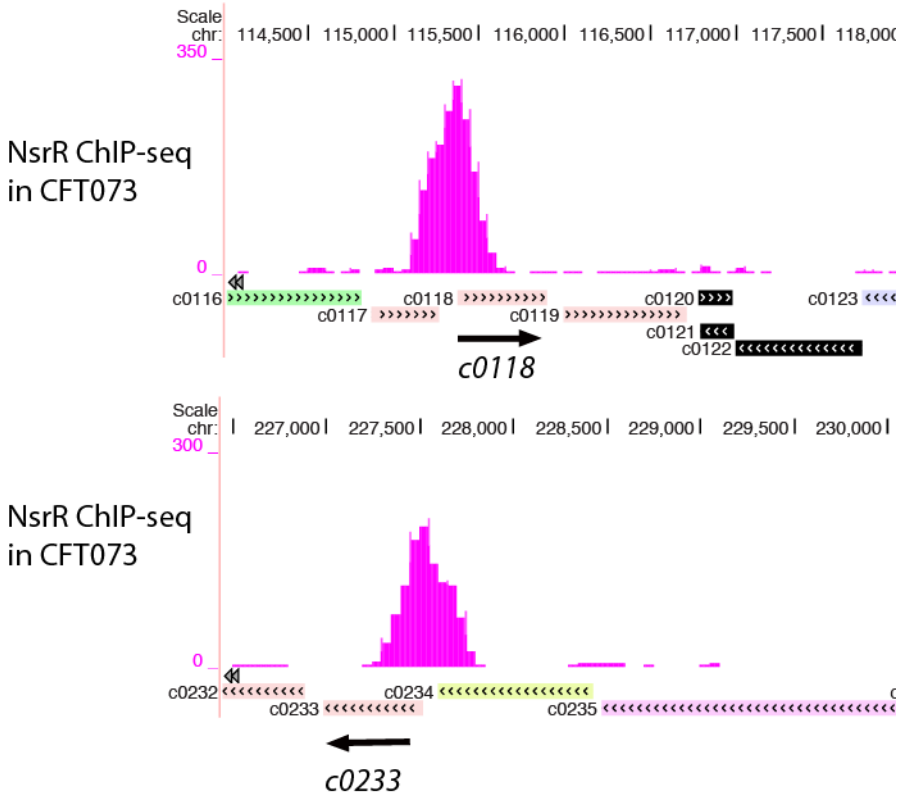

Figure 2.4. UCSC Genome Browser track view of *c0118* and *c0233* loci.

CHAPTER 3

TP53 ASSOCIATED LONG-RANGE INTERACTION PREDICTION

3.1 Materials and methods

3.1.1 Data sources

ChIA-PET data of pol2 in breast cancer MCF7 cells was obtained from (Li et al., 2012). In our study we used saturated MCF7 ChIA-PET dataset in table s3h where the count, p-value and FDR of each Paired-End Tag (PET) cluster were given. The TP53 ChIP-seq data with nutlin treatment in MCF7 was retrived from GSE30183 (Nikulenkov et al., 2012). The processed data files were based on genome build hg19. ChIP-seq data of H3K4me1, H3K4me3, H3K27ac and input control in MCF7 were from GSE57498 (Taberlay et al., 2014). MCF7 DNase data was from ENCODE project with the accession number of GSE32970 (Thurman et al., 2012; Natarajan et al., 2012). ChIP-seq data of p300, JUND and FOSL2 in MCF7 were also obtained from ENCODE project with accession number of GSE32465 (Gertz et al., 2013). Additional ChIP-seq data of H3K27ac and Gro-seq data before and after nutlin treatment were from GSE76657 (Verfaillie et al., 2016) and GSE53966 (Allen et al., 2014) respectively. TP53 ChIP-seq data, histone modification data (H3K27ac, H3K4me1 and H3K4me3) before and after nutlin treatment, and ATAC-seq data before and after nutlin treatment in IMR90 were obtained from GSE58740 (Sammons et al., 2015).

3.1.2 Mapping and Binding sites detection

Bowtie2 was used for the mapping of raw sequencing data with parameters of -k1 -N1. Model-based Analysis for ChIP-Seq (MACS) (Zhang et al., 2008) was used to identify genome-wide

binding sites of TP53. If the distances of peak summits were less than 200bp, then those peaks would be merged and new peak summit was assigned as the mid-point of those merged peaks. The lowest FDR among all the merged peaks is the FDR for newly merged peaks.

3.1.3 TP53 binding sites classification

TP53 peaks can be grouped into three classes after integrating with pol2 ChIA-PET data as shown in Figure 3.1. The first class is composed of TP53 binding sites overlapped with pol2 looping interactions whose both ends have overlap with TP53 peaks. Here we will assign a TP53 peak (FDR<0.01) into the first class if it overlaps with one end (<5kb) of a pol2 loop whose other end overlaps with another TP53 binding site (FDR not controlled). The strongest summit of TP53 binding sites will be selected if there exists more than one TP53 binding sites in the same anchor of a looping interaction. For the second class, if only one end of pol2 associated loops has overlap with TP53 binding site (FDR<0.01), then that binding site would be assigned to class II TP53 binding sites. In the end we have 327 Class I TP53 binding sites from 306 high confident interactions with both interacting anchors overlapping with TP53 peaks, 874 Class II TP53 binding sites overlap with interactions with only one end overlapping with TP53 peaks, and 2674 Class III TP53 binding sites that don't overlap with any pol2 looping interactions.

3.1.4 Histone modification, DNase-seq and Gro-seq profiles surrounding ChIP-seq binding peaks

TP53 peak summits were selected as center for signal alignments. Signals were plotted in a 4kb window surrounding the peak center with each window divided into 25 bp bins. Read coverage from histone modification ChIP-seq, DNase-seq and Gro-seq were calculated for each bin. For ChIP-seq data, reads were extended by 200bp in the 5'-3' direction, and read coverage of input was subtracted from that of ChIP data. For Gro-seq data, coverage from plus strand and minus strand were separated.

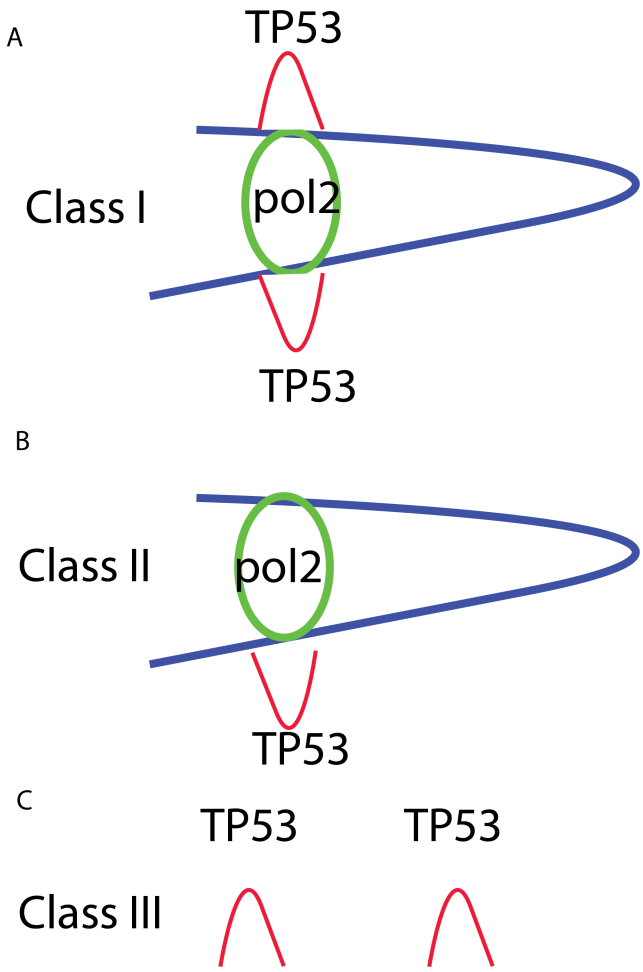

Figure 3.1. TP53 binding sites classification. A, Class I TP53 binding sites resulting from pol2 looping interactions with both interacting anchors overlapping with TP53 binding sites. B, Class II TP53 binding sites resulting from pol2 looping interactions with only one end overlapping with TP53 binding sites. C, Class III TP53 binding sites not associated with any pol2 looping interactions

3.1.5 Model to predict loop associated TP53 binding sites and loop associated TP53 binding interaction clusters

Figure 3.2 shows the workflow of predicting TP53 associated long-range interactions. It was done in two steps. In the first step, given all TP53 binding sites, model was built to

predict loop-anchor candidates. And in the second step model was built to predict which loop-anchor candidates from step 1 can form looping interactions.

For the first step, we used the scikit-learn Python package for logistic classification. We selected the following types of features for classification: 1) log2-transformed read counts for ChIP-seq data (TP53, JUND, FOSL2, P300 and multiple histone modification) or DNase-seq data in TP53 binding sites within a window size of 1000bp centered by the TP53 binding summits. 2) log2-distance between the TP53 binding site and its nearest neighboring binding site. 327 TP53 binding sites from Class I TP53 binding sites were selected as foreground training set and an equal number of TP53 binding sites that don't overlap with any pol2 looping interactions were randomly selected as background training set. After fitting logistic classifier with different features, we chose the non-redundant but significant features combinations that can discriminate positive and negative datasets to build the final classifier. Threshold was set to 0.1 to get putative loop associated TP53 binding sites which resulted in 3072 candidate anchors.

For the second step, we constructed all possible TP53 binding sites pairs with distance less than 1Mb using candidate TP53 anchors predicted from step 1. Out of all possible pair combinations, 101 TP53 peak pairs (one peak overlaps with one end of a pol2 loop and the other peak overlaps with the other end) with FDR < 0.01 were selected as foreground training set and equal number of candidate TP53 pairs that don't overlap with any pol2 or CTCF loops were randomly selected as background training set. We still used scikit-learn Python package to perform logistic classification. The following features were used to compute for each pair: 1) the sum of log2-transformed ChIP-seq read counts from TP53, JUND, FOSL2, P300 and multiple histone modification ChIP-seq or DNase-seq read counts for each TP53 binding sites pairs with window size of 400bp for each end. 2) log2-distance of each pair.

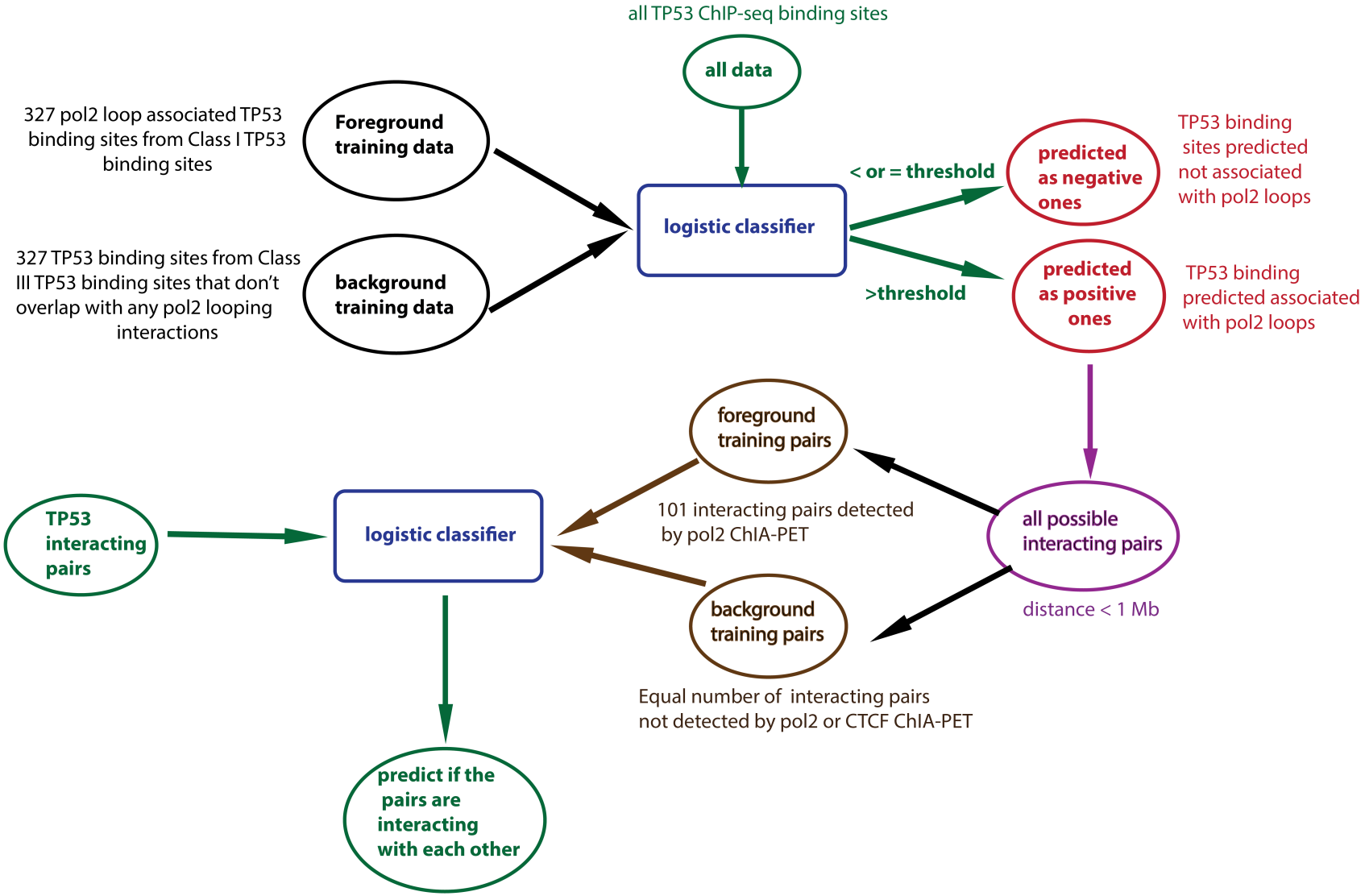

Figure 3.2. Workflow to predict loop associated TP53 binding sites and associated long-range interactions.

We evaluated our classifier using 5-fold cross-validation. Model parameters were trained on the whole training data set in MCF7. And the trained model in MCF7 was used to predict TP53-TP53 interactions in other cell lines where genome-wide chromatin interaction data is not available.

3.2 Results

3.2.1 Data integration is informative for identifying TP53 long-distance targets

If a transcription factor binds to a genomic region which has physical interaction with other genomic region, it is likely that ChIP-seq of the TF can detect binding signals on both regions. To investigate TP53 binding sites that are associated with putative chromatin looping interactions, we integrated TP53 ChIP-seq data and genome-wide chromatin interaction data from pol2 ChIA-PET. Two TP53 binding sites would be considered as a potential interacting pairs in our study, if they overlap with the two interacting anchors from pol2 ChIA-PET respectively. In the end, we got 306 pol2 mediated looping interactions with both ends overlapping with TP53 peaks (at least one peak has FDR less than 0.01), which results in 237 TP53 binding sites associated with pol2 looping interactions.

To find TP53 peaks that contain TP53 binding motif, we used HOMER (Heinz et al., 2010) to scan known TP53 motif across TP53 peaks with 400 bp around their binding summits. Among 306 interacting peak pairs, four have binding motifs on both ends and 85 have motifs on only one end. Thus the fact that some TP53 binding sites are absent of binding motif can be explained by the looping interactions with TP53 binding sites with motif.

Among the 306 interacting pairs, we identified some putative functional TP53 bound enhancers and their possible regulating targets which weren't studied before. Phosphoinositide-3-Kinase Interacting Protein 1(PIK3IP1), a negative regulator of Phosphatidylinositol-3-kinase (PI3K), was reported to suppress the development of hepatocellular carcinoma (He

et al., 2008). TP53 activation by nutlin elevates the the expression of PIK3IP1(Janky et al., 2014), but its detailed regulation by TP53 has not been elaborately studied yet. TP53 ChIP-seq data showed that there is TP53 binding signal near TSS of PIK3IP1 with promoter signature (Figure 3.3A, purple box) and another TP53 peak with enhancer signature is located 20kb upstream (Figure 3.3A, grey box). The promoter or enhancer state was obtained from GEO Accession GSE57498 where MCF7 ChromHMM chromatin state segmentation was performed (Taberlay et al., 2014). The looping interactions between those two peak regions were detected by pol2 ChIA-PET and both of regions have CTCF binding sites. We further confirmed enhancer state by looking at the enrichment of classic enhancer signals of p300 and H3K27ac. It was found that both of the signals can be detected in the enhancer peak region. But for another TP53 peak which is located between the interacting promoter peak and enhancer peak, neither of the signals was detected even it is closer to the promoter peak. Therefore, integration with chromatin interaction data can help us spot some potentially functional binding site and its putative long-distance targets. We also checked HOMER motif scan results for those two regions, TP53 motif is present in enhancer peak but absent in promoter peak which indicates that the binding of promoter region might result from looping interaction with enhancer peak.

For some other differentially expressed genes after TP53 activation but without TP53 binding signal at the promoter region, it is possible that they are indirect targets of TP53. And it is also possible that they are TP53 direct targets, but TP53 binds to some distant genomic region to regulate their expression through chromatin looping instead of binding to gene proximal promoters. False negatives from ChIP-seq may cause the failure of TP53 signal capture at promoter region. Krppel-like factor 4 (KLF4) is one such potential TP53 long-distance target that was not identified before. KLF4, on one hand can suppress TP53 expression and have a primary oncogenic function by repressing TP53-dependent apoptosis (Rowland et al., 2005). On the other hand, KLF4 can also serve as tumor suppressor

by physically interacting with TP53 to activate CDKN1A (p21) and lead cell cycle arrest (Zhang et al., 2000; Yoon et al., 2003). KLF4 is upregulated on TP53 activation by nutlin. Instead of detecting TP53 binding signal at KLF4 promoter region, two TP53 peaks with enhancer signatures were detect more than 200kb upstream of KLF4 (Figure 3.3B, grey boxes). Again, the their enhancer state from ChrommHMM chromatin state segmentation were further confirmed by the enrichment of classic enhancer signals of p300, H3K27ac and H3K4me1(Figure 3.3B). ChIA-PET of pol2 shows the existence of looping interactions between those two TP53 bound enhancers and KLF4 promoters. Therefore, KLF4 might be TP53 direct long-distance target regulated by TP53 through chromosome looping.

Collectively, integration of TP53 genome-wide binding ChIP-seq data and genome-wide chromatin interaction data from pol2 ChIA-PET can help us both identify some putative functional TP53 bound enhancers and also identify some putative TP53 direct long-distance targets.

3.2.2 Epigenetic features can discriminate loop and non-loop associated TP53 binding sites

Through the integration with TP53 ChIP-seq data, chromatin interactions detected by pol2 ChIA-PET can be grouped in to 3 classes, i.e. both ends of a loop overlapped with TP53 binding sites, only one end overlapped with TP53 binding sites and neither of the ends overlapped with TP53 binding sites. In our study TP53 binding sites from those three classes are denoted by Class I TP53 binding sites, Class II TP53 binding sites and Class III TP53 binding sites respectively. The detailed classification procedure was described in Method. TP53 binding sites associated with looping interactions may bear different epigenetic features due to their inclination to form larger protein complexes to link enhancers and promoters compared to simple protein binding sites. So here we checked different epigenetic markers of those three groups of TP53 binding sites.

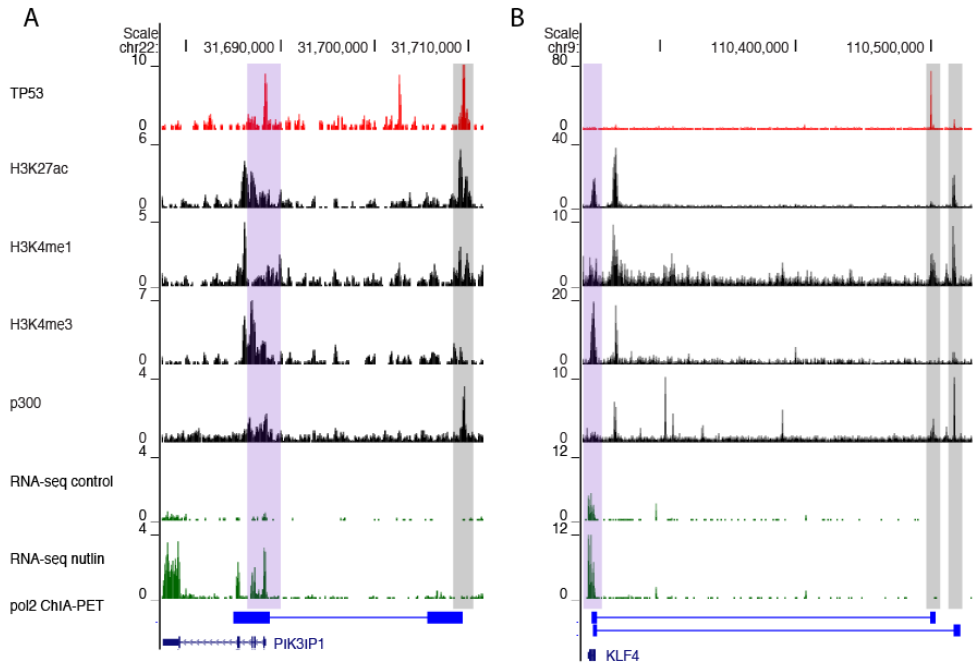

Figure 3.3. UCSC Genome Browser track view (A)PIK3IP1 locus and (B)KLF4 locus.Track for TP53 ChIP-seq is shown in red. Tracks for ChIP-seq of histone modifications (H3K27ac, H3K4me1 and H3K4me3) and p300 are shown in black. Tracks for RNA-seq without and with nutlin treatment are shown in green. Track for pol2 ChIA-PET is shown as connected blue boxes, and boxes represent interacting genomic regions. Different classes of peak types are in different color of shades. Purple shades represent promoters and gray shades represent enhancers.

We first examined TP53 binding signal itself. As expected, the average signal of Class I TP53 binding sites is higher than that of the other two groups (Figure 3.4A). There is no much difference between Class II and Class III binding sites (Figure 3.4A). We further divided TP53 binding sites according to the existence of TP53 binding motif. Out of 874 Class II and 2674 Class III binding sites, 36% and 69% of them have TP53 motif, but only 23% of Class I binding sites contain motif. TP53 motif is significantly enriched in non-loop associated TP53 binding sites with fisher exact p-value less than $4.1 \times e^{-58}$ when Class I and Class III binding sites were compared. It is possible that TP53 motif is responsible for the majority

of binding at non-loop associated sites, but for loop associated sites, motif is not necessary since binding could result from looping interactions with TP53 peak containing motif. We then further compared TP53 signals at Class I and Class III binding sites with or without TP 53 motif, and it was found that loop associated Class I TP53 binding sites with TP53 motifs have the highest TP53 binding signals (Figure 3.4B). It is reasonable because loop associated binding sites tend to have larger complex and higher TP53 concentration which could make DNA bound by TP53 easier to be pulled down and enriched by ChIP, especially for binding sites with TP53 motif which have higher binding affinities. The expected TP53 signal distribution in each group further confirmed the proper separation of TP53 binding sites according to pol2 ChIA-PET.

We then analyzed histone modification patterns of three different classes of TP53 binding sites. As mentioned before, loop associated Class I and Class II TP53 binding sites in our study are nutlin induced binding sites that overlap with pre-configured looping interactions detected by pol2 ChIA-PET. Individual studies in both human and *Drosophila* showed that TP53 acts on pre-existing chromatin interacting loops (Melo et al., 2013; Link et al., 2013). And epigenetic marks are associated with chromatin interaction sites (Li et al., 2012), thus in order to investigate chromatin environment associated with pre-configured loops we compared different epigenetic markers without TP53 activation at three classes of TP53 binding sites. As shown in Figure 3.5 and Figure A.1-A.5, loop associated Class I and Class II binding sites are more transcriptionally active with higher pol2 binding signal, higher enrichment of active histone markers (H3K27ac, H3K4me1 and H3K4me3), stronger DNase-seq signal and higher nascent transcription levels when compared to non-loop associated Class III TP53 binding sites. Interestingly for Class I and Class II binding sites, although both of them are from TP53 binding sites that are associated with pol2 looping interactions, the signals tested above tend to be stronger in Class I binding sites including pol2 signal itself. It is possible that for Class II binding sites, some looping interaction with only one end overlapped with

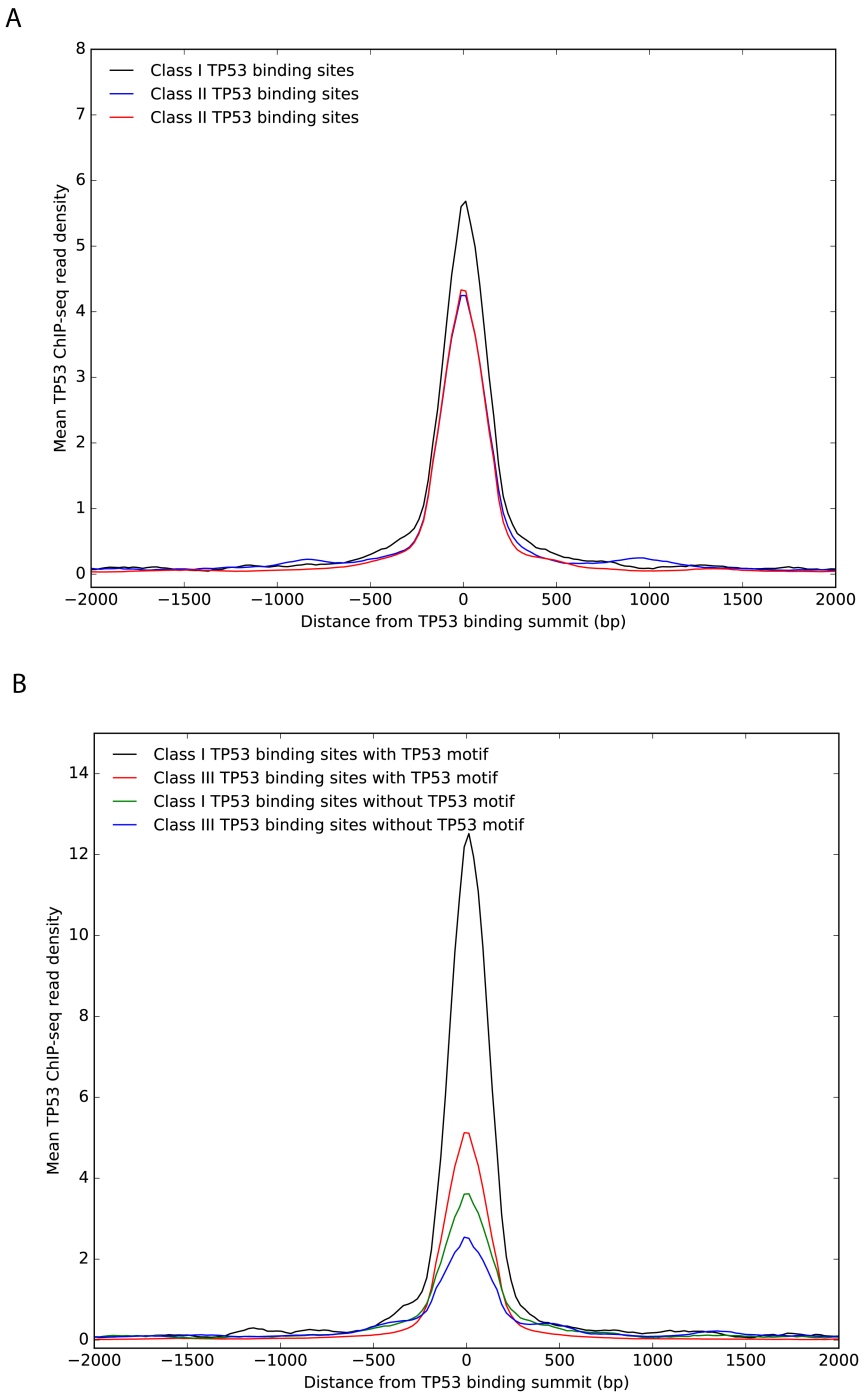

Figure 3.4. TP53 binding signal signal profile for different groups of TP53 binding sites.Average TP53 read coverage against control surrounding (A) three classes of TP53 binding sites (B) Class I TP53 binding sites with and without TP53 motif and Class III TP53 binding sites with and without TP53 binding sites

TP53 binding sites are false positives, and the others are true positives but TP53 binding signals failed to be detected due to false negatives from ChIP-seq. And Class II binding sites might be a mixture of TP53 binding sites that are associated with real looping interactions and looping interactions detected by ChIA-PET as false positives, therefore their signal levels are in between Class I and Class III binding sites whose population are purer. To better decipher the difference between loop associated TP53 binding sites and non loop associated TP53 binding sites, we mainly focused on the comparison between Class I and Class III TP53 binding sites in the following study.

To inspect the effect of TP53 activation, we included nutlin treated H3K27ac ChIP-seq and Gro-seq data which are only public available datasets in MCF7 with nutlin treatment. It was noticed that for both loop associated Class I TP53 binding sites and non-loop associated Class III TP53 binding sites, TP53 activation increased H3K27ac enrichment and changed nascent transcription to a certain extend, but the change is not dramatic (Figure 3.6A,B). They still have similar profiles as before activation and loop associated Class I TP53 binding sites still have higher H3K27ac enrichment and transcription level than non-loop associated TP53 Class III binding sites. Both enhancer transcript and H3K27ac marker are features correlated with enhancer activity and enhancer transcripts tend to be enriched at enhancers that can form looping with promoters or other enhancers (Shlyueva et al., 2014; Lin et al., 2012; Sanyal et al., 2012; Lam et al., 2014). From histone modification and transcription level, loop associated TP53 Class I binding sites in our study tend to be a mixture of active promoter and enhances which have high transcriptional activity even before TP53 activation (Figure 3.5A-F). And non-loop associated Class III TP53 binding sites are mainly located at inactivate genomic regions lacking active epigenetic marks. The fact that TP53 activation didn't dramatically increase the transcription activity of inactive non-loop associated Class III TP53 binding sites to a level as high as loop associated regions further supported previous study in IMR90 that TP53 bound to pre-established enhancers and TP53 binding would not

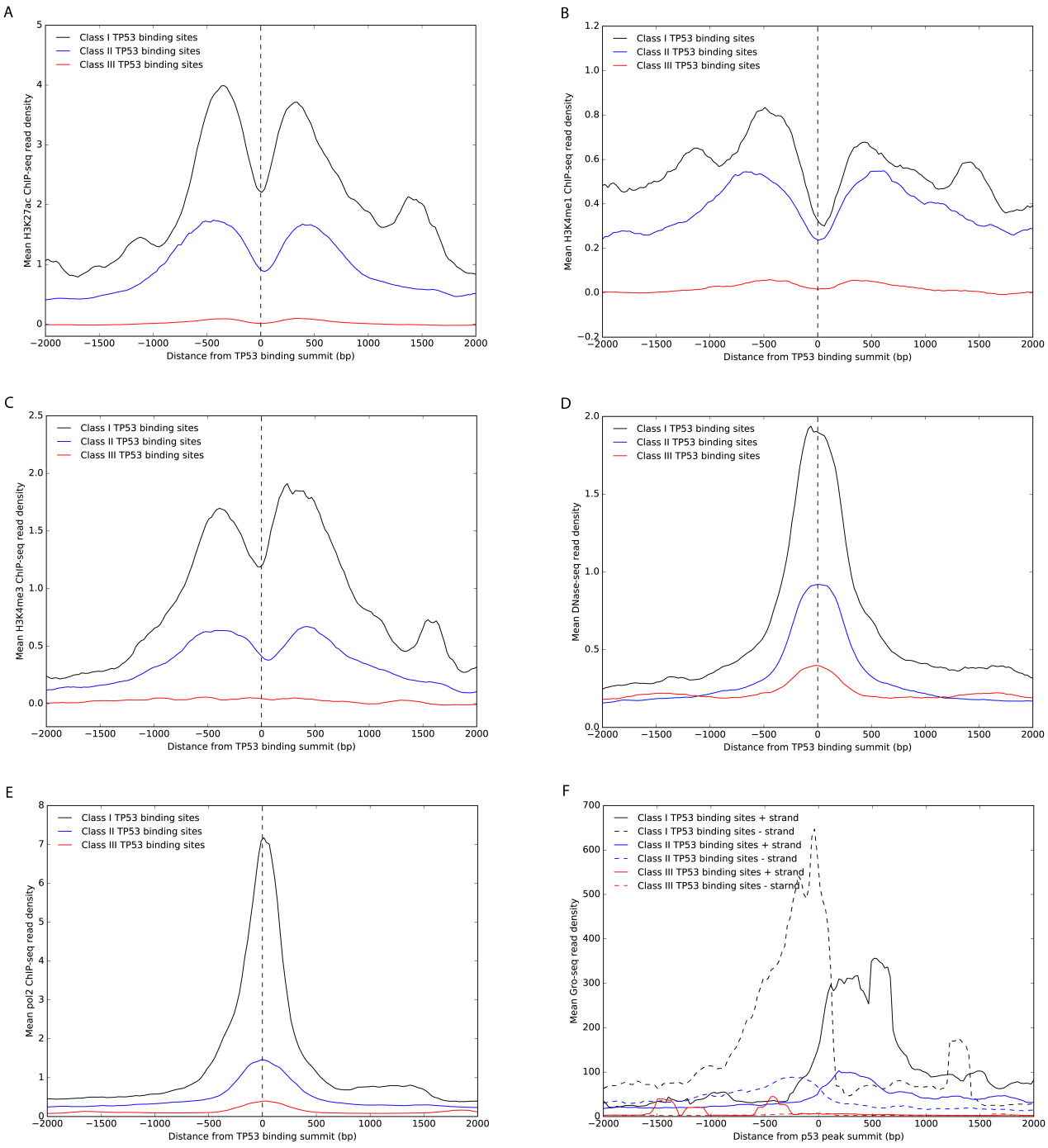

Figure 3.5. Histone modification, Dnase-seq and Gro-seq signal profiles for three classes of TP53 binding sites. Average H3K27ac(A), H3K4me1(B), H3K4me3(C) read-coverage against input surrounding three classes of TP53 binding sites. Average DNase(D), pol2(E), Gro-seq(F) read-coverage surrounding three classes of TP53 binding sites.

license new enhancer (Sammons et al., 2015). It's also consistent with previous study that TP53 is more likely to act on pre-existing loops instead of initiating new loops to regulate its targets (Melo et al., 2013; Link et al., 2013). The slight change and good correlation of epigenetic marks with and without TP53 activation (Figure 3.6C) makes it possible for us to learn transcriptionally activated loop associated TP53 binding sites using epigenetic marks in normal condition without TP53 activation which are usually public available for most cell lines. We will give more detailed description later in next chapter.

In addition, we found that histone modification signals were more likely to be depleted in the center of loop associated TP53 binding sites (Figure 3.7A-C. Figure A6-8). As mentioned before, loop associated TP53 binding sites in our study are more like to be promoters or enhancers with relatively higher transcriptional activity, therefore the enrichment of histone modification depletion in loop associated TP53 binding sites can be explained by nucleosome depletion in active promoter enhancer regions (Hon et al., 2009). And DNase-seq data further supported that nucleosomes were more likely to be depleted in loop associated TP53 binding sites (Figure 3.5D). We further divided TP53 binding sites by the presence of TP53 motif. We found that TP53 binding sites without TP53 motif tend to show more fraction of histone modification depletion than binding sites with motif (Figure A6-8). It is possible that for binding sites lacking TP53 motif, TP53 binding requires a more open chromatin structure for other transcription factors to bind and recruit TP53.

Collectively, Epigenetic markers can discriminate loop and non-loop associated TP53 binding sites. Loop associated TP53 binding sites, especially those without TP53 binding motif are more often associated with more open chromatin structure and higher transcriptional activity. TP53 activation won't dramatically change chromatin environment of TP53 binding sites especially for loop associated TP53 binding sites.

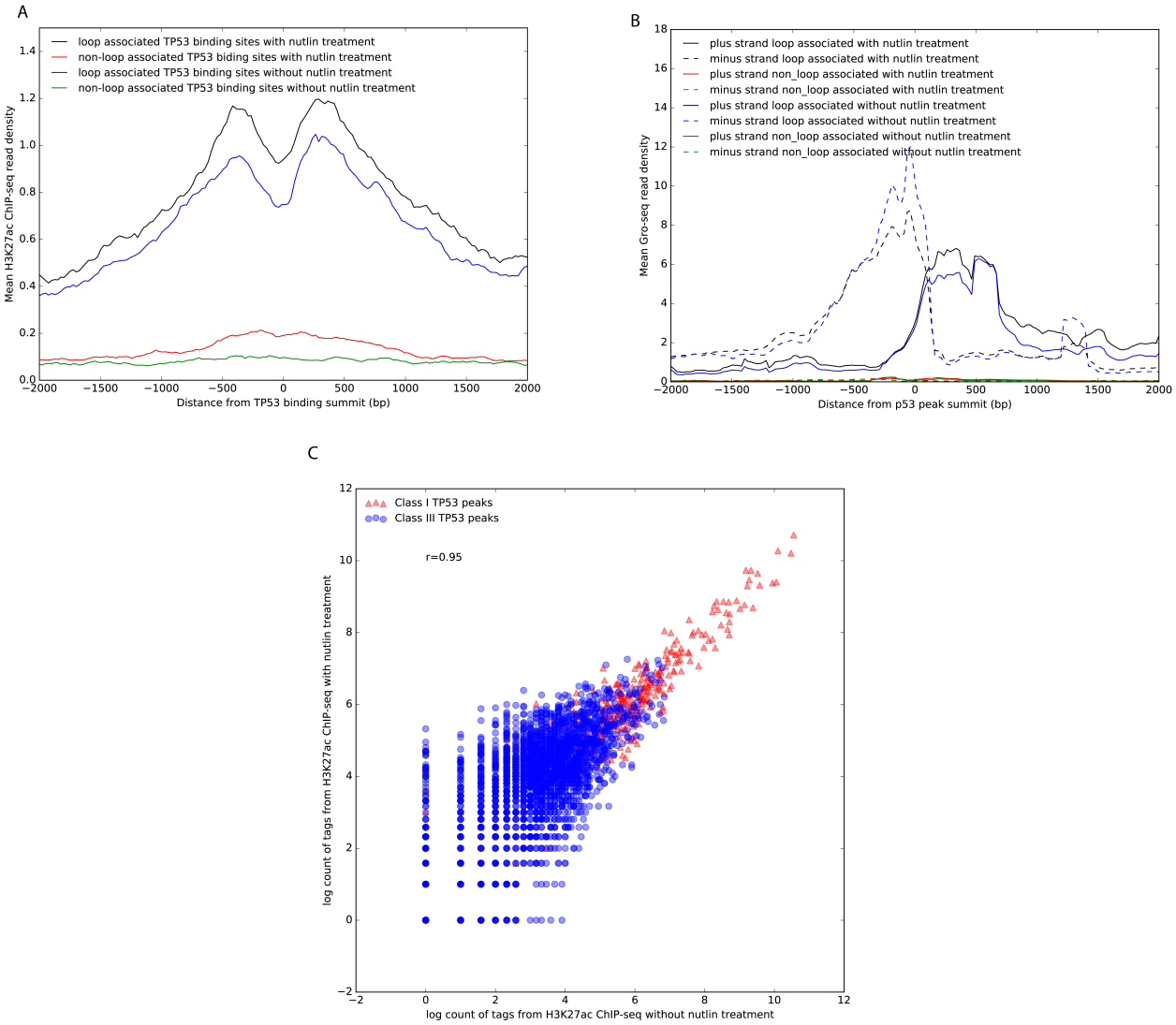

Figure 3.6. H3K27ac and Gro-seq profiles with and without nutlin treatment for Class I TP53 binding sites and Class III TP53 binding sites. Average H3K27ac (A) and Gro-seq (B) read-coverage with and without nutlin treatment surrounding Class I TP53 binding sites and Class III TP53 binding sites. (C) Correlation of H3K27ac read-coverage in the window of 1000bp surrounding peak summits before and after nutlin treatment

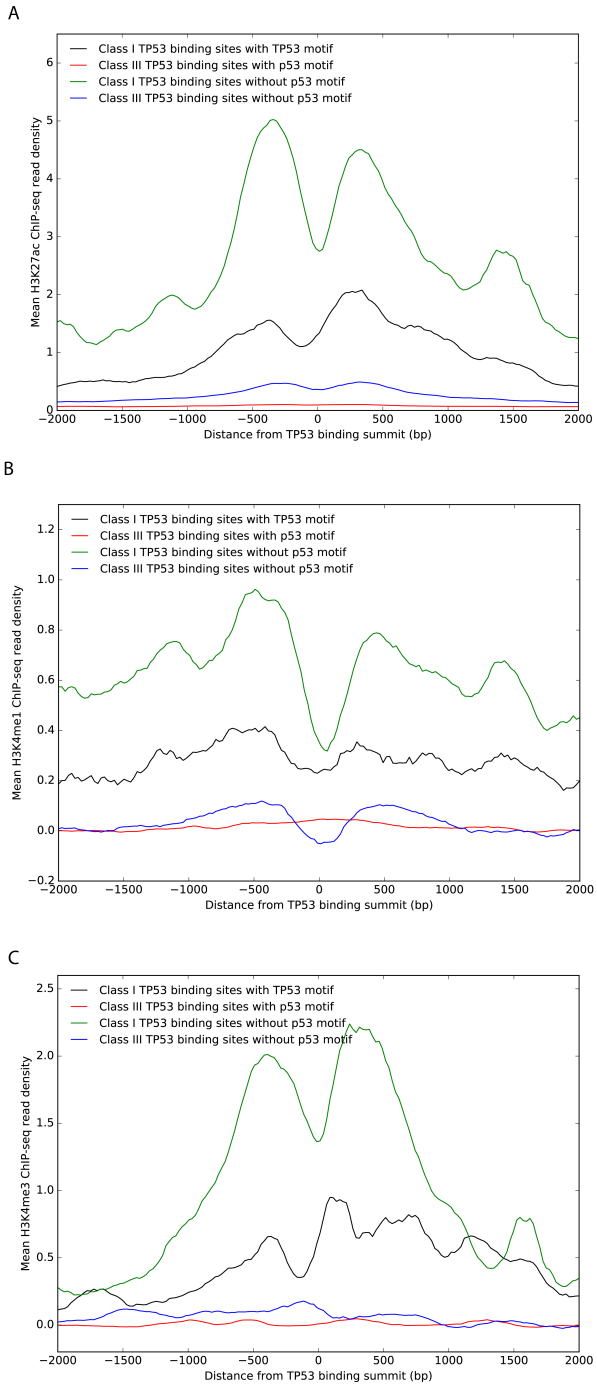

Figure 3.7. Histone modification profile for Class I TP53 binding sites with and without TP53 motif and Class III TP53 binding sites with and without TP53 motif. Average H3K27ac(A), H3K4me1(B), and H3K4me3(C) read-coverage surrounding Class I TP53 binding sites with and without TP53 motif and Class III TP53 binding sites with and without TP53 motif.

3.2.3 A logistic classifier can predict loop associated TP53 binding sites and associated long-range interactions

It can be seen from last section that different epigenetic features can discriminate TP53 binding sites that are associated with pol2 loops and that are not associated with pol2 loops. So here in this section, we were motivated to build a classifier using above features to distinguish those two classes of TP53 binding sites and predict TP53 interacting clusters in cell lines of which genome-wide chromatin interaction data is not available. First, we built a classifier to discriminate TP53 peaks that are associated with pol2 loops and that are not associated with pol2 loops. We selected 327 pol2 loop associated TP53 binding sites from Class I TP53 binding sites as foreground training set. An equal number of TP53 binding sites from Class III TP53 binding sites that don't overlap with any pol2 mediated interactions were randomly selected as background training set. A logistic classifier was used to separate foreground and background TP53 binding sites. We tested the prediction power of features from previous section and AUC for predictor with single feature was shown in Figure 3.8A. JUN/FOS family members were reported to overlap TSS and enhancer TP53 binding in IMR90, and they can mediate chromatin accessibility for other transcription factors (Sammons et al., 2015; Biddie et al., 2011), therefore we also checked the prediction power of two members (JUND and FOSL2) from JUN/FOS family using available ChIP-seq data from encode project and tested their prediction power. It was noticed that binding intensity of JUND and FOSL2 also show good predictive power in distinguishing foreground and background TP53 binding sites. ROC comparisons for different feature combinations were shown in Figure 3.8B. It was noticed that H3K27ac signal itself can give good prediction, addition of DNase-seq signal and neighboring TP53 distance can improve the prediction by some extend but not dramatically. Adding all the features didn't improve prediction power as just using the three features above. We evaluated the performance of the classifier using the three features above (H3K27ac ChIP-seq signal, DNase-seq signal and distance of nearest

neighboring TP53 binding sites) with 5-fold corss-validation and got the average true positive rate (TPR) at 81% and average false positive rate (FPR) at 10% at threshold of 0.5. To include as many as pol2 loop associated TP53 binding sites in our training set, we finally set the threshold to 0.1 at which the average TPR was as high as 0.99 and the relatively high rate of false positives (0.73) will be filtered out in the next step.

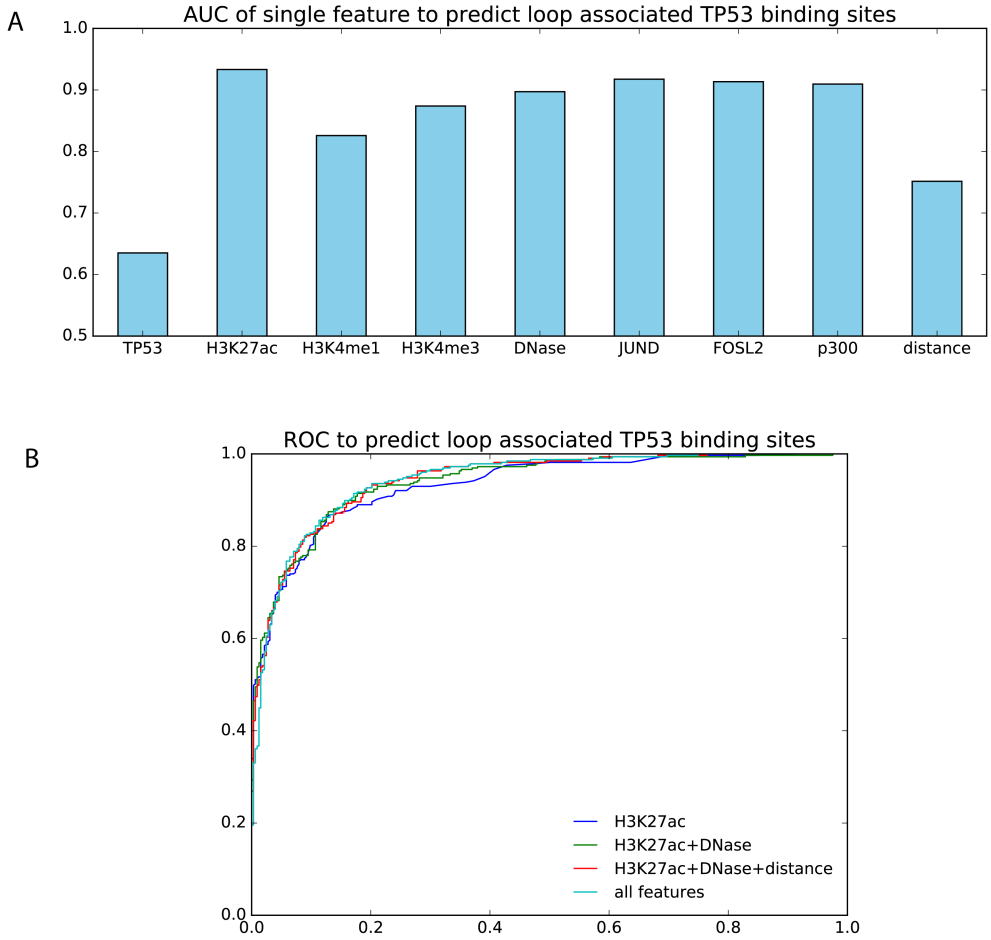

Figure 3.8. AUC and ROC curves for the predictors of loop associated TP53 binding sites (A) AUC for each feature to predict loop associated associated TP53 binding sites. Here AUC was computed as the average area under ROC curve for predictions with single feature for 5-fold cross-validation(B) ROC trained for H3K27ac ChIP-seq alone, H3K27ac ChIP-seq + DNase-seq, H3K27ac ChIP-seq + DNase-seq + distance of nearest neighboring TP53 binding sites and all the features.

In the second step, we tried to predict interactions between the putative pol2 loop associated TP53 binding sites obtained from the first step. From the above integration of pol2 ChIA-PET and TP53 ChIP-seq data, there are 103 pol2 looping interacting clusters (<5kb) with both ends overlapping with TP53 peaks (FDR <0.01). We checked the distance distribution between those interacting pairs and it was found that the distance of almost every interacting pair is less than 1Mb (Figure A9). We constructed all possible interacting pairs with distance less than 1Mb using predicted pol2 loop associated TP53 binding sites from step 1. 101 out of above 103 pol2 looping interactions with both ends overlapped with TP53 peaks can be found in all pair combinations and were selected as foreground training set. Equal number of candidate TP53 pairs that don't overlap with any pol2 looping interactions or CTCF looping interactions identified by CTCF ChIA-PET from the same study was randomly selected as background training set. Logistic classifier was still used to perform classification. We tested the prediction power of above features and AUC for predictor with single feature was shown in Figure 3.9A. ROC comparisons for different feature combinations were shown in Figure 3.9B. The intensity of H3K27ac and the distance between candidate pairs performed as the top features to distinguish foreground and background training data. Finally five features (i.e. H3K27ac ChIP-seq signal, distance between interacting pairs, H3K4me1 ChIP-seq signal, DNase-seq singal and H3K4me3 ChIP-seq signals) were selected to build final classifier. We evaluated the performance of our classifier using five-fold cross-validation and got average TPR at 73% and average FPR at 22% at the threshold of 0.5.

TP53 and multiple histone modification ChIP-seq data are available in IMR90 (GSE58740). But pol2 ChIA-PET data is not available. So, we were motivated to predict TP53 binding sites interactions in IMR90. And in IMR90 we used ATAC-seq data, an alternative method to DNase-seq, for prediction since DNase-seq data is not available. Then data were fitted into our classifier. Out of 4334 TP53 binding sites in IMR90, 2804 were predicted to be loop

associated in the first step of prediction when we set threshold to 0.1. And it resulted in 6487 possible pairs with distance less than 1Mb for cells. In the second step of prediction, 1335 out of 6487 pairs are predicted to interact with each other in cells.

Among those1335 predicted TP53 interacting pairs, we were able to identify the interaction validated by other study and can also predict putative interactions which were not studied before. As shown in Figure 3.10, there is a small peak in the IER5 promoter region (labeled in blue shade) called by MACS with FDR less than 0.01. Our predictor can detected its interactions with two other TP53 binding sites downstream with distance larger than 10kb. One of them (labeled in yellow shade) was validated by 4C-seq in previous study in human primary BJ fibroblasts (Melo et al., 2013). The other one (labeled in green shade) was not studied before and has classic enhancer signature of H3K27ac and H3K4me1. It might be another functional TP53 bound enhancer that can regulate IER5 gene.

3.3 Discussion

Long-range interaction can help us better understand gene regulation from chromatin 3D structures. Among all kinds of 3C-based methods, ChIA-PET can provide us the interactions genome widely associated with a given TF with relatively high resolution. It is a good tool for us to study promoter-enhancer interactions associated with some specific TF. But sometimes it is challenging to perform ChIA-PET for some TF. So here, we integrated TP53 ChIP-seq data and genome-wide interaction data detected by pol2 ChIA-PET to have an insight of looping interactions that TP53 potentially acts on and have a better understand of its regulatory function.

Our integrative analysis provided us some putative promoter-enhancer interactions associated with TP53 and also identified some putative TP53 direct targets which are regulated by TP53 through chromatin looping. It was found that TP53 binding sites that overlap with pol2 looping interactions are more transcriptionally active and have distinct epigenetic

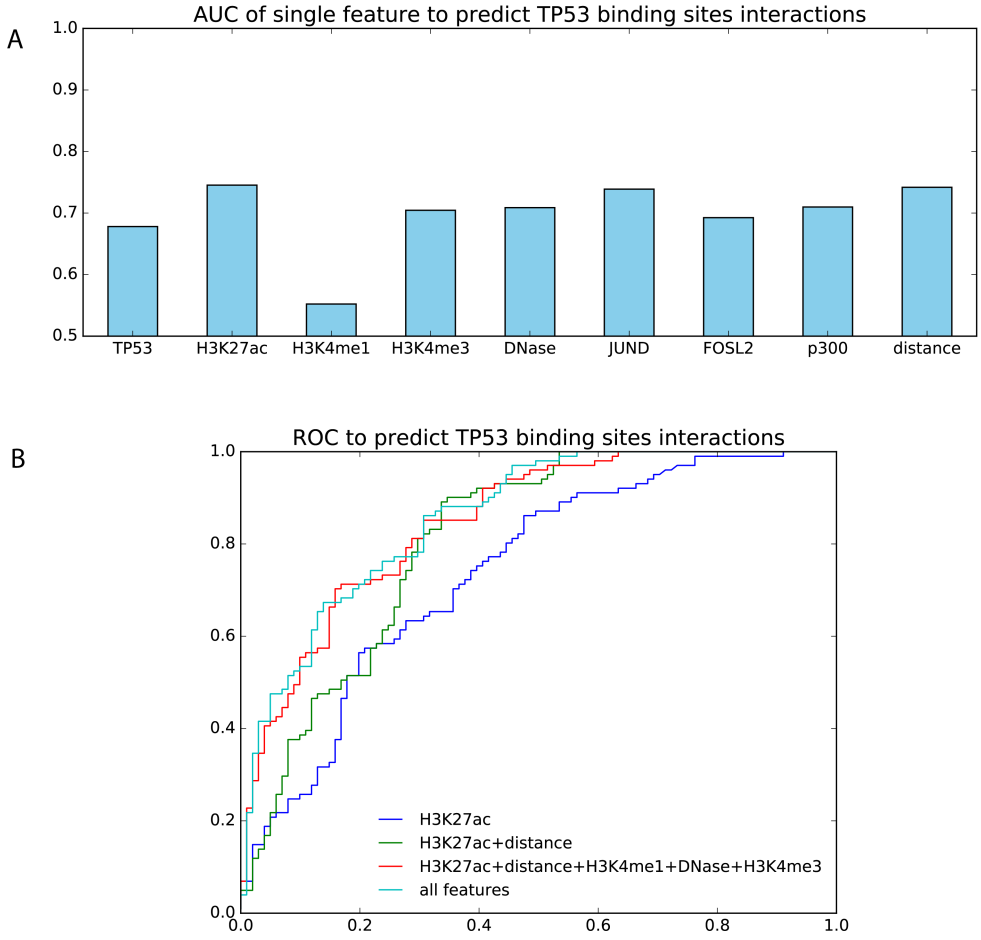

Figure 3.9. AUC and ROC curves for the predictors of TP53 binding sites interactions (A) AUC for each feature to predict TP53 binding sites interactions. Here AUC was computed as the average area under ROC curve for predictions with single feature for 5-fold cross-validation(B) ROC trained for H3K27ac ChIP-seq alone, H3K27ac ChIP-seq + distance, H3K27ac ChIP-seq+distance+H3K4me1 ChIP-seq+DNase+H3K4me3 ChIP-seq and all the features.

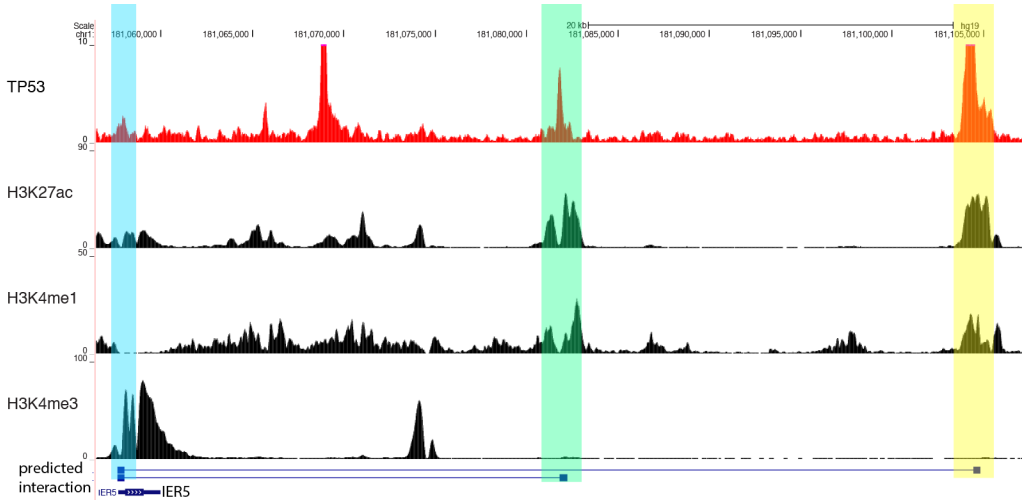

Figure 3.10. UCSC Genome Browser track view IER5 locus. Track for TP53 ChIP-seq is shown in red. Tracks for ChIP-seq of histone modifications (H3K27ac, H3K4me1 and H3K4me3) are shown in black. Track for predicted interactions is shown as connected blue boxes, and boxes represent TP53 peak regions. Blue shades represent P53 peak in IER5 promoter region and green shades and yellow shades represent TP53 peaks that more than 10kb downstream of IER5 and have predicted interactions with IER5 promoter peak. The interaction between the peak labeled by yellow shade and IER5 promoter was identified by in human BJ fibroblasts.

markers compared to TP53 binding sites that don't overlap with pol2 looping interactions. We developed classifiers to predict loop associated TP53 binding sites and interacting TP53 binding sites pairs through the extraction of features from multiple histone modification ChIP-seq and DNase-seq datasets.

However, there are still limitations of our classifiers, especially the classifier used to predict TP53 interacting pairs. It can only provide us average TPR of 73%. Although we can eliminate many false positives by increasing threshold and make sure there are as many as true positives in our predicted list, it also introduced more false negatives. Exploring more relevant features or trying other classification methods might further improve the performance of our classifier.

CHAPTER 4

CONCLUSION

Transcription factors are main regulators of gene transcription. Identifying their targets is important for understanding biological processes like stress responses and genetic cause of disease. In our study, we used integrative omics data analysis to identify TF targets in both simple prokaryotes and more complex mammalian system. Different omics data were integrated for TF targets identification in different systems. In prokaryotes, we integrated TF genome-wide binding data, expression data and sequence motif information to identify the targets of transcriptional repressor NsrR to better understand UPEC's response to NO stress. In mammalian system, we integrated TF ChIP-seq data, chromatin interaction data and different epigenetic data to identify long-distance targets of tumor suppressor TP53.

To identify genome-wide targets of NsrR, we performed ChIP-seq of NsrR and first identified NsrR regulon in UPEC CFT073. Forty-nine NsrR binding sites were located in putative promoter regions in CFT073 genome. Of those promoters bound by NsrR *in vivo*, 19 were identified in a previous ChIP-chip analysis of NsrR binding sites in *E. coli* K-12 (Partridge et al., 2009). Twenty of the remaining sites are associated with genes that have homologues in *E. coli* K-12, and 10 are specific to *E. coli* CFT073. Some of the new targets might be related to the virulence of CFT073. We integrated above NsrR genome-wide binding data from ChIP-seq with expression data from RNA-seq in a triple mutant strain of CFT073 lacking the three known NO detoxification system, Hmp, FIRd and NrfA with and without phsysiological source of NO. Nineteen out of 49 NsrR targets identified by ChIP-seq show differential expression with NsrR perturbation by physiological source of NO. We then did computational analysis of NsrR binding sites in CFT073, and it was found that the binding

motif of NsrR in CFT073 is the same as that in K-12. TF binding motif in prokaryotes is usually longer and more specific than that in mammalian system, so it is informative for TF targets identification in prokaryotes. We used the PSWM of NsrR to search throughout the genome of CFT073 as a complementary way of ChIP-seq to identify TF targets. Some of predicted targets were not detected by the ChIP-seq data but they are known NsrR targets in some other close related strains. They reason why ChIP-seq in CFT073 failed to detect them might be that growth conditions are different or epitope tags were blocked by other co-factors so that ChIP failed to pull down DNA. Therefore, TF motif could serve as a complementary method for ChIP-seq to identify potential TF targets missed by ChIP-seq.

To identify genome-wide long-distance targets of TP53 in mammalian system, we first integrated TP53 ChIP-seq data and genome-wide chromatin interaction data from pol2 ChIA-PET in cell line MCF7. Data integration that way helped us to identify some putative functional TP53 bound enhancers and their long-distance targets. We then divided TP53 binding site into group associated with pol2 mediated looping interactions and group that is not associated with pol2 mediated looping interactions. We checked epigenetic markers around TP53 binding sites in two groups and it was found that TP53 binding sites associated with pol2 loops tend to be more trancscriptionally active with stronger active epigenetic markers compared to TP53 binding sites that don't overlap with any pol2 looping interactions. Therefore, epigenetic markers is discriminative for those two groups of TP53 binding sites. In the end we built logistic classifiers using different epigenetic features to predict loop associated TP53 binding sites and TP53 interaction clusters, which can be used to predict TP53 associated looping interactions and TP53 long distance targets in cell lines where genome-wide chromatin data are not available.

APPENDIX

SUPPLEMENTARY FIGURES

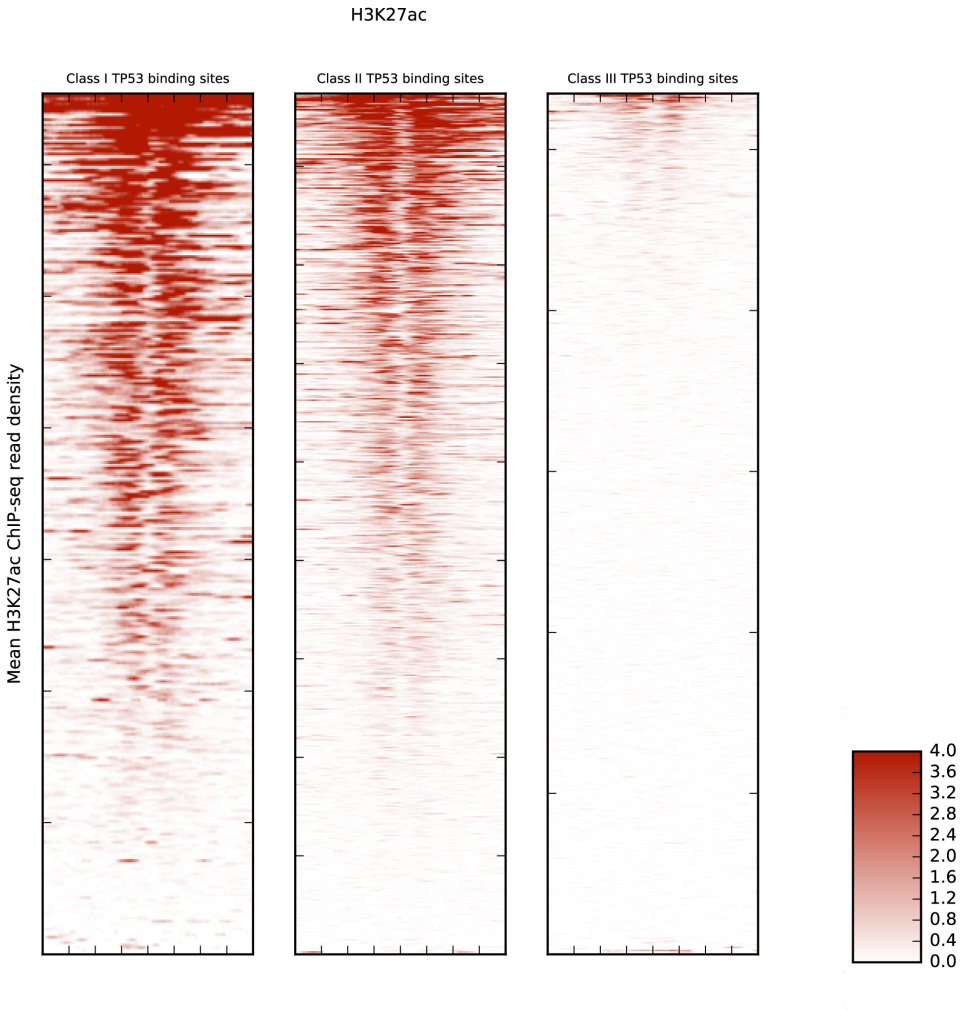

Figure A.1. Heatmap show of H3K27ac read coverage against input around summits of three classes of TP53 binding sites. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites.

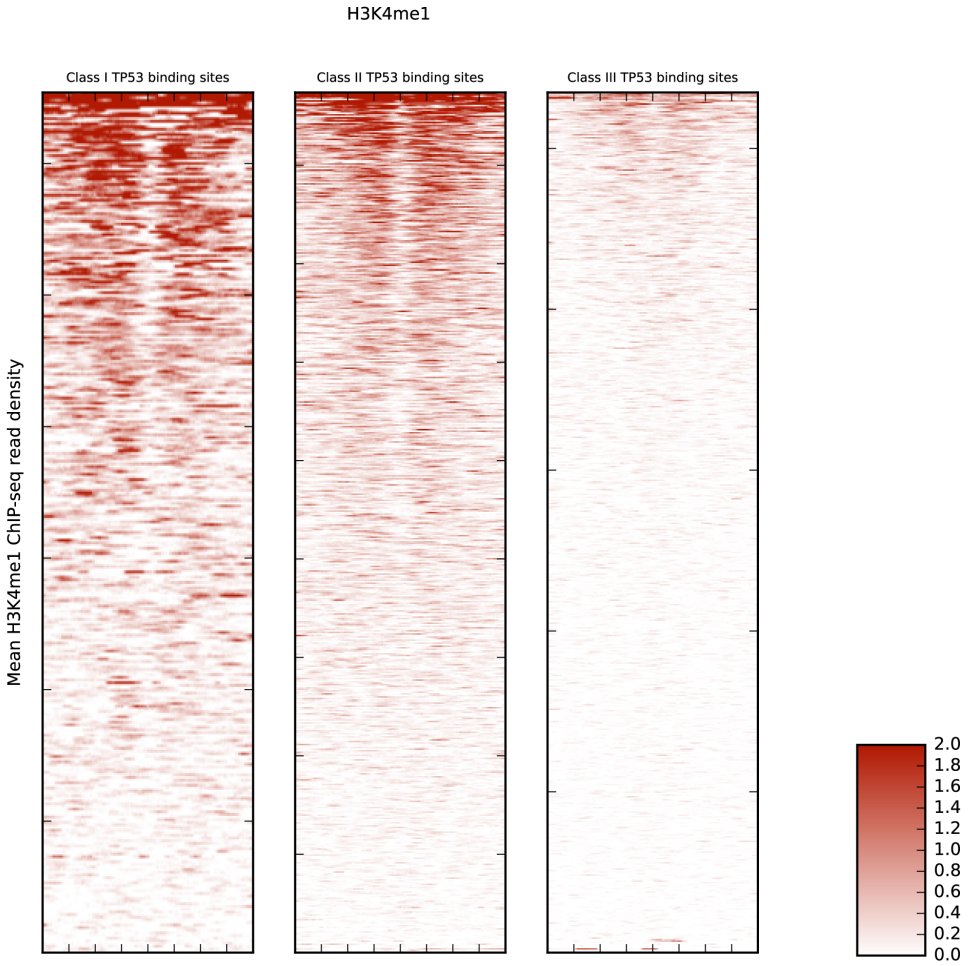

Figure A.2. Heatmap show of H3K4me1 read coverage against input around summits of three classes of TP53 binding sites. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites.

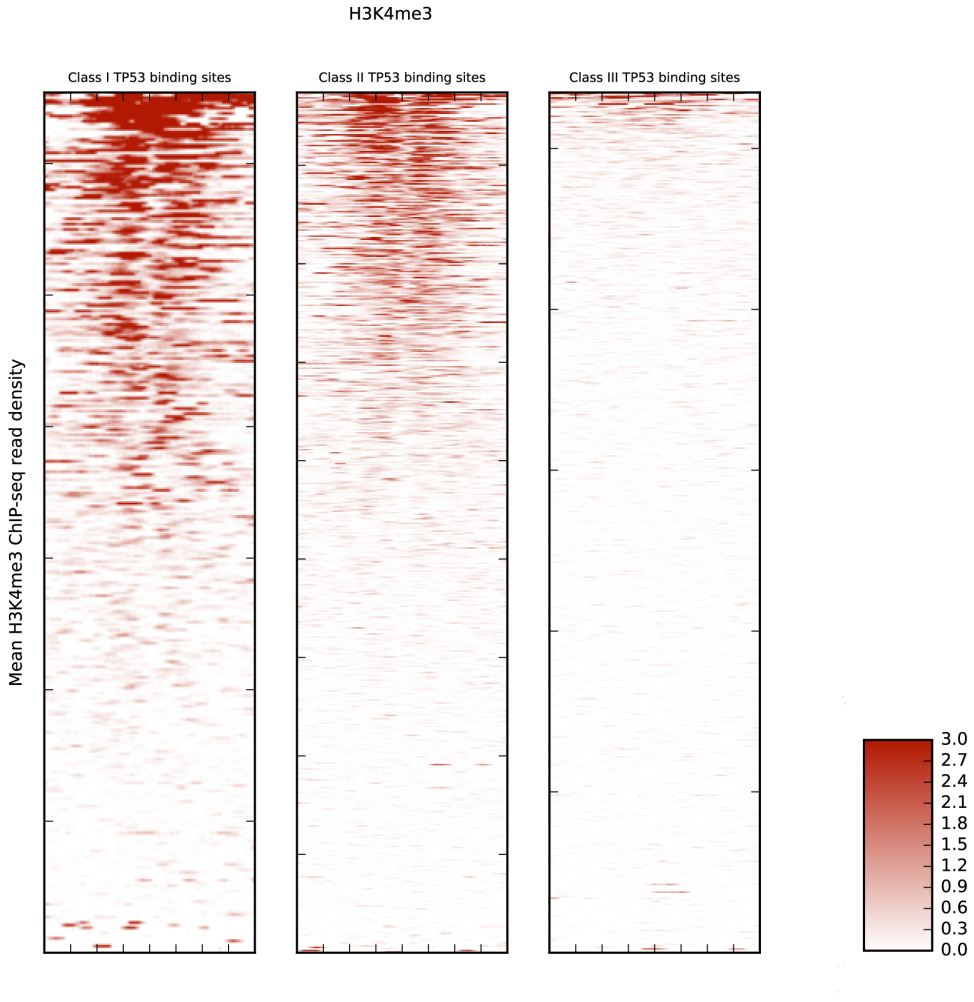

Figure A.3. Heatmap show of H3K4me3 read coverage against input around summits of three classes of TP53 binding sites. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites.

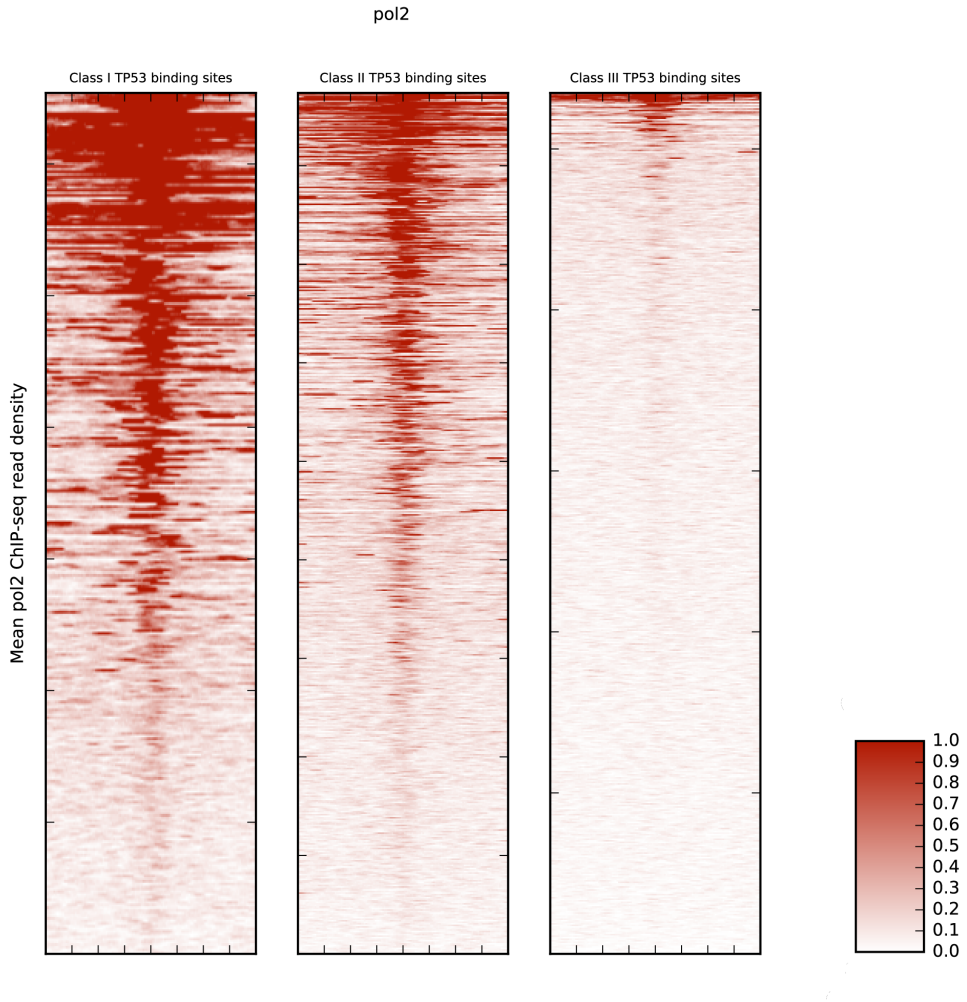

Figure A.4. Heatmap show of pol2 read coverage around summits of three classes of TP53 binding sites. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites.

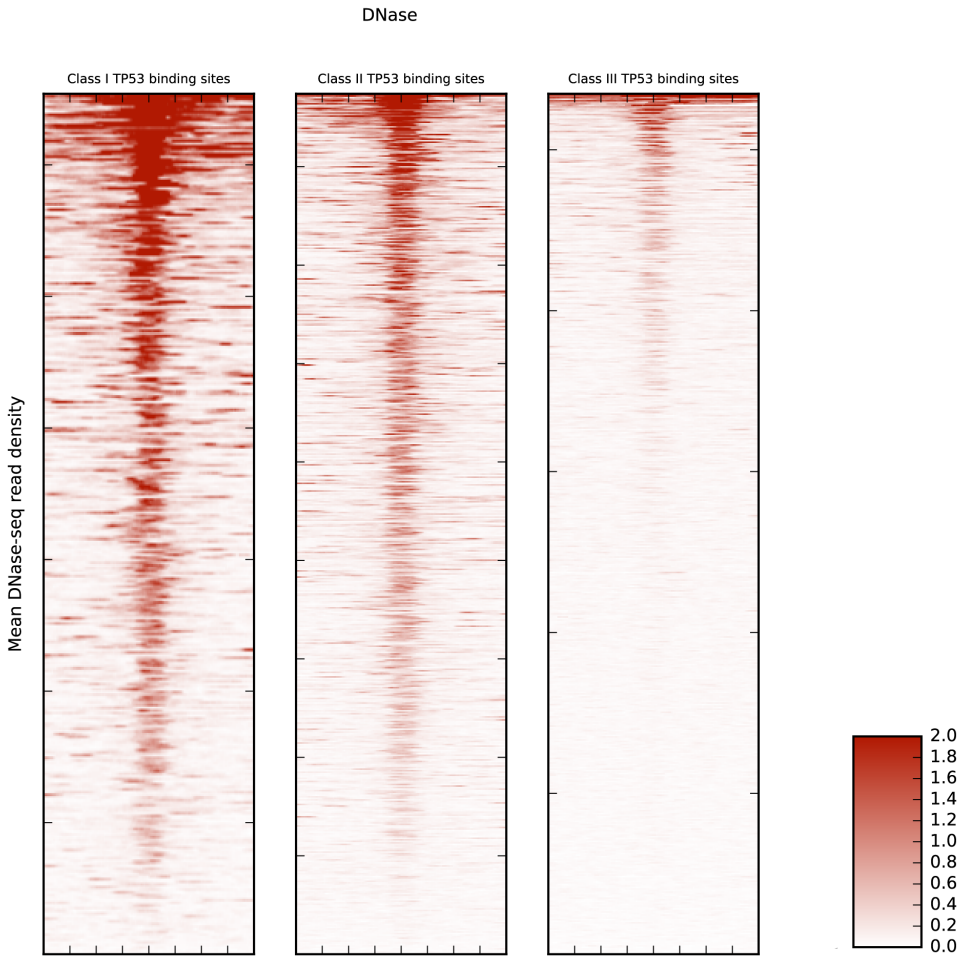

Figure A.5. Heatmap show of DNase read coverage around summits of three classes of TP53 binding sites.

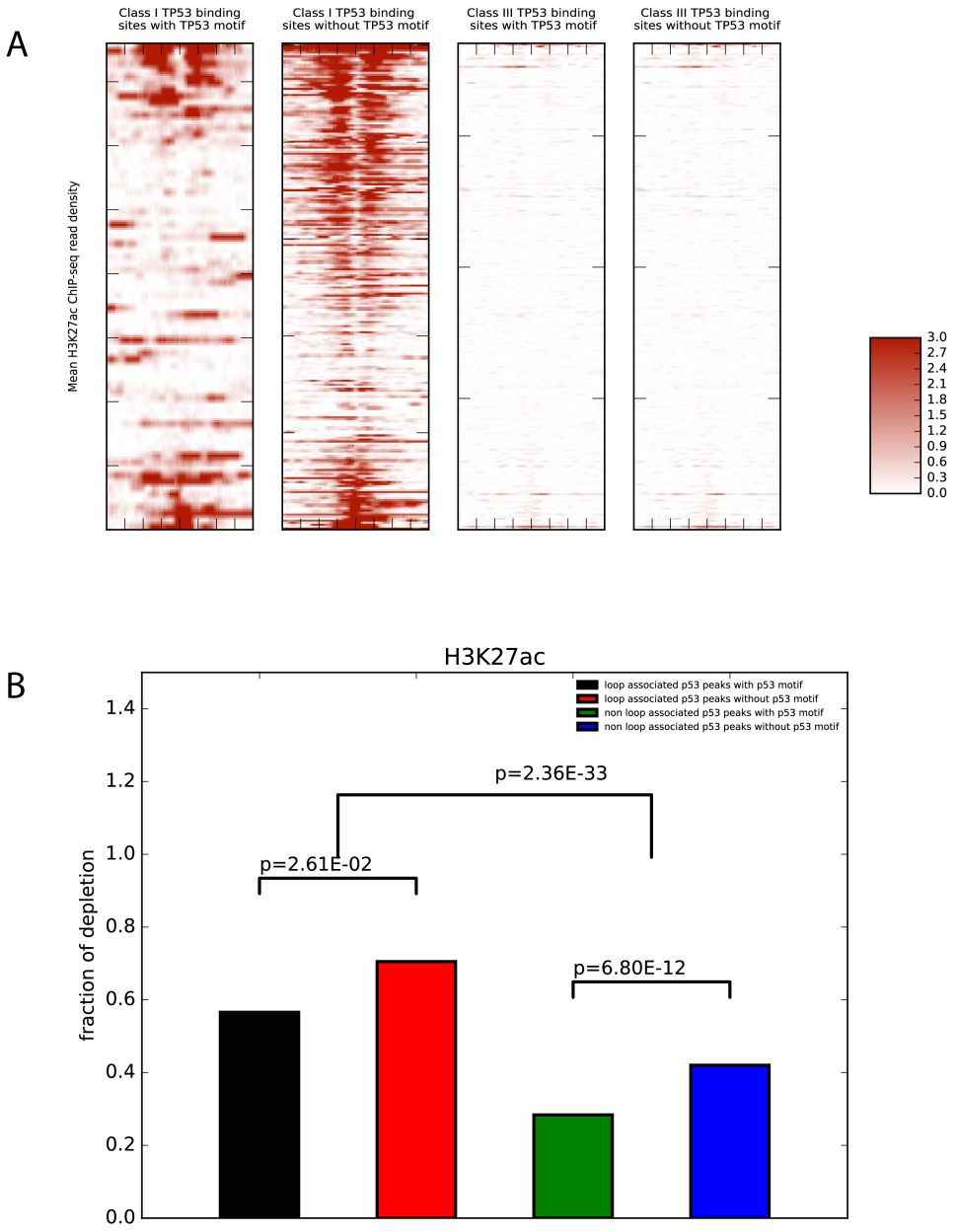

Figure A.6. (A) Heatmap show of H3K27ac signal around Class I and Class III TP53 binding sites with and without TP53 motif. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites (B) H3K27ac modification depletion in four groups of TP53 binding sites. To quantify central H3K27ac modification depletion level, we defined h_i as the difference between log2 transformed ratio of read-coverage against input in center region(±100bp region relative to peak summit) to average of read-coverage against input in two flaking regions (-600 bp to -400bp and +400bp to +600bp relative to peak summit) of TP53 binding site i for H3K27ac modification. Barplot shows the faction of depletion with $h_i < 0$.

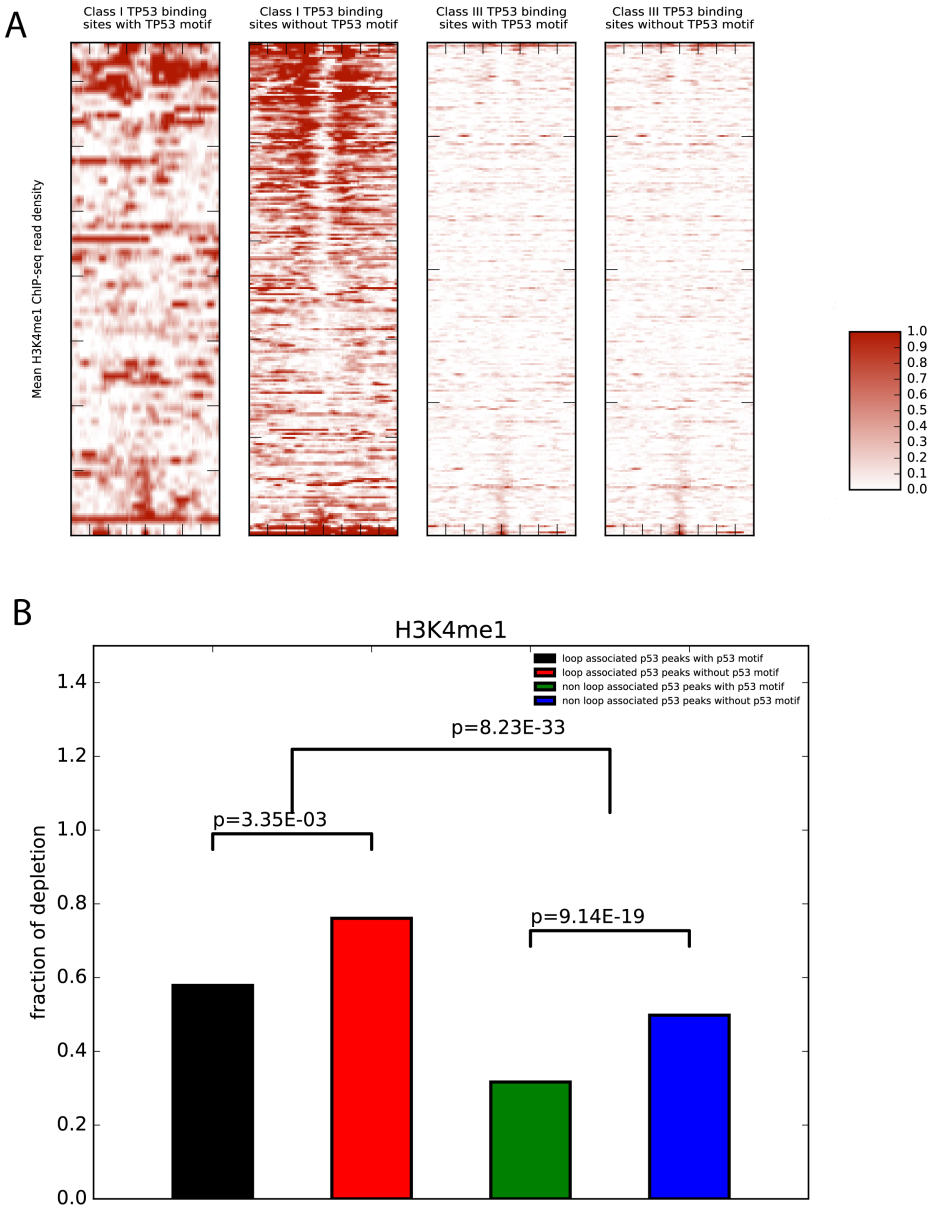

Figure A.7. (A) Heatmap show of H3K4me1 signal around Class I and Class III TP53 binding sites with and without TP53 motif. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites (B) H3K4me1 modification depletion in four groups of TP53 binding sites. To quantify central H3K4me1 modification depletion level, we defined h_i as the difference between log2 transformed ratio of read-coverage against input in center region(±100bp region relative to peak summit) to average of read-coverage against input in two flaking regions (-600 bp to -400bp and +400bp to +600bp relative to peak summit) of TP53 binding site i for H3K4me1 modification. Barplot shows the faction of depletion with $h_i < 0$.

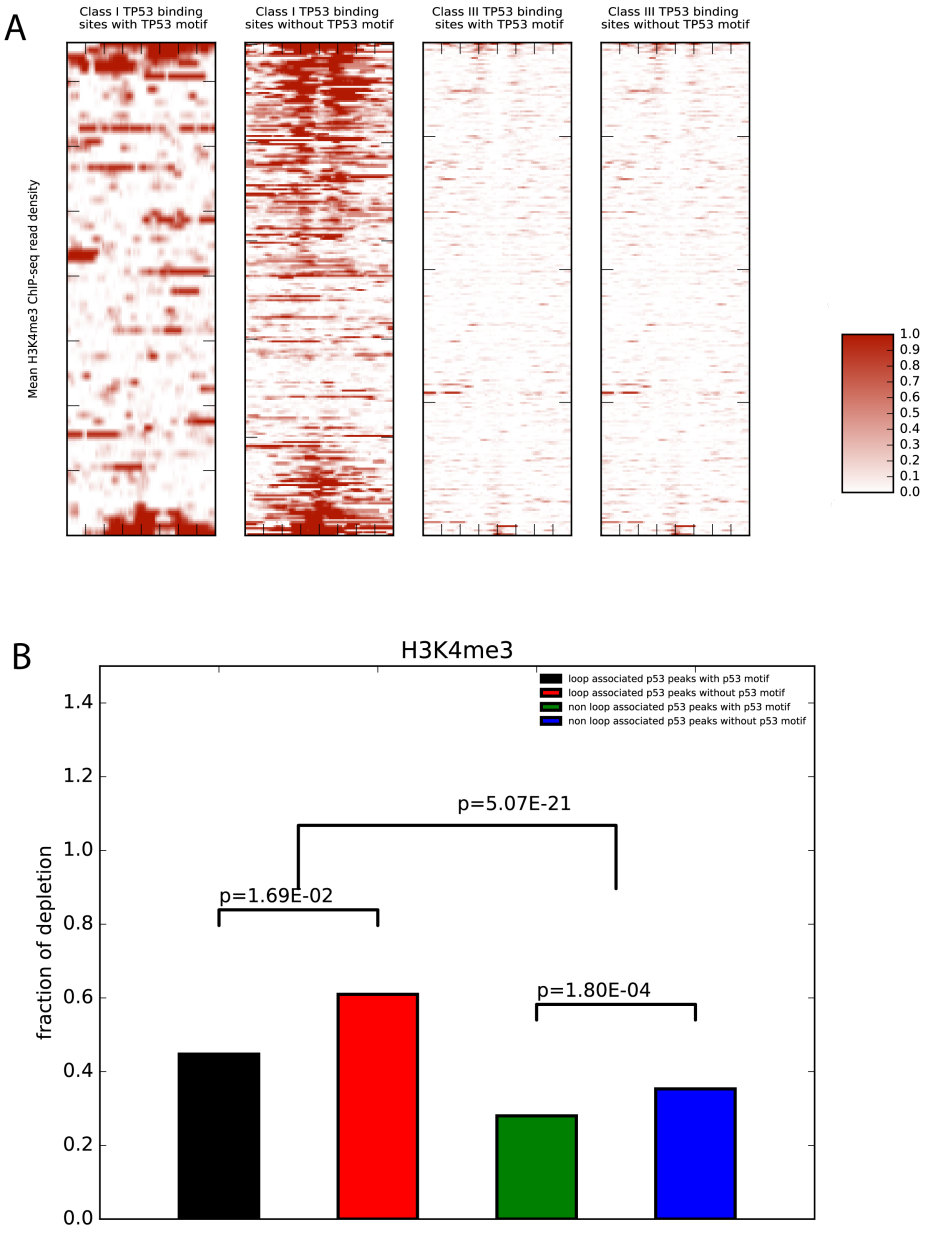

Figure A.8. (A) Heatmap show of H3K4me3 signal around Class I and Class III TP53 binding sites with and without TP53 motif. Each row represents signal strengths of ±2kb region around the summits of TP53 binding sites (B) H3K4me3 modification depletion in four groups of TP53 binding sites. To quantify central H3K4me3 modification depletion level, we defined h_i as the difference between log2 transformed ratio of read-coverage against input in center region(±100bp region relative to peak summit) to average of read-coverage against input in two flaking regions (-600 bp to -400bp and +400bp to +600bp relative to peak summit) of TP53 binding site i for H3K4me3 modification. Barplot shows the faction of depletion with $h_i < 0$.

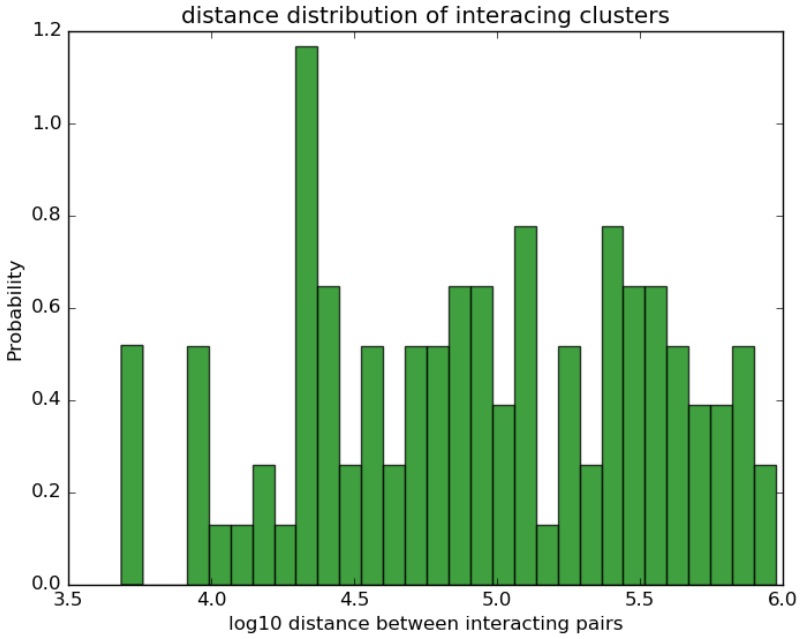

Figure A.9. Distance distribution of interacting clusters

REFERENCES

Allen, M. A., Z. Andrysik, V. L. Dengler, H. S. Mellert, A. Guarnieri, J. A. Freeman, K. D. Sullivan, M. D. Galbraith, X. Luo, W. L. Kraus, R. D. Dowell, and J. M. Espinosa (2014). Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. *Elife 3*, e02200.

Bailey, T. L. and C. Elkan (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol 2*, 28–36.

Biddie, S. C., S. John, P. J. Sabo, R. E. Thurman, T. A. Johnson, R. L. Schiltz, T. B. Miranda, M. H. Sung, S. Trump, S. L. Lightman, C. Vinson, J. A. Stamatoyannopoulos, and G. L. Hager (2011). Transcription factor ap1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell 43*(1), 145–55.

Bieging, K. T., S. S. Mello, and L. D. Attardi (2014). Unravelling mechanisms of p53-mediated tumour suppression. *Nat Rev Cancer 14*(5), 359–70.

Bodenmiller, D. M. and S. Spiro (2006a). The yjeb (nsrr) gene of escherichia coli encodes a nitric oxide-sensitive transcriptional regulator. *J Bacteriol 188*(3), 874–81.

Bodenmiller, D. M. and S. Spiro (2006b). The yjeb (nsrr) gene of escherichia coli encodes a nitric oxide-sensitive transcriptional regulator. *J Bacteriol 188*(3), 874–81.

Borggrefe, T. and X. Yue (2011). Interactions between subunits of the mediator complex with gene-specific transcription factors. *Semin Cell Dev Biol 22*(7), 759–68.

Bower, J. M., H. B. Gordon-Raagas, and M. A. Mulvey (2009). Conditioning of uropathogenic escherichia coli for enhanced colonization of host. *Infect Immun 77*(5), 2104–12.

Bower, J. M. and M. A. Mulvey (2006). Polyamine-mediated resistance of uropathogenic escherichia coli to nitrosative stress. *J Bacteriol 188*(3), 928–33.

Branchu, P., S. Matrat, M. Vareille, A. Garrivier, A. Durand, S. Crepin, J. Harel, G. Jubelin, and A. P. Gobert (2014). Nsrr, gade, and gadx interplay in repressing expression of the escherichia coli o157:h7 lee pathogenicity island in response to nitric oxide. *PLoS Pathog 10*(1), e1003874.

Crack, J. C., D. A. Svistunenko, J. Munnoch, A. J. Thomson, M. I. Hutchings, and N. E. Le Brun (2016). Differentiated, promoter-specific response of [4fe-4s] nsrr dna binding to reaction with nitric oxide. *J Biol Chem 291(16)*, 8663–72.

de Wit, E. and W. de Laat (2012). A decade of 3c technologies: insights into nuclear organization. *Genes Dev 26*(1), 11–24.

Dekker, J., K. Rippe, M. Dekker, and N. Kleckner (2002). Capturing chromosome conformation. *Science 295*(5558), 1306–11.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res 16*(10), 1299–309.

Efromovich, S., D. Grainger, D. Bodenmiller, and S. Spiro (2008). Genome-wide identification of binding sites for the nitric oxide-sensitive transcriptional regulator nsrr. *Methods Enzymol 437*, 211–33.

Fang, F. C. (1997). Perspectives series: host/pathogen interactions. mechanisms of nitric oxide-related antimicrobial activity. *J Clin Invest 99*(12), 2818–25.

Fang, F. C. and A. Vazquez-Torres (2002). Nitric oxide production by human macrophages: there's no doubt about it. *Am J Physiol Lung Cell Mol Physiol 282*(5), L941–3.

Filenko, N., S. Spiro, D. F. Browning, D. Squire, T. W. Overton, J. Cole, and C. Constantinidou (2007). The nsrr regulon of escherichia coli k-12 includes genes encoding the hybrid cluster protein and the periplasmic, respiratory nitrite reductase. *J Bacteriol 189*(12), 4410–7.

Fullwood, M. J., M. H. Liu, Y. F. Pan, J. Liu, H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H. S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K. V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H. G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature 462*(7269), 58–64.

Gardner, A. M. and P. R. Gardner (2002). Flavohemoglobin detoxifies nitric oxide in aerobic, but not anaerobic, escherichia coli. evidence for a novel inducible anaerobic nitric oxide-scavenging activity. *J Biol Chem 277*(10), 8166–71.

Gardner, A. M., R. A. Helmick, and P. R. Gardner (2002). Flavorubredoxin, an inducible catalyst for nitric oxide reduction and detoxification in escherichia coli. *J Biol Chem 277*(10), 8172–7.

Gertz, J., D. Savic, K. E. Varley, E. C. Partridge, A. Safi, P. Jain, G. M. Cooper, T. E. Reddy, G. E. Crawford, and R. M. Myers (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol Cell 52*(1), 25–36.

Giaccia, A. J. and M. B. Kastan (1998). The complexity of p53 modulation: emerging patterns from divergent signals. *Genes Dev 12*(19), 2973–83.

Godaly, G., G. Bergsten, L. Hang, H. Fischer, B. Frendeus, A. C. Lundstedt, M. Samuelsson, P. Samuelsson, and C. Svanborg (2001). Neutrophil recruitment, chemokine receptors, and resistance to mucosal infection. *J Leukoc Biol 69*(6), 899–906.

Hacker, J., G. Blum-Oehler, I. Mhldorfer, and H. Tschpe (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol 6*, 1089–97.

Hagan, E. C., A. L. Lloyd, D. A. Rasko, G. J. Faerber, and H. L. Mobley (2010). Escherichia coli global gene expression in urine from women with urinary tract infection. *PLoS Pathog 6*(11), e1001187.

Hausladen, A., A. Gow, and J. S. Stamler (2001). Flavohemoglobin denitrosylase catalyzes the reaction of a nitroxyl equivalent with molecular oxygen. *Proc Natl Acad Sci U S A 98*(18), 10108–12.

He, C., X. Wang, and M. Q. Zhang (2014). Nucleosome eviction and multiple co-factor binding predict estrogen-receptor-alpha-associated long-range interactions. *Nucleic Acids Res 42*(11), 6935–44.

He, X., Z. Zhu, C. Johnson, J. Stoops, A. E. Eaker, W. Bowen, and M. C. DeFrances (2008). Pik3ip1, a negative regulator of pi3k, suppresses the development of hepatocellular carcinoma. *Cancer Res 68*(14), 5591–8.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol Cell 38*(4), 576–89.

Hertz, G. Z. and G. D. Stormo (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics 15*(7-8), 563–77.

Hon, G. C., R. D. Hawkins, and B. Ren (2009). Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet 18*(R2), R195–201.

Hu, W., Z. Feng, and A. J. Levine (2012). The regulation of multiple p53 stress responses is mediated through mdm2. *Genes Cancer 3*(3-4), 199–208.

Hyduke, D. R., L. R. Jarboe, L. M. Tran, K. J. Chou, and J. C. Liao (2007). Integrated network analysis identifies nitric oxide response networks and dihydroxyacid dehydratase as a crucial target in escherichia coli. *Proc Natl Acad Sci U S A 104*(20), 8484–9.

Janky, R., A. Verfaillie, H. Imrichova, B. Van de Sande, L. Standaert, V. Christiaens, G. Hulselmans, K. Herten, M. Naval Sanchez, D. Potier, D. Svetlichnyy, Z. Kalender Atak, M. Fiers, J. C. Marine, and S. Aerts (2014). iregulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput Biol 10*(7), e1003731.

Kim, Y. M., H. A. Bergonia, C. Muller, B. R. Pitt, W. D. Watkins, and J. Lancaster, J. R. (1995). Loss and degradation of enzyme-bound heme induced by cellular nitric oxide synthesis. *J Biol Chem 270*(11), 5710–3.

Lam, M. T., W. Li, M. G. Rosenfeld, and C. K. Glass (2014). Enhancer rnas and regulated transcriptional programs. *Trends Biochem Sci 39*(4), 170–82.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol 10*(3), R25.

Li, G., X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell 148*(1-2), 84–98.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science 326*(5950), 289–93.

Lin, H. Y., P. J. Bledsoe, and V. Stewart (2007). Activation of year-yoag operon transcription by the nitrate-responsive regulator narl is independent of oxygen- responsive regulator fnr in escherichia coli k-12. *J Bacteriol 189*(21), 7539–48.

Lin, Y. C., C. Benner, R. Mansson, S. Heinz, K. Miyazaki, M. Miyazaki, V. Chandra, C. Bossen, C. K. Glass, and C. Murre (2012). Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate b cell fate. *Nat Immunol 13*(12), 1196–204.

Link, N., P. Kurtz, M. O'Neal, G. Garcia-Hughes, and J. M. Abrams (2013). A p53 enhancer region regulates target genes through chromatin conformations in cis and in trans. *Genes Dev 27*(22), 2433–8.

Luo, C., G. Hu, and H. Zhu (2009). Genome reannotation of escherichia coli cft073 with new insights into virulence. *BMC Genomics 10*, 552–61.

Mehta, H., Y. Liu, M. Q. Zhang, and S. Spiro (2015). Genome-wide analysis of the response to nitric oxide in uropathogenic escherichia coli cft073. *Microbial Genomics 1*.

Melo, C. A., J. Drost, P. J. Wijchers, H. van de Werken, E. de Wit, J. A. Oude Vrielink, R. Elkon, S. A. Melo, N. Leveille, R. Kalluri, W. de Laat, and R. Agami (2013). ernas are required for p53-dependent enhancer activity and gene transcription. *Mol Cell 49*(3), 524–35.

Melton, C., J. A. Reuter, D. V. Spacek, and M. Snyder (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics 47*, 71016.

Mowat, C. G., B. Gazur, L. P. Campbell, and S. K. Chapman (2010). Flavin-containing heme enzymes. *Arch Biochem Biophys 493*(1), 37–52.

Mukhopadhyay, P., M. Zheng, L. A. Bedzyk, R. A. LaRossa, and G. Storz (2004). Prominent roles of the norr and fur regulators in the escherichia coli transcriptional response to reactive nitrogen species. *Proc Natl Acad Sci U S A 101*(3), 745–50.

Myers, K. S., H. Yan, I. M. Ong, D. Chung, K. Liang, F. Tran, S. Keles, R. Landick, and P. J. Kiley (2013). Genome-scale analysis of escherichia coli fnr reveals complex features of transcription factor binding. *PLoS Genet 9*(6), e1003565.

Natarajan, A., G. G. Yardimci, N. C. Sheffield, G. E. Crawford, and U. Ohler (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res 22*(9), 1711–22.

Nikulenkov, F., C. Spinnler, H. Li, C. Tonelli, Y. Shi, M. Turunen, T. Kivioja, I. Ignatiev, A. Kel, J. Taipale, and G. Selivanova (2012). Insights into p53 transcriptional function via genome-wide chromatin occupancy and gene expression analysis. *Cell Death Differ 19*(12), 1992–2002.

Partridge, J. D., D. M. Bodenmiller, M. S. Humphrys, and S. Spiro (2009). Nsrr targets in the escherichia coli genome: new insights into dna sequence requirements for binding and a role for nsrr in the regulation of motility. *Mol Microbiol 73*(4), 680–94.

Poock, S. R., E. R. Leach, J. W. Moir, J. A. Cole, and D. J. Richardson (2002). Respiratory detoxification of nitric oxide by the cytochrome c nitrite reductase of escherichia coli. *J Biol Chem 277*(26), 23664–9.

Rankin, L. D., D. M. Bodenmiller, J. D. Partridge, S. F. Nishino, J. C. Spain, and S. Spiro (2008). Escherichia coli nsrr regulates a pathway for the oxidation of 3-nitrotyramine to 4-hydroxy-3-nitrophenylacetate. *J Bacteriol 190*(18), 6170–7.

Ren, B., N. Zhang, J. Yang, and H. Ding (2008). Nitric oxide-induced bacteriostasis and modification of iron-sulphur proteins in escherichia coli. *Mol Microbiol 70*(4), 953–64.

Rogozin, I. B., K. S. Makarova, D. A. Natale, A. N. Spiridonov, R. L. Tatusov, Y. I. Wolf, J. Yin, and E. V. Koonina (2002). Congruent evolution of different classes of non-coding dna in prokaryotic genomes. *Nucleic Acids Res 30(19)*, 426471.

Roos, V. and P. Klemm (2006). Global gene expression profiling of the asymptomatic bacteriuria escherichia coli strain 83972 in the human urinary tract. *Infect Immun 74*(6), 3565–75.

Rowland, B. D., R. Bernards, and D. S. Peeper (2005). The klf4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene. *Nat Cell Biol 7*(11), 1074–82.

Sammons, M. A., J. Zhu, A. M. Drake, and S. L. Berger (2015). Tp53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. *Genome Res 25*(2), 179–88.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker (2012). The long-range interaction landscape of gene promoters. *Nature 489*(7414), 109–13.

Seo, S. W., D. Kim, H. Latif, E. J. O'Brien, R. Szubin, and B. O. Palsson (2014). Deciphering fur transcriptional regulatory network highlights its complex role beyond iron metabolism in escherichia coli. *Nat Commun 5*, 4910.

Shlyueva, D., G. Stampfel, and A. Stark (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet 15*(4), 272–86.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat Genet 38*(11), 1348–54.

Snyder, J. A., B. J. Haugen, E. L. Buckles, C. V. Lockatell, D. E. Johnson, M. S. Donnenberg, R. A. Welch, and H. L. Mobley (2004). Transcriptome of uropathogenic escherichia coli during urinary tract infection. *Infect Immun 72*(11), 6373–81.

Spiro, S. (2007). Regulators of bacterial responses to nitric oxide. *FEMS Microbiol Rev 31*(2), 193–211.

Svensson, L., B. I. Marklund, M. Poljakovic, and K. Persson (2006). Uropathogenic escherichia coli and tolerance to nitric oxide: the role of flavohemoglobin. *J Urol 175*(2), 749–53.

Taberlay, P. C., A. L. Statham, T. K. Kelly, S. J. Clark, and P. A. Jones (2014). Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies dna methylation of enhancers and insulators in cancer. *Genome Res 24*(9), 1421–32.

The International Human Genome Sequencing Consortium, . (2004). *Finishing the euchromatic sequence of the human genome.*

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos (2012). The accessible chromatin landscape of the human genome. *Nature 489*(7414), 75–82.

Tucker, N. P., M. G. Hicks, T. A. Clarke, J. C. Crack, G. Chandra, N. E. Le Brun, R. Dixon, and M. I. Hutchings (2008). The transcriptional repressor protein nsrr senses nitric oxide directly via a [2fe-2s] cluster. *PLoS One 3*(11), e3623.

Tucker, N. P., N. E. Le Brun, R. Dixon, and M. I. Hutchings (2010). There's no stopping nsrr, a global regulator of the bacterial no stress response. *Trends Microbiol 18*(4), 149–56.

Verfaillie, A., D. Svetlichnyy, H. Imrichova, K. Davie, M. Fiers, Z. Kalender Atak, G. Hulselmans, V. Christiaens, and S. Aerts (2016). Multiplex enhancer-reporter assays uncover unsophisticated tp53 enhancer logic. *Genome Res 26*(7), 882–95.

Vousden, K. H. and D. P. Lane (2007). p53 in health and disease. *Nat Rev Mol Cell Biol 8*(4), 275–83.

Welch, R., V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and F. Blattner (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli. *Proc Natl Acad Sci U S A 99*, 17020–4.

Yoon, H. S., X. Chen, and V. W. Yang (2003). Kruppel-like factor 4 mediates p53-dependent g1/s cell cycle arrest in response to dna damage. *J Biol Chem 278*(4), 2101–5.

Young, R. A. (2011). Control of the embryonic stem cell state. *Cell 144*(6), 940–54.

Yukl, E. T., M. A. Elbaz, M. M. Nakano, and P. Moenne-Loccoz (2008). Transcription factor nsrr from bacillus subtilis senses nitric oxide with a 4fe-4s cluster (dagger). *Biochemistry 47*(49), 13084–92.

Zhang, W., D. E. Geiman, J. M. Shields, D. T. Dang, C. S. Mahatan, K. H. Kaestner, J. R. Biggs, A. S. Kraft, and V. W. Yang (2000). The gut-enriched kruppel-like factor (kruppel-like factor 4) mediates the transactivating effect of p53 on the p21waf1/cip1 promoter. *J Biol Chem 275*(24), 18391–8.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu (2008). Model-based analysis of chip-seq (macs). *Genome Biol 9*(9), R137.

VITA

Yuxuan Liu was born in Henan, China. She entered the School of Life Science at Shandong University in 2006 and earned a Bachelor's degree in Science in the year of 2010. Yuxuan Liu joined the graduate program in the Department of Biological Sciences, UT Dallas in August 2010, and received her Master of Science in Molecular and Cell Biology and Mathematics in 2013 and 2014.