The dissertation of Gue Su Chang was reviewed and approved* by the following:

B. Franklin Pugh  
Willaman Chair in Molecular Biology and  
Professor of Biochemistry and Molecular Biology  
Dissertation Advisor  
Chair of Committee

Ross Hardison  
T. Ming Chu Professor of Biochemistry and Molecular Biology

Istvan Albert  
Associate Professor of Bioinformatics and  
Biochemistry and Molecular Biology

Yu Zhang  
Associate Professor of Statistics

Peter Hudson  
William Professor of Biology  
Head of the Huck Institutes of the Life Sciences

*Signatures are on file in the Graduate School
How is gene expression controlled in eukaryotic organisms? Recently, the Encyclopedia of DNA Elements (or ENCODE) project has systematically mapped regulatory protein-DNA interactions including transcription-associated factors, epigenetic modifications, and 3D chromosome conformation. This research has uncovered unprecedented details on eukaryotic gene control, and these high-throughput genomic data have led to many challenges in genomics and bioinformatics. This dissertation presents an integrative genome-wide approach to understand transcriptional regulation in eukaryotes, by utilizing massively parallel sequencing and bioinformatics. In order to achieve this research goal, we performed functional, comparative, and statistical genomics analyses with high-throughput genomic data.

Gene and nucleosome organization in the Dictyostelium genome

Genome-wide mapping of nucleosomes has significantly expanded our understanding of chromatin structure and function in eukaryotic transcriptional initiation and regulation. We present the first high-resolution maps of in vivo nucleosome locations for the social amoeba Dictyostelium discoideum. Its exceptional A/T-richness (78%) enabled us to study the role of the extreme nucleotide usage in organizing nucleosomes, genes, and promoters across the genome. Our functional and comparative genomics analysis revealed a variety of functionally distinct polymeric A/T elements in the Dictyostelium genome. These tracts established the boundary of Dictyostelium genes, associated with nucleosome-free regions and precisely positioned TATA boxes in the promoter. Dictyostelium utilized polymeric-A/T elements for nucleosome placement. Moreover as situated in an earliest branch from the last common ancestor of all eukaryotes, the exceptional ability of D. discoideum in alternating unicellular and multicellular form provided us with an ideal system to understand which principle governs the evolution of the chromatin
structure across eukaryotes by emphasizing multicellular development. Surprisingly, *Dictyostelium* chromatin was organized as in higher multicellular eukaryotes. The phylogenetic linkage in the NELF homology and position of the first genic nucleosome across major eukaryotes implies that transcriptional regulation imposed on multicellularity has interplayed with eukaryotic chromatin over evolutionary time.

**Comprehensive and high resolution genome-wide response of p53 to UV-damage**

TP53 has been found the most frequently mutated among human cancers (≈50%) and intensively studied because of its tumor suppressive function. p53 transactivates a variety of genes in response to cellular signals and environmental stimuli. This activity requires DNA sequence-specific binding, which is notably degenerate in sequence requirement. Thus, a major challenge has been genome-wide characterization of interaction between p53 and its response elements (or REs). We employed ChIP-exo as a high-resolution genome-wide mapping assay, and comprehensively identified ~2,000 p53-bound REs across the human genome in response to UV-induced stress. Strikingly a characteristic 6-peak ChIP-exo pattern was associated with p53/RE binding, and a half-site overlapping spatial relationship was commonly found between REs. p53 regulates a subset of genes in each stress response, and how p53 achieves such specificity has been a long-standing question. Our systematic motif analysis revealed a stereotyped spatial arrangement of p53 REs with other nearby stress-response elements, such as AP1, NRF2, FOXO3. This result may provide a mechanistic basis for p53-mediated response specificity. p53 usually binds to distant REs for target gene activation, which leads to a challenge in identifying target genes. Our ChIP-exo mapping of the transcription preinitiation complex components, TFIIB and Pol II, provided not only a reliable way to identify transcription factor target genes, but also a genome-wide insight into p53-dependent transactivation mechanism. 154 genes activated by p53 in UV-stress response were identified (80% novel), which were involved
in various functions such as cell growth and death control, and DNA repair. This high-confidence target gene set greatly expanded our understanding of p53-regulated DNA repair and cell proliferation network. Recent GWAS evidence indicates that sequence variations in non-coding DNA can be a significant risk factor for disease, but their catalogue remains a challenge. Our comprehensive search for single nucleotide polymorphism (or SNP) present at p53-bound REs showed a strong association between a SNP at the UV-inducible p53 RE for the POLH gene and Xeroderma pigmentosum variant (or XPV). This result may exemplify how non-coding regulatory variants contribute to gene expression and human disease.
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................................................................. ix

LIST OF TABLES ................................................................................................................... x

ACKNOWLEDGEMENTS ....................................................................................................... xi

Chapter 1  Introduction ............................................................................................................ 1
  High-throughput sequencing technology in genomics .......................................................... 3
  Eukaryotic gene control and epigenetics ......................................................................... 5
  Regulatory diversification and organismal complexity ...................................................... 8
  Human genome variation and common disease .............................................................. 10
  Cancer genomics and personalized medicine ................................................................. 12
  References ..................................................................................................................... 15

Chapter 2  Unusual Combinatorial Involvement of Poly-A/T Tracts in Organizing Genes and Chromatin in *Dictyostelium* .............................................................................. 22
  Summary .......................................................................................................................... 23
  Introduction ..................................................................................................................... 23
  Results .............................................................................................................................. 26
    *Dictyostelium* start their genes with poly-T and end with poly-A ................................ 26
    Poly-T tracts in promoters are interrupted by TATA boxes ........................................... 29
    *Dictyostelium* nucleosomes have polymeric-A/T borders ....................................... 31
    *Dictyostelium* organizes its nucleosomes on genes similar to animals ....................... 33
    An evolutionarily conserved change in genic nucleosome spacing upon starvation. ..... 35
    *Dictyostelium* organizes genic nucleosomes into pairs ............................................. 37
    Evolutionary linkage between RNA polymerase II pausing and a downstream placement of the +1 nucleosome ................................................................. 37
  Discussion ........................................................................................................................ 40
    Is *Dictyostelium* moving towards a binary genomic code? ........................................ 40
    *Dictyostelium* organizes genic nucleosomes as in animals ....................................... 42
  Materials and Methods ................................................................................................ 43
    Growth and development of *D. discoideum* ................................................................ 43
    Preparation of nucleosomes ..................................................................................... 44
    Nucleosome mapping ............................................................................................... 45
    Transcriptome sequencing and analysis .................................................................... 46
    Data analysis ............................................................................................................. 48
  References ..................................................................................................................... 48
Chapter 3  A Comprehensive and High Resolution Genome-wide Response of p53 to UV-damage

Summary ..........................................................................................................................55
Introduction.....................................................................................................................55
Results.............................................................................................................................58
  A triple peak-pair is diagnostic of p53/RE binding.....................................................58
  UV induces p53 binding to weak REs..........................................................................61
  REs lack insertions.......................................................................................................64
  REs are spaced at 10 and 250 bp intervals.................................................................65
  Noncognate p53 crosslinking is dubious....................................................................67
  Potential enhancesome-like relationship between p53 and cellular factors..............67
  Linking p53 binding to target genes and UV-induced release of paused Pol II.........71
  p53 regulatory networks and the influence of SNPs....................................................75
Discussion .......................................................................................................................80
  p53-mediated transactivation ....................................................................................80
  Overall assessment of binding locations and RE....................................................81
  p53 regulatory network ............................................................................................82
Materials and Methods.................................................................................................83
  Cell culture and UV treatment ................................................................................83
  ChIP-exo ..................................................................................................................83
  Accession Numbers..................................................................................................84
References......................................................................................................................84

Chapter 4  Conclusion.........................................................................................................90

Summary of study ..........................................................................................................90
Major contribution and significance ............................................................................94
Future work....................................................................................................................98
References......................................................................................................................98

Appendix A  Supplemental Information for Chapter 2.....................................................103

Extended Methods and Materials.................................................................................103
  Data analysis .............................................................................................................103
  External data sources...............................................................................................104
  Data visualization and statistical significance test ...................................................105
  Supplementary Figures .........................................................................................107
  Supplementary Tables .............................................................................................125
  Supplementary References.......................................................................................129

Appendix B  Supplemental Information for Chapter 3.....................................................131

Extended Methods and Materials.................................................................................131
  Human genome annotation......................................................................................131
  External data sources...............................................................................................131
  Identification of p53 binding sites............................................................................132
  TFIIB and Pol II.......................................................................................................135
  p53 regulatory networks.........................................................................................137
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Dictyostelium gene organization with poly-T/poly-A enrichment</td>
<td>28</td>
</tr>
<tr>
<td>2-2</td>
<td>Frequency distribution of TATA elements as TATAAA(A/T)(A/T) midpoint locations around the 5’ and 3’ ends of Dictyostelium genes</td>
<td>30</td>
</tr>
<tr>
<td>2-3</td>
<td>Nucleosomal DNA properties of Dictyostelium nucleosomes</td>
<td>32</td>
</tr>
<tr>
<td>2-4</td>
<td>Nucleosome organization around the 5’ and 3’ end of the Dictyostelium genes</td>
<td>35</td>
</tr>
<tr>
<td>2-5</td>
<td>Comparison of inter-nucleosomal distances between five species in their nucleosome arrangement within the genic region</td>
<td>36</td>
</tr>
<tr>
<td>2-6</td>
<td>Chromatin structure around the 5’ end of genes evolutionarily conserved across major eukaryotes</td>
<td>39</td>
</tr>
<tr>
<td>3-1</td>
<td>Comprehensive genome-wide identification of p53-bound locations in response to UV DNA damage</td>
<td>60</td>
</tr>
<tr>
<td>3-2</td>
<td>Sequence assessment of a comprehensive set of REs</td>
<td>63</td>
</tr>
<tr>
<td>3-3</td>
<td>Spatial relationship of REs</td>
<td>66</td>
</tr>
<tr>
<td>3-4</td>
<td>Fixed orientation and distance of stress-response motifs relative to p53 REs</td>
<td>70</td>
</tr>
<tr>
<td>3-5</td>
<td>Positional relationship of p53 to the 5’ end of genes</td>
<td>74</td>
</tr>
<tr>
<td>3-6</td>
<td>p53 regulatory networks and SNPs</td>
<td>78</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 3-1. Categories of p53 target genes linked to p53-bound locations using ChIP-exo analysis

79
FOREMOST I WOULD LIKE TO THANK MY ADVISOR, DR. FRANK PUGH, WHO GUIDED ME THROUGH THIS CHALLENGE, WITH PATIENCE IN MY DEVELOPMENT OVER GRADUATE YEARS. HIS ENTHUSIASM IN RESEARCH WAS WHAT I LEARNED, HOW TO BECOME A QUALIFIED SCIENTIST. MANY THANKS TO MY COMMITTEE MEMBERS, DR. ISTVAN ALBERT AND DR. YU ZHANG WHO WERE NEVER TOO BUSY TO TOLERATE MY INTERRUPTIONS. AND SPECIAL THANKS TO DR. ROSS HARDISON AND DR. STEPHAN SCHUSTER WHO ALWAYS FOUND THEIR TIME FOR VALUABLE AND INSIGHTFUL DISCUSSION IN MY PROJECTS. ALONGSIDE MY COMMITTEE, I WANT TO EXPRESS MY SPECIAL THANKS TO DR. JOHN CARLSON AS MY MENTOR WHO LISTENED AND OFFERED MUCH HELP AND ENCOURAGEMENT AT MANY STAGES. DR. COODUVALLI SHASHIKANT, ANN MARIE DANIEL AND MICHAEL RADIS AS OUR IBIOS DIRECTORS DESERVING OF SPECIAL MENTION AND THANKS FOR THEIR NICE CARING. MANY THANKS TO JANICE KENNEDY WHO MADE MY STUDENT LIFE BETTER FOR HER KINDNESS AND RESOURCEFULNESS. MANY THANKS TO MY RESEARCH COLLEAGUES, WHO I HAVE BEEN SO FORTUNATE TO WORK WITH IN SUCH A COMMUNITY OF INTELLECTUALS, CHRISTINE WALSH, ELISSA WARD, AND TRAVIS MAVRICH, CIZHONG JIANG, OLA SODEINDE, KIRAN BATTA, KUANGYU YEN, BRYAN VENTERS, AND SHAUN MAHONY, HO SUNG LEE, ERIC SAMORODNITSKY, AND SUJANA GHOSH, VINESH VINAYACHANDRAN AND MATTHEW ROSSI, MEGHA WAL, GARAM HAN, ROHIT REJA, HAN ZHENG, AND YUNFEI LI. I AM PROUD OF OUR NICE COLLABORATORS, DR. ANGELIKA NOEGEL, DR. LUDWIG EICHINGER, AND DR. GERNOT GLÖCKNER, DR. YANMING WANG AND XIANGYUN AMY CHEN. ALSO, CHUNGOO PARK, MELISSA WILSON, GURU ANANDA, SUNANDO ROY, YOGESHWAR KELKAR AS MY FELLOW GRADUATE STUDENTS.

MY HEARTY THANKS TO MY PARENTS AND PARENTS-IN-LAW, NOT ONLY FOR THEIR CONSTANT INSPIRATION AND SUPPORT, BUT ALSO FOR THEIR UNFAILING LOVE. LAST BUT DEFINITELY NOT LEAST, MY WIFE, HYUN A KONG, AND MY LOVELY SONS, YOON SEO AND YOON JAE, WHO STOOD WITH ME THROUGH THIS CHALLENGE. I LOVE YOU AND LOOK FORWARD TO BEING ABLE TO THANK-YOU FOR MANY YEARS TO COME.
Chapter 1

Introduction

Recent advance in high-throughput profiling technology (e.g. ChIP-chip and ChIP-seq) is now spurring systematic efforts to understand the complex regulatory network that modulates gene transcription and chromatin structure in the eukaryotic genomes. Availability of complete genome sequences and improvement in high-performance computing and algorithms for fast high-throughput data process have made broad applications for such genomic data. This progress provides new routes to understand a yet unexplored aspect of the eukaryotic genomes, and continuously raises more intriguing questions on eukaryotic gene regulation. Furthermore, the challenge, how to integrate and interpret these complex high-throughput genomic data, still remains to be explored and under development, as massive data are being generated.

In this dissertation, I carried out two genome-wide studies with high-throughput sequencing data to research the mechanism of eukaryotic transcription control—(1) a question on chromatin organization and epigenetic gene regulation by genome-wide mapping of nucleosome locations utilizing micrococcal nuclease digestion (or MNase-seq), and (2) a question on transcription regulation mechanism in stress response by comprehensive characterization of interactions between transcription factors and the human genome, using chromatin immunoprecipitation and DNase digestion (or ChIP-exo).

First, I present the first high-resolution genome-wide nucleosome maps of the social amoeba, Dictyostelium discoideum. In particular, it has the highly A/T-rich genome (78%), surpassed only by the 81% A/T-richness of Plasmodium falciparum. It has been widely used as a model organism for multicellular development, because of its exceptional life cycle between a free-living unicellular cell and a multicellular form which is achieved by cellular aggregation.
Dictyostelium is one of the deepest-branching eukaryotic lineages in the life tree, diverged just after plants but before the divergence of fungi and animals (Eichinger and Noegel, 2003; Williams et al., 2005). We studied four major questions on eukaryotic gene and nucleosome organization with this organism which have implications for epigenetic gene regulation. (1) How does highly preferred usage of adenosine and thymine base contribute in organizing chromatin, genes, and regulatory elements across the Dictyostelium genome? (2) How does the A/T-richness underlie Dictyostelium nucleosome occupancy in vivo? (3) Does Dictyostelium package their genomic DNA into chromatin in the same way as observed from other eukaryotes? Does its nucleosome organization differ between the unicellular and multicellular state? (4) What kind of insight does Dictyostelium provide into the evolutionarily conserved chromatin structure across eukaryotes?

Secondly, I present a comprehensive and high-resolution map of p53-bound locations in the human genome. p53 as the guardian of the genome (Lane, 1992) is a sequence-specific transcription factor, activated by transcriptional, post-transcriptional, and post-translational control mechanisms in response to a variety of cellular stresses including DNA damage and oncogenic stimuli. It controls major pathways to prevent cells from malignant transformation by transactivating genes involved in cell cycle regulation, apoptosis, cellular senescence, and so on (Menendez et al., 2009; Riley et al., 2008). p53 is the most commonly mutated proteins among human cancers, and its sequence-specific DNA binding is known to be critical for p53 tumor suppressive activity (Crook et al., 1994; Donehower and Bradley, 1993; Muller and Vousden, 2013). Most genome-wide applications of the current high-throughput profiling technology have been limited to statistically enriched subsets that usually encompass board regions. Thus, comprehensive and definitive genome-wide identification of p53-bound REs has been a challenge because of the notably degenerate nature of p53 response element (or RE). Here, we applied the recently developed ChIP-exo genome-wide assay (Rhee and Pugh, 2011) to
accurately and comprehensively identify p53-bound locations across the human genome in response to UV-induced stress. We studied p53-regulated transcriptional program and gene regulatory network with four major questions. (1) How many REs are bound by UV-induced p53 proteins in the human genome? (2) How does DNA sequence intrinsically modulate p53 binding? (3) What provides a mechanism for p53 target selectivity and p53-mediated target gene transactivation? (4) Which genes are activated by p53 in response to UV stress, and what kind of insight do they provide into the p53 regulatory gene network?

As a bioinformatics challenge for high-throughput complex data analysis in genomics (Pepke et al., 2009), a big challenge in this dissertation was development of statistical, mathematical, and graphical methods and tools for genomic data analysis and biological knowledge discovery, for example by incorporating relevant genomic data and attributes in parallel via integrative visualization interface. They turned out successful as scientific discovery platform in functional, comparative, and evolutionary genomics. Another challenge was computational optimization for efficient high-throughput data analysis, especially for large mammalian genomes including human.

This introduction chapter intends to provide a background in eukaryotic transcription and epigenetics, with an introductory survey on recent advance in the next generation sequencing technology (or NGS). Major challenges were discussed in each topic. Finally, I discussed the current trend in Cancer Genomics as my next research goal, which is now intensively being studied with emphasis on personalized genomics and medicine.

**High-throughput sequencing technology in genomics**

DNA sequence information is a basic resource for genetic and biomedical research, by characterizing individual genomes, transcriptional states, and genetic variations in population.
Recent advance in massively parallel synthetic sequencing technology delivers individual human genomic information to high accuracy at low cost, enabling routine whole human genome resequencing and extraction of personal genetic background. This was previously not technically affordable. In parallel, new approaches for DNA sequencing are ongoing at various stages of development and commercialization, further aimed to increase sequencing throughput and reduce cost simultaneously (Bentley et al., 2008; Metzker, 2010). This notable progress will bring more benefits to individual medical care such as disease diagnosis and therapeutic strategy, as well as for understanding basic principles of biology and human disease (Green and Guyer, 2011).

Ultimately in order for translation of this technology into molecular diagnostics and clinical treatment, there are technical issues that need to be addressed. A robust and reliable method for representative sequencing requires not only accurate sequencing but also non-biased sampling of DNA material from the genome. Most commonly used NGS technologies rely on the templates immobilized to a solid surface or support, which are amplified from single DNA molecules. However, NGS platform-specific systematic bias and data variability in short read sequencing have been reported (Dohm et al., 2008; Harismendy et al., 2009; Hillier et al., 2008). Non-uniform read-depth coverage over the whole genome was found depending on local sequence contexts, for example G/C content. Base substitution error frequencies in base calling also varied dependent on nucleotide type. These platform-specific sequencing biases may affect accuracy of sequencing variant calls, as an example, and thus have implications on the interpretation of sequencing data for de novo sequencing and whole genome resequencing, genome-wide polymorphism discovery by strain-to-reference comparison, transcriptome sequencing analysis, etc. Besides, variation in the mappability of sequencing reads over the whole genome leads to a bioinformatics challenge in this technology. The large mammalian genomes with a higher portion of low complexity regions (such as repetitive sequences) caused intrinsic algorithmic difficulty in de novo short read assembly, and fortuitous read alignments which result
in aligned reads dramatically piled up at some specific genomic regions (seemingly owing to unique alignment threshold combined with repetitive element variants by SNPs). The technical limitation in the iterative sequencing-by-synthesis process was already known that basecall accuracy decreases as the sequencing cycle goes further, primarily due to dephasing noise during successive cycles (Voelkerding et al., 2009).

As the efficiency of high-throughput DNA sequencing improves at comparable cost and time, sequencing data continue to be generated, posing an immediate challenge to the infrastructure of information technology systems, for example data storage and transfer, and quality control. Billions of raw sequencing data files and subsequent data analysis files (such as aligned and assembled read data and quality assessment reports) require a systematic maintenance connecting to laboratory information management systems such as sample tracking (Metzker, 2010). Nonetheless, the utility of NGS demonstrated for genomics and biomedical research keeps attracting attention of more and more investigators. For example such high-throughput profiling technology coupled with chromatin immunoprecipitation or ChIP-seq has made remarkable progress in genome-wide characterization of DNA-associated protein binding, chromatin organization, and epigenetic histone modification. The current NGS technology, massively parallel sequencing of clonally amplified DNA molecules, is anticipated to more translation into clinical and molecular diagnostics, and also technical improvement in robustness and high-throughput, such as real-time single-molecule DNA-sequencing and nanopore-based sequencing.

**Eukaryotic gene control and epigenetics**

Packaging of eukaryotic DNA into chromatin *in vivo* is a key context to understand fundamental relationships between cis-acting regulatory elements, gene regulatory proteins, and epigenetic templates, which ultimately modulate the transcriptional plasticity of eukaryotic
Constitutive DNase I hypersensitive sites are often present in the promoter regions of eukaryotic genes, independent of gene expression, and associated with a wide variety of regulatory elements. This observation suggests that accessibility of such DNA sequences is required for binding of regulatory proteins (Gross and Garrard, 1988; Wu, 1980). A number of studies demonstrated that cis-regulatory elements incorporated into nucleosomes become accessible by ATP-dependent chromatin remodeling during gene activation (Fascher et al., 1990; Kwon et al., 1994; Tsukiyama and Wu, 1995; Wolffe and Pruss, 1996). The chromatin structure is subject to various modifications, which provides an epigenetic mechanism for heritable changes into daughter cells without changing the underlying DNA sequence (Goldberg et al., 2007; Schones and Zhao, 2008). Combination of chromatin immunoprecipitation with next generation sequencing (or ChIP-seq) provides with high-throughput profiling of DNA methylation, histone variants and post-translational modifications, nucleosome organization, and chromatin accessibility. The recent computational approach using machine learning for extracting biological insights from whole-genomic data is now allowing researchers to decipher the complex combinatorial codes underlying epigenetic phenomena (Ernst et al., 2011; Hoffman et al., 2012). Despite such endeavor and advance, many fundamental questions still remain unanswered, for example how is organismal development in a multicellular organism achieved with cells sharing an identical genotype? What provides a mechanism for transmission of epigenetic information during somatic cell divisions and contributes to the maintenance of cellular identity in multicellular organisms? (Margueron and Reinberg, 2010)

Transcriptional regulatory networks in eukaryotes involve RNA polymerases (or Pol), general transcriptional factors (or GTFs), sequence-specific DNA binding transcription factors (or TFs), and coregulatory factors, which are assembled as a general transcriptional machinery at the promoter (Roeder, 1991). The assembly process for the basal transcriptional machinery has been identified as a highly regulated step in regulation of gene expression (Orphanides et al., 1996),
and post-PIC (preinitiation complex) assembly process (i.e. the elongation stage of transcription including promoter-proximal pausing of Pol II) was recognized as a key step regulated to coordinate downstream events (Core et al., 2008; Sims et al., 2004). The holoenzyme of RNA polymerase involved in initiation and elongation is a central component of the transcription machinery (Lee and Young, 2000). For humans, 1,700 to 1,900 transcription factors, capable of sequence-specific DNA binding, were estimated to be involved in many biological processes, ranging from cell cycle progression and intracellular metabolism to cellular differentiation and stress response (Vaquerizas et al., 2009). The key control on eukaryotic transcription is by interacting of the basal transcription machinery with cis-/trans-acting transcription factors. Among the genes which have been well characterized by associating with human developmental disorders, one third were reported due to mutations in genes encoding transcription factors (Boyadjiev and Jabs, 2000). A major challenge here is characterization of the structure of eukaryotic core regulatory networks defined by highly interconnected interactions between various regulatory proteins including basal transcription machinery components and TFs, and regulatory DNA elements (Gerstein et al., 2012). Recent studies in systematic, large-scale rewiring of the transcriptional regulatory subnetworks, for example including 475 sequence-specific TFs across 41 diverse cell and tissue types (Neph et al., 2012a; Neph et al., 2012b), provide new insights into the circuitry, dynamics, and organizing principles of the transcriptional regulatory network.

The past decades have seen many remarkable studies which revealed that eukaryotic transcriptional regulation involves chromatin templates that can facilitate gene activation and repression. Many kinds of chromatin remodelers are reported, which influence chromatin structure, for example by affecting promoter accessibility in chromatin (Imbalzano et al., 1994). This knowledge prompted biochemical, structural and genetic studies to understand how general DNA packaging proteins contribute to gene regulation, ultimately the crucial step to recruit the
transcription machinery at promoters (Wolffe and Drew, 1989; Workman and Kingston, 1998). The importance of chromatin remodeling in transcription was emphasized by the discovery that the multiprotein complexes involved in transcription initiation and elongation contain chromatin-altering enzymes, for example the Gcn5 acetyltransferase of the SAGA complex (Belotserkovskaya and Berger, 1999; Brown et al., 2000; Brownell et al., 1996; Grant et al., 1997; Wittschieben et al., 1999). The demonstration that histone acetyltransferases (or HAT) complexes interact with sequence-specific activator proteins exemplified how they selectively affect gene expression (Cosma et al., 1999). Therefore, characterization of chromatin folding that represses the binding of sequence-specific activators, and mechanistic understanding in modulation of chromatin organization that enhances activator binding, are critical to reveal the regulation of transcription in eukaryotes (Felsenfeld and Groudine, 2003).

**Regulatory diversification and organismal complexity**

The genome encodes all genes required for building an organism. Recent comparative genomics showed that organismal complexity does not arise from a higher number of genes (Lander et al., 2001). Much recent evidence suggests more precise and complicated regulation of gene expression by utilizing a variety of transcriptional factors and their regulatory elements, flexible chromatin remodeling and modification can dramatically expand regulatory complexity, resulting in diversification of the transcriptional apparatus (Levine and Tjian, 2003b). Alternative promoter usage in mammalian genomes, as an example, expands spatial relationships and interactions of genes and regulatory elements, leading to regulatory diversity (Landry et al., 2003). This regulatory diversification may be essential for intricate control of interactions with various regulatory proteins and/or the transcriptional apparatus in a temporal and highly coordinated
manner. This may be crucial to higher metazoans and their multicellularity living under various cellular signals and environmental stimuli.

Differential gene expression is a key for proliferation and differentiation (or specialization) of cells in multicellular organisms, termed as tissue-specific which means to be expressed or present in a specific cell type or to differ among cell types. Recent genome-wide studies using high-throughput parallel sequencing are now producing massive data across diverse cell and tissue types, which have led to the re-emergence of questions about the molecular basis of multicellular organismal complexity (Bernstein et al., 2012). A recent system-wide analysis of genome-wide binding profiles of transcription-related factors including 119 sequence-specific, general and chromatin-acting factors revealed their distinct combinatorial co-associations are widespread in a highly context specific manner (i.e. at specific genomic locations), which may illuminate how human regulatory information is organized, in terms of connectivity and hierarchy (Gerstein et al., 2012). This result may show the basis on how human regulatory units have undergone diversification for tissue-specific gene expression (Natarajan et al., 2012; Wang et al., 2012). Still, mechanistic understanding for generating this complexity in multicellular organisms is a subject of great interest (Natarajan et al., 2012; Wang et al., 2012)—especially, tissue-specific spatial organization of genomes (Parada et al., 2004).

High order spatial organization of entire genomes is often nonrandom, which suggests that interactions between many genomic loci may be correlated with structural and conformational properties of chromosomes (Dekker, 2008). Recent evidence such as chromosome conformation capture (3C) and 5C (Dekker et al., 2002; Dostie et al., 2006) illuminated widespread long-range interactions between distant enhancers, promoters of genes, and epigenetic states. Their results have added a complex layer on the eukaryotic gene regulatory network, ultimately essential for highly coordinated gene expression. Major challenges now include deciphering the mechanisms on how the nonrandom spatial features of chromosomes are
conserved and tissue-specific among different types of cells. In addition, distant-acting transcriptional enhancers are of particular interest in decoding the non-coding sequence in the human genome, which control gene expression often in a temporally and spatially restricted manner (Visel et al., 2009b). These regulatory domains can be located at a long distance outside the acting transcription unit. Identifying of the distant-acting regulatory elements and defining of their roles in development, phenotypic diversity and human disease remain a challenge (Rada-Iglesias et al., 2011; Sanyal et al., 2012; Visel et al., 2009a). Taken together, such unprecedented-scaled system- and genome-wide approach towards understanding of the organization and function of the human genome reveals an unexpectedly extensive network of regulatory communication within and between human chromosomes, which will be important in the future personalized genomics and medicine.

**Human genome variation and common disease**

Understanding the relationship between genotype and phenotype in human population is one of the major goals in biomedical research. In particular, the role of inherited genetic variation in human disease has been a subject of intensive study and speculation in order to identify biomedically important genes for prevention, diagnosis and therapy of common diseases. Association of genetic variation with diseases and drug responses involves discovering of casual variants (with individuals affected and non affected by a disease) and determining of their genotype, statistical quantification of their contribution to disease susceptibility, and characterization of their roles in biological pathways for disorder (Donnelly, 2008). The public resource of human variation contained mostly common variant sites (present at >5% minor allele frequency, MAF) including 14 million single nucleotide polymorphisms (or SNPs, as in dbSNP 130) (Frazer et al., 2007; International HapMap, 2005; Sachidanandam et al., 2001). This
database showed an unexpectedly large extent of the source of human phenotypic diversity, which covers root causes of common human diseases. Despite such substantial progress in identifying genetic variants, a more complete spectrum is necessary to explain the hereditary risk for common disease. For comprehensive survey for genetic contribution to human phenotypes, it is required, also challenging, to examine all forms of human DNA sequence variation across the entire range of allele frequencies across human populations having a wide range of ancestry (Consortium, 2010).

The recent development of high-throughput sequencing platforms enabled faster and less expensive resequencing of individual human genomes, which now makes comprehensive mapping and genotyping of genetic variation possible with a large number of individuals (Consortium, 2010; Levy et al., 2007; Wang et al., 2008; Wheeler et al., 2008). These genome-wide resources are providing useful information to discover genotypes underlying phenotypes on an unprecedented scale, and extends genome-wide association studies (or GWAS) to identification of genomic regions associated with complex disease susceptibility (e.g. diabetes) and common traits (e.g. human height) (Cirulli and Goldstein, 2010). Remarkably, substantial progress has recently been made in identification of low frequency and rare variants (defined as 0.5% to 5% MAF and below 0.5% MAF respectively) in multiple human populations (Abecasis et al., 2012). Such systematic investigation with high-throughput genotyping of variant sites (such as genome-scanning array technology, comparative DNA-sequencing analysis) provides more accurate and complete catalogues of rare and common variation in the human genome (Alkan et al., 2011; Feuk et al., 2006). Major challenges are to discover the causal relationship between variants and diseases, and their quantitative correlation to disease susceptibility, and functional variation of associated genes.

Despite many successful association studies to Mendelian disorders and common diseases, the identified variants have accounted for only the limited portion of the heritability for
complex traits and diseases (usually of polygenic inheritance at many genetic loci). This result leads to the question, where is then the missing heritability? (Manolio et al., 2009) Maybe, an answer lies in incomplete catalogues of rare variants, structural or other forms of variation which are not captured by current studies (Craddock et al., 2010). Characterization of rare non-coding variants at regulatory elements, such as motif-disrupting changes at transcription factor binding sites, is still a daunting task (Abecasis et al., 2012). Nonetheless, in parallel with recent advance in understanding of human transcription regulation and gene regulatory network, combination with various types of recent genomic data (for example, high-resolution maps of epigenetic modification and transcription factor binding) will provide a more powerful framework to study the molecular connection of genomic DNA sequence to biological function including dysfunctional disease states. This progress will eventually uncover the fundamental of human phenotypic diversity and disease susceptibility. The future challenge would include how to translate these association results into designing a strategy for disease risk assessment and treatment, for instance personalized drug selection and dosing (Manolio, 2010).

Cancer genomics and personalized medicine

Cancer was reported to be responsible for one in eight deaths worldwide. Cancer is essentially a genetic disease in which changes in the DNA sequence of the genome occurring after birth are pathogenic (Vogelstein and Kinzler, 2004). Somatic mutations in some oncogenes and tumor-suppressor genes were well studied to be responsible for tumorigenesis, for example missense mutations at the residue that are essential for their gene activity. These examples include constitutive activation of the KRAS2 oncogene in pancreatic cancer whereas the wild-type one is normally inactivated by GTP (Maitra and Hruban, 2008) and inactivation of tumor-suppressor genes such as TP53 (Donehower and Bradley, 1993). However, it seems that no single
gene defect causes cancer. It is widely accepted that sequential accumulation of mutations in genes promotes tumorigenesis. Such somatic DNA alterations, as a result, cause dysregulation of core cellular signaling pathways (e.g. metabolic reprogramming for tumor growth) that stimulate cell birth, or inhibit cell death and cell-cycle arrest. This process eventually increases the number of neoplastic cells. A major challenge is to identify somatic driver mutations and pathogenetically relevant key genes for each cancer type.

Interplay of genomic changes (or mutations) and evolutionary process by natural selection molds the cancer genomes (Yates and Campbell, 2012). Cancers carry many types of somatic changes which can be classified by their consequences for cancer development, “drivers” or “passengers” (or bystanders) (Stratton et al., 2009). Driver mutations are positively selected during cancer development, because they confer a growth advantage on cell population, whereas passenger mutations are functionally neutral and accumulated by chance during cell division and clonal expansion, providing no advantage to tumor development. A key challenge by catalogues of somatic mutations in a cancer genome is how to distinguish driver mutations from passengers. An assumption is that driver mutations are accumulated in a certain genes, while passengers are distributed in random regardless of cancer type and developmental stage. Recent large-scale sequencing of cancer genomes has uncovered a surprising number of DNA alterations from cancer patients. However, the functional role and physiological effect of the majority of these genomic alterations on tumors are still poorly understood, and many bioinformatics methods are being developed to reveal their fundamental relationship to tumorigenesis (Akavia et al., 2010; Carter et al., 2009; Carter et al., 2012). Furthermore, other large-scale “non-genetic” strategies to identifying therapeutic targets in cancer are for further investigation, for example by analysis of the cellular signaling pathways corrupted in cancer. The large-scale phosphoproteomic analysis to classify lung cancers based on aberrant tyrosine kinase signaling not only revealed novel
oncogenic kinases, but also demonstrated its utility in identifying potential therapeutic targets for cancer treatment (Rikova et al., 2007).

How many cancer genes are found in human cancer? Identification of the mutated genes that are casually implicated in oncogenesis and confer a clonal advantage during cancer evolution (called “cancer genes”), is a major goal in cancer research (Futreal et al., 2004). As an example, one of the long-standing questions is about the metastatic stage of cancer colonization to distant organs. Which genes are involved in cancer metastasis? (Nguyen and Massague, 2007). In particular, germline mutations in a certain gene are known to cause inherited susceptibility of offspring to some tumor types, for example breast cancer susceptibility gene 1 (BRCA1) for breast and ovarian cancers (Rahman and Stratton, 1998) and Li–Fraumeni cancer susceptibility syndrome associated with germline mutations in one p53 allele that confers an increased risk of various cancers such as sarcomas and adrenocortical tumors (Evans et al., 1998). Natural variation in the human genome population has been identified, including from large, microscopically visible chromosomal anomalies to single-nucleotide polymorphisms. This DNA sequence variation is a potential contributor to elevated risks in certain types of cancer development (Redon et al., 2006). Thus, comprehensive catalogues of these cancer susceptibility genes among human populations will be a crucial resource for preventive and therapeutic tumor treatment for individuals.

Remarkably, multinational, collaborative initiative consortiums, the Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), were launched to coordinate large-scale cancer genome studies (Hudson et al., 2012; Network, 2008). Their ultimate goal is comprehensive and unbiased characterization of the cancer genome—for example by catalogues of chromosomal aberrations, nucleotide substitutions and epigenetic modifications found in human cancers. Especially, massively parallel sequencing technologies are generating high-resolution genomic and exon-specific transcriptomic profiling from cancer cell genomes on
an unprecedented scale. These systematic efforts for comprehensive genome characterization of various cancer types are expanding our understanding of tumorigenesis at the molecular level, including molecular-level categorization of cancer subtypes in the breast cancer (Ellis et al., 2012; Hodis et al., 2012; Network, 2012a, b, c; Wood et al., 2007). Moreover, association of such systematic cancer genome catalogues with clinical outcomes (for example relapse-free survival and chemotherapeutic treatment results of cancer patients) will improve optimal design for molecular markers of cancer diagnosis and treatment outcome prediction in various tumor types. This research will be translated into many benefits to cancer patients, in particular, as a revolutionary patient-orientated strategy by clinical implementation of molecular cancer diagnostics and personalized (or targeted) medicine according to individual genome-based drug responsiveness (Green and Guyer, 2011).

References


Chapter 2

Unusual Combinatorial Involvement of Poly-A/T Tracts in Organizing Genes and Chromatin in *Dictyostelium*

This chapter has been published in Genome Research (PMID: 22434426) and was formatted for that journal:

Summary

*Dictyostelium discoideum* is an amoebozoa that exists in both a free-living unicellular and a multi-cellular form. It is situated in a deep branch in the evolutionary tree, and is particularly noteworthy in having a very A/T-rich genome. *Dictyostelium* provides an ideal system to examine the extreme to which nucleotide bias may be employed in organizing promoters, genes, and nucleosomes across a genome. We find that *Dictyostelium* genes are demarcated precisely at their 5’ ends by poly-T tracts and precisely at their 3’ ends by poly-A tracts. These tracts are also associated with nucleosome-free regions, and are embedded with precisely positioned TATA boxes. Homo- and heteropolymeric tracts of A and T demarcate nucleosome border regions. Together these findings reveal the presence of a variety of functionally distinct polymeric A/T elements. Strikingly, *Dictyostelium* chromatin may be organized in di-nucleosome units, but is otherwise organized as in animals. This includes a +1 nucleosome in a position that predicts the presence of a paused RNA polymerase II. Indeed, we find a strong phylogenetic relationship between the presence of the NELF pausing factor and positioning of the +1 nucleosome. Pausing and +1 nucleosome positioning may have co-evolved in animals.

Introduction

Among the social amoebae, *Dictyostelium discoideum* is the most studied species. All of these organisms live in the forest soil as solitary cells, feed on bacteria and proliferate by binary fission (Chisholm and Firtel, 2004). Upon starvation, free-living cells aggregate to form multicellular structures that undergo morphogenesis and cell type differentiation. The major transition of gene expression during development occurs 4-6 hours after onset of starvation (Van
This transition not only comprises the turning on of developmental genes but also a shut down of basic cellular functions. At this stage the cells are committed to development and are no longer able to revert to the growth phase (Katoh et al., 2007). Development ends with the formation of a basal disc and stalk, which supports a fruiting body. Fruiting bodies contain spores that can survive starvation conditions.

*Dictyostelium* is located deep in the phylogenetic tree of life, diverging just after plants but before the divergence of fungi and animals (Eichinger and Noegel, 2003; Williams et al., 2005). Since fungi and animals organize their chromatin differently, *Dictyostelium* should provide some insight into this major evolutionary shift.

The 34 Mb and ~12,500 gene genome of *Dictyostelium* is unusual in that it is highly A/T-rich (78%), surpassed only by the 81% A/T-richness of *Plasmodium falciparum* (Eichinger et al., 2005). The evolutionary forces that drove these organisms to such an A/T-rich extreme are unclear. Analysis of the *Dictyostelium* genome structure may therefore provide insight into how highly preferred usage of adenosine and thymine bases leads to the organization of chromatin, genes, and regulatory elements.

Like most eukaryotes, *Dictyostelium* utilizes nucleosomes to package its nuclear genome into chromatin. A nucleosome is composed of a histone protein core with ~147 bp of DNA wrapped around its exterior. Typically, nucleosomes are spaced by linker DNA that ranges from 10 ~ 50 bp in length. Nucleosomes may regulate gene expression by allowing or blocking access of transcription factors to regulatory DNA elements such as TATA boxes (Fascher et al., 1990; Moreira and Holmberg, 1998). Across eukaryotic species, promoter regions and other transcriptional regulatory sites are generally nucleosome-free (termed NFR), whereas most genic regions are occupied with arrays of well-phased nucleosomes, especially just downstream of the transcription start site (or TSS) (Albert et al., 2007; Lantermann et al., 2010; Lee et al., 2004; Mavrich et al., 2008b; Schones et al., 2008; Valouev et al., 2011; Yuan et al., 2005).
As demonstrated in yeast (Ioshikhes et al., 2006; Kaplan et al., 2009; Lee et al., 2007; Tillo and Hughes, 2009; Zhang et al., 2009; Zhang et al., 2011b), nucleosome occupancy levels (but not positioning) are controlled in large part by the underlying DNA sequence. AA/TT dinucleotide periodicities and GC-richness have been proposed to contribute to nucleosome occupancy. An intriguing question, then, is whether nucleosome occupancy in the highly A/T-rich Dictyostelium genome utilizes similar principles, or instead has evolved different mechanisms.

Given the depth of the Dictyostelium branch on the tree of life and that it spends part of its life cycle in an independent unicellular state and part in a multicellular state, it is unclear as to whether Dictyostelium organizes its chromatin as a unicellular or a multicellular organism. The unicellular pattern has been defined by Saccharomyces cerevisiae and Schizosaccharomyces pombe, which are evolutionarily divergent by nearly a billion years. These fungi place the upstream border of the first genic nucleosome over the TSS (Albert et al., 2007), where it would be positioned to block access of the transcription machinery. This placement of nucleosomes over the TSS would suggest that loss of histone-DNA contacts around the TSS might precede transcription initiation.

In contrast, multicellular eukaryotes, defined largely by Drosophila and vertebrates, place their first nucleosome further downstream and away from the TSS (Mavrich et al., 2008b; Schones et al., 2008). As a result, the TSS resides in the NFR, which would appear to give unobstructed access of RNA polymerase II (Pol II) to the TSS. Possibly as a consequence of this, Pol II initiates transcription, elongates, then pauses just upstream of the +1 nucleosome, which is also about 20 ~ 50 bp downstream of the TSS (Gilmour and Lis, 1986; Lee et al., 2008; Rougvie and Lis, 1988). Recent genome-wide studies showed that this promoter-proximal pausing is widespread in animals (Gilchrist et al., 2010; Muse et al., 2007; Zeitlinger et al., 2007), and that it may contribute to cell-to-cell coordination/timing of developmental programs in animals (Levine,
Coordinated timing is critical for multicellular development, and is also a key element in the *Dictyostelium* life cycle.

Here we present the first high-resolution genome-wide nucleosome maps of an amoeba, whose genome is highly A/T-rich and representative of a deep branch in the phylogenic tree. First, we examined the nucleotide composition of *Dictyostelium* genes, promoter regions, and nucleosome positions. Second, we determined nucleosome organization around genes during its unicellular and multicellular life stages. Third, we addressed the evolutionary relationship between nucleosome organization and potential pausing of the transcription elongation complex.

**Results**

*Dictyostelium* start their genes with poly-T and end with poly-A.

As a prelude to nucleosome mapping, we first sought to characterize the extreme A/T-richness of the *Dictyostelium* genome. In particular we found tracts of poly-T or poly-A to be more prevalent across the *Dictyostelium* genome than expected by chance (*Supplemental Fig. 2-1A*). The same was seen for the highly A/T-rich genome of *Plasmodium falciparum*, and much less so in other genomes (*Supplemental Fig. 2-1B*). We therefore suspect that A/T-rich genomes impart function to long poly-T/poly-A tracts.

Next, we examined the distribution of poly-T or poly-A tracts around transcription start (or TSS) and end (or TES) sites. This necessitated the production of a high-resolution transcriptome map of *Dictyostelium*, which was accomplished by RNA-seq. Four observations regarding the distribution of T and A nucleotides surprised us. First, poly-T tracts were highly enriched on the sense (or non-template) strand near the start of genes, immediately upstream of the TSS (*Fig. 2-1*), as seen previously on a small number of *Dictyostelium* genes (Hori and Firtel, 2011).
Second, the 3’ ends of these poly-T tracts were tightly distributed ~2 bp upstream of the TSS. Third and fourth, poly-A tracts were highly enriched on the sense strand at the end of genes, and started ~1 bp downstream of the TES. Remarkably, poly-T and poly-A on the sense strand precisely demarcated the start and end of transcribed regions, respectively. The precision of poly-T and poly-A tracts with respect to the TSS and TES highlights the accuracy of the transcriptome map.

The lengths of these poly-T and poly-A tracts did not correlate with transcript abundance or with each other (Supplemental Fig. 2-2 and 2-3), the latter indicating that these complementary nucleotides were not quantitatively linked between the 5’ and 3’ ends of genes. Other organisms lacked an equivalent organization (Supplemental Fig. 2-4), although other patterns were evident, in accord with the literature (Wu and Li, 2010). Plasmodium lacked a strand-specific enrichment of poly-T/poly-A (Supplemental Fig. 2-5), and thus A/T-richness alone cannot explain the striking poly-T/poly-A demarcation of genes in Dictyostelium. The A/T-rich genome of Dictyostelium might be the product of a distinct evolutionary path towards A/T-richness, which Plasmodium did not encounter (Szafranski et al., 2005b).
Poly-T tracts in promoters are interrupted by TATA boxes.

The TATA box promoter element consists of an 8 bp consensus in *Saccharomyces*:
TATA(A/T)A(A/T)(A/G) (Basehoar et al., 2004). In *Dictyostelium*, the A/T-richness of the genome raises the question as to whether TATA boxes are sufficiently unique to specify the

**Figure 2-1. Dictyostelium** gene organization with poly-T/poly-A enrichment.

(A) Distribution of poly-T/poly-A tracts ($\geq 6$ bp) in 5,468 genes with the annotated TSS. Each track represents the DNA sequence of a gene, fetched from -500 to 2500 relative to the TSS. Genes were aligned by their TSS and their DNA sequence in the sense strand was oriented in the 5’ to 3’ direction. If any poly-T/poly-A tract of length 6 or more was found in the DNA sequences, the poly-T sequence was colored in green and the poly-A sequence in red, and all the others were drawn as white in the sequence. The genes were sorted in ascending order of length. The right panel shows poly-T tracts as green and coding sequences as black. A zoom-in screenshot demonstrates the positional details in poly-T enrichment at TSS and translation start sites.

(B) Composite distribution of the location of poly-T/poly-A tracts around the 5’ and 3’ end of *Dictyostelium* genes. The density distribution of poly-T/poly-A tracts ($\geq 6$ bp) was displayed as a function of the distance (bp) on the sense strand between the midpoint of each non-overlapping tract and the given TSS (or TES). The density of the occurrence of poly-T (green trace) and poly-A tracts (red trace) is shown in the y-axis, which was estimated by Gaussian kernel and a smoothing bandwidth of 5. The density curve was calculated within the range from -1500 to 1500 relative to the TSS (or TES) and only the -500 – 500 region shown in the figure.

(C) Transcription start (or TSS) and end (or TES) site are linked to high T and A enrichment, respectively in the adjacent intergenic sequence. Each track represents the sense strand of a gene, fetched from 200 bp intergenic to 100 genic relative to the TSS (or TES) of 5,468 (or 5,400) protein-coding genes. Genes are aligned by the TSS (left) or TES (right), in the 5’ to 3’ direction from left to right. Transcript abundance is shown for each gene in an adjacent column. Genes were ordered by descending T (left) or A (right) density of their extracted sequence (301 bp). Color codes for the four nucleotides are indicated. This trend was not observed with randomization simulations (not shown).

(D) Frequency distribution of A, C, G and T relative to the TSS and TES. Shown is a summation of columns from panel C, over the indicated distance, color-coded as in panel C. The same simulation was applied to a set of TSS or TES randomly positioned across the *Dictyostelium* genome (gray traces, shown for A and T).
location of the transcription initiation complex. If so, we would expect bona fide TATA boxes to be precisely positioned upstream of the TSS, as they are in other organisms (~30 bp in metazoans and ~60 bp in *Saccharomyces*). As in metazoans, we found a tight distribution of TATAAA(A/T)(A/T) elements centered ~29 bp upstream of *Dictyostelium* TSSs (Fig. 2-2A). The sequence directionality of these elements was strongly biased towards the sense strand. A highly focused enrichment at the canonical metazoan location, with a strong strand bias provides three lines of evidence that they are functional. As a fourth line of evidence, no strand-specific enrichment was found downstream of the TES at the 3’ ends of genes (Fig. 2-2B). In contrast, fortuitous TATA elements may be present on the genic side of the TSS-proximal poly-T (e.g., 5’-TTTTTATA-3’ is the reverse complement of a TATA box) and TES-proximal poly-A tracts (e.g. 5’-TATAAAAA-3’).
Dictyostelium nucleosomes have polymeric-A/T borders.

Nucleosomes were mapped across the Dictyostelium genome using MNase-seq (Supplemental Fig. 2-6), before and after the major transition in gene expression during the two distinct phases of its life cycle: as a free-living unicellular amoeba and as a multicellular aggregate. Dyad-enriched G/C and the presence of 10 bp repeats of WW-type (W=A/T) dinucleotides are important determinants of nucleosome formation. Indeed, Dictyostelium nucleosomes were relatively enriched with G/C near their midpoints (Fig. 2-3 and Supplemental Fig. 2-7). However, 10-bp periodic WW dinucleotides were not evident (Fig. 2-3B), despite these dinucleotides being among the most abundant in the genome (Supplemental Table 2-1). Similar observations were made for human nucleosomes (Tolstorukov et al., 2009), and auto correlation analysis of dinucleotides within nucleosomal sequences also indicated a general lack of such periodicities (data not shown).

Towards the nucleosome borders and linker region, we observed a high incidence of homo- and heteropolymeric-A/T (meaning any continuous combination of A or T along a strand) (Supplemental Fig. 2-8). Such nucleosome placement was observed in all classes of nucleosomes including the -1 and +1 nucleosomes that flank promoter regions (data not shown). Thus, Dictyostelium may utilize a combination of both homo and heteropolymeric-A/T near nucleosome peripheries and relative G/C-richness (albeit sparse on an absolute level) towards the...
dyad to promote nucleosome formation. This trend is similar, but perhaps a more extreme version of what is evident in fungi and metazoans.

**Figure 2-3.** Nucleosomal DNA properties of *Dictyostelium* nucleosomes

(A) 73,396 nucleosome dyad locations were grouped as -1 (N=3,230), +1 (N=7,285), and all other genic nucleosomes (N=63,643) (see diagram). For the three assigned nucleosome groups, the W (= A/T in black) and S (= G/C in red) nucleotide percentage at each position was...
Dictyostelium organizes its nucleosomes on genes similar to animals.

We next examined the organization of Dictyostelium nucleosomes around their transcription start and end sites and addressed the following questions: (1) Does Dictyostelium package the 5’ ends of its genes basically in the same way, by having canonical nucleosome positions relative to the TSS? (2) Are NFRs present at the 5’ and 3’ end of genes, as seen in other organisms? (3) Is the +1 nucleosome placed over the TSS, as in single-celled fungi, or downstream as in multi-cellular animals and plants? (4) Is nucleosome spacing uniform within genic nucleosomal arrays, as seen in other organisms? (5) Does nucleosome organization change between the unicellular and multicellular state?

Figure 2-4A displays a composite distribution of nucleosomes in vegetative amoebae and multicellular aggregates aligned by the transcriptional start site of 5,468 protein-coding genes. No canonical organization was seen upstream of the TSS, where the -1 nucleosome typically resides, although the region appeared rather depleted of nucleosomes compared to genic regions. A 5’ NFR, centered 54 bp upstream of the TSS, was evident. This region is where the TATA box and poly-T/poly-A tracts reside, and is the likely site of transcription complex assembly. At the ends of genes, a 3’ NFR was found as seen in other organisms (Mavrich et al., 2008a; Shivaswamy et al., 2008; Yadon et al., 2010). This region might be important for transcription termination. The
depletion of nucleosomes upstream of the TSS is unlikely to be due to an inability to extract those portions of the genome, because whole genome 454 sequencing of *Dictyostelium* does not reveal such bias (unpublished observations). Within genic regions, nucleosomes were indeed organized into a canonical arrangement relative to the TSS. The first genic nucleosome (+1) was centered at approximately +115 relative to the TSS.

Vegetative and aggregating cells showed a similar nucleosome organization in the genome, which indicates that the same overall chromatin structure predominates at *Dictyostelium* genes, and is maintained regardless of its uni- vs multi-cellular status (Fig. 2-4C, D). However, cells that developed for six hours (aggregation stage) appeared to have a higher nucleosome occupancy at the beginning (+1 position) and end of genes, as compared to the unicellular state. The +1 nucleosome position is among the most highly dynamic, at least in yeast (Rando and Ahmad, 2007; Rufiange et al., 2007). During cellular differentiation imposed by starvation, most genes become quiescent (Loomis and Shaulsky, 2011). Nucleosome acquisition at this position might therefore play a role in this process.

We further examined genes whose expression is up-regulated during the aggregation phase (measured by RNA-seq), and found that they displayed the same basic nucleosome organization as the average gene in the aggregation state. Thus, changes in gene expression are not generally associated with major or global changes in chromatin organization, although transient changes are not excluded (and are indeed likely). The main conclusion is that, like other organisms, *Dictyostelium* has a canonical nucleosome organization that is largely invariant, regardless of the activity of relevant genes and through at least two distinct life cycle states.
An evolutionarily conserved change in genic nucleosome spacing upon starvation.

Despite similarities in overall organization during the unicellular and aggregated states, we observed a modal shift in the spacing between genic nucleosomes in the two states (Fig. 2-5). Canonical spacing increased on average from 162 bp in unicellular state to 169 in the multicellular state (Supplemental Table 2-2B). This increase in spacing is reminiscent of a similar increase in spacing seen in yeast during a starvation response (Zhang et al., 2011a), and the increase seen in human at cells quiescent/repressed regions of the genome (Vaillant et al., 2010; Valouev et al., 2011). In Dictyostelium, a starvation response is essentially the basis for development into a multicellular organism, and a shutdown in gene expression. Thus, an increase in genic inter-nucleosome spacing may be an evolutionarily conserved response to starvation, which may be associated with transcriptional quiescence and the production of refractory chromatin.
Figure 2-4. Nucleosome organization around the 5’ and 3’ end of the Dictyostelium genes

(A,B) Composite distribution of in vivo nucleosome positions relative to the TSS and TES. The midpoints of all nucleosomal sequence reads were distributed around the TSS (or TES) of 5,468 (or 5,400) protein-coding genes, plotted as described in Fig. 2-1B and smoothed with a bandwidth of 15 bp. The free-living (vegetative) and multi-cellular (aggregation) stages are indicated by black and red traces, respectively.

(C,D) Nucleosome organization around the TSS of 325 developmentally upregulated genes (black trace) in the vegetative and aggregation stages, compared to all nucleosomes (gray fill). Traces were smoothed with a 30-bp bandwidth.

Figure 2-5. Comparison of inter-nucleosomal distances between five species in their nucleosome arrangement within the genic region. The canonical inter-nucleosomal distance (bp) between two neighboring nucleosomes (for example +1 and +2) was calculated as the peak-to-peak distance, which was measured in the composite distribution of in vivo nucleosome locations. The nucleosome distribution of each species was produced as described in the Method. The nucleosomal repeat length (as an averaged distance between neighboring nucleosomes) is reported in Supplemental Table 2-2B.
**Dictyostelium organizes genic nucleosomes into pairs.**

We examined nucleosome spacing more quantitatively in Figure 2-5, by plotting the canonical inter-nucleosomal distance between the +1 and +2 positions, between +2 and +3, and so on. We also did the same in other organisms whose nucleosomes had been mapped at high resolution. Strikingly, *Dictyostelium* nucleosomes appeared to be arranged into di-nucleosome units where +1 and +2 are closely spaced, whereas +2 and +3 were more separated. Similarly, +3 and +4 were closely spaced compared to their flanking nucleosomes. These differences were not due to a position-specific alteration in the size or sequence composition of the nucleosomal DNA (Supplemental Fig. 2-9). Such di-nucleosome patterns were not evident in yeast and other organisms, suggesting that *Dictyostelium* undergoes an unusual organizational pattern involving pairs of closely spaced nucleosomes (150 ~ 155 bp), which are separated from adjacent pairs by a somewhat greater distance (170 ~ 180 bp).

**Evolutionary linkage between RNA polymerase II pausing and a downstream placement of the +1 nucleosome.**

One of our driving questions concerning *D. discoideum*, was whether the placement of the first genic nucleosome (or +1) resembled that of multicellular eukaryotes or that defined in unicellular fungi. The placement of the first nucleosome has implications with respect to unicellular versus multicellular gene regulatory strategies. A downstream placement has been linked to the existence of an adjacent upstream paused RNA polymerase (Mavrich et al., 2008b; Valouev et al., 2011). In Figure 2-6, we compare the canonical genic nucleosome organization across the tree of life, with particular emphasis on the +1 nucleosome position. Seven species were chosen in this study, which come from four major phylogenetic branches of eukaryotes (Baldauf et al., 2000) and which range widely across the evolutionary spectrum—*S. cerevisiae*, *S.*
pombe, human, D. melanogaster, C. elegans, A. thaliana, and also D. discoideum. For D. discoideum, genic nucleosomal arrays started at approximately the same distance from the TSS as seen in other multicellular organisms, with the apparent exception of C. elegans. The canonical pattern in C. elegans was rather weak. Most of the TSS used for analyses in C. elegans are actually ORF starts, since C. elegans undergoes trans-splicing, which obfuscates the location of the TSS (Blumenthal, 1995; Valouev et al., 2008; von Mering and Bork, 2002). Our results suggest that chromatin organization in Dictyostelium is more like that of multicellular organisms rather than unicellular fungi.

We therefore wondered whether placement of a canonical +1 nucleosome in the “downstream” position is associated with the presence of a paused RNA polymerase II. To address this possibility, we searched for genes encoding the NELF pausing factor. NELF is a complex of proteins that is responsible for pausing RNA polymerase II at a fixed distance immediately downstream of the TSS (Gilchrist et al., 2008; Lee et al., 2008; Wu et al., 2003). As reported (Narita et al., 2003), this type of pausing has been documented only in metazoans, and not in fungi, although a retention of RNA polymerase II in the promoter region has been observed (Venters and Pugh, 2009).

As expected of a role of NELF in pausing, NELF homologs were found in metazoans (Supplemental Fig. 2-10, and summarized in Fig. 2-6). However, NELF homologs were not found in fungi (Saccharomyces or Schizosaccharomyces), and also surprisingly not detected in C. elegans. Importantly, NELF homologs were found in Dictyostelium (Supplemental Fig. 2-10), which suggests that Dictyostelium is likely to undergo RNA polymerase II pausing, and is consistent with the downstream placement of the +1 nucleosome. Thus, there appears to be a strong linkage between having NELF (presumed to reflect the existence of pausing) and having a +1 nucleosome positioned downstream, next to where polymerase is expected to pause. Remarkably, this relationship did not appear to have a phylogenetic continuum, suggesting that
the positioning of the +1 nucleosome and its linkage to RNA polymerase pausing may have shifted multiple times during evolution.

NELF was also not found in plants, despite plants seemingly having a downstream placement of the +1 nucleosome. However, plants reside on a more distant branch of the evolutionary tree compared with all other multicellular eukaryotes tested, and thus a NELF homolog may be present but sufficiently diverged to be undetectable by our homology search algorithm. Taken together these findings provide evidence for a deep evolutionary linkage between placement of the +1 nucleosome in a downstream position and the presence of an adjacent upstream paused polymerase that perhaps arose as part of a genomic gene regulatory mechanism that is imposed by the requirements of multicellularity. This is consistent with the notion that RNA polymerase pausing is a means of coordinating the developmental timing between adjacent cells in multicellular eukaryotes (Levine, 2011).

Figure 2-6. Chromatin structure around the 5' end of genes evolutionarily conserved across
major eukaryotes. The data of genome-wide nucleosome positions in vivo were curated from the literature and the composite distributions of nucleosome locations were shown for each species in the right panel, aligned by the TSS. The life tree was adapted from the kingdom-level phylogenetic tree of eukaryotes (Baldauf et al., 2000) and simplified in the figure. A red circle denotes those having a substantial portion of the nucleosome over the TSS, and a blue circle denotes those where the TSS is located in the NFR. See Supplemental Table 2-2A for consensus positions.

Discussion

Is Dictyostelium moving towards a binary genomic code?

The unusual A/T-richness of the Dictyostelium genome makes it well-suited to evaluate the constraints placed on the utilization of G/C vs A/T nucleotides. Most organisms have a more-or-less balanced usage of the four nucleotides to encode not only genetic information but also for the regulation of the maintenance and expression of that information. For Dictyostelium we see some very biased and unusual usages of A/T to the point where it raises the question as to whether this organism is evolving towards a binary code. Inasmuch as G or C is critical for certain codons, this seems unlikely to ultimately occur. Nonetheless, the extent to which different combinations of A and T can be used for different purposes is remarkable.

The diverse usage of A/T nucleotides is illustrated in the following examples. First, we find that the 3’ ends of poly-T tracts, located on the sense strand upstream of ORF starts, demarcate transcriptional start sites. Second, the 5’ ends of poly-A tracts, when located on the sense strand downstream of ORF ends, demarcate transcript end sites. Third, embedded within these poly-T tracts resides the TATA box, which most typically has the sequence TATAAA(A/T)(A/T). In general, the TATA box is located 25 to 30 bp upstream of the transcription start site, and together with the general transcription factors helps determine TSS
selection (Fairley et al., 2002). Fourth, further upstream of the TATA box resides poly-T and poly-A tracts, which disfavor nucleosome formation, and may help maintain nucleosome-free promoter regions. Fifth, as in most other organisms, AATAAA near the 3’ ends of genes signifies the site of transcript cleavage, and sixth, to some extent in some organisms, periodic AA/TT dinucleotides may promote nucleosome formation.

We do not know to what extent poly-T tracts define promoter regions, and why poly-T rather than poly-A is preferred on the sense strand. Certainly we would expect such tracts to exclude nucleosomes around the TATA box. This suggests that nucleosomes may not be as influential in controlling TATA box accessibility and pre-initiation complex formation in Dictyostelium as might occur in more GC-rich genomes. One model of transcription initiation is that RNA polymerase II moves away from the preinitiation complex in search of a start site (Giardina and Lis, 1993; Kuehner and Brow, 2006; Li et al., 1994; Pardee et al., 1998). Such movement might be accompanied by abortive transcription, where short RNA transcripts are released until a bona fide start site is found. After the start site, the nascent transcript may be more stable in an RNA/DNA hybrid to allow productive transcript elongation. Thus, it is conceivable that abortive transcripts consisting of poly-rU are particularly dissociable and thus preferred just upstream of the TSS. On the other end of the transcript, higher RNA/DNA stability imparted by a tract of poly-rA might slow down RNA polymerase, making it more susceptible to termination mechanisms. While these ideas are speculative, studies on RNA/DNA hybrid stability support these ideas. In particular, poly-rU/poly-dA hybrids (akin to those occurring just upstream of the TSS) are the least stable, and poly-rA/poly-dT hybrids (akin to the downstream arrangement) are the most stable of all RNA/DNA hybrid tracts consisting of only A or T (or U) (Kraeva et al., 2007; Sugimoto et al., 1995).

While G/C-richness near the nucleosome dyad or AA/TT dinucleotide periodicities may be contributory toward nucleosome assembly, we only observed G/C-enrichment near the dyad of
Dictyostelium nucleosomes, with little or no AA/TT periodicity. The latter is surprising given the extreme A/T-richness of the genome, but such periodicities also appear to be weak or nonexistent in other organisms (Mavrich et al., 2008a; Tolstorukov et al., 2009; Valouev et al., 2011). Dyad-enriched G/C was also weak on an absolute scale. However, homo and heteropolymeric A/T tracts were enriched near nucleosome borders, which raises the question as to what extent these sequence preferences contribute to nucleosome formation.

**Dictyostelium organizes genic nucleosomes as in animals.**

Nucleosome organization in the Dictyostelium genome is most similar to that of metazoans, where the first nucleosome downstream of the TSS places the TSS in the NFR. As such, the transcription machinery would be generally free to initiate transcription without direct influence of the +1 nucleosome. This contrasts with fungi where one edge of the +1 nucleosome resides over the TSS, and thus nucleosome displacement is expected to be a prerequisite to initiation. A downstream placement of the +1 nucleosome raises the question as to whether RNA polymerase II pauses as it approaches that nucleosome. More generally, the question arises as to whether RNA polymerase II pausing and downstream placement of the +1 nucleosomes have co-evolved, perhaps originating from constraints imposed by multicellularity occurring at some stage of the life cycle. We examined this by comparing whether downstream placement of the +1 nucleosome is evolutionarily linked to the presence of NELF-encoding genes. NELF is a key determinant of pausing. Consistent with the link between pausing and nucleosome placement, fungi (*S. cerevisiae* and *S. pombe*, representing deep phylogenetic branches in fungi) lacked both, whereas metazoans and Dictyostelium contained both. This linkage ties in and agrees well with the recent notion that RNA polymerase II pausing functions to coordinate the timing of
multicellular development (Levine, 2011), and this would include *Dictyostelium*. One rationale is that many events are needed for assembly of the transcription machinery at promoters, and the combination of these events may have some stochastic elements. Thus, response to cell-cell signaling in development may be less coordinated among cells if left solely to the regulation of pre-initiation complex assembly. In contrast, a paused polymerase may only need a release factor to promote the expression of the gene, and in principle this would be a more effective mechanism to ensure coordination amongst cells.

**Materials and Methods**

**Growth and development of *D. discoideum***

Strain AX2-214 (Bloomfield et al., 2008; Harloff et al., 1989) was grown axenically in liquid nutrient medium containing 1.8% maltose (Watts and Ashworth, 1970). Shaking was done at 160 rpm (22 °C) until the cells were in the exponential growth phase and had reached a density of ~4 × 10^6 cells/ml. Crosslinking with formaldehyde (1% final concentration) was done for 15 minutes at 22 °C while shaking. 2.5 M glycine solution was added to 0.125 M final concentration to quench fixation and shaking continued for 5 min. After quenching cells were kept on ice. Cells were pelleted (2 min, 2000 rpm) and gently washed two times with Soerensen phosphate buffer (17 mM Na^+/K^+-phosphate buffer, pH 6.0). They were then resuspended in lysis buffer (50 mM Tris/HCl, pH 7.4, 100 mM NaCl, 5 mM MgCl_2, 250 mM sucrose, 1 mM benzamidine, 1 mM PMSF, aprotinin, leupeptin, pepstatin at 10 g/ml each) and lysed by the addition of 1% Triton X-100. Lysis was microscopically controlled. Equivalents of 5 × 10^8 cells were distributed into tubes, pelleted at 16,000 × g, the lysis buffer discarded, the samples snap frozen in liquid nitrogen and kept at minus 80°C until further use.
To isolate nucleosomes from cells in a specific developmental stage we chose the aggregation stage in which cells homogeneously acquire new cell surface markers, change their cell shape and become sensitive to cAMP. This is associated with changes in gene expression. Cells were grown in shaking suspension as described above, harvested, washed twice in Soerensen phosphate buffer and resuspended in the same buffer at a density of $1 \times 10^7$ cells/ml and shaken at 160 rpm and 22 °C for six hours when they had formed aggregates as monitored by microscopy. Crosslinking and lysis was done as above.

**Preparation of nucleosomes**

Preparation of mononucleosomal DNA is similar to what has been previously described (Albert et al., 2007; Mavrich et al., 2008b). The chromatin pellet from $1 \times 10^9$ crosslinked and lysed cells (equivalent to 200 ml of cell culture) was washed once in 4 ml NP-S Buffer and spun down at 16,000 $\times$ g for 5 minutes at 4°C. The supernatant was discarded, and the chromatin pellet was resuspended in 2 ml NP-S Buffer + 1 mM β-mercaptoethanol. The sample was treated with 2,000 units of Micrococcal Nuclease for 20 minutes at 37 °C to digest the chromatin down to predominantly mononucleosomal size, and the digestion was quenched by adding EDTA to a final concentration of 10 mM and chilling the sample in ice for 10 minutes. MNase has known sequence bias, with A/T-rich dinucleotides being preferred. However, at a high degree of digestion this bias is greatly diminished, since even unfavorable sites are cleaved. Correction for this bias does not alter nucleosome patterns (Albert et al., 2007). Nonetheless, as the *Dictyostelium* genome is highly A/T-rich we expect even less potential bias.

The sample was treated with Proteinase K at 65 °C overnight and nucleosomal DNA was extracted twice with phenol:chloroform:isoamyl alcohol, treated with RNase for 2 hr, extracted...
once more with phenol:chloroform:isoamyl alcohol, precipitated with 100 % ethanol, and resuspended in TE. Nucleosomal DNA from $1.5 \times 10^7$ cells (equivalent to 64 ml of cell culture) was separated by electrophoresis in a 2 % agarose gel for approximately 1.5 hr at 135 V, and using a long wavelength UV transilluminator DNA fragments ranging in size from 100 to 200 bp were excised. DNA was purified from the agarose using a QIAquick Gel Extraction Kit and was subject to pyrosequencing using the Roche GS20/FLX in accordance with the manufacturer’s instructions.

**Nucleosome mapping**

Due to the A/T-richness of the Dictyostelium genome, and thus potentially lower sequence complexity of its DNA, and to inherent sequencing biases of some deep sequencing platforms (Hoeijmakers et al., 2011; Kozarewa et al., 2009), we utilized Roche/454 long-read sequencing technology to sequence entire nucleosomal DNA fragments. This allowed both ends of each nucleosomal DNA fragment to be identified and mapped to the Dictyostelium genome with the highest accuracy.

Nucleosome locations on the genome were identified by aligning sequencing reads to the reference genome. Sequencing reads were mapped to the Dictyostelium discoideum AX4 reference genome which was released from dictyBase (Kreppel et al., 2004) in February 2008. We used SHRiMP version 1.3.2 (SHort Read Mapping Package) (Rumble et al., 2009b) for read mapping. The local alignments produced with the sequencing reads were retained, whose alignment length covers at least 90% of the length of the sequencing read and whose percent identity was 90% or more. Only uniquely aligned reads were used for nucleosome maps. Candidate nucleosome dyads were determined by calculating the coordinate of each sequenced read midpoint. We found that the same specific loci have an abnormally high number of reads in
both the vegetative and aggregation read mapping data. Their read numbers were as extreme statistical outliers in terms of the distribution of the mapping read number per each genomic coordinate. Since we detected no special genomic feature surrounding them, we concluded they might be artifacts owing to biased PCR amplification by unknown reasons and filtered them out in the subsequent data analysis.

We collected 274,046 nucleosomes in 246,513 positions across the genome—i.e. 88,265 nucleosomes in 84,441 unique positions from vegetative Dictyostelium and 185,781 nucleosomes in 162,072 unique positions from multicellular aggregates. These numbers represent a genome-wide sampling, rather than complete coverage. We assumed that the digested DNA fragment is centered at the nucleosome midpoint (or dyad) and examined the composite distribution of single nucleotides and dinucleotides of sequenced fragments around each nucleosome midpoint.

Transcriptome sequencing and analysis

In order to identify transcription start sites (or TSS) and transcription end sites (TES), transcriptome sequencing with the Roche GS20/FLX platform was employed in this study. As the transcribed regions are defined by independent experimental measurements, it seems unlikely that the patterns in Fig. 2-1 arose from systematic errors. Moreover, it seems unlikely the reverse transcriptase used in mapping the RNA TSS prematurely stopped at poly-T tracts that randomly occur throughout the transcribed region since other mapped features (e.g. TATA box, ORF start, and nucleosomes) all showed a canonical and non-random distribution relative to the measured TSS.

For D. discoideum several axenic laboratory strains are available. Most common are AX4 and AX2. AX4 was the strain which was sequenced. It is characterized by a 750 kb duplication on chromosome 2, which might influence expression at least in this duplicated region. We
therefore decided to use the wild type strain NC4 from which the sequenced AX4 strain as well as AX2 were derived. NC4 grows only in association with bacteria. We used *Klebsiella planticola* as food source.

NC4 cells were mixed with a culture of these bacteria in logarithmic growth phase and plated onto NA medium of the following composition: a mixture of 0.5 g glucose, 0.5 g peptone, 20 ml 50× phosphate buffer pH6.5, and diluted with 1000 ml H₂O, then 20 ml phosphate buffer, 99.86 g KH₂PO₄, 17.8 g NaHPO₄·2H₂O, and 1000 ml H₂O added again. When the bacteria were consumed the amoebae were recovered from the plates and washed twice with the phosphate buffer to remove residual bacteria. The amoebae were then plated onto phosphate agar plates (buffered plates without nutrients). The plating time point was taken as the initiation of the developmental cycle. Cells were harvested and RNA was extracted using the RNaseq Mini Kit from Qiagen at 0, 4, 8, 12, 16, 20, and 24 hours during development. The RNA was converted to cDNA using the Evrogen Mint Kit. The sequencing libraries were prepared from this cDNA and subject to pyrosequencing by Roche GS20/FLX system according to the instructions of the manufacturer on experimental setup and design.

The resulting sequencing reads were sorted into “gene bins” according to the annotated genome of *D. discoideum* AX4 (http://dictybase.org, released in February 2008) using BLAT (Kent, 2002), and we mapped the positions of 5’ and 3’ UTRs of protein-coding genes. In order to find intron/exon boundaries the best hits were aligned to the genome using exalin (Zhang and Gish, 2006). To obtain an unbiased location of transcripts, all seven RNA-seq data of 0, 4, 8, 12, 16, 20, and 24 hours were compiled together, and the 5’ and 3’ most base of the matching region of a gene were defined as TSS and TES, respectively. As a result, the TSS of 5,468 protein-coding genes and the TES of 5,400 protein-coding genes were annotated in this study (Supplemental Table 2-3), and were compared to TSS locations that pre-exist in the literature (Supplemental Table 2-4). Furthermore, the read counts for each gene at the indicated time point
of development were taken as a proxy for expression levels. Normalization of the data was done with the following formula: \((\text{number of mapped reads to an individual CDS ÷ (all RNAseq reads \times \text{gene length})}) \times 10^8\). Differential expression was assumed if the read numbers between start point of development and 4 and/or 8 hours into the cycle were at least 5 fold different. Among the 5,468 genes with the annotated TSS, 325 genes were finally grouped as developmentally upregulated genes, which were highly expressed during the first 8 hours of starvation (Supplemental Table 2-3).

**Data analysis**

Detailed data analyses methods can be found in the Supplemental Information. The raw nucleosomal sequencing reads are available at NCBI Trace Archives (Accession number SRA044979.3). The raw RNA sequencing reads are available at NCBI Trace Archives (Accession number SRA045580).

**References**


Chapter 3

A Comprehensive and High Resolution Genome-wide Response of p53 to UV-damage

This chapter was submitted and now in peer-review process, and was formatted for that journal.
Summary

The identification of transcription factor bound sites across a human genome by ChIP-seq has been limited to statistically-enriched subsets that encompass broad binding regions, and include false positives (wrong locations), false negatives (missed locations), and imprecision. This has limited our understanding of transcription factor binding mechanisms and the genes they target. We employed ChIP-exo to accurately and comprehensively identify ~2,000 p53-bound recognition elements (RE) across the human genome in response to UV-induced stress. In addition to a characteristic 6-peak ChIP-exo pattern, we present evidence that p53 binds essentially to a 20 bp unsplit RE, and is not convincingly detectable at noncognate locations. At REs, p53 might form an enhancesome-like structure with other stress-response regulatory elements/factors that co-occur at fixed distances and orientation. We have linked p53 to 154 target genes, of which about 80% are novel. At these genes, UV-treatment leads to p53 binding and release of a pre-existing paused RNA polymerase II into an elongation state, suggesting that p53 regulates pausing and not initiation. The initiation machinery was present before and after polymerase release. The identified target genes greatly fill out and expand on our understanding of p53-regulated DNA repair and cell proliferation networks, including potentially disease-associated SNPs in REs.

Introduction

p53, the “guardian” of the genome, is a sequence-specific transcription factor that, along with other factors, regulates genes involved in stress responses, including UV-induced DNA damage (Menendez et al., 2009; Muller and Vousden, 2013; Riley et al., 2008). p53 is one of the
most commonly mutated proteins in cancers (Donehower and Bradley, 1993). Knowing exactly where p53 binds across a genome, its interacting partners, and the genes it regulates, has been a critical limitation in defining how p53 elicits its protective functions. The dysfunction of p53 has largely been attributed to defects in p53’s DNA binding surface (Freed-Pastor and Prives, 2012; Vogelstein and Kinzler, 1992). However, mutations in a p53 response element (RE) could also impair p53’s ability to regulate its target gene (Bandele et al., 2011; Naqvi et al., 2010).

Efforts have been made to identify p53 REs, p53-bound locations, and p53-regulated genes (Botcheva et al., 2011; Cawley et al., 2004; Smeenk et al., 2011; Smeenk et al., 2008; Wei et al., 2006; Yu et al., 1999; Zhao et al., 2000). But these attempts are limited by assay sensitivity (signal:noise) that tends to detect the most highly occupied regions, and assay resolution that assigns p53 binding events over several hundred base-pair intervals rather than pin-pointing its exact location. Importantly, current RE descriptions lack sufficient uniqueness to confidently identify all but the most robust REs. A p53 RE is comprised of two 10-bp half-sites that have the highly degenerate consensus sequence \((\text{RRRCWWGYYY})_2\), where RWY = A/G, A/T, and C/T, respectively. However, deviations from this consensus are common, and REs reportedly tolerate 1-13 bp insertions between half-sites, deletion of half-sites, and RRRCW quarter-sites in multiple orientations (El-Deiry et al., 1992; Funk et al., 1992; Riley et al., 2008). These criteria, if correct, allow for many millions of potential REs across a genome. Since p53 binding may be enriched in nucleosomal regions (Lidor Nili et al., 2010), invoking chromatin occlusion of REs may not provide an adequate explanation as to the restriction of p53 binding to only a small fraction of putative sites. Thus, a comprehensive and definitive genome-wide identification of p53-bound REs has not been possible with ChIP-seq technology.

Here we apply the recently developed ChIP-exo genome-wide assay to p53 in human U2OS cells before and 6 hours after UV treatment (Rhee and Pugh, 2011). ChIP-exo involves formaldehyde crosslinking of proteins to their native genomic sites in vivo, then chromatin
immunoprecipitation under denaturing conditions. A 5’-3’ exonuclease digests each DNA strand up to the point of crosslinking, which is detected genome-wide by deep sequencing. A crosslinking point is manifested as a pair of peaks, with the second peak located at a fixed distance (8-12 bp) in the 3’ direction and on the opposite strand relative to the first. Since a pattern of peak-pairs is characteristics of a particular protein, we sought to identify such patterns for p53 REs and for sites of potential interactions with other proteins. With a comprehensive and accurate set of p53 binding locations we more accurately delineate parameters surrounding p53 site utilization and the genes it may target. Our data suggest simplicity in site binding, which raises a cautionary flag in interpreting lower resolution ChIP-seq data.

Associating transcription factors like p53 with the genes it regulates has been problematic due to potentially large distances between the two along linear chromosomes, and the potential for these factors to regulate as-yet undiscovered transcription or chromatin events. p53 responds to acute cellular stress, and such stress like UV-induced DNA damage can cause rapid turnover of mRNA, thereby confounding attempts to link changes in p53 occupancy to changes in mRNA levels. To overcome this limitation, we measured changes in gene expression upon UV treatment via ChIP-exo mapping of components of the core transcription machinery TFIIB and RNA polymerase (Pol) II. This allowed us to more comprehensively link p53 with its transcription event regardless of transcript stability. Consequently, we achieve a more comprehensive p53 gene regulatory network, and a refined ability to associate noncoding SNPs in p53 REs with potential phenotypic consequences.
Results

A triple peak-pair is diagnostic of p53/RE binding.

The distribution of ChIP-exo tags, six hours after UV treatment, for the well-studied p21/CDKN1A (El-Deiry et al., 1993) and GADD45A (Kastan et al., 1992) genes are shown in Figure S3-1A. Exonuclease stop sites surrounded their known p53 elements, as expected. Taking advantage of the resolution offered by ChIP-exo peak-pairing, we identified 1,825 putative p53-bound regions across the human genome. As expected, MEME analysis identified the p53 RE to be the most enriched motif. However, unlike other studies, our primary goal was to unambiguously and comprehensively link each p53 binding event to its cognate RE, to the extent that it exists. The reported extensive variations in RE structure and sequence degeneracy lead us to apply only minimal criteria for putative RE sequence identification (see Methods), but then apply a highly restricted search distance.

Figure 3-1A displays strand-separated (blue and red) exonuclease stop sites for all 1,825 putative p53-bound regions (rows) aligned by the midpoint of an identified RE, which was typically less than a few bp away, or aligned by peak-pair midpoints where an RE was not detected. A four-color plot of each resident RE sequence and its deviations from the consensus is shown in Figure 3-1B. 1,572 locations (86%) were classified as Group 1 regions that contain a recognizable 20 bp RE. About half of these regions were constitutively bound by p53, whereas binding to the other half was UV induced. Constitutive and induced p53 binding has been previously reported (Ceribelli et al., 2006; Espinosa et al., 2003; Kaeser and Iggo, 2002; Shaked et al., 2008). Group 2 (n=146 or 8%) included those locations having a putative RE, but with a 1 bp deletion or a 1-13 bp insertion between half-sites. These sites were subject to further scrutiny.
below. Binding locations that failed to meet our minimal RE criteria were placed into Group 3 (n=107 or 6%).

The set of ChIP-exo determined p53 locations substantially overlapped with nine pre-existing datasets of locations determined by ChIP-chip, ChIP-PET, and ChIP-seq (Figure S3-1B, C). However, 36-95% of the locations in these other datasets did not overlap and thus may be specific to the different cell types or genotoxic stresses employed or they might be false positives. A similar percentage of all locations determined by ChIP-exo were missed in prior studies (potential false negatives), and these tended to be inducible sites. The high sensitivity and resolution of ChIP-exo leads us to conclude that the current dataset provides the most comprehensive and the most accurate identification and characterization of p53 REs, particularly those that are UV-inducible, to date.

Strikingly, a triple peak-pair pattern was evident around Group 1 REs (Figure 3-1C). Group 2 displayed a more diffuse pattern that resembled a triple-peak (Figure S3-1D). Group 3 display a single peak pair pattern. Group 1 contained three peaks on each strand, which were offset by 8 bp in the 3’ direction on opposite strands. Peaks on the same strand were 10 bp apart. Our interpretation of the triple-peak pattern is that each p53 dimer/RE half-site is flanked by formaldehyde crosslinks, as this is the region that most frequently contains available (non-bonded) and closely juxtaposed primary amines on both the protein and the DNA (Figure 3-1D). Inasmuch as p53 binds to an RE as a dimer of dimers (Kitayner et al., 2006; Tidow et al., 2007), we infer that a crosslink flanks both sides of each dimer. Since two dimers sit immediately adjacent to each other, the crosslinking points between two dimers merge into a single peak (shown as the highest one), resulting in three peaks per strand for a p53 tetramer. The detection of robust DNA crosslinking in the regions between two p53 dimers indicates accessibility between the two bound half sites. This unique distribution of ChIP-exo tags provides a robust diagnostic for asserting p53 binding across a genome.
Figure 3-1. Comprehensive genome-wide identification of p53-bound locations in response to UV DNA damage
Surprisingly, only 162 (10%) of all Group 1 p53-bound REs contained a perfect match to the degenerate 20 bp RE consensus. Thus, an exact match to a consensus is not a comprehensive means of site identification. p53-bound REs typically contained one half-site that matched the consensus (shown as the left-most half-site in group 1 of Figure 3-1B), and a second half-site that deviated from the consensus by a limited degree. This observation fits with prior studies (El-Deiry et al., 1992; Funk et al., 1992), and with the notion that p53 initiates binding at one half-site then completes binding at a second half site (McLure and Lee, 1998). We did not observe peak-
pair patterns around isolated half-sites, indicating that the functionally active form of p53 is a tetramer that binds to a full site (Hupp and Lane, 1994).

Low-occupancy p53 locations were at REs that deviated the most from an RE consensus, suggesting that RE sequence quality is one determinant of p53 occupancy (blue in Figure 3-2A). However, the same low quality REs tended to be the most UV-inducible (Figure 3-2B), and thus may be important determinants in stress responses.

Deviations from the consensus quarter-site (R1R2R3CW) followed the trend: R1>R2>R3=W>C, with the central two quarter-sites being the least variable overall (Figure 3-2C and S3-2). These observations suggest that basal and constitutive p53 site occupancy at specific REs is tuned in part through site deviation from the consensus, with different nucleotide positions in the RRRCW quarter site and inner vs. outer quarter-sites having characteristic contributions. Thus, R1,2, R3W, and C nucleotide positions potentially provide fine, medium, and coarse tuning, respectively. These sequence properties do not exclude additional positive or negative contributions from other proteins. Indeed, a wide range of occupancy levels was observed for sites of similar quality, indicating that other factors contribute to p53 binding. However, some of this differential is likely due to differences in crosslinking efficiency imparted by the local DNA sequence.
Figure 3-2. Sequence assessment of a comprehensive set of REs.

(A) Relationship between p53 site quality and occupancy level. Group 1 sites were binned according to occupancy levels (percentages) in uninduced cells, then color-coded according site quality according to Menendez et al. (Menendez et al., 2009). Blue is further from the consensus.

(B) Relationship between p53 site quality and changes in occupancy ($\Delta$occ) after UV treatment. Left panel color code is as in panel A. The two panels represent two metrics of site quality from Menendez et al. (2009).

(C) Base composition at each position in an RE quarter site (RRRCW), separated out by occupancy levels (percent ranks) in uninduced control cells.
REs lack insertions.

p53 is widely viewed as binding REs that contain 1-13 bp insertions between their half-sites (El-Deiry et al., 1992; Funk et al., 1992; Jordan et al., 2008; Qian et al., 2002; Riley et al., 2008). However, the vast majority of p53-bound REs detected here lacked an insertion (Group 1 in Figure 3-1A). Putative REs with insertions (Group 2) were not only rare (146 locations or 8%), but they also had very low occupancy and had many more nucleotide deviations from the consensus than Group 1 (Figure 3-1B, 3-3A). Moreover, they displayed a composite peak-pair pattern that looked very similar to Group 1, when aligned by the stronger resident RE (Figure S3-1D). Collectively, the data suggest that Group 2 is enriched with low-level p53 binding to Group 1 REs that may have just missed the criteria for being an RE outright, rather than p53 binding to sites with insertions. Others in Group 2 may belong to Group 3 (but fortuitously having two nearby degenerate half-sites that placed them in Group 2), as described below.

A no-insertion requirement fits well with the tetrameric p53/RE crystal structure (Kitayner et al., 2006), and suggests that its dimer-dimer interface is physically constrained, but accessible, in vivo. These conclusions are in line with a study placing similar constraints on p53/REs that were tested in functional assays in yeast (Tokino et al., 1994). These results are in apparent conflict with in vitro binding studies demonstrating p53/DNA interactions with REs having insertions (Wang et al., 1995). However, purified in vitro systems having relatively high levels of p53 may be less discriminating than in vivo. In vivo, putative alternative sites would be unable to compete for limiting amounts of p53 that are bound at canonical high affinity sites. The binding of p53 to weak sites in vivo may be through stress-induced stabilization/activation of p53, through cooperative interactions with other factors, or through tetramer stacking across multiple sites (Stenger et al., 1994).
REs are spaced at 10 and 250 bp intervals.

We examined the spatial relationship of one p53-bound RE in close proximity (<50 bp) to another. Remarkably, a large number of closely-spaced REs overlapped by 10 bp (n=267 pairs, Figure 3-3B). Intriguingly, this affords p53 with two binding sites at the cost of evolving three half-sites. Overlapping sites had the same distribution of sequence quality as non-overlapping sites (data not shown). However, ½ overlapping sites tended to be highly occupied and constitutively bound rather than inducible (Figure 3-3C). Thus, overlapping sites may provide for stronger constitutive p53 binding. Taken together, considering REs that are isolated, overlapping, or adjacent, we estimate that there are ~2,000 detectable p53-occupied REs in UV-treated cells, and they all exist essentially in the 20 bp RE form.

A second mode of RE site clustering was observed at distances of ~250 bp (Figure 3-3D). This distance is sufficient to place these REs on two separate nucleosomes (one and a half nucleosomes cover ~250 bp of DNA). The function of such spacing is unclear, as this pairing is not particularly enriched or depleted at promoter regions.
Figure 3-3. Spatial relationship of REs.

(A) Quality metrics of p53 bound to putative REs with varying insertion lengths between two half sites. 1,718 full sites of Group 1 and 2 were included in the table and the average occupancy level (tag counts) in response to UV treatment and averaged number of nucleotide deviations from the 20-bp p53 consensus sequence (El-Deiry et al., 1992) are shown.

(B) Frequency distribution of pairwise distances between p53-bound REs. 2,130 REs in 1,572 Group 1 regions were subjected to autocorrelation analysis.

(C) Left panel, frequency distribution of fold changes in p53 occupancy for those 267 pairs of p53 REs that overlapped by 10-bp (half-site overlap), compared to the non-overlapping set. Data were plotted by Gaussian kernel and a smoothing bandwidth of 0.5. Right panel shows a box plot of the occupancy level of nonoverlapping and ½ overlapping sites before UV induction.

(D) Frequency distribution of pairwise distance between 1,825 p53-occupied regions. Within each region, the midpoint of the strongest RE was used as the reference point for calculating their pairwise distance. Multiple REs within the same region were treated as a single location. Data were binned in 20 bp intervals and smoothed using a 3-bin moving average. Forty paired regions were <600 bp apart.
Noncognate p53 crosslinking is dubious.

Group 3 is represented by a single set of peak pairs that were ~30 bp apart. It stands in striking contrast to the triple peak pairs of REs that were offset by 8 bp (Figure S3-1D), and is therefore not likely to reflect direct binding of p53 to DNA. This larger peak-pair size may reflect one or more other proteins that were crosslinked to DNA. We therefore looked for motifs present between Group 3 peak-pairs. We also looked at Group 2 since some may have been improperly parsed there due to our low-threshold initial definition of an RE. Two motifs were enriched among Group 3 (38% had Motif 1 or 2, Figure S3-3A). Motifs 3, detected within Group 2, was highly similar to Motif 1 (Figure S3-3A). This confirmed that Group 2 contained a mixture of locations that belonged to either Group 1 or Group 3, and thus further diminished any confidence that split REs exist. Moreover, binding to the discovered motifs was not appreciably UV-inducible, and their tag distributions were rather diffused in most cases. Very few peak-pair locations were centered on motifs 1-4, as opposed to be near them, and thus may not be the cognate motif for those peak pairs. In sum, only ~14% of all 1,825 locations lacked a 20 bp unsplit RE, of which 33% contained a CT-rich element. CT-richness is characteristic of so-called “blacklisted” regions of the genome, which tend to recur in ChIP experiments independent of antibody specificity. Therefore, we find no convincing evidence for constitutive or UV-induced p53 binding to any DNA sequence in the genome, directly or indirectly, other than to 2,130 unsplit 20 bp REs.


p53 interacts with sequence-specific transcription regulators and the core transcription machinery (Laptenko and Prives, 2006). We were therefore surprised by the general lack of
detection of indirect p53 crosslinking to other DNA sites containing presumed p53 targets. This lack of detection might be due to the inherent inefficiency of formaldehyde crosslinking, which would require both protein-protein and protein-DNA crosslinking. As an alternative approach, MEME identified highly enriched sequence patterns within 50 bp of p53-bound REs (Bailey and Elkan, 1994) (Figure S3-3B). Other search distances yielded essentially the same results. TomTom analysis associated with the JASPAR database (Gupta et al., 2007; Sandelin et al., 2004) suggested that these sequences may represent a complex mixture of factor recognition motifs, some of which might be fortuitous. We therefore down-selected those motifs having prior reports of linkages to p53 using iHOP (Hoffmann and Valencia, 2004), and searched for their occurrence within 1 kb of p53-bound REs, using FIMO (Grant et al., 2011). We further down-selected to present those motifs having an apparent nonrandom local distance relationship with p53 REs. Such candidates might have structural constraints with p53, as exemplified by the enhancesome model (Thanos and Maniatis, 1995). Structural constraints may be manifested in two ways: 1) by a fixed distance relationship between two motifs, and 2) an orientation constraint for non-palindromic motifs.

We plotted the distribution of p53-bound REs around each candidate motif, while keeping a fixed orientation of any asymmetric motif. We selected 9 known and 1 unknown motifs (but perhaps related to Xyr1) having restricted locations or orientation relative to p53 (Figure 3-4 and Figure S3-3C). Approximately 1,000 binary RE/motif relationships were identified. Several motifs are recognized by factors involved in environmental stress and oncogenesis such as AP1, GATA1, NRF2, FOXO3 and SPI1, and pluripotent stem cell maintenance such as Oct4, Sox2, Klf4, and Myc. Since at least some of these factors exist within highly related gene families, we examined their expression in our U2OS system (Figure S3-3D), and confirmed the expression of all factors except GATA1, Oct4, SPI1, and Klf4. For GATA1, Oct4 and Klf4, we instead detected expression of GATA4, POU2F1 (an Oct4 paralog) and Klf5.
The robust distance and orientation restriction of the motif is evident by p53 peak enrichment primarily on one side of a directional motif. For example, p53-bound REs displayed a peak of enrichment 30 bp to the right of WGATAR (midpoint to midpoint, Figure 3-4A). Similarly, REs were directionally distributed 50 bp upstream of Klf4/5 sites (GGGYGKGG, Figure 3-4B), suggesting that p53 might be in direct contact with this protein, as demonstrated in vitro (Brandt et al., 2012), and in vivo (Yoon et al., 2003; Zhang et al., 2000). When the aligned sequence of individual locations was examined, recurring sequence patterns often extended beyond the identified cores (see blocks of vertical colored stripes in the right panels of Figure 3-4). This may reflect the existence of additional factor binding sites, sequence constraints, or repetitive elements. Indeed, when we examined the distribution of repetitive elements, endogenous retroviral long terminal repeats (LTR) stood out as being highly enriched at p53-bound REs, and were positionally-restricted near other factor binding sites (Figure S3-3E). AP1 was an exception. These findings are entirely consistent with and expand upon the concept that endogenous LTRs act as enhancer vehicles (Cohen et al., 2009). Together, the distance and orientation constraints of the motifs suggest that a spatially constrained complex of proteins exist between many p53/RE complexes and the factors that bind to the identified motifs.
Figure 3-4. Fixed orientation and distance of stress-response motifs relative to p53 REs.

(A,B) MEME logos of motifs enriched near the 1,572 p53-bound locations of Group 1, along
with the number of occurrences are indicated to the left. The distribution of p53/REs around each motif, orientated according to the sequence logo, is shown as a solid colored trace, along with the underlying nucleosome pattern (gray fill) determined in T-cells (Schones et al., 2008). Four color plots of the nucleotide sequence for each instance of the indicated motifs (GATA1 and Klf4) are shown to the right, in which p53 REs are shaded.

**Linking p53 binding to target genes and UV-induced release of paused Pol II.**

A major function of p53 is to activate genes that respond to DNA damage. Part of this response involves increased levels of active p53. Perhaps a naïve simplification is that p53 binds near its target genes, as has been demonstrated in select cases (Stenger et al., 1994; Thut et al., 1995). However, activators are well known to act at a distance with respect to the linear sequence between its binding site and the transcription start site (Levine and Tjian, 2003a). Consequently, p53 may not be linearly placed near to its target genes. Distally-bound p53 complexes might engage their targets by looping out the intervening chromatin (Dean, 2011). In principle, such loops could contain other genes or transcription units, and so p53 may not necessarily target the next or closest gene defined by the linear DNA sequence. Moreover, some fraction of p53 may not regulate transcription at all, or at least not directly.

Looping assays like “3C” are designed to detect long-distance interactions. They rely on a chain of DNA-protein-protein-DNA crosslinks to noncognate locations. Since we were unable to convincingly detect p53 crosslinking to anything but an RE, it seems unlikely that a 3C-type assay would be successful in identifying p53-gene connections. We therefore initially characterized annotated genes as candidates if they were within 10 Mb of a detected p53 location (n=751), which may be a nominal maximal distance for looping (**Table S3-2**). We identified genes using a combination of annotated transcription start sites (TSSs) and by performing ChIP-exo assays on the transcription pre-initiation complex (PIC) component TFIIB, and Pol II under
normal and UV-induced conditions. At these genes, and for genes in general, TFIIB occupancy remained essentially unchanged in response to UV treatment (peak distribution centered over zero in Figure S3-4A, left panel). Pol II occupancy in promoter regions, on the other hand, decreased substantially at nearly all genes (Figure S3-4A, right panel). Thus, in response to UV damage there was widespread loss of Pol II, leaving partial PICs (containing at least TFIIB, but depleted of Pol II) in its wake. These findings are entirely consistent with similar observations reported at selected p53-regulated genes (Espinosa et al., 2003). In addition to dissociation, loss of Pol II at promoter regions can also occur via release of Pol II from a paused state into a transcription elongation state as observed at the p21 gene (Gomes et al., 2006).

When Pol II occupancy in gene bodies was examined, a substantial subset of the 751 genes displayed elevated Pol II levels in response to UV, despite having the same or lower levels of Pol II at the promoter region (Figure S3-4B). The effect of UV treatment on TFIIB/Pol II retention at promoter regions predominated more where p53 was closer to the promoter (Figure 3-5A). This correlation suggests that in response to UV damage, p53 binds within 15 kb of target promoters, and promotes the release of a transcriptionally paused Pol II into an elongation state. These observations are consistent with established interactions of p53 with elongation factors (Lew et al., 2012; Shinobu et al., 1999), which includes c-Myc (Rahl et al., 2010). The analysis here does not address whether p53 is the direct effector of Pol II release from the paused state or indirectly operates through other proteins that co-occur with p53.

Since the UV-induced transcriptional response was more prevalent where p53 was detected nearer to the promoter, we further focused on p53 binding within 15 kb of a TSS (269 genes). p53 binding was monotonically distributed across this region, except for substantial enrichment within 1 kb of the TSS of 48 mRNA, 14 ncRNA, and 6 tRNA genes (Figure 3-5B). This promoter-proximal p53 generally had REs that deviated more from the consensus than distal REs (Figure 3-5C), which is a property of more-inducible REs.
While the data thus far point to a statistical enrichment of p53 connectivity to the core promoter region, <4% (65/1825) of all p53/RE binding is <1 kb from a detectable core promoter (15% or 281/1825 are <15 kb from a promoter). This presented a challenge in assigning p53 to an appropriate gene, assuming that proximity implies regulation. We addressed this problem by linking UV-induced fold increases in p53 occupancy to fold increases of pol II occupancy within gene bodies, for all p53-TSS distances that were <15 kb apart (Figure 3-5D). We identified 154 genes meeting these criteria (Table 3-1 and Table S3-3). Furthermore, we retained a lower confidence set of 100 genes having UV-induced p53 binding >15 kb away, that displayed increased Pol II binding in the gene body (Table S3-4). Based upon the observations thus far a composite model for p53 regulation of genes is illustrated in Figure 3-5E.
Figure 3-5. Positional relationship of p53 to the 5' end of genes.

(A) Fold changes in TFIIB or Pol II occupancy, upon UV treatment, are shown as a function of distance between TFIIB/TSS locations and the closest p53. Profiles were smoothed using LOWESS with default parameters.

(B) Distribution of p53-bound locations within 15 kb of an annotated TSS that also contained the highest TFIIB levels in the 30 kb interval. Since the PIC locations for tRNA genes were not determined, the nearest tRNA genes to p53 were selected among all genes. Rows were sorted by
p53-TSS distance. Each p53 location was color-coded according to its changes in occupancy in response to UV treatment, and all TFIIB/TSS locations were shown in blue.

(C) Site quality of p53 full sites in relation to distance from TSS. 252 Group 1 p53-bound REs were binned according to their distance to the attached TSS, then the fraction of sites were color-coded according to site quality, as in Figure 2A.

(D) Bar graph of changes in pol II occupancy in gene bodies in response to UV treatment. Any of the 1,825 p53-bound locations that were <2 kb apart were first aggregated to produce 1,776 broad regions. This was done to minimize over-weighting of the same gene linked to multiple p53 locations. Genes having significant TFIIB/TSSs were searched within 10 Mb of these 1,776 regions, then rank ordered by distance from the nearest p53. Bars represent the closest gene (1st), second-closest gene (2nd), and so on. The median fold changes in Pol II occupancy across gene body upon UV treatment are shown for each gene rank. Distances of their TFIIB/TSS locations from the nearest p53 are indicated in the top panel.

(E) Model of p53-mediated transcription regulation upon UV DNA damage.

*p53 regulatory networks and the influence of SNPs.*

The 154 high confidence genes in Table S3-3 include 33 previously reported p53 target genes, while 121 genes were not previously identified as being p53 regulated, which demonstrates the utility of the ChIP-Exo in identifying new transcription factor target genes. Several new targets genes include long intergenic RNAs (lincRNAs) and unannotated transcripts (Figure S3-5A). Several lincRNAs have been previously identified as p53 targets (Barsotti and Prives, 2010; Huarte et al., 2010). They play a key role in p53-mediated repression of genes on a broad scale. Newly assigned p53 target genes are involved in a variety of functions, including cell growth, cell cycle, cell signaling, differentiation, apoptosis, cell adhesion and motility, and intracellular transportation (Table 3-1 and Table S3-3).

IPA gene network analyses found the DNA replication, recombination, and repair gene network with the highest enrichment score (p-value ≈ 10^{-48}) (Figure 3-6A). Of the 25 genes involved in this network, 14 of them are newly discovered p53 targets. The functions of these
genes implicate p53 in regulating multiple steps of the DNA damage response after UV irradiation. CDKN1A, GADD45, and PLK3 control the cell cycle, allowing time to repair damaged DNA before the cell cycle progresses further. XPC and DDB2 recognize damaged DNA and recruit the DNA damage repair complex. POLH, PCNA, and UBR5 repair the damaged DNA. FTL and RRM2B can participate in the synthesis of deoxyribonucleotides for DNA repair. Several factors regulating p53 posttranslational modifications (e.g., PLK3, MDM2, MTBP, and OGT) are present in this network, suggesting that p53 feedback loops are activated to repair DNA damage. Notably, we identified RAD51C as a novel p53-activated target gene (Figure S3-5A, Table S3-3). RAD51C is related to RAD51, an important mediator of homologous DNA recombination repair. RAD51 is reportedly p53-repressed (Arias-Lopez et al., 2006), while RAD51C is induced, which suggests differential usage of these recombinases during a UV stress response. As such, this network regulates p53 itself, the cell cycle, nucleotide excision repair and homologous DNA recombination repair.

Another highly scored network regulates cell death and survival, with 12 out of the 22 genes in this network being new p53 targets (Figure S3-5B). Growth factors and their receptors, such as FGFR2, VGF, PGF, and JAG1, were identified. Notably, p53 may also directly regulate FOSL1, an AP1 family transcription factor that was recently discovered as a new target of JQ1 in lung cancer cells (Lockwood et al., 2012). This represents a new link of p53 with cell growth and survival regulation. Meanwhile, genes with growth inhibition functions were also identified, such as DUSP7, BTG2 and BTG3. In general, genes within this network regulate cell growth, cell death and survival, consistent with p53 being important in maintaining tissue homeostasis after DNA damage recovery. The role of p53 in metabolism has been more studied recently (Vousden and Prives, 2009). Within the metabolism network, 15 out of the 17 genes were newly identified p53 target (Figure S3-5B). These include genes regulating RNA stability and degradation (e.g., EXOSC4, DUSP11, and PUM2), protein degradation (e.g., FBOX15 and FBOX22), and
mitochondrial functions (e.g., COX6A1, MRPL36, and PANK2). Notably, PANK2 is a master regulator of CoA synthesis in mitochondria, the genetic alteration of which is linked with neurodegenerative diseases (Zhou et al., 2001).

With a comprehensive and accurate map of p53-bound REs, we sifted through the NCBI dbSNP database (build 130 having 13,864,001 SNPs) (Sherry et al., 2001) in search for any that overlapped with a p53 RE. We found one SNP (rs9333500) that was of particular interest because it resided in a UV-inducible p53-bound RE that we associated with the polH gene (Figure 3-6B). UV treatment resulted in increase Pol II occupancy in its gene body. This gene encodes the DNA-directed DNA polymerase η, and has been reported to be upregulated by DNA damage in a p53-dependent manner (Liu and Chen, 2006). PolH conducts trans-lesion DNA synthesis through UV-induced pyrimidine dimers. Defects in this gene result in XPV Xeroderma Pigmentosum, displaying hypermutability after exposure to UV irradiation and resulting in skin malignancy (Loeb and Monnat, 2008; Masutani et al., 1999). The rs9333500 SNP results in a change from G:C to T:A at position 7. This is the “C” position in the RRRCW quarter-site that is internal to the RE, which is the least variable position within the least variable quarter site. Surprisingly, the other quarter-site within the same half-site contained a nonconsensus nonSNP “G” at position 4 (in RRRCW). Thus, this RE is expected to be intrinsically weak and inducible. The rs9333500 SNP is expected to eliminate p53 binding, altogether. rs9333500, which is homozygous in ~2% of the human population (PDR90), is expected to render these individuals sensitive to UV-induced DNA damage.
**Figure 3-6.** p53 regulatory networks and SNPs.

(A) Pre-existing IPA DNA repair network of gene/protein (nodes) interactions (edges). Nodes are colored shades of red or green to reflect fold changes (increased and decreased, respectively) in p53 occupancy upon UV treatment. Gray indicates that p53/RE was not assigned to their genes. Relative change in Pol II occupancy in the gene body is marked by yellow/cyan-shaded (up/down) halos encompassing the nodes.

(B) Browser shot of the POLH gene, displaying UV-induced p53 binding at an RE, that contains a recorded SNP. The lower panel displays a blow-up of the bound p53 RE, identifying the location of the rs9333500 SNP at 7th G position.

**Table 3-1.** Categories of p53 target genes linked to p53-bound locations using ChIP-exo analysis.

<table>
<thead>
<tr>
<th>Apoptosis Cell cycle</th>
<th>DNA Repair Chromatin Modifier</th>
<th>Cell growth Differentiation</th>
<th>Transcription Regulation</th>
<th>Signal Transduction</th>
<th>Cell Adhesion Mobility</th>
<th>Biosynthesis Metabolism</th>
<th>Transportation Membrane Homeostasis</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRMS1</td>
<td>ASC3</td>
<td>ACYP2</td>
<td>ATF3</td>
<td>AKAP10</td>
<td>ACTA2</td>
<td>CAPN2</td>
<td>ABCD3</td>
<td>C11orR24</td>
</tr>
<tr>
<td>BAX</td>
<td>DDB2</td>
<td>DSBK3</td>
<td>CSDA</td>
<td>ARVC</td>
<td>ASTN2</td>
<td>CES2</td>
<td>ACBD5</td>
<td>C20orI30</td>
</tr>
<tr>
<td>BTG2</td>
<td>GBDK</td>
<td>FBXO15</td>
<td>DDT3</td>
<td>CIDorF10</td>
<td>MTBP</td>
<td>COX6a1</td>
<td>ABC</td>
<td>C20orI72</td>
</tr>
<tr>
<td>BTG3</td>
<td>GPD1</td>
<td>FBXO22</td>
<td>FHL2</td>
<td>D10P11</td>
<td>PLK2</td>
<td>DCP1B</td>
<td>BLOC1S2</td>
<td>CCDC51</td>
</tr>
<tr>
<td>CDCC7</td>
<td>KOMSC</td>
<td>GIH</td>
<td>FOSL1</td>
<td>DUSP7</td>
<td>PVT1</td>
<td>DGAT1</td>
<td>C18orf55</td>
<td>CCDC58</td>
</tr>
<tr>
<td>CCNG1</td>
<td>UMNA</td>
<td>JERG</td>
<td>MED21</td>
<td>FGFR2</td>
<td>RAF2B</td>
<td>EXOSC5C</td>
<td>C30orf52</td>
<td>FM200A</td>
</tr>
<tr>
<td>CDKN1A</td>
<td>METTL8</td>
<td>KRT17</td>
<td>RARA</td>
<td>HRH1</td>
<td>GATC</td>
<td>FTL</td>
<td>IKA4</td>
<td>04A0247</td>
</tr>
<tr>
<td>EPHA2</td>
<td>POLH</td>
<td>MRPL13</td>
<td>SERTAD1</td>
<td>IP09</td>
<td>GPAA1</td>
<td>GCC2</td>
<td>RBM128</td>
<td></td>
</tr>
<tr>
<td>GADD45A</td>
<td>RAD51C</td>
<td>MRPL14</td>
<td>T5S1</td>
<td>JAG1</td>
<td>GTPB5P</td>
<td>SAD1</td>
<td>TP53T51</td>
<td></td>
</tr>
<tr>
<td>OGT</td>
<td>RBM29</td>
<td>MRPL51</td>
<td>ZNF219</td>
<td>KLBG1</td>
<td>LASS5</td>
<td>SEC31A</td>
<td>Novel protein (2)</td>
<td></td>
</tr>
<tr>
<td>PCNA</td>
<td>SETD1A</td>
<td>PUM2</td>
<td>ZNF337</td>
<td>LRR2C8</td>
<td>LRP1</td>
<td>SLC12A4</td>
<td>IncRNAs (3)</td>
<td></td>
</tr>
<tr>
<td>PGF</td>
<td>TET2</td>
<td>SE62</td>
<td>ZNF488</td>
<td>RBM14</td>
<td>MDMD2</td>
<td>SNAP47</td>
<td>ncRNAs (16)</td>
<td></td>
</tr>
<tr>
<td>PHLD3A</td>
<td>XPC</td>
<td>ZNF65S</td>
<td>S100A16</td>
<td>MRPL36</td>
<td>TSNARE1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLK3</td>
<td>SDC4</td>
<td>NTSM</td>
<td>VP524</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTEN1</td>
<td>SERT</td>
<td>SEPRN1</td>
<td>OPLA4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S100A6</td>
<td>TBC1D2B</td>
<td>PANK2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T6AR</td>
<td>TBC1</td>
<td>RPLP2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAF4</td>
<td>TM7SF3</td>
<td>SARS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAP2</td>
<td>TMEM115</td>
<td>SDF4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMEM183A</td>
<td>UBR5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TMEM68</td>
<td>UTP14A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRIM5</td>
<td>UTP3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VGF</td>
<td>WD40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2C3H7B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Genes were selected from Table S3. Previously known p53 targets are underscored.
### Discussion

**p53-mediated transactivation**

Our results point to the assembly of about 1,800 p53/RE complexes spread across the human genome when cells have been damaged by UV light. ChIP-exo detects p53/RE interactions as a defined triple peak-pair pattern that may serve as a useful definitive diagnostic for p53 binding.

About half of p53-bound sites are constitutively occupied in the absence of UV. Their occupancy levels appear to be controlled in part by sequence diversion from an RE consensus. Many sites overlap by 10 bp, whereas others are concentrated at about 250 bp apart (illustrated in Figure 3-5E). Split REs, having 2-13 bp spacers between two half-sites likely do not exist, but instead are part of low-affinity degenerate REs that are not split. Our findings suggest that p53 activates at least 154 genes in response to UV damage by promoting release of Pol II from a transcriptionally paused state.

Remarkably, most p53/RE interactions could not be linked directly to a gene based on proximity or on correlated responses to UV treatment. However, those within 15 kb of a gene start could be reliably connected. Conceivably, the vast majority of p53/RE interactions, particularly constitutive interactions, may have other genome functions or might play an indirect role. For example, p53 might help keep chromatin regions constitutively open, but not directly plug into the transcription machinery. Another possibility is that a large number of REs simply serve as a reservoir of p53 that is stable to turnover.

We found no convincing evidence for p53 binding to sites other than to an RE. Other sites were indeed detected, but were infrequent and low occupancy. They lacked definitive
properties, including well-defined peaks, and a well-defined motif centered between peak pairs. We suspect that the CT-rich motifs that were enriched in the vicinity may be non-specific.

**Overall assessment of binding locations and RE**

As we expect will be typical of sequence-specific DNA binding proteins, we found no deterministic rules governing genome-wide occupancy of p53 at REs. Instead, p53 bound the genome at a wide-range of occupancy levels, being detected over three orders of magnitude. We suspect that an even greater continuum exists and would be evident with greater depth of coverage. This continuum may be interpreted by the cell as tapered regulation by p53.

We observed a continuum of RE site quality that included nucleotide deviations in frequency, type, and positions from the RE consensus, and perhaps indels of a single bp between two half sites. We observed a continuum of distances between sites, although there were fixed distance preferences. Deviations from the norm were typically accompanied by decreased site occupancy. As binding reached the lower limit of detection so did the resemblance of the putative RE to a consensus. Genome-wide binding may therefore not be considered as an absolute number, but rather a fixed number at an arbitrary minimal occupancy level, and qualified as to the types of interactions (direct RE contact vs interactions with other proteins).

If the data are of sufficient quality, arbitrary occupancy thresholds may be well above the minimal acceptable threshold for false discovery (i.e. 95% confidence interval), and thus few if any false positives may be present. Thresholding based on occupancy may discard low occupancy locations that could be biologically important. The challenge is to evaluate the degree of biological importance of each site. This too may reflect a continuum, whereby low occupancy sites contribute little, but not zero, to biological function. Importantly, low affinity p53 sites tend to be UV-inducible, thereby invoking importance to intrinsically low affinity interactions. This
concept has ramifications in all other inducible systems, including developmental processes, where motif analysis may be limited to high-confidence high-affinity sites. Site degeneracy may be an important biological tool to promote inducibility.

**p53 regulatory network**

In response to various upstream activation signals (e.g., UV light, DNA breaks, or stresses), p53 is known to turn on the transcription of distinct sets of target genes, which in turn regulate the cell cycle, cell death, metabolism and apoptosis (Vogelstein et al., 2000; Vousden and Prives, 2009). The set of genes activated by p53 is highly context dependent: low levels of genotoxic stress induce mainly cell cycle arrest gene expression, while high levels of genotoxic stress induce cell death gene expression. In agreement with UV irradiation inducing extensive DNA damages, p53 significantly activates the DNA replication, recombination, and repair network genes. Therefore, p53 appears to be an accurate sensor of the cell stress types, and appropriately activates the set of genes suitable for the cell’s need. Strikingly, this network calls upon p53 target genes involved in all steps of the DNA damage repair, ranging from halting the cell cycle, regenerating the cellular pools of dNTPs, and repairing DNA by both nucleotide excision repair and homologous DNA recombination pathways.

Many feedback loops composed of p53 and its covalent modification enzymes were found, indicating a fine-tuning of the p53 function within this network. In addition, cell death and survival together with metabolism are the top gene networks enriched with p53 target genes after UV irradiation, demonstrating that p53 coordinates multiple pathways in response to a stress. The final outcome is cell survival if successful DNA damage repair is achieved, or cell death if there is a failure to repair damaged DNA. A cell’s decision to live or die may be in part regulated by over one hundred p53 target genes impinging on interconnected networks. Given the complexity
of the p53 network (Vousden and Prives, 2009), future challenges lie with understanding how
other signaling cascades and stress-response transcription factors impinge upon p53 regulation.
The findings here and elsewhere hint at possible roles for AP1, GATA4, NRF2, FOXO3 and SPI1
as well as stem cell maintenance factors in modulating p53 activity via a direct physical
relationship with p53 bound at an RE.

Materials and Methods

Cell culture and UV treatment

Human osteosarcoma U2OS cells were grown in Dulbecco modified Eagle medium
supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin in a 5% CO2
incubator. For UV irradiation, cells were exposed to 50 J/m2 of UVC light in a Spectrolinker XL-
1000 UV cross-linker (Spectronics Inc.). p53 induction was monitored by Western analysis using
the anti-p53 monoclonal antibody.

ChIP-exo

ChIP-exo experiments were carried out essentially as described with minor alterations
(Rhee and Pugh, 2012). With prepared sonicated chromatin, chromatin immunoprecipitation
(ChIP) was performed on p53 (mouse monoclonal anti-p53 derived from BP53-12, P5813;
Sigma), TFIIB (rabbit polyclonal IgG, SC-225; Santa Cruz Biotechnology, Inc.), and Pol II
(rabbit polyclonal IgG, SC-899; Santa Cruz Biotechnology, Inc.). Sequencing was performed
using Applied Biosystems SOLiD (for p53) and Illumina HiSeq 2000 (for TFIIB and Pol II).
Sequence tags were mapped to the human genome (hg18, NCBI build 36.1) using SHRImp (v.
1.3.2) (Rumble et al., 2009a) for p53, and BWA (version 0.5.9-r16) (Li and Durbin, 2009) for TFIIB and Pol II. Uniquely aligned tags were retained and filtered to remove those from heterochromatin and so-called black-listed regions (provided by the ENCODE project).

**Accession Numbers**

Raw data were deposited at NCBI Sequence Read Archive (Khosravi et al.) (http://www.ncbi.nlm.nih.gov/sra) under accession numbers p53 ChIP-exo (U2OS): SRA065035; TFIIB ChIP-exo (U2OS): SRA065037; Pol II ChIP-exo (U2OS): SRA065038.

**References**


transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116, 499-509.


Chapter 4

Conclusion

Eukaryotic cells respond to many different extracellular signals and environmental stimuli. The regulation of gene expression is central to their response. The mechanisms involved in this cellular control have been the most intensely studied, and transcriptional regulation is one of the key mechanisms. Differential transcription regulation is also fundamental in multicellularity that requires tissue-specific gene expression (Kleinjan and van Heyningen, 2005). Recent advance in next-generation sequencing technology has led to a new era in genomics, including targeted genomic sequencing, metagenomics, transcriptome sequencing, ChIP-sequencing, etc. (Voelkerding et al., 2009). These high-throughput genomic data started to reveal the core regulatory network in eukaryotic transcription regulation, consisting of highly interconnected interactions between transcriptional machinery, cis/trans-regulatory domains, DNA-binding proteins, chromatin conformation, and epigenetic modification. This dissertation demonstrated how massively parallel sequencing data are utilized to address biological questions about eukaryotic transcription control, and how computational and statistical tools in functional and comparative genomics are applied to high-throughput genomic data analysis for these questions.

Summary of study

This dissertation presented a genome-wide data analysis for understanding transcription regulation and chromatin structure in eukaryotes. High-throughput sequencing data were produced by formaldehyde crosslinking and DNA digestion that includes micrococcal nuclease
(or MNase) and λ-exonuclease. This genome-wide approach utilizing DNase and next-generation sequencing technology successfully mapped genome-wide binding locations of regulatory DNA-binding proteins at a high resolution. The results in this dissertation demonstrated broad applications of such NGS platforms in functional and comparative genomics.

In Chapter 2, strain AX2-214 was grown and mononucleosomal DNA materials were prepared in the free-living (vegetative) and multicellular (aggregation) stage. After treated with formaldehyde, cells were lysed, and mononucleosomes were prepared using MNase, followed by deep sequencing using the 454 Life Sciences (Roche) GS20/FLX sequencer. The long-read sequencing capability of this pyrosequencing was utilized to improve mappability for a low complexity genome (i.e. highly A/T-rich). The high-resolution nucleosome maps obtained from *D. discoideum* showed that *Dictyostelium* packages their genomic DNA in the same way as other eukaryotes, showing evident nucleosome-free regions (or NFRs) at the 5’ and 3’ end of genes, and canonical arrangement of well-phased nucleosomes in the genic region relative to the transcription start site (or TSS). The similar nucleosome organization from vegetative amoebae and multicellular aggregates indicated that the same chromatin structure is maintained at *Dictyostelium* genes regardless of its unicellular vs. multicellular state. Strikingly, *Dictyostelium* organizes its nucleosome at the 5’ ends of its genes similar to multicellular eukaryotes, in a way that the placement of the first genic nucleosome (or +1) is just downstream of the transcription start site (or TSS) (Albert et al., 2007; Mavrich et al., 2008b). Furthermore, the strong phylogenetic relationship between the homology of the NELF pausing factor and the positioning of the +1 nucleosome across major eukaryotes implies that the presence of paused RNA polymerase II may be associated with the +1 nucleosome shift downstream of the TSS. This linkage agrees with the recent notion that RNA polymerase II pausing functions to coordinate the timing of multicellular development which includes *Dictyostelium* (Levine, 2011). This result implies that pausing Pol II and +1 nucleosome positioning may have coevolved in animals for a
more effective mechanism to coordinate cell-cell signaling in multicellular development. The analysis of nucleosomal core DNA sequences showed that the A/T-rich *Dictyostelium* genome places nucleosomes with polymeric-A/T sequences near nucleosome peripheries and G/C-rich sequences toward the nucleosome dyad. However, 10-bp periodic WW (W=A/T) dinucleotides were not evident for the rotational positioning of *Dictyostelium* nucleosomes. The distribution of poly-T and poly-A tracts around the 5' and 3' end of *Dictyostelium* genes showed that these tracts precisely demarcated the transcription start and end site of transcribed regions. These elements were also associated with NFRs upstream of TSS, embedded with precisely positioned TATA boxes. Our comparative genomics analysis with other eukaryotes including *Plasmodium* showed that such poly-T/poly-A demarcation of *Dictyostelium* genes is not simply owing to the A/T-richness. The A/T-rich property in the *Dictyostelium* genome might be the product of a distinct evolutionary path toward A/T-richness, which *Plasmodium* did not experience.

In Chapter 3, human osteosarcoma U2OS cells were grown and exposed to UV irradiation for p53 pathway activation. This p53 induction was monitored by Western analysis using an anti-p53 monoclonal antibody. After proteins were crosslinked to their interacting DNA by formaldehyde in vivo, sonicated chromatin materials were prepared before and 6 hours after UV treatment. Subsequent chromatin immunoprecipitation (or ChIP) and 5'→3' exonuclease digestion provided target DNA materials for the next deep sequencing (Applied Biosystems SOLiD and Illumina HiSeq 2000 platform used). This ChIP and DNase digestion coupled with high-throughput sequencing (or ChIP-exo) successfully provided a comprehensive and accurate map of 1,825 p53-bound locations across the human genome in response to UV-induced stress. Strikingly, a triple peak-pair pattern was found to be associated with p53 binding to its REs. This pattern was used as a diagnosis of p53/RE binding. The accurate ChIP-exo genome-wide identification of p53-bound REs confirmed the following p53 binding properties to its DNA recognition elements. First, our comprehensive p53 binding map demonstrated that the vast
majority of p53-bound REs lacked an insertion. This result suggested that a 20-bp unsplit RE is a bona fide site for tetrameric p53 proteins, whose degenerate unsplit REs may not be detected around split REs. In addition, no evident peak-pair pattern around isolated half-sites was not found, indicating that the functionally active form of p53 is a tetramer that binds to a full site (Hupp and Lane, 1994). Secondly, p53 RE sequence quality was one determinant of p53 occupancy in a way that higher quality REs tended to be more occupied by p53. The same low quality REs tended to be the most UV-inducible, which suggests that they may be an important player in UV-induced stress response. Interestingly, approximately 20% p53-occupied REs displayed a spacing relationship at 10 bp interval. This spatial arrangement implies a particular evolutionary constraint in that the three half-sites of each palindromically related motif coevolve for p53-mediated stress response. Our further analysis showed that these half-site overlapping sites are usually highly occupied and constitutively bound by p53 rather than inducible by UV. Next, we discovered that other transcription factor motifs involved in environmental stress and oncogenesis co-occur at a nonrandom distance from p53-bound REs. Their structural constraint imposed by their non-palindromic motif orientation strongly suggested that p53 is in a stereotyped contact with those proteins and forms a spatially constrained complex, for example an enhanceosome, which was exemplified by the virus-inducible INF-β gene (Thanos and Maniatis, 1995). This analysis result may provide a mechanism for how p53 transactivates a specific set of target genes in various stress responses. We performed a ChIP-exo assay on the transcription pre-initiation complex (or PIC) component TFIIB and Pol II under the same normal and UV-induced conditions with U2OS. This assay revealed widespread loss of Pol II at the promoter-proximal Pol II pausing region in response to UV treatment, as leaving partial PICs. Our Pol II occupancy measurement in the gene body indicated that the UV-induced transcriptional response was more prevalent where p53 binding was detected closer to the promoter. This result implies that p53-mediated transcriptional regulation promotes the release of a transcriptionally paused Pol II into
an elongation state. Finally, we identified p53 target genes by linking UV-induced fold increases in p53 occupancy to fold increases of Pol II occupancy within the gene body. As a result, 154 high-confidence target genes were identified within 15 kb of each p53-bound location, and 80% were novel p53 targets. They were involved in various cellular functions, such as cell growth, cell cycle, cell signaling, differentiation, cell adhesion and motility, and so on. These data dramatically expanded our understanding of p53-regulated gene networks involving DNA replication, recombination, repair, cellular metabolism and cell proliferation. We furthermore discussed potential disease susceptibility associated with single nucleotide polymorphisms (or SNPs) at p53-bound regulatory elements.

**Major contribution and significance**

The research presented in this dissertation provides the following contributions.

1. The advent of high-throughput genomic profiling technology allows genome-wide mapping of nucleosomes for understanding the role of chromatin in eukaryotic transcription initiation and regulation. By utilizing massively parallel DNA sequencing, we successfully produced the first high-resolution genome-wide maps of nucleosomes in the A/T-rich genome of the social amoeba, *D. discoideum*. This organism is widely used as a model organism for multicellular development and cell motility. The comparison analysis with the nucleosome maps between vegetative cells and multicellular aggregates revealed that *Dictyostelium* has a canonical nucleosome organization that is stable and invariant, regardless of the activity of relevant genes and through different life cycle states.

2. Our comparative genomics analysis of the evolutionarily conserved chromatin structure across eukaryotes led to an insight into a general principle governing the evolutionary constraint on chromatins by transcriptional regulation in multicellular development. The social
amoebae alternate between unicellular and multicellular forms, and its gene number was estimated closer to that of multicellular organisms (Eichinger et al., 2005). Our result demonstrated that *Dictyostelium* organizes genic nucleosomes as in higher multicellular eukaryotes, and downstream placement of the +1 nucleosome was evolutionarily linked to the homology of NELF-encoding genes found across major eukaryotes including *D. discoideum*. This result may reflect that the chromatin structure and the regulatory transcription mechanism imposed by multicellularity coevolve in metazoa.

3. Our high-resolution transcriptome maps revealed the presence of a variety of functionally distinct polymeric A/T elements in the A/T-rich *Dictyostelium* genome. The unusual usage of A/T nucleotides in organizing *Dictyostelium* promoters and genes was surprising. This nucleotide bias in the A/T-rich genome includes poly-A/T tracts highly enriched at the 5' and 3' ends of genes, which were associated with NFRs. These homopolymeric A/T elements intrinsically disfavor nucleosome formation and may help maintain nucleosome-free promoter regions (Struhl, 1985).

4. Nucleosome occupancy levels are governed in part by the underlying DNA sequence, for example G/C content correlated with nucleosome occupancy as reported in yeast (Tillo and Hughes, 2009). Our analysis of the *Dictyostelium* nucleosomal DNA sequences showed that nucleosome placement *in vivo* associates with polymeric-A/T richness towards the nucleosome borders, and G/C-richness (of a relatively higher level) toward the nucleosome dyad that promotes *Dictyostelium* nucleosome formation as observed in other eukaryotes.

5. p53 tumor suppressive function requires DNA sequence-specific interaction, especially noteworthy of degeneracy in its sequence requirement (El-Deiry et al., 1992). This has led to a challenge in characterizing their interactions with DNA in both targeted and genome-wide studies. We applied the recently developed ChIP-exo assay and generated the comprehensive and accurate map of p53 binding in response to UV-induced stress. 1,825 p53-bound locations were identified
across the human genome. Our unbiased comparison with other p53 binding datasets by using p53 motif demonstrated the high-resolution mapping ability of our ChIP-exo approach. Ultimately, this accurate set of p53-bound locations enabled us to delineate parameters surrounding p53 site utilization and discover a common spatial relationship between p53-bound REs.

6. p53 transactivates a particular subset of genes in each distinct stress response, and a long-standing question is how p53 achieves such response specificity. Our systematic approach to identification of other stress-response regulatory elements co-occurring with a nearby p53-bound RE demonstrated that p53 may form an enhanceosome-like structure with others, such as AP1, NRF2, Myc, and FOXO3. This result suggests that other stress-response elements create diverse composite elements including p53 REs, which may provide a mechanistic basis for p53 response specificity. Their common association with endogenous retroviral long terminal repeats supports this model.

7. Our ChIP-exo genome-wide assay with the transcription pre-initiation complex component TFIIB and Pol II showed that the effect of UV treatment on TFIIB/Pol II retention predominated more where p53 binding was closer to the promoter regions. This correlation suggests that UV-induced p53 binding promotes the release of a pre-existing paused RNA polymerase II into an elongation state. This result provides an insight into the mechanism of p53-mediated transcriptional activation of genes in that p53 regulates pausing and not initiation.

8. p53 binds to REs for target gene transactivation, but its distant binding sites from the target have been a challenge in identifying p53 target gene in the human genome. Our novel approach to assigning p53 binding to target genes by incorporating the ChIP-exo mapping of TFIIB and Pol II successfully identified 154 target genes involved in a variety of cellular functions such as cell growth and cycle control. This analysis led to the discovery of 80% novel genes regulated by p53, including non-coding RNA genes and unannotated transcripts. This result
demonstrated the powerful utility of our novel approach by ChIP-exo in identifying new transcription factor target genes.

9. Our gene network analysis by incorporating the target genes (identified in this study) remarkably expanded our understanding of p53 regulatory networks in DNA repair, cellular metabolism, and cell survival and cycle control. Especially, the DNA replication, recombination and repair gene network was filled out by many genes transactivated by p53. This result indicated that UV-induced p53 binding is implicated in regulating multiple steps of the DNA replication, recombination, and repair pathway, associating each other by autoregulatory feedback loop that regulates the activity of the p53 protein.

10. Recent genome-wide association studies suggest that sequence variation in functional non-coding DNA is a significant risk factor for human diseases, but defining these features in this non-coding portion of the human genome remains a challenge (Bernstein et al., 2012). Our systematic search for single nucleotide polymorphisms (or SNPs) present at a p53-responsive element suggested a strong association between a SNP at the p53/RE for POLH and Xeroderma pigmentosum variant (XPV), which is characterized by hypersensitivity to sunlight (Loeb and Monnat, 2008). POLH encodes the DNA-dependent DNA polymerase η which makes a trans-lesion DNA synthesis through UV-damaged pyrimidine dimers. The rs9333500 SNP is expected to eliminate p53 binding for POLH transactivation, resulting in hypersensitivity to UV irradiation or skin malignancy by UV-induced DNA damage. This result demonstrated that the high-resolution ChIP-exo map of transcription factor binding locations provides a powerful framework for GWAS. It may also exemplify a mechanism by which non-coding regulatory variants contribute to gene expression and disease.
Future work

In Chapter 2, our functional and comparative genomics analysis indicated that the exceptional A/T-richness in the Dictyostelium genome implicates a distinct constraint in the evolution of *D. discoideum*, which is different from another A/T-rich genome of *Plasmodium falciparum* (Szafranski et al., 2005a). *D. discoideum* is included in Mycetozoa. Mycetozoa have three distinct groups, the cellular (dictyostelid, e.g. *D. discoideum*), acellular (myxogastrid, e.g. *Physarum polycephalum*), and protostelid (e.g. *Planoprotostelium aurantium*) slime molds. The molecular phylogenetic analysis with the elongation factor-1α (EF-1α) showed the Mycetozoa as a monophyletic group which is branched more closely to the animal-fungal clade than plants (Baldauf and Doolittle, 1997). Further comparative genomics investigation on nucleosome organization with this clade would illuminate the deep root of evolutionarily conserved chromatin structure constrained by multicellularity. Another A/T-rich organism of interest is human pathogen *Entamoeba histolytica*. It is also an amoebozoan (like *D. discoideum*), situated in one of the earliest branches from the last common ancestor of all eukaryotes (Eichinger and Noegel, 2005). In addition to other A/T-rich species in dictyostelid slime molds (e.g. *Dictyostelium lacteum*), it would provide a further insight into the role of highly biased nucleotide usage in organizing nucleosomes, genes, and regulatory elements.

Our evolutionary genomics analysis for the nucleosome organization across the major eukaryotes implied that eukaryotic chromatins and transcription regulation have coevolved in animals. Many archaea contain histones whose primary amino acid sequence is conserved with those of eukaryote nucleosome core histones, although they lack a nuclear membrane. This archaeal nucleosome is known as structurally related to the eukaryotic nucleosome (Pereira et al., 1997; Reeve et al., 1997). Archaeal transcription initiation complies with eukaryotic chromatin context (Wilkinson et al., 2010). Archaea (including Euryarchaeota and Crenarchaeota) and
Eukarya (including Animalia, Plantae, Fungi) are sister lineages, sharing a common ancestor (Woese et al., 1990). A systematic genome-wide nucleosome mapping with Archaea would offer a clue to the evolution and specialization of eukaryotic chromatin structure, since it is situated in the deepest branch diverged just from Bacteria.

In Chapter 3, our ChIP-exo study with the tumor suppressor p53 provided genome-wide new insights into UV-induced and constitutive p53/RE interaction, p53-mediated transactivation mechanism, and p53-regulated gene regulatory network. Our novel approach to identifying p53 target genes included measurement of UV-induced transcriptional response by utilizing elongating Pol II across the gene body. Deep-sequencing-based transcriptome analysis or RNA-seq has been successful and now widely used for whole transcription analysis (Tang et al., 2010). However, such RNA-seq approach did not work for differential gene expression in this study. We speculated that UV irradiation (employed in this study) damages RNA molecules, which blocks the polymerase-mediated reaction step in preparing sequencing library for RNA-seq. Nonetheless, further validation of the experimental basis of Pol II ChIP-exo analysis for transcriptional response would strengthen our approach. This is also connected with our conclusion that p53 transcriptionally regulates genes by releasing paused Pol II in the promoter-proximal region. These following works are considered as our next challenge. (1) Utilization of ChIP with two antibodies for measuring UV-induced transcriptional response, targeting for RNA polymerase II (or Pol II) phosphorylated on Serine 5 of the C-terminal domain upon transcription initiation, and specific for phosphorylation on Serine 2 found on productively elongating and terminating Pol II (Phatnani and Greenleaf, 2006), (2) Use of the isogenic cell lines, p53⁺/⁺ (p53-wildtype) and p53⁻/⁻ (p53-null) HCT116 colon carcinoma cells for evaluating p53-mediated transcription response upon UV treatment, (3) p53 induction by a different mechanism which does not include UV irradiation, for example Nutlin treatment. In detail, the cellular level of p53 is tightly controlled by the p53-Mdm2 autoregulatory feedback loop (Wu et al., 1993). Nutlin is a small
chemical which inhibits Mdm2/p53 complex formation, thus activating the p53 pathway in cells (Poyurovsky et al., 2010; Vassilev et al., 2004). Moreover, the Nutlin approach raises an interesting question on UV-activated p53 protein. This Mdm2 antagonist inhibits Mdm2 interaction with p53, thus p53 does not require post-translational modification on the N-terminal transactivation domain in order to prevent Mdm2-mediated p53 proteolysis degradation (Thompson et al., 2004). This analysis may give us further understanding on how UV-induced DNA damage molds genome-wide p53 binding across the human genome, ultimately connecting to the question of how p53 maintains genome stability.

Previous genome-wide association results have raised a question, where is the missing heritability? (Manolio et al., 2009) Recent progress in identifying rare genetic variations across diverse human populations now provides a rich resource for understanding the role of inherited variants in human diseases (Abecasis et al., 2012). Especially, characterizing noncoding variants at DNA regulatory elements is necessary for our complete understanding of the genetic contribution to human phenotypic diversity, for example by identifying causal genetic variants to phenotypes and disease susceptibility. Our GWAS result for p53 included the public database that contains mostly common variant sites, but not low frequency and rare variants. The much more complete catalogue of human DNA variation would enrich our GWAS approach on noncoding variants at p53 responsive elements, which also becomes a pilot for the future challenge of personalized genome sequencing and medicine.

References


Appendix A

Supplemental Information for Chapter 2

Extended Methods and Materials

Data analysis

The consensus sequence of TATA element, TATAAA[AT][AT], for *D. discoideum* was determined based on the transcriptome mapping data in this study, which allows us to analyze most significant TATA-containing DNA sequences. In order to identify over-represented TATA sequence motifs in core promoters, we initially used the *D. discoideum* consensus TATA box TATAAA[T/A]A, which was reported in 15 genes (Kimmel and Firtel, 1983). We located TATA boxes with this consensus as a preliminary screen and aligned their locations by the TSS of 5,468 genes annotated in this study. We extracted the DNA sequences in the core promoter region from -37 to -21 bp upstream of the TSS, where bona fide TATA boxes are highly likely to be located. Since the purpose of this analysis was to identify the most authentic TATA sequence, we considered the TATA sequences only in the sense (or sense) strand (Basehoar et al., 2004). We looked for over-represented motifs with those sequences using the MEME software (Bailey and Elkan, 1994). A significantly over-represented TATA motif was identified by the MEME algorithm for *D. discoideum* and its consensus TATA sequence, TATAAA[AT][AT], was used in this study.

In accordance with the systematic nomenclature of nucleosome positions in the study of Jiang and Pugh (Jiang and Pugh, 2009), we demarcated the border of the -1, NFR, and +1 zones around transcription start sites by using the level of nucleosome occupancy. As a result, we
categorized nucleosome positions relative to the TSS for both life stages, where the nucleosome dyad is defined as +1 nucleosomes if positioned between +60 and +187 and -1 nucleosomes from -294 to -110, and thus the boundary of 5’ NRF is defined as the region from -109 (as the right border of -1) to +59 (as the left border of +1). The nucleosomes between the +1 nucleosome and the transcription end site of each gene are grouped as other genic nucleosomes such as +2, +3, etc. The TES of some genes was not available, and the ORF end site was used instead. Since the nucleosome occupancy of the -1 nucleosome was moderate, a wider range was chosen by our visual inspection (compared with Yeast (Jiang and Pugh, 2009), i.e. from -307 to -111). As a result, 73,396 nucleosomal core particle DNAs were categorized as -1, +1, and the other genic nucleosomes as of 5,468 protein coding genes with the annotated TSS.

In order for the NPS (nucleosome-positioning sequence) analysis with nucleosomal core particle DNA sequences, the frequency of 10 dinucleotide sets (Satchwell et al., 1986) was compiled as a function of the distance from the nucleosome dyad. Each dinucleotide is defined by reading their sequence in the 5’ to 3’-end direction, and all plots in this study were prepared according to what has been described elsewhere (Albert et al., 2007; Mavrich et al., 2008a; Mavrich et al., 2008b). Furthermore, the expected frequency of each dinucleotide was estimated by random sampling of any DNA sequence in the entire Dictyostelium genome.

**External data sources**

In order for the nucleosome organization of major eukaryotes, 7 species of human, *D.melanogaster, S.pombe, S.cerevisiae, A.thaliana, and C.elegans* were chosen by considering availability of their nucleosome mapping data in public repositories and the evolutionary position in the tree of life (Dacks and Doolittle, 2001). For human, mononucleosome mapping data by MNase digestion was obtained in Schones et al. (Schones et al., 2008), which is for CD4+ T cells
without activation by TCR signaling. We used 3,859 TSSs of the expressed genes (courtesy of Dr. Schones) that were screened by microarray experiments in the study and produced the composite distribution of nucleosome positions at the 5’ end of genes.

The nucleosome occupancy data of *S. pombe* was obtained from Lantermann et al. (Lantermann et al., 2010) (courtesy of Dr. Korber), which were determined at 20-bp resolution by Affymetrix *S. pombe* Tiling 1.0FR array. The nucleosome organization was obtained according to the method described in the study with the TSS of around 4,000 genes. The high resolution maps of nucleosome positions for *S. cerevisiae* and *D. melanogaster* were obtained in our previous publications (Albert et al., 2007; Mavrich et al., 2008a; Mavrich et al., 2008b). The composite distributions of nucleosome positions for these species were produced using the same TSS annotation sets.

For *A. thaliana*, the nucleosome mapping data was obtained as of shoots of three-week old *Arabidopsis* plants in Chodavarapu et al. (Chodavarapu et al., 2010). The nucleosome distribution around the 5’ end of genes was derived with the TSS annotation curated by The *Arabidopsis* Information Resource (TAIR). Also, we obtained the mapping data for *C. elegans* in mixed stages from Valouev et al. (Valouev et al., 2008). 5-pile data set was used to show the nucleosome organization among the full data set that was downloaded from the UCSC genome browser.

**Data visualization and statistical significance test**

Data plots were generated using the statistical package of R (version 2.11.1) and perl scripts with GD graphics library. The density for the locations of specific features (e.g. poly-T/poly-A tracts in the sense strand) were calculated by Gaussian kernel function with bandwidths described, which is implemented in R. For the statistical significance test of each pattern in the
plots (e.g. Figs. 1 and 3), the expected frequency of mononucleotide or dinucleotide was estimated by randomly sampling the same number of DNA sequences at any location in the Dictyostelium genome. This simulation assumed hypothetical genomic features (e.g. nucleosomal DNA, TSS, or TES) which are randomly positioned in the sequence-independent manner. Two randomization simulations were independently performed and shown in all plots.
Supplementary Figures

Supplementary Figure 2-1. Overrepresentation of poly-T/poly-A tracts in the *Dictyostelium* genome.

**A** Frequency distribution of poly-T/poly-A tracts in the *Dictyostelium* genome. The occurrence of non-overlapping poly-T/poly-A tracts of length 2 or more (n ≥ 2) were counted in all 6 chromosomes of *D. discoideum* (total 33,929,503-bp genome size), and their relative frequencies were displayed as a function of the tract length (n) in the figure. The observed relative frequency of each tract of a given length was calculated as a ratio of their count number to the genome size, which is the number of observed tracts normalized to the genome size. For instance, n=10 was observed to occur once every 1,810 bp in the genome, whereas the expected frequency of n=10 was only once every 17,400 bp. The expected frequency of the occurrence of a non-overlapping poly-T/poly-A tract was calculated as a zero-order Markov chain (Dechering et al., 1998).

**B** Frequency distribution of poly-T/poly-A tracts in five eukaryotic genomes. The observed relative frequency of poly-T/poly-A tracts was shown as a function of the tract length after normalization by genome size, as described. As a result, the high incidence of long poly-T/poly-A tracts is unique to A/T-rich genomes, i.e. *D. discoideum* and *P. falciparum*. For example, 90,201 (n ≥ 12) tracts were found in the *Dictyostelium* genome, which means n ≥ 12 was observed to occur once every 376 bp in average. In *P. falciparum*, 34,798 found for n ≥ 12, thus once every 669 bp. Moreover, 1,117 were found, occurring on average once every 10,807 bp in *S. cerevisiae*, 15,617 were found once every 7,708 bp in *D. melanogaster*, and 97,541 found once every 31,581 bp in the human genome.
Supplementary Figure 2-2. Lack of significant correlation of transcript abundance and the enrichment of poly-T/poly-A tracts at the ends of Dictyostelium genes.

Correlation analysis between poly-T tracts at the 5’ end (from -200 to TSS) of 5,468 protein-coding genes and their mRNA expression level (A) and between poly-A tracts at the 3’ end (from TES to 200) of 5,400 protein-coding genes and their expression level (B). Each scatter plot represents the number of T’s at the 5’ end and A’s at the 3’ end of those Dictyostelium genes versus the mRNA expression level. The mRNA transcript abundance of a gene was calculated in the transcriptome analysis after data normalization. As shown with $R^2 \approx 0.000$ and $R^2 \approx 0.000$, no significant correlation was found between poly-T/poly-A tracts and transcript abundance of Dictyostelium genes.
Supplementary Figure 2-3. Lack of significant correlation between poly-T enrichment near the TSS and poly-A enrichment near the TES in the *Dictyostelium* genome.

(A,B) DNA sequence track views around the 5’ (or TSS) and 3’ (or TES) end of *Dictyostelium* genes. Both TSS and TES were identified for 3,265 protein-coding genes of *D. discoideum* in this study. Two track views were produced akin to Figure 2-1C of the main text, and each paired tracks were linked at the 5’ and 3’ end. Plots were primarily sorted by “T” nucleotide content in the 5’ region (from -100 to TSS) in panel A, and by “A” nucleotide content in the 3’ region (from TES to +100) in panel B.

(C) Correlation analysis between poly-T at the 5’ end (from -100 to TSS) and positionally-linked poly-A at the 3’ end (from TES to +100). As a result, 3,265 data points were used to extrapolate the linear regression line in the figure, and the coefficient of determination (R^2) was 0.004. When a 20-bp region was chosen near the TSS and TES instead of 100-bp, no significant correlation was detected (R^2 = 0.003, data not shown).
Supplementary Figure 2-4. Distribution of poly-T/poly-A tracts around the 5’ and 3’ end of Pol
II-transcribed genes of S. cerevisiae and D. melanogaster.

(A,B) Distribution of the locations of poly-T/poly-A tracts in the DNA sequence of all genes with the known TSS. The figures were generated by the same protocol described in Figure 2-1A.

(C-F) Composite distribution of the location of poly-T/poly-A tracts around the TSS (C,E) and TES (D,F) of all annotated genes of Saccharomyces (C,D) and Drosophila (E,F). The density of poly-T/poly-A tracts (n ≥ 6) was estimated as a function of the distance (bp) between the middle location of each tract and the given TSS as described in Figure 2-1B, except the smoothing bandwidth of 10 bp used. Tracts were counted from the sense (non-template) strand. The composite distributions of nucleosome positions were added as gray-filled traces (Albert et al., 2007; Mavrich et al., 2008b).
Supplementary Figure 2-5. Comparison of the DNA content around the 5’ and 3’ end of protein-coding genes between *P. falciparum* and *D. discoideum*. We did not have high resolution maps of TSS and TES for *Plasmodium*, and so could only examine distributions around ORF start and end sites. Therefore we also generated similar ORF-centered maps in *Dictyostelium*. In comparison to *Dictyostelium*, similar A and T nucleotides biases did not exist in the expected regions upstream of *Plasmodium* ORF start sites.

(A,C) DNA sequence track view around the 5’ and 3’ end of *Plasmodium* (A) or *Dictyostelium* (C) ORFs. We curated 5,491 protein-coding genes that were annotated in *Plasmodium* Genomics Resource (7.2 release) (Aurrecoechea et al., 2009). Genes were aligned by the ORF start/end and sorted by the T or A content in the whole fetched DNA sequences (301 bp).

(B,D) Frequency distribution of A, C, G and T nucleotide at every position across the DNA sequence including the start/end of *Plasmodium* (B) or *Dictyostelium* (D) ORFs. The same frequency distribution of each nucleotide was obtained for the *Plasmodium* genome, as described previously except that the DNA sequences were fetched in the region from -60 to 40 relative to the ORF start (upper panel) and from -40 to 60 relative to the ORF end of 5,491 genes (lower panel). Also, two random simulations were performed to estimate the expected frequency of T and A nucleotides, shown as two gray traces in the figures.
**Supplementary Figure 2-6.** Isolation of nucleosome core particle DNA from *D. discoideum*.

(A) Ethidium bromide stained agarose gel of samples in mononucleosome preparation for sequencing. After crosslinked and lysed, omatin pellets were treated with MNase and digested to predominantly mononucleosomal size. Finally, resulting nucleosomal DNAs were size selected by gel purification. The nucleosomal DNA sample prepared in the vegetative stage was shown in the figure.

(B) Frequency distribution of the length of sequencing reads for nucleosomal DNA. The frequency distribution was calculated with nucleosome core particle DNA prepared and sequenced from multicellular *D. discoideum* aggregates. This figure included all sequencing reads and the mode of the read length was 144 bp. Also, a 133-bp mode was calculated in the vegetative stage.
Supplementary Figure 2-7. Frequency distribution of the indicated dinucleotide pairs and W, S nucleotides at every position across the nucleosomal DNA.

(A) Frequency distribution of the indicated dinucleotide pairs at every position across the 147-bp nucleosomal DNA and flanking regions. Each dinucleotide is defined in the 5’ to 3’-end direction, for example 5’-AT-3’. The number of each dinucleotide was counted by reading both strands in the 5’ to 3’ direction. Only 10 unique frequency distributions exist from 16 dinucleotides—AA (= TT), GG (= CC), AT, GC, TA, CG, AG (= CT), TG (= CA), AC (= GT), and TC (= GA)(Albert et al., 2007; Satchwell et al., 1986). 246,513 positions of nucleosome dyads were identified in the two genome-wide nucleosome maps of vegetative and aggregating cells. Their nucleosomal DNA sequences were aligned by their nucleosome dyads and extended to include the linker DNA (i.e. from -150 to 150 relative to the dyad). The frequency count of each dinucleotide was shown in the y axis after smoothing (see the Methods). The schematic bar in the upper part of each plot represents the rotational orientation of the major groove against the histone octamer surface. For the significance test of each pattern, the expected frequency of each dinucleotide was calculated by randomly picking up the same numbers of DNA sequences at any location from the Dictyostelium genome. This simulation assumed the dinucleotide frequency of hypothetical nucleosomes, which are randomly positioned in a sequence-independent manner. Two computational simulations were independently carried out and shown as two gray traces in each plot.

(B) Frequency distribution of the W and S single nucleotides at every position across the 147-bp nucleosomal DNA. The frequency distribution of W (A or T) and S (G or C) was calculated by reading the 246,513 nucleosomal sequences. The count number of a given nucleotide was obtained in the forward strand, which was finally summed up from all the sequences in a composite manner and smoothed using a three-base moving average algorithm. The frequency of W was shown in black and S in red in the figure, and their expected frequencies were also calculated described as in A, displayed as gray for W and light red for S. As shown, in vivo nucleosome placement was associated by being G/C-rich near the nucleosomal dyads.
Supplementary Figure 2-8. In vivo nucleosome placement with polymeric-A/T sequences near nucleosome peripheries.
(A) Distribution of the locations of heteropolymeric-A/T tracts across nucleosomal core particle DNAs. 246,513 nucleosomal DNA sequences were sorted by G/C content that varied from 2% to 59%. Each track represents an individual nucleosomal DNA sequence, in which the dyad of the 147-bp nucleosomal core DNA was centered as shown in the yellow box. All DNA sequences in the forward strand were aligned by their dyad and shown in the 5’ to 3’ direction. They included linker DNAs by being extended from -150 to 150 relative to the nucleosome dyad. The G/C content was measured by taking the whole 301-bp DNA sequence, and all tracks were displayed in ascending order of G/C content in the figure. If any heteropolymeric-A/T tracts (n ≥ 12) were found along the nucleosomal DNA sequence, all the bases were colored as black in the tract. The other bases were colored in white. The homopolymeric stretches of A or T (called as poly-T/poly-A tracts) were not included as heteropolymeric-A/T.

(B) Composite distribution of the locations of polymeric-A/T tracts across nucleosomal DNAs. Heteropolymeric-A/T sequences were counted per each position across the 246,513 nucleosomal DNA sequences. The density of the occurrence of heteropolymeric-A/T tracts (n ≥ 12) was estimated using Gaussian kernel with the smoothing bandwidth of 5, and plotted as a function of the distance between the tract and dyad. This relative distance was calculated by the location of the center of each heteropolymeric-A/T sequence and their frequency was calculated in the forward strand, shown as a black trace in the figure. Also, the frequency distribution of homopolymeric-A/T sequences (n ≥ 12) was obtained in the same manner and displayed as grey in the figure. The density of polymeric-A/T tracts towards the nucleosome border is relatively higher than in the dyad region, which indicates the higher incidence of polymeric-A/T tracts in the linker DNA.
Supplementary Figure 2-9. Nucleosomal DNA properties of Dictyostelium intergenic nucleosomes and genic nucleosomes.

(A,B) Distribution of the size of nucleosome protected DNA fragments which are grouped based on nucleosome position. The sequencing reads mapped to the reference genome were grouped by their dyad (or middle) locations, as -1, +1, +2, +3, +4, +5, and all other genic nucleosomes (such as +6, +7, …). The distribution of the length of reads is displayed for each group, and plotted in the vegetative Dictyostelium (A) and in the multicellular aggregate (B) separately. The frequency of read length was binned at 10-bp interval for the vegetative stage and 5-bp interval for the aggregation stage, and shown for -1 in red, +1 in orange, +2 in yellow, and so on.

(C) Distribution of the sequence composition for nucleosome protected DNA for each position. 73,396 nucleosome dyad locations were grouped as -1 (N=3,230), +1 (N=7,285), +2 (N=8,288), +3 (N=8,792), +4 (N=7,605), +5 (N=7,082), and all other genic nucleosomes (N=31,922), which were combined from the data of the vegetative and aggregation stage. The nucleosomal DNA sequences were fetched as in 301-bp length including linker DNA, and the W (A or T) nucleotide
percentage was calculated at every position on the sense (nontemplate) strand (from 5’ left to 3’ right with the same directionality of transcription), and the frequency of W was plotted by the relative distance term from the dyad and smoothed using a three-bp moving average algorithm.
Supplementary Figure 2-10. Multiple sequence alignment with NELF-B and NELF-D subunit to generate the NELF homology model between human, Drosophila, and Dictyostelium.

NELF-B and NELF-D subunits in the NELF complex were reported as evolutionarily conserved
between human, fruit fly, and mouse (Narita et al., 2003). Structure-based sequence alignments of homologous COBRA1 domains (Cofactor of BRCA1, Pfam accession number: PF06209) in NELF-B (A) and TH1 domain (Trihydrophobin 1, Pfam accession: PF04858) of NELF-D (B) are shown. The amino acid sequence of NELF-B was obtained as NCBI Accession number NP_056271 (human), NP_572402 (D. melanogaster), and XP_638637 (D. discoideum), and their multiple alignment was produced by using PROMALS (Pei and Grishin, 2007; Pei et al., 2007) with default parameters. The first line in each block is conservation indexes associated with each position (an integer between 0 and 9, with 9 corresponding to highest conservation (Pei and Grishin, 2001)). A predicted secondary structure was reported as a colored sequence (red: alpha-helix, blue: beta-strand), and together the consensus secondary structures is shown in the last line (h: alpha-helix, e: beta-strand). Also, the amino acid sequence of the TH1 domain was fetched as NP_945327 (human), NP_573123 (D. melanogaster), and XP_646857 (D. discoideum) and their multiple alignment is shown in B. Furthermore, we have tried to detect homology of the B and D subunit with the proteomes of A. thaliana, C. elegans, S. cerevisiae, and S. pombe by using PSI-BLAST, but no sequence-level homology was detected from those eukaryotes.
Supplementary Table 2-1. Dinucleotide compositions in five eukaryotic genomes including A/T-rich *D. discoideum* and *P. falciparum* genome.

The dinucleotide compositions were calculated based on all omosome DNA sequences of 5 eukaryotes. If the reverse complementary sequence is considered for each of all possible 16 dinucleotides, several dinucleotides are equivalently assigned, for example AA=TT, GG=CC, AG = CT, etc., which results in unique 10 dinucleotide sets (or steps). The percent composition (%) of the 10 dinucleotide sets was calculated per each species in the table. Compared with budding yeast, human, and fruit fly, the most common dinucleotide in the A/T-rich *Dictyostelium* and *Plasmodium* genome was WW. (WW: AA, TT, AT, TA, SS: GG, CC, GC, CG, SW: GA, GT, CA, CT, WS: AG, AC, TG, TC).

<table>
<thead>
<tr>
<th></th>
<th><em>D. discoideum</em></th>
<th><em>P. falciparum</em></th>
<th><em>S. cerevisiae</em></th>
<th>Human</th>
<th><em>D. melanogaster</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>AA + TT</td>
<td>36.4</td>
<td>32.3</td>
<td>21.6</td>
<td>19.6</td>
<td>20.2</td>
</tr>
<tr>
<td>GG + CC</td>
<td>3.8</td>
<td>2.9</td>
<td>7.8</td>
<td>10.4</td>
<td>9.4</td>
</tr>
<tr>
<td>AT</td>
<td>13.4</td>
<td>17.5</td>
<td>8.9</td>
<td>7.7</td>
<td>8.1</td>
</tr>
<tr>
<td>GC</td>
<td>1.1</td>
<td>0.9</td>
<td>3.7</td>
<td>4.3</td>
<td>5.7</td>
</tr>
<tr>
<td>TA</td>
<td>10.9</td>
<td>15.9</td>
<td>7.3</td>
<td>6.6</td>
<td>6.3</td>
</tr>
<tr>
<td>CG</td>
<td>0.7</td>
<td>0.7</td>
<td>2.9</td>
<td>1.0</td>
<td>4.2</td>
</tr>
<tr>
<td>AG + CT</td>
<td>6.6</td>
<td>6.4</td>
<td>11.7</td>
<td>14.0</td>
<td>10.8</td>
</tr>
<tr>
<td>TG + CA</td>
<td>10.7</td>
<td>8.7</td>
<td>13.0</td>
<td>14.5</td>
<td>13.8</td>
</tr>
<tr>
<td>AC + GT</td>
<td>7.7</td>
<td>7.0</td>
<td>10.5</td>
<td>10.1</td>
<td>10.4</td>
</tr>
<tr>
<td>TC + GA</td>
<td>8.7</td>
<td>7.7</td>
<td>12.5</td>
<td>11.9</td>
<td>11.1</td>
</tr>
<tr>
<td>WW</td>
<td>60.8</td>
<td>65.8</td>
<td>37.9</td>
<td>33.9</td>
<td>34.5</td>
</tr>
<tr>
<td>SS</td>
<td>5.6</td>
<td>4.5</td>
<td>14.5</td>
<td>15.7</td>
<td>19.3</td>
</tr>
<tr>
<td>SW + WS</td>
<td>33.7</td>
<td>29.8</td>
<td>47.7</td>
<td>50.5</td>
<td>46.1</td>
</tr>
</tbody>
</table>

* The ambiguity code “N” in DNA sequences was excluded in the content.
**Supplementary Table 2-2.** (A) Consensus location of nucleosome positions downstream of the TSS of seven eukaryotes. We curated *in vivo* mononucleosome mapping data of human, *D. melanogaster, S. pombe, S. cerevisiae, A. thaliana,* and *C. elegans* from the literatures and produced the composite distributions of nucleosome locations at the 5' end of genes. According to the systematic naming of each predominant nucleosome position (Jiang and Pugh, 2009), the consensus location of each nucleosome position (such +1, +2, and so on) was obtained as the location of the highest nucleosome occupancy based on the composite distribution of nucleosome locations relative to the TSS. Some nucleosome positions were not clearly identified from the composite distributions due to lack of high resolution maps of TSS.

(B) Averages of inter-nucleosomal spacing intervals of 7 species. The spacing distances between genic nucleosomes (i.e. +1, +2, +3, +4, and +5) were obtained as the peak-to-peak distance (bp) from the composite distribution of nucleosome positions each species, for example the distance between +1 and +2, and between +2 and +3. These spacing intervals were averaged in the table. These averages are consistent with the literatures. For example, (Lantermann et al., 2010) reported nucleosome repeat lengths as 154 bp. However, we were not able to calculate robust estimates for *A. thaliana* and *C. elegans* due to lack of high resolution map of TTS annotation. In case of *A. thaliana,* (Chodavarapu et al., 2010) computed approximately 175-bp spacing between inter-nucleosomes, and (Valouev et al., 2008) reported that *C. elegans* nucleosomes are uniformly distributed at the 175-bp interval.

**A. Consensus location of the nucleosome dyads relative to the TSS of genes**

<table>
<thead>
<tr>
<th>Species</th>
<th>+1</th>
<th>+2</th>
<th>+3</th>
<th>+4</th>
<th>+5</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. discoideum</em> (vegetative)</td>
<td>116</td>
<td>265</td>
<td>435</td>
<td>588</td>
<td>762</td>
</tr>
<tr>
<td><em>D. discoideum</em> (aggregation)</td>
<td>109</td>
<td>264</td>
<td>441</td>
<td>601</td>
<td>786</td>
</tr>
<tr>
<td>Human</td>
<td>120</td>
<td>308</td>
<td>488</td>
<td>663</td>
<td>841</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>130</td>
<td>306</td>
<td>473</td>
<td>649</td>
<td>832</td>
</tr>
<tr>
<td><em>S. pombe</em></td>
<td>60</td>
<td>210</td>
<td>360</td>
<td>510</td>
<td>660</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>56</td>
<td>225</td>
<td>391</td>
<td>558</td>
<td>724</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>126</td>
<td>256</td>
<td>413</td>
<td>575</td>
<td>736</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>61</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

* Not determined in our composite distribution plots.
B. Averaged spacing interval (bp) of inter-nucleosomes

<table>
<thead>
<tr>
<th>Species</th>
<th>Interval (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>D. discoideum</em> (vegetative)</td>
<td>162</td>
</tr>
<tr>
<td><em>D. discoideum</em> (aggregation)</td>
<td>169</td>
</tr>
<tr>
<td>Human</td>
<td>180</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>176</td>
</tr>
<tr>
<td><em>S. pombe</em></td>
<td>150</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>167</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>175*</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>175*</td>
</tr>
</tbody>
</table>

* The spacing distances estimated in the literatures.

**Supplementary Table 2-3.** Annotation of *D. discoideum* transcription start sites (or TSS) and transcription end sites (or TES). The reference genome sequence of *Dictyostelium discoideum* AX4 was obtained from dictyBase (http://dictybase.org), which was distributed in February 2008. This annotation data was generated by the transcriptome sequencing and analysis (described the Method) based on this genome build.

Genome Research at [http://genome.cshlp.org/content/22/6/1098.long](http://genome.cshlp.org/content/22/6/1098.long)

Thesis-Chang, Gue Su-Tables.xls
**Supplementary Table 2-4.** Comparison of literature TSS with RNA-seq TSS

<table>
<thead>
<tr>
<th>ID</th>
<th>Chr</th>
<th>Str</th>
<th>COS start</th>
<th>COS end</th>
<th>TSS</th>
<th>TES</th>
<th>UTR S' (bp)</th>
<th>UTR 3' (bp)</th>
<th>5' UTR</th>
<th>3' UTR</th>
<th>literature/notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sequence of described UTR not entirely the</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>same sequence as from genome project</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>polyA stretch</td>
</tr>
<tr>
<td>DOBO191498</td>
<td>3</td>
<td>-</td>
<td>2570624</td>
<td>2570981</td>
<td>2570652</td>
<td>2568934</td>
<td>28</td>
<td>147</td>
<td>93</td>
<td>65</td>
<td>J. Gen Microbiol (1993), 330:94-3052; Na</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>experiment described in this paper</td>
</tr>
<tr>
<td>DOBO1915010</td>
<td>3</td>
<td>-</td>
<td>4543011</td>
<td>4543912</td>
<td>4543905</td>
<td>4540889</td>
<td>294</td>
<td>23</td>
<td>134</td>
<td>-160</td>
<td>Mechanisms of Development (1994) 45:59-72; This gene has an</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>extraordinarily long and complex promoter</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>region with at least 3 initiation sites. Adjacent</td>
</tr>
<tr>
<td>DOBO191479</td>
<td>4</td>
<td>-</td>
<td>1595735</td>
<td>1956975</td>
<td>1956901</td>
<td>1956917</td>
<td>1166</td>
<td>58</td>
<td>050</td>
<td>-516</td>
<td>gene is far away.</td>
</tr>
<tr>
<td>DOBO193156</td>
<td>5</td>
<td>-</td>
<td>2344889</td>
<td>2344895</td>
<td>2344841</td>
<td>2342704</td>
<td>142</td>
<td>92</td>
<td>142</td>
<td>0</td>
<td>J. Biol Chem (1992) 267:27, 19935-19964</td>
</tr>
<tr>
<td>DOBO191974</td>
<td>4</td>
<td>-</td>
<td>3931286</td>
<td>3929775</td>
<td>NA</td>
<td>3929722</td>
<td>ND</td>
<td>53</td>
<td>487</td>
<td>ND</td>
<td>Devel Biol (2000) 255:373-382</td>
</tr>
</tbody>
</table>

1. 5' UTR length (bp) from literature.
2. Bases upstream (-) or downstream (+) of TSS in literature.
Supplementary References


Appendix B

Supplemental Information for Chapter 3

Extended Methods and Materials

Human genome annotation

Human gene annotation was obtained from the Vertebrate Genome Annotation (or VEGA) Database (Release 45, http://vega.sanger.ac.uk) (Wilming et al., 2008) and the UCSC Genome Bioinformatics (http://genome.ucsc.edu) (Dreszer et al., 2011). The reference human genome of NCBI36/hg18 (Mar. 2006) was downloaded from the UCSC genome annotation database, to which sequencing reads were aligned. Gene and transcript classification (called as biotypes) annotated by Human and Vertebrate Analysis and Annotation (or HAVANA) team were used in this study. 622 human tRNA genes of hg18 screened by tRNAscan-SE were downloaded from the Genomic tRNA Database (http://gtrnadb.ucsc.edu) (Lowe and Eddy, 1997). Annotation of human repeating elements, created by using RepeatMasker with the Repbase update (Jurka, 2000), were obtained from the UCSC genome annotation databases.

External data sources

Published genome-wide datasets of p53 binding were obtained from ChIP-PET (Wei et al., 2006), ChIP-seq (Botcheva et al., 2011; Smeenk et al., 2011), ChIP-chip (Cawley et al., 2004; Shaked et al., 2008; Smeenk et al., 2008), and ChIP-chip for p63 (Yang et al., 2006), which were used for comparison with the p53-bound list identify by ChIP-exo in this study. The list of
experimentally validated p53-response genes was obtained from Riley et al. (Riley et al., 2008), which were chosen by the four sets of experimental criteria to categorize as up/down-regulated genes by p53. Human SNPs (NCBI dbSNP Build 130, http://www.ncbi.nlm.nih.gov/projects/SNP/) (Sherry et al., 2001) were used to detect a new association between common disease and known genetic variation in the p53 REs identified in this study. GWAS resources were from PheGenI (Phenotype-Genotype integrator, http://www.ncbi.nlm.nih.gov/gap/PheGenI) (Mailman et al., 2007), OMIM (Online Mendelian Inheritance in Man, OMIM®, http://www.omim.org/) (McKusick, 2007), and the NHGRI GWAS Catalog (A Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/gwastudies/) (Hindorff et al., 2009). TOMTOM (Gupta et al., 2007) associated with the JASPAR database (Sandelin et al., 2004) were used for motif identification of potential p53 cofactors around p53-bound regions. FIMO (Grant et al., 2011) was employed to search for binding sequence candidates of each motif, and their position-specific scoring matrices (or PSSM) were obtained from the literature, JASPAR, or by MEME motif analysis. Furthermore, scientific literature was searched via iHOP (http://www.ihop-net.org/UniPub/iHOP/) (Hoffmann and Valencia, 2004) to find out any reported interaction or associated function of potential p53 cofactors with p53. For human nucleosome data, mononucleosome mapping data, prepared by MNase digestion, were obtained from (Schones et al., 2008) for CD4+ T cells without/with activation by TCR signaling, and these mapping data were combined for use in this study.

Identification of p53 binding sites

**Rationale.** In an attempt to determine a comprehensive set of p53 binding locations, as opposed to only the highest confidence subset, we applied a minimal set of criteria for defining candidate p53-bound locations that included reproducible tag clustering on both strands. We
recognized the possibility that p53 may crosslink to regions of the genome that may not directly involve RE-p53 interactions (e.g. looping or DNA binding stabilization via other DNA-bound factors). We also did not want to exclude the possibility that p53 might bind to half-sites, split sites, and clustered sites. For this reason, we sought to identify the maximum number of potential p53-occupied regions, then separate them based upon predominant patterns of tag distributions. Ultimately, we used the underlying DNA sequence to validate or question putative binding locations.

**p53-occupied regions.** To identify a common set of p53-occupied regions, data from all six replicates (three each for 0 h and 6 h) were combined (after demonstrating the reproducibility of the data patterns). For this particular processing, the strand identity of each tag was removed. Only the coordinate corresponding to the 5' end of each tag was processed. The set of p53-occupied regions were defined as genomic intervals that lacked a >60 bp gap of tags, and also had a reproducible peak pair as defined below. The purpose of creating this data intersection was to ensure that peak-pairs and any surrounding tags were captured. A total of discrete 1,825 p53-occupied regions which contained at least one reproducible p53-bound location were determined as follows.

**Peak determination and peak-pairing.** Only the 5' end of each tag was used in data processing. Individual biological replicates were subject to peak calling using GeneTrack (Albert et al., 2008) (sigma = 2, exclusion zone = 5). Peaks were then paired if they existed on opposite strands and were offset in the range of -10 to +45 bp (a positive value indicates the offsets are in the 3’ direction). Peak-pairs were retained if they were identified in at least 2 out of 3 biological replicates. The top 500 peak pairs in 6h, selected as the most occupied p53-bound location per each occupied region, were subjected to MEME analysis, and the top-scoring motif was the unsplit p53 RE consensus (Hoh et al., 2002; Wei et al., 2006). The PSSM (position-specific scoring matrix) obtained with these top p53-bound locations was used in Figure S1C.
**RE identification.** We searched for half-site candidates (consensus RRRC\(\_\)WWG\(\_\)YYY) by allowing up to three mismatches from 1,825 occupied regions identified above. Next a search for a second half-sites was performed allowing up to 4 mismatches, but requiring that any found half-site be 9-23 bp away (midpoint to midpoint, or -1 to 13 bp indel between the half-sites). The second half site was allowed to have up to four mismatches. Only one mismatch was allowed the C\(\_\) or the G\(\_\) position in the paired half-sites (or full-sites), and the total number of mismatches in full-sites could not exceed seven. The reason that we chose seven as the limit is that the number of new RE instances near p53-bound locations fell to near zero at this limit. Only 19 of the top 1000 occupied regions were at this limit, see Table S3-1. Only paired half-sites that had at least one half-site with reproducible tag count > 0 on each strand in the templated region (described under the *Occupancy determination* section, below) were called as p53-bound. If multiple sites met the criteria for a bound region, the strongest RE with the shortest insertion size (i.e. 0 or unsplit RE) was taken. The reason is that unsplit REs (or full-site without insertions between half-sites) were predominantly found near the p53-bound locations. For all other bound regions that did not meet this criterion (most of which were lowly bound by p53), we searched for an RE (no insertion) within 25 bp of a measured peak-pair midpoint. Where multiple REs were bound by p53 in an occupied region, the one with the closest match the p53 consensus was displayed in Figure 3-1A, 3-1B. The total 1,825 regions were grouped preliminarily into Group 1 (n=1,453) and Group 2 (n=265), with the latter containing only those with -1, and 1~13 bp indels. 107 regions did not meet the RE criteria and were designated as Group 3.

Since an unsplit RE was the most predominant species, and that the degeneracy of an RE might allow the same motif to simultaneously appear as a split and an unsplit motif, we opted for a second sweep through the initial Group 2 set using MEME to identify any over-represented motif. An unsplit RE was returned (E-value = 1.6e-97). We then applied this Group 2 RE PSSM and FIMO (p < 10-3) to Group 2 and found 119 instances of unsplit REs within 120 bp interval
centered at Group 2 RE midpoint. These were transferred to Group 1 (bottom set of Group 1 in Figure 3-1A), resulting in a final total of 1,572 Group 1 locations. This group set was used for analysis in this study. The remaining 146 Group 2 were subsequently analyzed in Figure S3-3A.

*Occupancy determination.* The general purpose here was to use the peak pattern around an RE to quantify p53 occupancy levels. This was achieved for Group 1 and 2 by centering an 18-nucleotide window over the 5’ end of each half site and summing the number of the 5’ end of tags present on the same strand within that window. This 18-nucleotide window length was set to encompass the two canonical peak locations per half site on each strand, as shown in Figure 3-1C. For occupancy measurement for a full site, the occupancy of each half site were combined and summed up. Group 3 locations placed a 28 bp window centered at the peak-pair midpoint for occupancy measurement. The p53 occupancy levels were normalized by globally with a constant background assumption between the 0 and 6 hour conditions. Thus, three biological replicates were combined for each time point, and all tag counts scaled such that the total tag count in the background regions (i.e., excluding p53-bound regions) were the same between the 0 and 6 hour time point. The ratio of 6 hour occupancy level to 0 hour occupancy level for each p53-bound location was then log2-transformed as a fold change (Δocc) in occupancy in response to UV-induced stress.

**TFIIB and Pol II**

*Promoter-proximal TFIIB-enriched locations.* For mapping the transcription pre-initiation complex component TFIIB, our standard ChIP-exo data analysis protocol was employed (Rhee and Pugh, 2011). Peaks were called with each biological replicate (3 replicates per each condition) by GeneTrack (sigma = 20, exclusion zone = 100). Peaks that were on opposite strands and -50 to 100 bp apart were paired, and 2 out of 3 biological replicates at each location were
required to be present and <100 bp apart. This resulted in 1,313 TFIIB-bound locations (included in Figure S3-4A). Of these, 751 were <50 bp upstream and <150 bp downstream of known protein-coding and noncoding TSS (by the Human and Vertebrate Analysis and Annotation or HAVANA) and thus represents a high confidence set of TFIIB/TSS locations, included in Figures 3-5A and S3-4B,C, and Table S3-2. These TFIIB/TSS locations were the basis for calculating fold changes in TFIIB and Pol II occupancy upon UV induction in the promoter-proximal region (described in the following section). If a gene had multiple TSSs, the TSS with the highest TFIIB occupancy at its TFIIB/TSS location was chosen as the reference TSS for that gene. The rationale behind this TSS selection is that the majority of known genes have a single region (within 100 bp in length) that is highly enriched with TFIIB slightly downstream of the TSS (except for divergent transcription loci).

Since the strict TFIIB/TSS criteria and low coverage of some replicate sequencing data resulted in a small number of TFIIB/TSSs, more relaxed criteria were applied to identify additional promoter-proximal TFIIB-enriched locations for certain analyses. First, the 5’ end of all sequencing tags were shifted in the 3’ direction by 20 bp for TFIIB, then tags on both strands were combined from all biological replicates of both 0h and 6h prior to peak calling. Peak calling was conducted using GeneTrack (sigma = 20, exclusion zone = 100). Singleton peaks (defined as a peak with reads only at a single coordinate) were filtered out. Then, these peaks (not peak-pairs) were searched for the nearest known TSS, meeting the criteria <50 bp upstream and <150 bp downstream of known protein-coding and noncoding TSS. The minimum tag count criterion (>100 TFIIB tags) was applied to determine significant TFIIB/TSS locations. Any genes of less than 600 bp in length were not included, which turned out to be a small set of annotated genes (n < 30) and most of them were highly expressed histone clusters such as HIST2H2AB. As a result, 7,132 genes as protein-coding and noncoding met the criteria and were used to identify genes around p53-bound locations (Figure 3-5B and Table S3-3,S3-4).
TFIIB and Pol II occupancy. The occupancy of TFIIB and Pol II in promoter-proximal regions was measured as the number of tag 5’ ends residing 100 bp upstream and downstream of TFIIB peak pair midpoint, after shifting the tag 5’ ends in the 3’ direction by 20 bp for TFIIB and 18 bp for Pol II. These shifts were also employed when graphically displaying tag locations (Figures S3-3D, S3-4D, S3-5A, and S3-5C). The combined tags of all replicates in each condition were counted. Occupancy changes were measured as the ratio of the TFIIB and Pol II occupancy at 6 versus 0 hour after normalizing to background as described above for p53, then log2-transformed. For each gene with a TFIIB/TSS location, Pol II occupancy across the gene body was measured by using the transcription end site (TES) of the attached transcript, where a gene body is defined as the genomic region from 100-bp downstream of the TFIIB/TSS midpoint to the TES. The upstream boundary therefore excluded paused Pol II.

p53 regulatory networks.

The purpose of this analysis was to ascertain the extent to which p53 is physically linked to genes that exist within the same functional network (e.g. DNA repair). 154 genes identified as p53 targets (Table S3-3) were analyzed, by incorporating the fold changes in occupancy of their associated p53 binding upon UV treatment. This gene set was uploaded into IPA® software developed by Ingenuity Systems, Inc. (Redwood City, CA, USA. http://www.ingenuity.com/), and their prevalence among the genes represented in pre-established networks were indicated by red (increase in p53 binding) or green (decrease in p53 binding) nodes, and gray nodes are the genes which are not in our input list. Pre-existing interactions among genes/proteins are shown as edges.
<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>Cell</th>
<th>Treatment</th>
<th>N</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>FPR (%)</th>
<th>Mod (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This study</td>
<td>ChIP-seq</td>
<td>p53</td>
<td>U2OS</td>
<td>2408</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>59.9</td>
</tr>
<tr>
<td>Wei et al. (2006)</td>
<td>ChIP-FET</td>
<td>p53</td>
<td>HCT116 5-fluorouracil</td>
<td>542</td>
<td>369</td>
<td>1,466</td>
<td>195</td>
<td>36.0</td>
<td>8.9</td>
</tr>
<tr>
<td>Smeerik et al. (2011)</td>
<td>ChIP-seq</td>
<td>p53</td>
<td>U2OS Actinomycin D</td>
<td>2,132</td>
<td>1,027</td>
<td>798</td>
<td>1,113</td>
<td>52.2</td>
<td>27.6</td>
</tr>
<tr>
<td>Smeerik et al. (2011)</td>
<td>ChIP-seq</td>
<td>p53</td>
<td>U2OS Etoposide</td>
<td>2,920</td>
<td>1,253</td>
<td>832</td>
<td>1,646</td>
<td>56.4</td>
<td>25.3</td>
</tr>
<tr>
<td>Bottchera et al. (2011)</td>
<td>ChIP-seq</td>
<td>p53</td>
<td>IMR90 5-fluorouracil</td>
<td>743</td>
<td>382</td>
<td>1,443</td>
<td>370</td>
<td>49.8</td>
<td>21.5</td>
</tr>
<tr>
<td>Casley et al. (2004)</td>
<td>ChIP-chip</td>
<td>p53</td>
<td>HCT116 bleomycin</td>
<td>14</td>
<td>6</td>
<td>1,819</td>
<td>8</td>
<td>57.1</td>
<td>7.1</td>
</tr>
<tr>
<td>Smeerik et al. (2008)</td>
<td>ChIP-chip</td>
<td>p53</td>
<td>U2OS Actinomycin D</td>
<td>1,946</td>
<td>254</td>
<td>1,571</td>
<td>1,292</td>
<td>83.6</td>
<td>12.4</td>
</tr>
<tr>
<td>Shakled et al. (2008)</td>
<td>ChIP-chip</td>
<td>p53</td>
<td>various genotoxic agents</td>
<td>603</td>
<td>362</td>
<td>1,463</td>
<td>241</td>
<td>40.0</td>
<td>6.6</td>
</tr>
<tr>
<td>Yang et al. (2006)</td>
<td>ChIP-chip</td>
<td>p53</td>
<td>ME180 (-) Actinomycin D</td>
<td>5,794</td>
<td>289</td>
<td>1,536</td>
<td>5,517</td>
<td>95.2</td>
<td>7.2</td>
</tr>
<tr>
<td>Yang et al. (2006)</td>
<td>ChIP-chip</td>
<td>p53</td>
<td>ME180 (+) Actinomycin D</td>
<td>3,677</td>
<td>285</td>
<td>1,540</td>
<td>3,403</td>
<td>92.5</td>
<td>7.0</td>
</tr>
</tbody>
</table>
Figure S3-1. Genomic binding properties of p53 (related to Figure 3-1).

(A) Screen shot of two genomic spans that contain CDKN1A (Cyclin-dependent kinase inhibitor 1A, known as p21/WAF1/Cip1 gene) and GADD45A (Growth arrest and DNA-damage-inducible, alpha). p53, TFIIB, and Pol II occupancy levels are shown 0 an 6 hours after UV treatment. The 5' end of all sequencing tags were shifted in the 3' direction by 5 bp for p53, by 20 bp for TFIIB, and Pol II, then tags on both strands were combined, binned in 50 bp intervals, and smoothed using a 3-bin moving average each track. Each p53-bound location was zoomed in and shown as the sequencing read number per each genomic coordinate, separated by each strand, which were 5-bp binned and smoothed using a 3-bin moving average here. Three p53/RE-bound locations are shown around CDKN1A and one p53-bound location in GADD45A. Cognate 20-bp RE sequences are shown.

(B) Comparison of genome-wide identification of p53-binding locations by ChIP-exo with previous studies—ChIP-PET (Wei et al., 2006), ChIP-seq (Botcheva et al., 2011; Smeenk et al., 2011), ChIP-chip (Cawley et al., 2004; Shaked et al., 2008; Smeenk et al., 2008), and ChIP-chip for p63 (Yang et al., 2006). The left set of panels are from Figure 3-1A. All binding intervals (i.e. in 1,825 rows displayed) identified by the other studies were aligned by the motif midpoint identified by ChIP-exo. Note the x-axis scale is quite distinct for ChIP-exo compared to the others. For the ChIP-chip study of Smeenk et al. (2008), the p53 binding sites identified from their binding intervals were shown.

(C) Summary comparison table indicating the statistics of the accuracy in measurement of p53 binding locations in the nine studies. Metrics are the total number in each dataset (N), true positives (Frazer et al.), false negatives (FN), and false positives (FP) in relation to the set of p53-bound locations identified by ChIP-exo. A match was counted when any published interval was within 100 bp centered at the 1,825 p53-bound locations identified in this study. The percentage of each dataset (N) which contain a p53 binding site within 30 bp centered at each interval was shown (as the midpoint-to-midpoint) in the last column. The p53 binding sites were identified by FIMO (p-value < 10^-3 cutoff) with the PSSM of unsplit p53 full-site which was obtained by our MEME motif discovery analysis in this study. A similar result was obtained with the PSSM
trained with the other study of Wei et al. 2006.

(D) Composite distribution of the location of fine-grain peaks (called separately in both strands in blue and red by GeneTrack using the parameter $\sigma=2$ and $D=5$), determined for Group 2 and 3 binding locations as described for Group 1 in Figure 3-1C. For Group 2 ($n=146$), the frequency of peaks was plotted as a function of their distance (bp) from the end of the stronger RE half-site (oriented such that the weaker half-site is to the right). For Group 3 ($n=107$), the x-axis reference point is the peak pair midpoint.
Figure S3-2. Base composition at each position in an RE quarter site (RRRCW) of Group 1 (related to Figure 3-2).

(A-D). Panels are the same as in Figure 3-2C, except separated out by the four quarter sites indicated above the panels (such as A, B, C, and D). Four arrows represent p53 RE consisting of two half sites, each of which has two quarter sites in a head-to-head orientation. The half site of higher site quality (based on the rules of Menendez et al. for transactivation potential) among two in a full site was assigned to positions AB.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Genomic Location</th>
<th>Description</th>
<th>Transcriptional Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP11</td>
<td>chr11</td>
<td>47,345,000 - 47,350,000</td>
<td>&lt;&lt; SP1 (spleen focus forming virus proviral integration oncogene sp1)</td>
<td>Pol II TFIIB</td>
</tr>
<tr>
<td>Oct6</td>
<td>chr6</td>
<td>31,245,000 - 31,250,000</td>
<td>&lt;&lt; POU5F1 (POU class 5 homeobox 1)</td>
<td>Pol II TFIIB</td>
</tr>
<tr>
<td>POU2F1</td>
<td>chr1</td>
<td>165,455,000 - 165,480,000</td>
<td>&gt;&gt; POU2F1 (POU class 2 homeobox 1)</td>
<td>Pol II TFIIB</td>
</tr>
<tr>
<td>KLF4</td>
<td>chr9</td>
<td>109,280,000 - 109,290,000</td>
<td>&lt;&lt; KLF4 (Kruppel-like factor 4)</td>
<td>Pol II TFIIB</td>
</tr>
<tr>
<td>KLF5</td>
<td>chr13</td>
<td>72,530,000 - 72,540,000</td>
<td>&gt;&gt; KLF5 (Kruppel-like factor 5)</td>
<td>Pol II TFIIB</td>
</tr>
</tbody>
</table>

**Transcriptional Profile:**
- Pol II TFIIB indicates transcriptional activity over time.
Figure S3-3. Non-cognate motifs within the p53 data, and motifs that co-occur near p53/REs.
Crosslinking at non-cognate sites. Genomic regions encompassing Group 3 and Group 2 p53 locations were subjected to motif discovery using MEME. The top four significantly enriched motifs are shown—motif 1 and 2 discovered in Group 3 and motif 3 and 4 from Group 2, along with their e-values. The binding sites of each motif were determined in the interval from peak to peak for Group 3, and a 100-bp interval for Group 2, using FIMO with the 10-4 p-value cutoff. Sequences are shown as four-color plots, and tag 5’ ends are shown (blue on same strand as motif sequence, red on opposite).

Identification of motifs enriched around p53 using the de novo motif discovery algorithm MEME. 50 bp DNA sequences were retrieved from the left and right border of 1,572 REs of Group 1, and subject to MEME, in which the top 30 motifs were analyzed further. Motifs with similarity to a p53 RE were removed. The top 10 motifs having more than 45 sites are shown along with their e-values.

“Unknown 1” is a highly enriched motif near p53/RE that was not linked to an known factor via TomTom query of the JASPAR database. All sequences around p53/RE were analyzed for matches to Unknown 1 PSSM using FIMO with 10-4 p-value cutoff. The distribution of p53/REs around those sites is plotted using a Gaussian kernel and a smoothing bandwidth of 20.

ChIP-exo mapping of TFIIB and Pol II in U2OS cells, around the promoter regions of genes whose product recognizes motifs that co-occur with p53/REs. Each track of 0h and 6h displays the normalized tag count per 50-bp genomic bin, smoothed by a 3-bin moving average.

Distribution of repetitive elements around the nine motifs that co-occur with p53/REs. Motif logos (and number of occurrences) are shown to the left. The middle panel displays four-color DNA sequence plots ±0.5 kb from each instance of the motif, and sorted by distance from p53/RE (shaded colors). Sequences are oriented according to the logo. The right panel covers the same region but color coded according to the type of repetitive element or low complexity DNA present. Their percentage distribution in these regions and throughout the genome is also shown as pie charts. The number of repetitive element classes (such as SINE, LINE, LTR) were counted if they overlapped with the motif site and p53/RE. The majority of the enriched LTRs belongs to the ERV1 (endogenous retroviral sequence 1) family.
Figure S3-4. Cluster analysis of fold changes in Pol II occupancy following p53 induction in response to UV treatment (related to Figure 3-5).

(A) Distribution of fold changes in occupancy of TFIIB and Pol II at 1,313 TFIIB-bound locations in response to UV treatment. Gray-filled plots reflect the distribution of fold changes in
occupancy of TFIIB and Pol II, whereas the black trace shows the distribution for those genes whose TFIIB binding is <15 kb from a p53-bound location. The occupancy level of TFIIB and Pol II was measured as the number of tags within 100 bp of a TFIIB bound location. The density of fold changes in occupancy was estimated by Gaussian kernel function with a 0.4 smoothing bandwidth.

(B) Cluster plot of 751 genes that have a p53/RE – TFIIB/TSS relationship (see Methods) were grouped by log2 fold changes in promoter-proximal and gene body Pol II occupancy upon UV induction. Group designations are D = down (<-0.58 log2 change), N = no change, U = up (>0.58). Rows were sorted by Pol II occupancy changes in the gene body. p53/RE – TFIIB/TSS distances are indicated using heat map colors (red indicates a short distance). Literature reports (Riley et al., 2008) of up- (red) and down-regulated (green) genes are also indicated.

(C) Composite distribution of ChIP-exo sequencing tags of Pol II around the TFIIB/TSS locations for the seven gene clusters in panel B. The frequency distribution of distances between the tags and TFIIB/TSS locations is shown. Data were binned in 10 bp intervals and plots smoothed using a 3-bin moving average. Gray filled plots represent 0 hour, and black traces represent 6 hr after UV.
Cell death and survival network

Carbohydrate network
Supplementary Tables

**Table S3-1** (related to Figure 3-1). 1,825 locations identified by p53 ChIP-exo in U2OS cells treated with UV irradiation. Corresponding RE sequences, insertion size, and tag counts are reported.

Online article supplementary material
Thesis-Chang, Gue Su-Tables.xls
Table S3-2 (related to Figure 3-5). HAVANA-annotated genes containing nearby p53/RE and TFIIB/TSSs in UV-treated U2OS cells. Gene names, TFIIB/TSS locations and occupancy, p53/RE distance, and Pol II promoter and gene-body occupancy are reported.

Online article supplementary material
Thesis-Chang, Gue Su-Tables.xls

Table S3-3 (related to Figure 3-5). Set of 154 high confidence genes having p53/RE <15 kb away and UV-induced Pol II occupancy. The subset of genes from Table S3-2 having a p53/RE <15 kb from a TFIIB/TSS and a log2 fold change in Pol II occupancy >0 in its gene body are reported. Additional information includes TFIIB/TSS location, gene-body Pol II occupancy change, and p53 locations and occupancy changes.

Online article supplementary material
Thesis-Chang, Gue Su-Tables.xls

Table S3-4 (related to Figure 3-5). A set of 100 lower confidence genes having p53/RE 15-1000 kb away and UV-induced Pol II occupancy. Genes were retrieved that had all of the following conditions: 1) contains a >2 fold increase in UV-induced p53 occupancy, 2) contained least 32 p53 tags, 3) p53/RE located between 15-1000 kb from a TFIIB/TSS location, 4) log2 fold change in gene-body Pol II occupancy >0 and among the top 100 (out of 585 meeting the criteria thus far), and 5) not present in Table S3-3.

Online article supplementary material
Thesis-Chang, Gue Su-Tables.xls
Supplementary References


VITA (Gue Su Chang)

EDUCATION

**Doctor of Philosophy**, Integrative Biosciences in Bioinformatics and Genomics Option (2013)

The Pennsylvania State University, University Park, PA, USA

Dissertation: Integrated Genomic Approach to Uncovering Gene Regulation and Epigenetics Across Eukaryotes

**Master of Science**, Biotechnology (1998)

Yonsei University, Seoul, South Korea

Dissertation: Gas Chromatographic Profiling and Screening for Biogenic Polyamines from Complex Samples

**Bachelor of Science**, Biotechnology (1996)

Yonsei University, Seoul, South Korea

PUBLICATIONS


SELECTED PRESENTATIONS & PROCEEDINGS

Chang GS. Gene and nucleosome organization in the *Dictyostelium* genome. Invited lecture at The 2nd Bioinformatics and Genomics Retreat. Sep 16-17, 2011. The Pennsylvania State University, University Park, USA.

