TRANSCRIPTIONAL REGULATION OF ERYTHROPOIESIS VIA
HIGH RESOLUTION CHIP-EXO

A Dissertation in
Integrative Biosciences

by

Garam Han

© 2015 Garam Han

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2015
The dissertation of Garam Han was reviewed and approved* by the following:

B. Franklin Pugh  
Evan Pugh Professor  
Willamen 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

Anton Nekrutenko  
Associate Professor of Biochemistry and Molecular Biology

Debashis Ghosh  
Colorado School of Public Health  
Chair of Biostatistics and Informatics

Peter Hudson  
Director, Huck Institutes of the Life Sciences  
Willaman Professor of Biology

*Signatures are on file in the Graduate School
ABSTRACT

Regulation of gene expression is important for cell differentiation and growth in eukaryotic cells. Precise and complex temporal and spatial regulation enables cells to express various transcriptomes, thus function differently despite having identical genomes. Numerous proteins interact with DNA in the genome or other proteins to regulate this process. Thus, determining the precise locations of protein-DNA interaction is important to understand how they regulate gene expression. Erythropoiesis, differentiation of multiprogenitor cells to red blood cells, serves as an excellent system to study the process to cellular differentiation, which requires precise and intricate transcriptional regulation. GATA1 and TAL1 are two key erythroid transcription factors (TFs) that form a multicomponent complex to regulate this process. However, the precise positional organization of these TFs in the genome remains unclear. Here we applied high resolution ChIP-exo to GATA1 and TAL1 to study their positional organization and determinants of their binding during GATA1-dependent development in mouse. Two complementary methods, MultiGPS and peak pairing, were implemented to determine confident binding locations. While ChIP-exo and ChIP-seq showed substantial overlap in their binding locations, high resolution of binding reveals the homotypic clusters of GATA1 and TAL1 binding locations and the precise positional organization within the protein complex, which cannot be detected using ChIP-seq. We show that the presence of TG (half E-box) 7~8 bp upstream of WGATAA is a strong determinant for co-occupancy of GATA1 and TAL1. Furthermore, we confirm differential GATA1 binding along the time points of erythroid differentiation. This differential recruitment of GATA1 influences gene expression positively when these binding events are nearer to the transcription start sites.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ vi

LIST OF TABLES ........................................................................................................... xi

ACKNOWLEDGEMENTS ............................................................................................... xii

Chapter 1  Introduction ................................................................................................. 1

  Eukaryotic Gene Regulation and Transcription Factor Binding .............................. 1
    Eukaryotic gene regulation ...................................................................................... 1
    High-throughput sequencing in genomics & ENCODE ........................................ 5
    Characterization of protein-DNA interaction ....................................................... 6
    ChIP-Exo .................................................................................................................. 9

  Transcriptional regulation during erythropoiesis .................................................. 12
    Key Transcriptional factors involved in Erythropoiesis: GATA1 and TAL1 ........ 15
    GATA1-dependent erythroid development ........................................................ 18

  Summary of transcriptional regulation during erythropoiesis via ChIP-exo ............ 20

Chapter 2  ChIP-exo Analysis Pipeline ........................................................................ 22

  Alignment of reads to reference genome ............................................................... 23
  Reproducibility assessment ................................................................................... 25
  Identification of binding locations in multiple conditions ..................................... 27
    Union of peak pairs ............................................................................................... 27
    MultiGPS ............................................................................................................... 31
    Intersection of MultiGPS and peak pairing ......................................................... 32
  Bioinformatics analysis for mammalian ChIP-exo ................................................. 34

Chapter 3  Genome-wide organization of GATA1 and TAL1 determined at high
resolution ..................................................................................................................... 37

  Abstract .................................................................................................................... 38
    Background ............................................................................................................ 38
    Results ................................................................................................................... 38
    Conclusions .......................................................................................................... 38

  Background ............................................................................................................. 39

  Results .................................................................................................................... 42
    GATA1 and TAL1 locations determined by ChIP-exo ........................................... 42
    A structural model for the genome-wide co-binding of TAL1 and GATA1 ............ 48
    TG is enriched 7-8 bp upstream of WGATAA at GATA1 and TAL1 co-
    occupied locations ............................................................................................... 54
    Structural relationship between points of TAL1/GATA1 crosslinking and DNA
    sequence ................................................................................................................ 58

  ChIP-exo versus ChIP-seq in location detection ...................................................... 59
  Dynamics of GATA1 and TAL1 binding during erythroid development ................ 63
  Conclusions ............................................................................................................. 66
LIST OF FIGURES

Figure 1-1: Classes and features of cis-regulatory modules (CRMs). Box 1 from (Hardison and Taylor 2012).................................................................................................. 4

Figure 1-2 Experimental methods used for mapping of functional elements in ENCODE project Figure 1 of (Consortium 2011). Image credit: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI). ........................................................................... 6

Figure 1-3: Comparison of protein-DNA interaction assays: ChIP-seq of DNA binding protein and histone modifications, DNase-seq. Figure from (Furey 2012)............... 7

Figure 1-4: Illustration of DNase hypersensitive sites (DHS) footprinting.......................... 8

Figure 1-5: A) Aggregated raw Reb1 signal distribution around all 791 instances of TTACCCG in the yeast genome. The ChIP-seq and ChIP-exo datasets included 2,938,677, and 2,920,571 uniquely aligned tags, respectively. Figure from (Rhee and Pugh 2011), B) Scheme for ChIP-exo (Rhee and Pugh 2012)......................................................... 10

Figure 1-6: The hierarchy of hematopoietic cells (Larsson and Karlsson 2005). LT-HSC, long-term repopulating HSC; ST-HSC, short-term repopulating HSC; MPP, multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP, granulocyte-macrophage progenitor........................................... 13

Figure 1-7: Stages of erythropoiesis (courtesy of Ross Hardison), along GATA1-dependent erythroid development using G1E and G1E-ER4 system......................... 14

Figure 1-8: Chromatin landscape and involvement of transcription factors during erythroid development (figure courtesy of Ross Hardison), including GATA1 and TAL1 complex.................................................................................................. 15

Figure 2-1: Analysis pipeline workflow (Peak pairing and MultiGPS)................................. 23

Figure 2-2: Methods to assess reproducibility among biological replicates...................... 25

Figure 2-3: GeneTrack parameters (sigma, exclusion zone)................................................ 29

Figure 2-4: Overlap of binding locations from two methods, Peak pairing and MultiGPS... 33

Figure 2-5: Example browser shot of binding locations detected by one method (Peak pairing or MultiGPS)................................................................................. 34

Figure 2-6: Parallel Computing .......................................................................................... 36

Figure 3-1. Browser shot of GATA1 ChIP-seq and ChIP-exo tags and binding locations. Browser shot of ChIP-exo tag 5’ ends for GATA1 measured by ChIP-seq (CS) and ChIP-exo (XO) on chromosome 1 from coordinate 36,977,009 to 36,977,009. Tag locations were smoothed (20 bp moving average). Tag density on the lower
(negative) strand is shown as an inverted plot. Locations of WGATAA and E-box motifs (p-value < 10^{-3}) and RefSeq genes are shown below through IGV browser.  

Figure 3-2. Overlap of GATA1 union peak pairs and MultiGPS binding locations. Venn overlap of GATA1 binding locations determined by the union of peak pairs and MultiGPS. MultiGPS binding locations positioned within 100 bp window from the union of peak pairs were determined as final binding locations and percentages (n=10,290). The most enriched MEME motif and their occurrence frequencies from intersection and outersects of two methods are shown below.

Figure 3-3. GATA1 and TAL1 binding locations in clusters and non-clusters. Percentage of GATA1 and TAL1 binding locations in homotypic clusters from the total binding locations. A cluster is defined as adjacent binding locations that are <500 bp apart. In case of GATA1, 13.7% (1,414) binding locations form 663 clusters, while 64.4% (9,796) TAL1 binding locations form 2,905 clusters.

Figure 3-4. Expression change of GATA1 and TAL1 clusters and non-clusters. Expression change of the closest gene to a homotypic cluster or non-cluster for GATA1 and TAL1. Expression change is calculated as log_{2} fold change of gene expression between G1E-ER4 induction of 30 hour relative to 0 hour.

Figure 3-5. Genome-wide structural organization of GATA1/TAL1 complexes. Venn diagram showing the overlap of TAL1 and/or GATA1 binding locations further classified to TAL1 only, GATA1/TAL1 co-occupied, and GATA1 only binding locations. Co-occupancy was defined as having their midpoints within <40 bp distance apart.

Figure 3-6. Genome-wide structural organization of GATA1/TAL1 complexes. Composite distribution of TAL1 (black trace) and GATA1 (red filled trace) ChIP-exo tag 5' ends around the WGATAA motif (the underlined “A” is set to zero) at GATA1 binding locations. Tags that map to the reverse strand are displayed as inverted traces. Data were plotted as moving average of 5 bp. Above the plot are X-ray structure images of TAL1 and GATA1 bound to cognate DNA sites (PDB accession code 3VEK for GATA1 (Matthews, J.M., unpublished), and 2YPB for TAL1 [17]). The “X” denotes the deduced site of crosslinking of GATA1 (red) and TAL1 (gold).

Figure 3-7. Distribution of 5' ends of GATA1, TAL1 ChIP-exo sequencing tags around 7,927 WGATAA sites in GATA1 binding locations, comprised of 2,997 GATA1-enriched WGATAA sites (rows) having TAL1, and 4,930 sites lacking TAL1. Both GATA1 and TAL1 co-bound WGATAA sites (Top), and GATA1-only bound WGATAA sites (Bottom) were sorted by GATA1 occupancy in merged data of all time points. Occupancy is calculated by the sum of total tags around motif reference point (WGATAA) from 25 bp upstream to 25 bp downstream, and 40 bp upstream to 30 bp downstream for GATA1 and TAL1, respectively. A four colored plot (Left) reports the nucleotide composition of 50-bp surrounding regions of WGATAA motif: A (Red), C (Blue), G (Gold), T (Green).
Figure 3-8. Distance of TG (half E-box) to WGATAA sites in GATA1, TAL1 co-bound locations. Frequency histogram of distances of TG relative to WGATAA sites in GATA1 binding sites with high TAL1 occupancy (Left) and in GATA1 binding sites with no/low TAL1 occupancy (Right).......................... 56

Figure 3-9. Pie chart for nucleotide composition of E-box portion of composite motif [TG(N7-8)WGATAA]. When denoting the E-box motif as ZZNNTG, binding locations were first classified by their proportion of ZZ (Left). Then the most frequent CANNTG locations were further classified by their NN content (Right)........ 57

Figure 3-10. Venn Diagram overlap of ChIP-exo (XO) and ChIP-seq (CS) binding locations for GATA1 (G1E-ER4 0 h, 3 h, 24 h) and TAL1 (G1E, G1E-ER4 24 h). Upper numbers correspond to the number of ChIP-seq binding intervals that overlapped with at least one ChIP-exo peak-pair location (window size = 80 bp), while the number below indicates the number of ChIP-exo peak pair locations overlapping with ChIP-seq intervals................................................................. 60

Figure 3-11. Scatter plot with Spearman’s correlation coefficient of log10 transformed ChIP-exo and ChIP-seq tag counts (window size = 400bp) around all binding locations of ChIP-exo and ChIP-seq at GATA1 G1E-ER4 3 h. Red to blue color indicates the high to low density of data points in the scatter plot. Similar results were obtained for the other time point datasets (not shown)......................................................... 61

Figure 3-12. Composite distribution of G1E-ER4 3 h ChIP-exo reads (smoothing = 20 bp) around occupied segments detected only by ChIP-seq (n= 14,071). Similar results were obtained for other time points........................................ 63

Figure 3-13. Heatmap of GATA1, TAL1 total tag count normalized tags (strand merged) plotted around GATA1 binding locations which were clustered by differential occupancy of GATA1 between time points (left). Differential occupancies between time points (G1E-ER4 3h/0h, 24h/0h, and 24h/3h) were determined as log2 fold change of ratio of GATA1 tags over input control between time points using EdgeR and MultiGPS. For each binding location, a heatmap of Log2 expression fold change of the closest gene between G1E-ER4 3h/0h is shown (right) where increased, unchanged, decreased expression are depicted as yellow, black, and blue, respectively. .......................................................... 64

Figure 3-14. Box plot of expression change (G1E-ER4 30h/0h) (left), and distance to the closest transcription start site (right)............................................................................. 65

Supplementary Figure 3-1. Workflow of ChIP-exo data analysis. Confident binding locations were determined by obtaining the intersection of union of significant peak pairs from multiple time points (Left), and binding events of MultiGPS (Right). Motif discovery analysis using MEME and MEME-ChIP showed enrichment of WGATAA, E-box, and composite half E-box/WGATAA motifs on these stringent binding locations validated by both methods. To classify binding locations by presence of motifs, further motif occurrences and downstream analyses were conducted........................................... 74
Supplementary Figure 3-2. Binding locations detected by only one method, only peak pairing or MultiGPS. Browser shot of 5’-end of TAL1 ChIP-exo tags surrounding the binding locations detected by only peak pairing (Top left panel) or MultiGPS (Top Right panel). Binding locations called by union of peak pairing and MultiGPS are shown in the middle panel, while binding locations in intersec (Blue) and outersec (Red) of two methods are highlighted. WGATAA and E-box motifs are shown in black. ................................................................. 75

Supplementary Figure 3-3. TAL1 and GATA1 co-binding around WGATAA sites .......... 76

Distribution of 5’ ends of GATA1, TAL1 ChIP-exo sequencing tags around 7927 WGATAA sites in GATA1 binding locations. Rows are linked across all panels and sorted by TAL1 occupancy of merged data of all time points. TAL1 occupancy is calculated by the sum of tags from 40 bp upstream and 30 bp downstream from motif reference point (WGATAA). The far-left panel reports the nucleotide composition of 25 bp upstream and downstream regions from GATA motif reference point. ........................................................................ 76

Supplementary Figure 3-4. Tag distribution of GATA1 and TAL1 sorted by occupancy of individual peak of 4 peak patterns. GATA1 and TAL1 tag distribution around WGATAA sites sorted by the occupancy of individual peak of GATA1 four-peak locations. Tags in the same strand of motif (Blue) and opposite strand (Red) and sequence composition plot were sorted by occupancy of sense peak 1 (defined as 0 to 25 bp upstream of WGATAA reference point, Top left), sense peak 2 (0 to 25 bp downstream, Top right), antisense peak 1 (25 bp upstream and 5 bp downstream, Bottom left), and antisense peak 2 (5 to 25 bp downstream, Bottom right). ............... 78

Supplementary Figure 3-5. Sequence composition around crosslinking sites when sorted by GATA1 crosslinking level in individual peak of 4 peak patterns. Relationship of nucleotide composition in-18 bp to +3p from WGATAA reference point and GATA1 crosslinking level in individual peak of GATA1. For each position, the nucleotide compositions in low to high (Left to right) crosslinking of individual peak are shown through 4 adjacent bars................................................................. 79

Figure 4-1: Summed tag distribution around GATA1 motif. Multiple peak patterns are observed around motif midpoints. ........................................................................................................ 82

Figure 4-2: (A) ChIP-exo tags around Reb1 binding sites centered by the motif midpoint. Figure 2A from (Rhee and Pugh 2011). (B) ChIP-exo sequencing tags around CTCF-bound locations centered by the motif midpoint. Figure 6B from (Rhee and Pugh 2011). (C) Model for presence of exonuclease blockage sites (two peak pairs) for CTCT binding locations. Figure 6C from (Rhee and Pugh 2011). (D) Composite tag distribution around p53-bound regulatory elements (RE). Figure 1B from (Chang, Chen et al. 2014) ........................................................................................................ 83

Figure 4-3. Diagram explaining the multiple crosslinking and exonuclease digestion from GATA1 pattern ........................................................................................................ 84

Figure 4-4. Diagram illustrating GATA1 peak patterns through multiple peak patterns....... 85
Figure 4-5. Goal of examining multiple peak pattern of ChIP-exo. ChIP-exo provides higher precision compared to ChIP-seq. More comprehensive list of binding locations may be detected using multiple peak pattern compared to conventional peak pairing.

Figure 4-6. Workflow of determining binding locations using multiple peak pattern.

Figure 4-7. (Left) Quad “template” is determined from the summed tag distribution around the motif. (Right) Quad template is overlaid on the reference in two possible ways. Sp= peak on the same strand of reference peak, op= peak on the opposite strand of reference peak.

Figure 4-8. Overlap of ChIP-seq peaks with ChIP-exo peaks from quad method.

Figure 4-9: Possible peak patterns from two crosslinking sites of protein binding.

Figure 4-10: Algorithm of Peak Pair cluster (PPC) method. Step 1 - Solid red and blue peaks indicate the peaks that were paired after C-W peak pair distance calculation. Dashed C peaks indicate the possible peaks that were not chosen as peak pairs. Step 2 – The distance between crosslinking sites is calculated using each as a reference crosslinking site (green circle in bold). In this diagram, the arrow was calculated as the mode of doublet pair distance. Final Outcome – Gray box indicates the region with the paired crosslinking points that forms a four-peak pattern (PP2). The final reference coordinate of the binding location is indicated as the inverted black triangle.

Figure 4-11: Tag distribution around GATA1 four peak (PP2) and three peak (PPO) reference points. Motif indicates the MEME logo from all peak pair clusters, including all possible 2~4 peaks. Also, depending on the orphan peak location, there are 4 possible PPO configurations. Therefore, when PPO are combined, the peak pattern seems like a four-peak pattern, when it is a combination of various orphan peak orientation.

Figure 4-12: Tag distribution of GATA1 and TAL1 sorted by strength of individual peak of four peak patterns.

Figure 4-13: Sequence composition around crosslinking sites when ordered by occupancy level of GATA1 in individual peak among four peaks. (Left) Sorted by occupancy of sense 1 peak, (Right) sorted by occupancy of sense 2 peak.

Figure 5-1: Integration of multi-omic data: Multiple types of –omic data allows deeper understanding of biological mechanisms, such as transcriptional regulation. Study in Chapter 3 incorporated various types of data (italicized in figure), including ChIP-exo, RNA-seq, histone modification, and DNase-seq (generated by ENCODE).
LIST OF TABLES

Table 1-1: List of key histone modification marks and variants. Table 2 from (Consortium 2012)........................................................................................................................................ 3

Table 2-1: List of various short-read aligners (Furey 2012).................................................................................. 24

Table 3-1. Sequencing statistics. Sequencing statistics including total reads, uniquely mapped read and percentage yield of all biological replicates and time points of GATA1 and TAL1........................................................................................................................................ 69
ACKNOWLEDGEMENTS

Foremost, I would like to thank my thesis advisor, Dr. Frank Pugh, for his guidance and patience during my development as a graduate student. His rigorous pursuit and passion for science has been a great inspiration, so that I can follow after him to become an independent scientist. I express my thanks to my committee members, Dr. Ross Hardison, Dr. Anton Nekrutenko, and Dr. Debashis Ghosh, for supporting me throughout my graduate studies with suggestions and advice on career development. I would like to thank the collaborators in Hardison lab, for their expertise and support for the cells that went into my project, and thank Ross for the opportunity to work on this collaboration project. His kindness toward students will be an inspiration as I aspire to mentor and teach students in the future.

Many thanks to members of Pugh Lab, past and present, for many scientific discussions that strengthened my studies. Special thanks to soon-to-be-doctor Rohit Reja and Dr. Vinesh Vinayachandran for not only their scientific feedback, but also for their support and friendship that made every day in the lab more enjoyable. I also thank Dr. Cizhong Jiang, who helped me to write my very first python code. I also thank Dr. Shaun Mahony and Akshay Kakumanu for their help with MultiGPS and scientific suggestions. I express my thanks to the members of Center for Eukaryotic Gene Regulation (CEGR), especially Jason Miller for his friendship throughout many years, and Dr. Robert McGinty for his openness and his helpful and timely advice.

I also thank the support from the Bioinformatics and Genomics (BG) program, especially to Shashi (Dr. Cooduvalli Shashikant) for regularly meeting up with students and providing guidance. I also thank all the people who gave me the opportunity to start my exciting journey in Bioinformatics, including Dr. John Carlson, Masume Assaf, Dr. Mike Radis and Janice Kennedy. Many thanks to my BG buddies, Dr. Jihye Park (9 year buddy at Penn State), Dr. Christopher Morrissey, Dr. Sushant Kumar, Dr. Sridhar Ranganathan, and soon-to-be-doctors Nathaniel
Cannon and Qingyu Wang. Many thanks to the support from all the staffs in the Huck Institute Office, and the Korean bioinformatics community, especially Dr. Chungoo Park, Dr. Dokyoon Kim, Dr. Bongsoo Park and Dr. Donghwan Shim. Also, special thanks to Dr. Seong Jo Kim who has taught me so much about computer programming, especially optimizing algorithms to be more suitable for larger mammalian genomes.

Many thanks to my friends who reached out me throughout many years despite the distance and kept the friendships going: Jenny Jou, Deborah Kim, June Lee, Annabel Hughes, Jenn Mun, Dr. Soyoung Bae, Dr. Matt Moll, Calvin Kim, Minhyung Joo, Susan Oh and the whole Oh family, Daniel Chun, Michael and Myro Joy Lee, Eunhye Kim, Christine Chun, Yuyu Bu, Stephanie Tang, Hyun Park, Dr. Stephen Sier, Rebecca Boon for your prayers and support. I thank Dr. Curtis Swagler, Kevin Gilliam and Dr. Kimberly Trainer for watching out for my physical health and encouraging me throughout my studies. I give thanks to Dr. John Riew for taking the time to teach me about writing. I give special thanks to Harvest Global Mission Church and State College Korean Church, for spiritually supporting me for many years.

I cannot thank enough my family, especially my parents, Dr. Myunghee Yun and Dr. Geunjo Han, my one and only twin sister, Dr. Boram Han, and my brother-in-law, Dr. Daniel Kim, and my baby nephew Christian Kim, and all my extended families. Growing up while seeing great examples like my parents was a great privilege. Their passions for research and teaching, and their sincere care for students have inspired me to follow their footsteps. Although my beloved dad, grandpa and grandma all have passed away during my recent few years of grad school, I am sure they are proud of me for persevering and finishing strong. I dedicate this work to my lovely dad and mom.

Last but not least, I thank God who showed me that there is a greater purpose in life, and that even through every trials and hardships, this gift of life is worth living for.
For I am convinced that neither death nor life, neither angels nor demons,
neither the present nor the future, nor any powers, neither height nor depth, nor anything else in
all creation, will be able to separate us from the love of God that is in Christ Jesus our Lord.

- Romans 8:38, 39
Chapter 1

Introduction

Eukaryotic Gene Regulation and Transcription Factor Binding

Eukaryotic gene regulation

Under the central dogma, hereditary information in DNA is converted to RNA by DNA-dependent RNA polymerase through the process called transcription (Crick 1970). These RNA messages are further read and converted to proteins through translation. Temporal and spatial regulation of gene expression is critical for various cells, organelles and organisms to undergo various physiological conditions and development.

The first step in the regulation of gene expression is transcription. However, rather than occurring on naked DNA, transcription must occur in the context of chromatin. In order to package the eukaryotic genome consisting of billions of base pairs of DNA inside a small nucleus, cells package their genomes in a polymeric complex called chromatin. This packaging mechanism uses highly conserved proteins called histones, which can be further classified as core histones (H3, H4, H2A and H2B) and linker histones (H1 and H5). Genomic DNA is wrapped ~1.65 turns around the histone octamer (two sets of core histones) forming a “beads on a string” nucleosomal array structure (Kornberg and Lorch 1999, Richmond and Davey 2003). These nucleosomes and linker histones form a higher ordered chromatin fiber, which further forms a mitotic chromosome through supercoiling. The structure and mechanisms underlying this high level of compaction are still unclear.
For many years, DNA compaction and higher order chromatin structure have been considered merely as packaging units. It has become clearer during recent years that these structures are critical for gene regulation and for protection of hereditary information from mutations. Chromatin can be classified into highly condensed, gene-poor and less transcribed heterochromatin, and less compacted, gene-rich and highly transcribed euchromatin (Csink and Henikoff 1996).

Cells have developed mechanisms to unwrap DNA from chromatin and making it more accessible for various DNA-templated processes, such as transcription, replication and DNA repair. These mechanisms for unwrapping DNA include chromatin remodeling, histone modifications, DNA methylation and more. Based on the combinatorial effect of histone modifications, binding of chromatin remodelers and transcription factors, transcriptional output is determined.

Histone modifying enzymes, the writers of histone marks, make covalent modifications largely on the N-terminal tails of core histones (Strahl and Allis 2000, Jenuwein and Allis 2001, Millar and Grunstein 2006). These modifications include phosphorylation, acetylation, methylation, SUMOylation and ubiquitination. Histone modifications can be marks of gene activation or repression (Table 1.1). Then readers, chromatin remodelers and transcription factors, interpret these histone marks.
Chromatin remodelers are multisubunit protein complexes that utilize energy of ATP hydrolysis to enable access to nucleosomal DNA, through sliding, looping and histone eviction (Saha, Wittmeyer et al. 2006, Cairns 2007). Chromatin remodelers contain an ATPase subunit that belongs to a SNF2 superfamily of helicase-related proteins. They can be classified into SWI/SNF (mating type SWItching/Sucrose Non-Fermenting), ISWI (Imitation SWItch), INO80 (INOsitol), and CHD (Chromodomain Helicase DNA-binding) families based on the presence of conserved ATPase domains (Clapier and Cairns 2009).

Table 1-1: List of key histone modification marks and variants. Table 2 from (Consortium 2012)

<table>
<thead>
<tr>
<th>Histone modification or variant</th>
<th>Signal characteristics</th>
<th>Putative functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2A.Z</td>
<td>Peak</td>
<td>Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>Peak/region</td>
<td>Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts</td>
</tr>
<tr>
<td>H3K4me2</td>
<td>Peak</td>
<td>Mark of regulatory elements associated with promoters and enhancers</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Peak</td>
<td>Mark of regulatory elements primarily associated with promoters/transcription starts</td>
</tr>
<tr>
<td>H3K9ac</td>
<td>Peak</td>
<td>Mark of active regulatory elements with preference for promoters</td>
</tr>
<tr>
<td>H3K9me1</td>
<td>Region</td>
<td>Preference for the 5’ end of genes</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>Peak/region</td>
<td>Repressive mark associated with constitutive heterochromatin and repetitive elements</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Peak</td>
<td>Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts</td>
</tr>
<tr>
<td>H3K27me3</td>
<td>Region</td>
<td>Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes</td>
</tr>
<tr>
<td>H3K36me3</td>
<td>Region</td>
<td>Elongation mark associated with transcribed portions of genes, with preference for 3’ regions after intron 1</td>
</tr>
<tr>
<td>H3K79me2</td>
<td>Region</td>
<td>Transcription-associated mark, with preference for 5’ end of genes</td>
</tr>
<tr>
<td>H4K20me1</td>
<td>Region</td>
<td>Preference for 5’ end of genes</td>
</tr>
</tbody>
</table>
Transcription requires recruitment of RNA polymerase II (PolII) and general transcription factors (GTFs) to the promoters of genes. Directing PolII to the promoters, mostly in nucleosome free regions (NFR), is the key role of transcription factor binding. The regions where transcription factors bind to regulate gene expression are called cis-regulatory modules (CRMs).

In addition to promoters, transcription factors bind to other types of CRMs, including enhancers and insulators (Figure 1-1) (Hardison and Taylor 2012). Transcription factor binding in the enhancers, mostly located further away from their target genes, can recruit PolII leading to enhancer transcription. Considering the important roles of transcription factors, knowing their precise binding locations will shed more light into understanding their role in gene regulation.

Figure 1-1: Classes and features of cis-regulatory modules (CRMs). Box 1 from (Hardison and
High-throughput sequencing in genomics & ENCODE

The field of genomics has advanced rapidly since the human genome project started in 1990. The finished version of human genome sequence was completed in 2003, and the first Encyclopedia of DNA Elements (ENCODE) paper was published in 2007 (Consortium, Birney et al. 2007). Ever since, the development of high-throughput next generation sequencing (NGS) technology has revolutionized the field of genomics (Koboldt, Steinberg et al. 2013). While previous studies have relied on single genes studies, microscopy and fluorescence studies, NGS increased the speed and scale of genomic discoveries. Owing to NGS, much effort has been invested in mapping the functional elements of multitude of genomes and cell types, leading to development of bioinformatics tools. Many researchers, including the Encyclopedia Of DNA Elements (ENCODE) consortium, helped promote our understanding of elements acting at DNA and RNA levels (Consortium 2012). A variety of NGS techniques have been incorporated into the ENCODE project (Figure 1-3); including RNA-seq for the transcriptional landscape (Djebali, Davis et al. 2012), ChIP-seq for the genomic locations of proteins and histone modifications (Park 2009), and DNase-seq for accessibility of the chromatin landscape (Park 2009, Neph, Vierstra et al. 2012). Similar effort has also been made in other organisms, including mouse (Mouse ENCODE) (Mouse, Stamatoyannopoulos et al. 2012) and model organisms (modENCODE) (mod, Roy et al. 2010), and cancer genomes (The Cancer Genome Atlas, also known as, TCGA) (Brennan, Verhaak et al. 2013).
Characterization of protein-DNA interaction

To understand the gene regulation through binding of transcription factors, obtaining a high resolution of protein-DNA binding locations is crucial. Among the various methods to examine transcription factor binding locations, chromatin immunoprecipitation (ChIP)-based assays and open chromatin assays are most commonly used (Figure 1-4).

Figure 1-2 Experimental methods used for mapping of functional elements in ENCODE project Figure 1 of (Consortium 2011). Image credit: Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI).
Chromatin immunoprecipitation-based assays detect genome-wide binding locations of a protein of interest. Chromatin immunoprecipitation (ChIP) followed by microarray hybridization (ChIP-chip) (Ren, Robert et al. 2000) or high-throughput sequencing (ChIP-seq) (Margulies, Egholm et al. 2005, Barski, Cuddapah et al. 2007, Johnson, Mortazavi et al. 2007) are widely used to examine the interaction of transcription factors with DNA or genomic locations of chemical modification of histone proteins. During ChIP assays, proteins are crosslinked to the DNA, then immunopurified after chromatin was fragmented through sonication. ChIP-chip and ChIP-seq methods are limited by their low resolution and high background. While transcription factor footprints on DNA are 6-20 bp (Furey 2012), the resolutions of these methods are limited.
to ~200-300 bp long fragmented chromatin size. Libraries are mixed with DNA fragments without proteins of interest bound, which increases the background and makes detection of binding locations challenging.

Open chromatin assays are another common method to characterize transcription factor binding locations. DNase hypersensitive sites (DHS) are regions that are sensitive to the degradation by the DNase I nuclease. These correspond to the nucleosome-depleted open chromatin regions where the exposed DNA is prone for any transcription factors to bind, and regions where genes are transcriptionally active. Within a DHS, there are smaller DNase I protected regions, often referred as DNase I footprints (Figure 1-5). Motif analysis on individual DNase I footprint is conducted to discover the short DNA sequence recognized by DNA-binding domain of transcription factors (Hardison and Taylor 2012) and to infer the specific transcription factors that maybe bound.

Figure 1-4: Illustration of DNase hypersensitive sites (DHS) footprinting
However, studies have shown that intrinsic DNase I cleavage bias (He, Meyer et al. 2014) and sequence bias (Koohy, Down et al. 2013) maybe driving the high-resolution cleavage pattern, rather than the binding of transcription factor to the DNA.

Although these technologies allowed advancement in our knowledge of protein binding while providing great resources to the scientific community, they are limited due to the low resolution of ChIP-seq and the bias from sequence specificity of DNase I enzyme in DNase-seq (Koohy, Down et al. 2013). With these limitations in mind, ChIP-exo has been developed (Rhee and Pugh 2011).

**ChIP-Exo**

ChIP-exo allows genome-wide mapping of the binding locations of proteins that interact with DNA with high resolution (Rhee and Pugh 2011). The strength of ChIP-exo stems from incorporating 5’ to 3’ directional lambda (λ) exonuclease, which allows the DNA to be digested to the borders of protein-DNA crosslinking points. When sequences are mapped to the reference genome, the enrichment of reads on the positive and negative strands demarcate the borders of the protein binding locations. ChIP-exo allows the detection of borders of binding locations, which often span around 20-30bp. Furthermore, the exonuclease removes background reads where no proteins were bound, which increases the signal-to-noise ratio. ChIP-exo experiment procedure involves formaldehyde crosslinking treatment on cells or tissues, fragmentation through sonication, immunoprecipitation with specific antibody for the protein of interest, lambda (λ) exonuclease treatment, then library preparation for sequencing (Rhee and Pugh 2012) (Figure 1-6B).
To obtain high quality results of protein binding locations through chromatin immunoprecipitation-based assays, both ChIP-seq and ChIP-exo requires a few important steps.

Figure 1-5: A) Aggregated raw Reb1 signal distribution around all 791 instances of TTACCCG in the yeast genome. The ChIP-seq and ChIP-exo datasets included 2,938,677, and 2,920,571 uniquely aligned tags, respectively. Figure from (Rhee and Pugh 2011), B) Scheme for ChIP-exo (Rhee and Pugh 2012)
experimental considerations (Furey 2012). First, validation of highly specific antibodies is critical. Chromatin immunoprecipitation-based assays achieve the profiling of binding locations for specific protein of interest through the use of antibodies. Therefore, the quality of the results heavily depends on the specificity and quality of antibodies. Strict validation is required since antibody quality may even vary between different lots of the same antibody (modENCODE, Egelhofer et al, 2011). Secondly, obtaining a sufficient number of cells is important, especially for mammalian cells. Chromatin immunoprecipitation requires a population of cells to identify the binding locations of transcription factors. Typically, for mammalian genomes, such as human or mouse, this requires 10-20 million cells per experiments, which may limit the number experiments on valuable samples. Therefore, careful planning for each experiment is required depending on the size of the genome and the protein of interest. To overcome this limitation, methods requiring smaller numbers of cells have been developed, such as Nano-ChIP-seq (Adli and Bernstein 2011) or the use of single-tube linear DNA amplification (LinDA) (Shankaranarayanan, Mendoza-Parra et al. 2011). For example, LinDA avoids the GC content bias, which is common in PCR-based methods. When PCR amplification are biased toward high AT, this method allows high GC content DNA sequences to be amplified to same extent as other sequences. Lastly, adequate sequencing depth and library complexity is crucial. The required amount of sequencing data varies depending on the size of the genome and protein of interest. Sufficient sequencing depth is desired to allow detection of binding locations including the lowly enriched reads. Examination of library complexity, the amount of unique DNA molecules, is important. Low library complexity from repeated PCR-amplification of same molecule may lead to high false positive rate (Furey 2012). ChIP-exo will perform its best when these points are taken into considerations.

ChIP-exo is a powerful method that can be applied to study DNA-protein interactions involved in numerous biological processes. For my Ph.D. thesis project, I set out to study the
transcriptional regulation in erythropoiesis, the production and differentiation of red blood cells from hematopoietic stem cells. ChIP-exo was applied at various stages of differentiation to study the changes in chromatin structure and gene regulation.

Transcriptional regulation during erythropoiesis

Hematopoiesis, and differentiation to specific blood cell lineages, such as erythropoiesis, is an ideal model to study stem cell development and differentiation. Hematopoiesis is the process by which hematopoietic stem cells (HSC) undergo self-renewal, or differentiation into multiple blood lineages, including erythroid (red blood) cells, lymphocytes (T-cells, B-cells), and myelocytes (granulocytes, megakaryocytes, macrophages) (Figure 1-6). Hematopoietic stem cells (HSCs) develop from pluripotent progenitors, to multipotent progenitor (MPP) cells, to bipotential cells, and finally commit to a unique lineage. Decision making of hematopoietic stem cells to commit to a particular lineage versus the others requires a precise regulation of gene expression patterns and transcription programs, which is a combined effect of chromatin structures and interactions with proteins. Many players regulate this process, such as transcription factors, extracellular signals (i.e. erythropoietin), pausing of RNA polymerase II (PolII), epigenetic changes, and non-coding RNA, such as miRNA.
Among these many pathways in hematopoiesis, the scope of my research focuses on the process that blood stem cells matures and differentiates into red blood cells, also called erythropoiesis. This process involves the hematopoietic stem cells (HSC) to first mature into common myeloid progenitors (CMP), which can become one of the two possible bipotential progenitors, megakaryocyte-erythroid progenitors (MEP) or granulocyte-macrophage progenitor (GMP) cells (Figure 1-7). Then MEP undergoes lineage commitments and progressively matures to BFU-e (burst forming unit erythroid), CFU-e (colony forming unit erythroid), then sequentially undergo stages of erythroblasts (from proerythroblast, basophilic, polychromatic and
orthochromatic erythroblasts), to become enucleated, fully differentiated red blood cells that can deliver oxygen from lungs to every part of the body.

This section highlights the transcriptional regulation during erythropoiesis, especially the role of GATA1 and TAL1 binding, then discusses the G1E and G1E-ER4 system that are used to study in vivo synchronous erythroid development.

Figure 1: Stages of erythropoiesis (courtesy of Ross Hardison), along GATA1-dependent erythroid development using G1E and G1E-ER4 system.
Key Transcriptional factors involved in Erythropoiesis: GATA1 and TAL1

Erythroid development involves many transcription factors including GATA1, TAL1/SCL, LMO2, LDB1, KLF1 and more. The importance of these transcription factors has been shown through emerging of anemia and leukemia when they are misregulated. GATA1 and TAL1 are the key components of the multiprotein complex that bind to DNA, which are essential for in vivo erythroid cell production and differentiation (Suzuki, Moriguchi et al. 2009, Lausen, Pless et al. 2010).

Figure 1-8: Chromatin landscape and involvement of transcription factors during erythroid development (figure courtesy of Ross Hardison), including GATA1 and TAL1 complex.
**GATA1**

GATA1 is a zinc finger transcription factor that binds to the WGATAR motif. It is critical for the production of hemoglobin-containing and mature red blood cells, while its absence leads to inhibition of erythroid development (Pevny, Simon et al. 1991). GATA1 is expressed in erythrocytes, megakaryocytes, mast cells, eosinophilic, and in multipotential progenitor (MPP) cells, but not in other blood cell lineages, or in non-hematopoietic cells (Pevny, Simon et al. 1991), whereas GATA2 is present in early hematopoietic progenitors, mast cells and megakaryocytes (Weiss, Yu et al. 1997). When GATA1 is absent, erythroid cell precursors are arrested at proerythroblast stage of development and undergo apoptosis (Weiss, Yu et al. 1997).

To better understand the role of GATA1 in erythroid regulation, several ChIP-seq studies have been performed on GATA1 (Cheng, Wu et al. 2009, Fujiwara, O'Geen et al. 2009, Yu, Riva et al. 2009, Soler, Andrieu-Soler et al. 2010) and TAL1 (Tripic, Deng et al. 2009, Kassouf, Hughes et al. 2010, Soler, Andrieu-Soler et al. 2010) in erythroid cell lines. However, the low resolution and sensitivity (signal:noise) of the assay placed limits on the conclusions. Indeed the broad binding regions of ChIP-seq, which span over 100bp or even 1kb wide, makes it challenging to decipher the driving motif when numerous motifs are present within the region. It is challenging to decipher the precise distances between GATA1 and TAL1 binding locations and their positional organization at co-occupied locus. There has been disagreements in the number of binding locations, where three studies identified 4,000–6,000 in vivo binding sites for GATA-1 in mouse erythroleukemia (MEL) cells expressing a tagged form of GATA-1 (Yu, Riva et al. 2009, Soler, Andrieu-Soler et al. 2010) or human K562 erythroleukemia cells (Fujiwara, O'Geen et al. 2009), while another study identified >15,000 sites occupied in a GATA1-inducible G1E and G1E-ER4 cell system in mouse (Cheng, Wu et al. 2009). Mouse ENCODE has reported ~12,000 GATA1 and ~2,000-8,000 TAL1 binding locations in G1E and G1E-ER4 cells after 24 hours of
GATA1 induction (Mouse, Stamatoyannopoulos et al. 2012). One challenge with these studies is setting a threshold for binding, where if it is set too high, many binding locations may be missed (false negatives), while if it is set too low, many putative locations may be background (false positives). While differences in number of bound locations may be due to different thresholds, other factors such as peak calling methods and data quality may contribute to the differences. Majority (~85%) of GATA1 was bound in the distal enhancer regulatory regions that are enriched with H3K4me1 (Heintzman, Stuart et al. 2007, Cheng, Wu et al. 2009), while remainders were bound in proximal promoter (Fujiwara, O'Geen et al. 2009, Yu, Riva et al. 2009). Upon GATA1 activation, approximately 5000 genes were differentially expressed. Among the differentially expressed genes, almost equivalent numbers of genes were upregulated and downregulated. The mechanism of upregulation mostly involves co-occupancy of GATA1 with TAL1/SCL (Cheng, Wu et al. 2009, Fujiwara, O'Geen et al. 2009, Tripic, Deng et al. 2009, Yu, Riva et al. 2009, Kassouf, Hughes et al. 2010), while the involved partners and the mechanism of gene repression remain unclear.

**TAL1**

TAL1/SCL is a basic helix-loop-helix (bHLH) factor, expressed in erythroid cells, megakaryocytes and mast cells, that binds to the E-box motif. While general E-box motif is degenerate (CANNTG), the E-box motif for TAL1 has been reported to be CAGVTG (Wu, Morrissey et al. 2014). Noteworthy also, TAL1 is known to form a complex with its heterodimeric partner, ubiquitous bHLH protein E2A (Cantor, Katz et al. 2002). The multiprotein complexes are connected through the interactions with LIM domain containing cofactors, LMO2 and LDB1 (Wadman, Osada et al. 1997). TAL1 forms a multiprotein complex with GATA1
binding to a WGATAR/E-box composite motif spaced 9~11 bp apart (Wadman, Osada et al. 1997). Previous gel electrophoresis mobility shift assays (EMSA) have provided evidence for a pentameric complex, composed of GATA1, TAL1/SCL, LMO2, LDB1, KLF1 and more (Wadman, Osada et al. 1997, Cantor and Orkin 2002). DNA site selection studies, it may be noted also, have shown that TAL1 complex preferentially binds to a composite motif, made of a GATA motif and a neighboring E-box motif. (Wadman, Osada et al. 1997, Kim and Bresnick 2007). While ChIP-seq studies are helpful for identifying the co-occupied locations of GATA1 and TAL1, the resolution is limited to detecting protein bound locations spanning over 100bp.

**GATA1-dependent erythroid development**

Examining the role of transcription factors binding during a particular stage during differentiation can be a challenge, since the cells are a mixed population of cells in different stages. Therefore, a system that allows synchronous differentiation, such as G1E and G1E-ER4, serves as a powerful tool.

G1E and G1E-ER4 system was designed to study erythroid development in a synchronous population of cells. G1E cells are proerythroblasts that lack GATA1 and are arrested in their maturation. G1E was constructed through targeted disruption of X-linked GATA1 gene in a male (XY) murine embryonic stem cell line (Weiss, Yu et al. 1997). The wild-type GATA1 gene was disrupted through homologous recombination with targeted vector, resulting in a mutated GATA1. The vector includes a poly(A) addition site, which disabled any expression of GATA1 protein. This was supported by not observing any GATA1 protein by a band-shift assay in extract of the mutant ES cells (Pevny, Simon et al. 1991). G1E system can be used as a control, where GATA1 protein binding is absent when studying the role of GATA1 binding during erythroid development. The high correlation between RNA-seq transcript profile of G1E and
primary erythroid progenitor cells (Pilon, Ajay et al. 2011) supports the case for using G1E to study early erythroid progenitor stage.

G1E-ER4, a model for differentiating erythroblasts, was derived from G1E cells by stably introducing GATA1 that is fused to ligand binding domain of estrogen receptor (GATA1-ER) (Briegel, Lim et al. 1993). Treatment of cells with estradiol or tamoxifen led to synchronous erythroid maturation and cell cycle arrest (Briegel, Lim et al. 1993). Then reexpression of GATA1 triggered an extensive program of gene activation and repression (Kerenyi and Orkin 2010). Gene expression profile from RNA-seq revealed high correlation between G1E-ER4 and erythroblasts (Pilon, Ajay et al. 2011). These studies showed that GATA1 is required for normal differentiation of red blood cells, and that other GATA-binding proteins cannot substitute its role (Pevny, Simon et al. 1991). Therefore, G1E-ER4 system is an excellent system to study GATA1-dependent erythroid development with synchronous cells.

While binding of GATA1-ER to DNA is expected to be absent before induction with estradiol, a considerable amount of GATA1 binding was detected in uninduced G1E-ER4 cells, also referred as “leakiness” (Jing, Vakoc et al. 2008, Tripic, Deng et al. 2009). Also, GATA1-ER levels were shown not to change during the course of induction through western blots and no substantial change of GATA1 amount was observed in the whole cell and nuclear extract during G1E-ER4 0h to 30 hour. Though GATA1-ER movement from cytoplasm to nucleus was shown through immunofluorescences, the amount of movement seems to be low compared to the substantial amount of GATA1-ER that are already present in the nucleus. A possible reason for GATA1-ER presence before induction may be attributable to the presence of basal level of GATA1-ER transcription and protein in the nucleus.

To circumvent the limitation of GATA1 binding in uninduced G1E-ER4 cells, many studies have compared the induced G1E-ER4 cells with G1E (Jing, Vakoc et al. 2008, Tripic, Deng et al. 2009). However, the regions with occupancy change relative to uninduced GATA1-
ER prior to estradiol induction (GATA1-ER4 0h) may allow us to decipher GATA1 binding that contributes to differentiation of erythroid progenitor cells to erythroblast, and their phenotypic and morphological changes. Therefore, both cell types, G1E and uninduced G1E-ER4 0h, can be utilized as controls for different analyses. Taking these into consideration, ChIP-exo was performed on G1E and G1E-ER system to study the role of GATA1 and TAL1 binding in GATA1-dependent red blood cell development with synchronous cells.

**Summary of transcriptional regulation during erythropoiesis via ChIP-exo**

Transcriptional regulation in erythroid development has been well studied, however, the precise organization of binding of erythroid factors and the determinants for their binding is still unknown. To better understand the interplay of GATA1 and TAL1 forming the multiprotein complex and to decipher the sequence determinants for their co-occupancy, we performed high resolution ChIP-exo on GATA1-dependent system, G1E and G1E-ER4 of multiple time points.

Chapter 1 provided a review on the field of eukaryotic gene regulation and transcription factor binding, and an introduction to ChIP-exo, which overcomes the limitations of low resolution and high background of previously available protein-DNA interaction and open chromatin assays. Then more specifically, it delved into transcriptional regulation and key transcription factors involved in erythropoiesis, and into the use of GATA1-dependent erythroid development system G1E and G1E-ER4.

In the upcoming Chapter 2, we discuss bioinformatics and computational methods for determining robust and confident transcription factor binding locations using ChIP-exo.

In Chapter 3, we seek to understand the role of transcription factor binding in gene regulation during red blood cell development employing a mouse model. To address this, key
erythroid factor (GATA1 and TAL1) and general transcription factors (TFIIB and RNA polymerase II (PolII)) were mapped in G1E and G1E-ER4 cells using ChIP-exo.

Chapter 4 discusses the use of multiple peak pattern of ChIP-exo to determine comprehensive binding locations, and to classify binding locations by their crosslinking patterns.

Finally in Chapter 5, we discuss the significance and contributions of the present work, and possible future works.
Chapter 2

ChIP-exo Analysis Pipeline

ChIP-exo is a technology that allows mapping of transcription factor binding sites with high resolution and low background. This technology requires sophisticated methods and unique considerations to analyze the high-resolution data. Goal of ChIP-exo analysis described in chapter 2 is to determine significant and reproducible ChIP-exo binding locations for sequence specific factors in mammalian genome across multiple conditions, and to classify the detected binding locations by genomic features. This chapter covers from the alignment of reads to the reference genome, assessment of reproducibility and quality of the data, identification of confident binding locations in multiple conditions using intersection of peak pairing and MultiGPS, and further classification of binding locations using various genomic features. Throughout this chapter, examples will be provided using erythroid transcription factors GATA1 and TAL1 in multiple conditions, from G1E cells to G1E-ER4 cells after treatment of estradiol.
Alignment of reads to reference genome

Before sequencing, typically multiple samples are multiplexed (also referred as pooled) into a single sequencing lane using barcoded adapters for reduction of cost. Prepared ChIP-exo samples are typically sequenced using Illumina (or SOLiD) sequencing. Then sequenced reads from the same lane needs to be split into individual samples by their unique barcode identifiers. Subsequently, barcode-splitted reads are aligned to the appropriate reference genomes using BWA (Li and Durbin 2009) or Bowtie (Langmead, Trapnell et al. 2009) for Illumina, and Short Read Mapping Package (SHRiMP) (Rumble et al., 2009) for SOLiD, while allowing 2~3 bases of
mismatch. Depending on the user’s needs, optimal short-read aligner can be chosen among numerous available mappers (Table 2-1).

Table 2-1: List of various short-read aligners (Furey 2012)

<table>
<thead>
<tr>
<th>Software tool</th>
<th>Web address</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Short-read aligners</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BWA</td>
<td><a href="http://bio-bwa.sourceforge.net">http://bio-bwa.sourceforge.net</a></td>
<td>Fast and efficient; based on the Burrows–Wheeler transform</td>
</tr>
<tr>
<td>Bowtie</td>
<td><a href="http://bowtie-bio.sourceforge.net">http://bowtie-bio.sourceforge.net</a></td>
<td>Similar to BWA, part of suite of tools that includes TopHat and Cufflinks for RNA-seq processing</td>
</tr>
<tr>
<td>GSNAP</td>
<td><a href="http://research-pub.gene.com/gmap">http://research-pub.gene.com/gmap</a></td>
<td>Considers a set of variant allele inputs to better align to heterozygous sites</td>
</tr>
<tr>
<td>Wikipedia list of aligners</td>
<td><a href="http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment">http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment</a></td>
<td>A comprehensive list of available short-read aligners, with descriptions and links to download the software</td>
</tr>
<tr>
<td><strong>Peak callers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MACS</td>
<td><a href="http://liulab.dfci.harvard.edu/MACS">http://liulab.dfci.harvard.edu/MACS</a></td>
<td>Fits data to a dynamic Poisson distribution; works with and without control data</td>
</tr>
<tr>
<td>PeakSeq</td>
<td><a href="http://info.gersteinlab.org/PeakSeq">http://info.gersteinlab.org/PeakSeq</a></td>
<td>Takes into account differences in mappability of genomic regions; enrichment based on FDR calculation</td>
</tr>
<tr>
<td>ZINBA</td>
<td><a href="http://code.google.com/p/zinba">http://code.google.com/p/zinba</a></td>
<td>Can incorporate multiple genomic factors, such as mappability and GC content; can work with point-source and broad-source peak data</td>
</tr>
<tr>
<td><strong>Differential peak calling</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edgeR</td>
<td><a href="http://www.bioconductor.org/packages/2.8/bioc/html/edgeR.html">http://www.bioconductor.org/packages/2.8/bioc/html/edgeR.html</a></td>
<td>Uses negative binomial distribution to model differences in tag counts; uses replicates to better estimate significant differences</td>
</tr>
<tr>
<td>DESeq</td>
<td><a href="http://www.huber.embl.de/users/anders/DESeq">http://www.huber.embl.de/users/anders/DESeq</a></td>
<td>Also uses negative binomial distribution modelling, but differs in the calculation of the mean and variance of the distribution</td>
</tr>
<tr>
<td>baySeq</td>
<td><a href="http://www.bioconductor.org/packages/release/bioc/html/baySeq.html">http://www.bioconductor.org/packages/release/bioc/html/baySeq.html</a></td>
<td>Uses empirical Bayes approach to identify significant differences; assumes negative binomial distribution of data</td>
</tr>
<tr>
<td>SAMSeq</td>
<td><a href="http://www.stanford.edu/~juni07/research.html#SAM">http://www.stanford.edu/~juni07/research.html#SAM</a></td>
<td>Based on the popular SAM software; a non-parametric method that uses resampling to normalize for differences in sequencing depth</td>
</tr>
</tbody>
</table>

BWA, Burrows–Wheeler Aligner; DESeq, analysis of high-throughput sequencing to detect differential expression; FDR, false-discovery rate; GSNAP, Genomic Short-read Nucleotide Alignment Program; MACS, Model-based Analysis for ChiP-seq; RNA-seq, high-throughput RNA sequencing; SAM, significance analysis of microarrays; ZINBA, Zero-Inflated Negative Binomial Algorithm.

Among all mapped reads, reads that are uniquely mapped (meaning reads that are mapped to only one location the genome while allowing a number of mismatches) to reference genome are used for further analysis, while reads aligning to multiple locations are filtered out. The percentage of uniquely mapped reads among total reads is often used as a metric to examine the quality of sequencing result. Number of reads and percentage of uniquely mapped reads may vary depending on the genome size of organism (mammalian vs. yeast) and the type of factor, for
example, Sequence-specific factor versus RNA Polymerase II. Then the quality and reproducibility of individual samples and biological replicates are further examined.

**Reproducibility assessment**

When biological conclusions are being made from experiments, it is essential to ensure high consistency and reproducibility in multiple replicates. Having multiple replicates not only increases precision, but also detects the variability among experiments. There are two types of replicates, biological replicates and technical replicates. Biological replicates are defined as experiments produced from independent cell cultures, embryo pools or tissue samples, while technical replicates are to ensure consistency in technical aspect during experiments. To assess the reproducibility of ChIP-exo biological replicates, PCA analysis, Venn Overlap, or scatter plot and calculation of correlation among biological replicates are commonly used.

![Methods to assess reproducibility among biological replicates](image)

**Figure 2-2:** Methods to assess reproducibility among biological replicates
First, principal component analysis (PCA), a method to reduce dimensions and variables, can be applied to examine reproducibility among replicates. For example, occupancy of individual replicate in a common set of binding locations can be used for PCA, as shown in the left panel of Figure 2-2, which is a PCA analysis output from DESeq(Anders and Huber 2010). When data is projected on a two-dimensional space, the data is spanned across PC1 and PC2, the direction where data shows the highest and second highest variability, respectively. For highly reproducible biological replicates, experiments from various conditions span across PC1, while PC2 shows groups of biological replicates. When examining time series experiments, early to later time point often gradually span across PC1.

Secondly, Venn overlap of binding locations of biological replicate can be a useful metric for reproducibility (Figure 2-2, Middle), where higher consistency will be represented as higher overlap. However, it should be noted that a Venn diagram might underestimate the reproducibility between data sets. When raw reads of the sequencing data were processed with certain peak calling threshold to be called as binding locations. Therefore, there maybe highly correlated data points in the outersection of the Venn diagram, while low reads which were below the peak calling threshold.

To overcome the caveat of Venn overlap, estimating the correlation coefficient between replicates and visualizing scatter plots are useful methods to validate reproducibility (Figure 2-2, Right). Through this, binding locations with low reads below the threshold of peak calling can be included for calculation of correlation coefficient, which maybe a more accurate estimate of reproducibility. Using one of more of these three metrics maybe useful for selecting highly reproducible biological replicates for further ChIP-exo analyses.
Identification of binding locations in multiple conditions

Key question for ChIP-exo data analysis is to determine the binding locations. Depending on whether the goal is to determine comprehensive or robust binding locations, the analysis may differ. When the goal is to determine and report a comprehensive list of binding locations, relaxed filtering criteria maybe applied. However, the limitation of this approach is including many false positives. On the other hand, stringent criteria can be applied to determine robust binding locations. When defining rules and making biological interpretation on the characteristics of protein binding locations, selecting for confident binding locations would be important.

The goal of my project is to characterize binding locations by various genomic features. Therefore, the later approach was chosen to ensure that robust biological conclusions derived from stringent binding locations. To accomplish this, final binding locations were detected when they were validated by two independent methods; union of peak pairs and MultiGPS (Mahony, Edwards et al. 2014).

Union of peak pairs

The high resolution of ChIP-exo stems from the use of 5’ to 3’ directional lambda exonuclease, where the exonuclease digestion stops at the border of crosslinking site. Therefore, the conventional method to detect ChIP-exo binding locations is by identifying a pair of peaks on opposite strands within a distance range. This peak pairing method has been applied for individual cell types, time points for each factor. Then to collect a comprehensive list of binding locations across multiple time points for a factor, binding locations in one or more time points were combined. The specifics of each step will be discussed below.
**Peak Calling**

Applying the conventional ChIP-seq pipeline was not able to take advantage of the high resolution of the data. Especially, popular peak calling softwares for ChIP-seq, such as MACS (Zhang, Liu et al. 2008), was not able to capture the specific binding locations while taking advantage of the high resolution of ChIP-exo. Therefore, GeneTrack (Albert, Wachi et al. 2008), a peak caller specifically developed for ChIP-exo was applied.

First, sequence read distribution are used to identify strand-specific peaks using the peak-calling algorithm of GeneTrack (Albert, Wachi et al. 2008), which allows data smoothing using Gaussian smoothing, fitting and peak detection. GeneTrack plots the genomic locations of 5’ end of sequencing reads throughout the mouse genome, and calls peaks separately on each Watson and Crick strand, by converting each read with a probabilistic distribution of occurrence of each read. User-defined parameters, sigma and exclusion zone (D), can be specified during peak calling. Sigma (i.e. the standard deviation of a distribution) determines the smoothening resolution of the distribution, where lower sigma shows higher resolution. Any secondary peaks that are within the exclusion zone (D) region from the first identified highest peaks were prevented from being identified. Example below depicts when sigma and exclusion zone is varied. While the recommended parameters for ChIP-exo data of sequence-specific factors are $s=5$, $D=10$, while $s=20$, $D=147$ for nucleosomes, parameters can be optimized through iterative process.
Genomics experiments using next-generation sequencing often produce artifact signal in certain region, also known as the “blacklisted genomic regions” (Consortium 2012). These regions were found to be in particular types of repetitive regions, including telomeres, centromeres and satellite repeats. Therefore, it is important to filter these regions out when conducting further analyses of ChIP-exo, ChIP-seq, DNase-seq and more. To obtain a comprehensive set of blacklisted regions, blacklisted regions generated by the ENCODE group and Hardison lab have been combined. Subsequently, peaks that lie within these blacklisted regions were filtered out.

Black list regions of Hardison lab (n=4397) have been generated from MACS output peaks from ChIP input of all available cell lines in the lab. The regions were determined when the ratio of average signal in a 10kb-window over the average signal in the genome was greater than 1.5 fold. These regions that highly overlap with repetitive elements, such as centromeric and...
telomeric repeats or satellites have high non-uniquely mapped (multi-mapped) to uniquely mapped reads ratio and large fluctuation in mappability.

**Peak Pairing**

Remaining strand separate peaks, positive (W) and negative (C) peaks, are paired if the negative peak with the highest signal was within the distance range from the nearby positive peak. After calculating the frequency of C-W distance genome wide as a histogram, the mode of the histogram is a unique feature of transcription factor, which indicates the size of the transcription factor binding. Peak pairing is separately performed for each transcription factors in different conditions and cell type. When examining reproducibility between replicates, C-W distance mode is another feature to examine.

**Determination of peak pairs significant over input**

To reduce false positives and to select robust binding locations that are significant over control data, peak pairs with read counts that are minimum 2 fold enrichment over ENCODE input controls (Mouse, Stamatoyannopoulos et al. 2012) in 40-bp window around specified binding events were determined as significant peak pairs. The algorithm normalizes signal over control using median scaling, and assesses the significance of signal over control enrichment using a Binomial test (q-value < 0.05).
Union of significant peak pairs

To examine the dynamics of transcription factor binding and their change of occupancy across multiple conditions, obtaining a comprehensive list of binding locations that includes both time point-specific and constitutively bound locations across multiple time points ("the union" in mathematical term) is necessary.

MultiGPS

MultiGPS is designed to detect binding locations across multiple conditions while characterizing differential binding between conditions (Mahony, Edwards et al. 2014). MultiGPS performs peak calling through GPS, another peak calling method independent from GeneTrack that can be incorporated into the ChIP-exo analysis pipeline. MultiGPS addresses limitations of two common strategies that are used to determine binding locations consistently across multiple conditions (Mahony, Edwards et al. 2014). First common method is analyzing each experiment independently, then combining binding events by using a fixed-window threshold. However, this may falsely combine distinct adjacent binding events (Mahony, Edwards et al. 2014). Second common method is to merge the reads, then call binding events from the merged data set, and to quantify the occupancy in individual experiments. The danger of this method is that noise level in one data set can overpower the low binding in another data set, and that final result maybe biased to particular data sets with higher coverage or antibody quality (Mahony, Edwards et al. 2014).

MultiGPS addresses these limitations by detecting binding locations significant over control across multiple conditions while iteratively scanning for the distribution shape that was built using machine learning. It finally detects both joint binding events (constitutively bound events) and condition-specific events. When binding locations are constitutively bound nearby
across multiple conditions, MultiGPS deconvolves them by encouraging the binding locations of multiple conditions to align. Furthermore, MultiGPS internally runs motif finding software MEME and differential binding software EdgeR, and furthermore, provides the option whether to center binding locations on the motif, and to correct for PCR artifacts. Also, it accounts for the background noise by allowing tags to be assigned to appropriate binding event.

**Intersection of MultiGPS and peak pairing**

After examining the overlap of binding locations from MultiGPS and peak pairing, MultiGPS binding locations that overlap with peak pairing were determined as final confident binding locations. Although GATA1 binding locations in the outersect regions were enriched with the WGATAA motif and GATA1 tags, motif from the intersecting regions showed higher significance (E-value) (Figure 2-4).
When examining the binding locations determined by only one method, peak pairing detected peaks that were enriched in only one-coordinates (so called “singleton peaks”), which MultiGPS ignored (Figure 2-5). Also, many of them were lowly occupied peak pairs that are “shadows” of a nearby robust binding location, which MultiGPS accounted as a single binding location. Furthermore, these regions had tags enriched on both strands while not fitting the distribution model that MultiGPS scanned for, therefore was detected only by peak pairing. On the contrast, many locations that were detected only by MultiGPS were enriched in a single strand, which was not detected by peak pairing.

Figure 2-4: Overlap of binding locations from two methods, Peak pairing and MultiGPS
The goal of detecting binding locations is to further derive biological conclusions and characteristics of transcription factor binding. Thus, determining confident binding locations while removing possible false positives is critical. To achieve this, binding locations that were validated by both MultiGPS and peak pairing were selected for further analyses.

**Bioinformatics analysis for mammalian ChIP-exo**

To accurately and efficiently design and execute bioinformatics analyses pipeline on next generation sequencing (NGS) data, there are various principles to consider, which I learned over the years of my graduate studies.

Standardization of input and output file formats (such as GFF or CDT format) is critical, in order to efficiently conduct series of bioinformatics analyses. It is important to design and write scripts so that output of an upstream analysis script can be used as an input file for the next
downstream analysis. Furthermore, standardization of unique ID format (chr:start-end) for each data point is important. For example, when sorting or filtering binding locations, identical ID format makes tracing the data points and sorting various data sets in identical order possible.

Secondly, automation is crucial when repeating similar analyses on numerous data sets. Automation can be achieved through batch processing using shell scripting or looping through the numerous files in the same directory. For example, MEME-ChIP (Machanick and Bailey 2011) can be run in batch for all the FASTA files within the directory through looping. Also, multiple sequencing tags files within a directory can be aligned to numerous reference points (such as transcription start sites, transcription end sites, or motif sites).

Lastly, parallel computing is a powerful method when performing bioinformatics analysis, especially on big data such as mammalian genomes. An analysis that can take several minutes for small genome (such as S. cerevisiae) can require a day for mammalian genome. Therefore, efficient and fast computation is critical, but is a challenge. To overcome the computational challenge, an intensive job can be splitting into smaller jobs, then submitted in parallel to the computing clusters using PBS, then merged for final output. As illustrated in Figure 2-6, a 3200-hour job that aligns tags to the reference points using serial computing can be split into 840 smaller jobs (split by chromosomes) and ran in parallel, reducing the jobs down to 2 hours.
Figure 2-6: Parallel Computing
Chapter 3

Genome-wide organization of GATA1 and TAL1 determined at high resolution

This chapter has been submitted for publication as:
Garam Celine Han, Vinesh Vinayachandran, Alain R. Bataille, Ka Yim Chan-Salis, Cheryl A Keller, Maria Long, Shaun Mahony, Ross C Hardison, B. Franklin Pugh* 2015. Genome-wide organization of GATA1 and TAL1 determined at high resolution. *Epigenetics and Chromatin
Abstract

Background

Erythroid development and differentiation from multiprogenitor cells to red blood cells requires precise transcriptional regulation. Key erythroid transcription factors, GATA1 and TAL1, cooperate, along with other proteins, to regulate many aspects of this process. How GATA1 and TAL1 are positionally organized with respect to each other and their cognate DNA binding site across the mouse genome remains unclear.

Results

We applied high resolution ChIP-exo to GATA1 and TAL1 to study their positional organization across the mouse genome during GATA1-dependent maturation. Two complementary methods, MultiGPS and peak-pairing, were used to determine high confidence binding locations by ChIP-exo. We identified ~10,000 GATA1 and ~15,000 TAL1 locations, which were essentially confirmed by ChIP-seq. Of these, ~4,000 locations were bound by both GATA1 and TAL1. About three-quarters of these were tightly linked (<40 bp away) to a partial E-box located 7-8 bp upstream of a WGATAA motif. Both TAL1 and GATA1 generated distinct characteristic ChIP-exo peaks around WGATAA motifs, that reflect on their positional arrangement within a complex.

Conclusions

We show that TAL1 and GATA1 form a precisely organized complex at a compound motif consisting of a TG 7-8 bp upstream of a WGATAA motif across thousands of genomic locations.
Background

Hematopoietic stem cells undergo self-renewal and differentiation to many blood cell lineages: erythroid (red blood cells), lymphocytes and myelocytes (including megakaryocytes) (Orkin and Zon 2008). Differentiation to red blood cells, referred to as erythropoiesis, requires several transcription factors, such as GATA1, SCL/TAL1, LMO2, LDB1, FOG1 and KLF1 (Kerenyi and Orkin 2010). GATA factors are essential for hematopoiesis, as shown by the anemic phenotypes of mouse knockout mutations and the leukemias and lymphomas associated with mutations of the human genes (Crispino 2005). GATA1 is a master regulator of differentiation, proliferation, and apoptosis of red blood cells and megakaryocytic cells (Shimizu, Engel et al. 2008). Mutations in or misregulation of GATA1 leads to transient myeloproliferative disorder and acute megakaryoblastic leukemia in infants with Down syndrome (Shimizu, Engel et al. 2008). TAL1, a basic helix-loop-helix hematopoietic transcription factor (formerly SCL), is required for multiple functions in hematopoiesis, including terminal differentiation of red blood cells (Kassouf, Hughes et al. 2010). TAL1 interacts with GATA1. An important missing component in our understanding of how GATA1, TAL1 and their cognate DNA recognition motif function is their precise spatial organization across the genome.

Several ChIP-seq studies have been performed on GATA1 (Cheng, Wu et al. 2009, Fujiwara, O'Geen et al. 2009, Yu, Riva et al. 2009, Soler, Andrieu-Soler et al. 2010) and TAL1 (Tripic, Deng et al. 2009, Kassouf, Hughes et al. 2010, Soler, Andrieu-Soler et al. 2010) in erythroid cell lines and primary cells as a first step toward understanding the genome-wide binding properties of GATA1 and TAL1. However, the low resolution of the assay places confidence limits on binding locations. Indeed the broad binding regions of ChIP-seq, which span over 100 bp, makes it challenging to decipher the relevant motif when numerous motifs may be present within an occupied segment; it also may not allow the difference in binding patterns and
distances between GATA1 and TAL1 binding locations at co-occupied locations to be distinguished. Estimates of the number of binding locations vary considerably among ChIP-seq studies. Three studies identified 4,000–6,000 in vivo binding sites for GATA1 in mouse MEL erythroleukemia cells expressing a tagged form of GATA1 (Yu, Riva et al. 2009, Soler, Andrieu-Soler et al. 2010) or human K562 erythroleukemia cells (Fujiwara, O'Geen et al. 2009). Another study identified >15,000 sites occupied in mouse G1E-ER4 cells, where the GATA1 genes has been knocked out mouse embryonic stem cells (G1E), then restored under the artificial control of estradiol (Cheng, Wu et al. 2009). Recently, ENCODE has reported ~12,000 GATA1 and ~2,000-8,000 TAL1 binding locations in the mouse G1E-ER4 cell system after 24 hours of GATA1 activation, and 24,000 to 60,000 sites bound by GATA1 in MEL cells (Mouse, Stamatoyannopoulos et al. 2012, Yue, Cheng et al. 2014, Jain, Mishra et al. 2015). While some differences in the number of bound locations may result from occupancy level thresholding, other factors such as cell type, antibody quality, peak-calling methods, and data quality might also contribute to differences.

GATA1 recognizes the WGATAA motif, whereas TAL1 and its heterodimeric partner, E2A, recognize the E-box (CANNTG) (Begley, Aplan et al. 1989, Chen, Yang et al. 1990). Complexes containing both GATA1 and TAL1 tend to have their cognate motifs spaced 9 bp apart (Wadman, Osada et al. 1997, Cantor and Orkin 2002). However, the determinants of binding may be more complex than captured by their individual consensus motifs. First, the highly conserved WGATAA consensus sites are insufficient to accurately predict in vivo GATA1 binding (Bresnick, Martowicz et al. 2005, Zhang, Wu et al. 2009). Second, TAL1/E2A complexes have been suggested to bind to DNA with GATA1 without the need for an E-box (El Omari, Hoosdally et al. 2013), although another study found CTG upstream of WGATAA, [CTG(N7–8)]WGATA], in peaks co-occupied by GATA1 and TAL1 (Soler, Andrieu-Soler et al. 2010). Third, the in vivo developmental functions of TAL1 do not require its DNA binding region
Fourth, DNA site selection studies show that in complex with other proteins including GATA1, the TAL1/E2A complex prefers to bind GATA1/E-box composite motifs, rather than an E-box alone (Wadman, Osada et al. 1997). Moreover, the GATA binding site motif is a stronger determinant of TAL1 occupancy than is the E-box (Wadman, Osada et al. 1997, Cantor and Orkin 2002, Kassouf, Hughes et al. 2010, Wu, Morrissey et al. 2014). How these erythroid transcription factors are positionally organized around their cognate motif remains unclear.

GATA1 activates or represses transcription depending on the context with other transcription factors. GATA1-TAL1/E2A complexes induce gene expression, while GATA1 without TAL1 represses (Wozniak, Keles et al. 2008, Cheng, Wu et al. 2009, Tripic, Deng et al. 2009, Soler, Andrieu-Soler et al. 2010). However, GATA1 and KLF1 co-occupancy leads to gene activation, and may be TAL1-independent. This was indicated by the low overlap between GATA1/KLF1 and GATA1/TAL1 regions ascertained in that study (Tallack, Whittington et al. 2010). How the interplay and co-occupancy of various proteins in GATA1 multiprotein complexes activate or repress transcription, and therefore regulate the erythroid differentiation program, is not well-understood (Cantor and Orkin 2002).

To obtain a more detailed, highly resolved view of the genomic organization of GATA1 and TAL1 binding during erythroid development, we employed G1E cells (Pevny, Simon et al. 1991, Weiss, Yu et al. 1997). Reintroduction of GATA1 fused the estrogen binding domain provides a cell system (G1E-ER4) for synchronous and homogeneous erythroid maturation in response to estradiol treatment (Welch, Watts et al. 2004). We comprehensively mapped the genomic binding locations of GATA1 and TAL1 at near bp resolution using ChIP-exo in G1E (GATA1+) and rescued G1E-ER4 mouse erythroid cells before and after 3 and 24 hours of estradiol treatment. This very high-resolution genomic view of GATA1 and TAL1 binding to
their cognate sites during erythroid differentiation revealed a very precise positional organization of TAL1 next to GATA1.

**Results**

**GATA1 and TAL1 locations determined by ChIP-exo**

We conducted ChIP-exo with the intent of obtaining higher resolution binding locations than offered by ChIP-seq (Table 3-1). In ChIP-exo, the 5’ ends of the sequencing tags correspond to a point about six bp upstream (5’) of a protein-DNA crosslink, where movement of the lambda single-stranded exonuclease is blocked. For a single protein-DNA crosslink, a block occurs on both strands resulting in two peaks of 5’ ends, located on opposite strands and separate by about 12 bp in the 3’ direction. Pairing of these peaks provides a potential resolution of a few bp (Rhee and Pugh 2011). In ChIP-seq, 5’ ends of sequenced tags correspond to random break points in sonicated solubilized chromatin, and have less resolution than ChIP-exo. Unlike ChIP-exo, it cannot resolve individual points of crosslinking within an individual binding location. Resolution is essential where proteins bind in closely clustered locations. For example, at many genomic locations, ChIP-seq detects GATA1 as a broad peak, whereas ChIP-exo reports many individual peaks, of which some correspond to separate WGATAA (IUPAC consensus) motifs (Figure 3-1). Even when the signal track for ChIP-seq suggests multiple binding sites, peak calling algorithms may combine them into a single peak.
As common with ChIP assays, the signal intensity at many locations may be relatively low and not entirely discernible from background, resulting in substantial false positives. These false positives can be reduced by placing tighter positional constraints on relative peak locations. However, we wanted to allow some positional flexibility so as to identify alternative modes of binding. To balance these opposing stringencies, we implemented and compared two complementary approaches to identify binding locations: peak-pairing and MultiGPS [23] (Figure S3-1). In an effort to examine comprehensive binding locations across all time points, we considered all significant peak pairs arising from 0, 3, and 24 hours of GATA1 activation. MultiGPS uses integrated machine learning to call binding events that are consistent with the

Figure 3-1. Browser shot of GATA1 ChIP-seq and ChIP-exo tags and binding locations. Browser shot of ChIP-exo tag 5’ ends for GATA1 measured by ChIP-seq (CS) and ChIP-exo (XO) on chromosome 1 from coordinate 36,977,009 to 36,977,009. Tag locations were smoothed (20 bp moving average). Tag density on the lower (negative) strand is shown as an inverted plot. Locations of WGATAA and E-box motifs (p-value < 10^-3) and RefSeq genes are shown below through IGV browser.

As common with ChIP assays, the signal intensity at many locations may be relatively low and not entirely discernible from background, resulting in substantial false positives. These false positives can be reduced by placing tighter positional constraints on relative peak locations. However, we wanted to allow some positional flexibility so as to identify alternative modes of binding. To balance these opposing stringencies, we implemented and compared two complementary approaches to identify binding locations: peak-pairing and MultiGPS [23] (Figure S3-1). In an effort to examine comprehensive binding locations across all time points, we considered all significant peak pairs arising from 0, 3, and 24 hours of GATA1 activation. MultiGPS uses integrated machine learning to call binding events that are consistent with the
learned binding pattern across multiple time points and biological replicates. MultiGPS allows the determination of binding events that are shared across conditions or are condition-specific using the binding pattern profile. Both methods required ChIP-exo signals to be at least two-fold enriched over the input control, and a q-value < 0.05 (q-values are multiple hypothesis testing adjusted p-values from binomial tests).

Peak-pairing yielded ~47,000 initial GATA1 candidate crosslinking sites in mouse G1E-ER4 cells, whereas MultiGPS yielded ~13,000 initial locations. Often distinct peak pairs were very close together (<20 bp), and MultiGPS modeled them as a single location. Consequently, ~10,000 (80%) MultiGPS locations contained ~20,000 peak-pairs (Figure 3-2). This intersect was highly enriched with the WGATAA motif (motif p-value threshold = 10⁻³), thereby providing a general validation of the binding locations. Approximately 80% (7,927/10,290) contained the WGATAA motif. Thus, we detected and further analyzed ~8,000 GATA1-bound WGATAA binding sites in differentiating mouse G1E-ER4 cells. These locations were detected by peak-pairing and MultiGPS and contained a WGATAA motif. The remaining ~2,000 that were detected by both methods but lacked a WGATAA motif likely include non-cognate DNA interactions, interactions with degenerate WGATAA motifs that fell below our bioinformatics detection threshold, and false positives.
When comparing the binding locations separately detected by MultiGPS and by peak-pairing, the outersects consisting of calls made by only one method were less enriched with WGATAA motifs (60% and 46% vs 82%). Thus, while highly enriched with true positives, those called by only one method have a higher false positive rate or may involve noncanonical crosslinking patterns. Representative examples of how calls at WGATAA sites can be made by only one method and not the other are shown in Figure S3-2 (for TAL1). Peak-pairing detected simple one-coordinate peaks in each pair, which MultiGPS discounted. In contrast, many binding locations that were detected only by MultiGPS were predominantly enriched with tags on only
one strand, or had sparsely distributed tags, thereby precluding peak-pairing. Moreover, low-occupancy peak pairs often occurred at noncognate locations in the “shadows” of robust cognate binding events, which MultiGPS rolled into a single location. Also, tags may be piled up on both strands, which are detected by peak-pairing, while not by MultiGPS since it does not fit the distribution shape that MultiGPS deems to be a consensus.

The intersection of the two location-calling methods resulted in ~15,000 candidate TAL1 locations detected in one or more GATA1 activation time points. Therefore, MultiGPS peaks that overlapped with peak-pairing were taken as higher confidence binding locations, resulting in 10,290 GATA1 binding locations and 15,209 TAL1 binding locations.

One advantage of ChIP-exo is its ability to detect two closely bound factors that would be called as the same binding location by ChIP-seq. We therefore examined whether GATA1 locations occurred in clusters. Among 10,290 GATA1 binding locations, 14% (1,414 GATA1 binding locations) were less than 500 bp away from each other, forming 663 GATA1 clusters (Figure 3-3) with approximately two GATA1 locations per homotypic cluster. Surprisingly, a substantial portion (64% or 9,796) of TAL1 binding locations formed 2,905 TAL1 clusters (<500 bp) (Figure 3-3).
On average approximately three TAL1 locations formed a homotypic cluster. The nearest genes to a cluster showed a slightly higher average gene expression change upon GATA1 activation compared to non-homotypic clusters for both GATA1 and TAL1 (Figure 3-4), indicating that clustering either has little effect on gene expression or that it affects relatively few genes. We observed little difference in the location of clusters vs non-clusters relative to the nearest gene transcriptional start site.

Figure 3-3. GATA1 and TAL1 binding locations in clusters and non-clusters. Percentage of GATA1 and TAL1 binding locations in homotypic clusters from the total binding locations. A cluster is defined as adjacent binding locations that are <500 bp apart. In case of GATA1, 13.7% (1,414) binding locations form 663 clusters, while 64.4% (9,796) TAL1 binding locations form 2,905 clusters.
A structural model for the genome-wide co-binding of TAL1 and GATA1

Though GATA1 and TAL1 have been known to work together, a high-resolution view of their precise positioning within a complex on DNA has not been examined on a genomic scale in vivo. Further, although co-occupancy of GATA1 and TAL1 are known to activate gene

Figure 3-4. Expression change of GATA1 and TAL1 clusters and non-clusters. Expression change of the closest gene to a homotypic cluster or non-cluster for GATA1 and TAL1. Expression change is calculated as log2 fold change of gene expression between G1E-ER4 induction of 30 hour relative to 0 hour.
expression, the recognition motif for their co-occupancy still remains poorly identified. To this end we compared the ChIP-exo binding locations of GATA1 and TAL1 (Figure 3-5) to determine the regions of co-binding and analyze their underlying DNA sequence. Within a defined occupancy threshold, we identified 3,736 GATA1-bound GATA motif locations that also contained 4,245 TAL1 binding locations within 40 bp. The remaining 6,554 GATA1-bound GATA motif locations contained either no or low (i.e., subthreshold) levels of TAL1. Similarly, the remaining 10,964 TAL1 locations contained either no or subthreshold levels of GATA1. These “TAL1-only” locations were enriched with a compound motif containing a half E-box adjacent to a WGATAA motif (not shown), suggesting that another GATA factor might be bound instead of GATA1.
In an effort to explore the genomic organization of GATA1 and TAL1 around the WGATAA motif, we plotted GATA1 ChIP-exo tag 5’-ends around all 7,927 WGATAA motifs (motif p-value < 10^{-3}) that were enriched in the 10,290 GATA1 binding locations (Figure 3-6). Notably, GATA1 displayed a double peak-pair pattern around the WGATAA motif, with a peak-pair being located at each end of the motif. This is similar to what we have seen with many other

Figure 3-5. Genome-wide structural organization of GATA1/TAL1 complexes. Venn diagram showing the overlap of TAL1 and/or GATA1 binding locations further classified to TAL1 only, GATA1/TAL1 co-occupied, and GATA1 only binding locations. Co-occupancy was defined as having their midpoints within <40 bp distance apart.
proteins including CTCF [31] and p53 [32], where crosslinking typically occurs at the edges of protein/DNA complexes. For GATA1, a major and a minor peak-pair were observed, with the two pairs being 16 bp apart. These points of crosslinking align very well with structural models of the DNA binding domain of GATA1 in complex with DNA (shown in Figure 3-6, Top), which indicates these models likely reflect the binding structure of GATA1 bound throughout the genome. Crosslinking to the left side of WGATAA was stronger than on the right side, which likely reflects differential reactivity between an appropriate crosslinkable amino acid with the DNA on the two sides.

TAL1 also displayed two major peak-pairs, but the pairs were situated to the left of WGATAA sites. Their midpoints were 21 and 13 bp upstream (more 5’) of the WGATAA midpoint (Figure 3-6). The inferred two points of crosslinking were about 8 bp apart, which agrees with the modeled structure of TAL1/E47 with an E box [17]. A minor level of TAL1 crosslinking was observed on the right side of the WGATAA motif. This crosslinking did not match with the double peak pattern of GATA1, and thus largely rules out TAL1 crosslinking indirectly to DNA via crosslinks to GATA1. Instead, we suspect that TAL1, in complex with GATA1 and other proteins, may also be in close proximity to DNA on the distal side of GATA1 (in addition to its main proximal-side interactions).
Figure 3-6. Genome-wide structural organization of GATA1/TAL1 complexes. Composite distribution of TAL1 (black trace) and GATA1 (red filled trace) ChIP-exo tag 5’ ends around the
The positioning of GATA1 and TAL1 crosslinking points remained consistent relative to WGATAA motifs, across most locations (Figure 3-7, Figure S3-3), including whether or not GATA1 was associated with TAL1. Hence, the genome-wide average was not a skewed representation of a few sites having high tag counts. Locations often did not contain a full complement of GATA1 tags at each of the consensus four-peak locations (i.e., two peak pairs), but rather had substantial tag count variation. This was evident when sorting by signal strength of each of the four peaks separately (Figure S3-4). Each peak in a peak-pair reflected two distinct measurements of the same crosslink, and so should have had roughly equivalent tag counts. We attribute the variance in tag counts at each of the four peak locations to statistical sampling, low library complexity and non-uniform amplification during library construction and PCR.

We also addressed whether potential nucleotide bias in crosslinking efficiency might account for the observed range of GATA1 occupancy. The nucleotide sequence in the vicinity of the major GATA1 crosslinking point (-8 position from WGATAA midpoint) deviated little from the overall average, when comparing high versus low GATA1-occupied sites (Figure S3-5). The small amount of deviation towards higher G+C frequency at highly occupied sites was also evident in surrounding regions, indicating that it was not specific to the site of crosslinking. Therefore, we conclude that the detection of different binding site occupancy levels is not substantially influenced by putative base-specific differences in crosslinking efficiency.
TG is enriched 7-8 bp upstream of WGATAA at GATA1 and TAL1 co-occupied locations

We examined the DNA sequence underlying the co-occupancy of GATA1 and TAL1 at all 7,927 identified GATA1-bound WGATAA motifs. Motif locations were first grouped as to whether they contained or lacked (subthreshold levels) of TAL1, then sorted by GATA1 occupancy (Figure 3-7). Where GATA1 and TAL1 co-occupy the same location, their occupancy levels were positively correlated (R=0.42). A “TG” dinucleotide motif was enriched upstream (more 5’) of the WGATAA motif (green/yellow vertical stripe in Figure 3-7, upper panel) in the GATA1 and TAL1 co-occupied sites, while no TG enrichment was observed in TAL1 lowly bound regions.

Figure 3-7. Distribution of 5’ ends of GATA1, TAL1 ChIP-exo sequencing tags around 7,927
WGATAA sites in GATA1 binding locations, comprised of 2,997 GATA1-enriched WGATAA sites (rows) having TAL1, and 4,930 sites lacking TAL1. Both GATA1 and TAL1 co-bound WGATAA sites (Top), and GATA1-only bound WGATAA sites (Bottom) were sorted by GATA1 occupancy in merged data of all time points. Occupancy is calculated by the sum of total tags around motif reference point (WGATAA) from 25 bp upstream to 25 bp downstream, and 40 bp upstream to 30 bp downstream for GATA1 and TAL1, respectively. A four colored plot (Left) reports the nucleotide composition of 50-bp surrounding regions of WGATAA motif: A (Red), C (Blue), G (Gold), T (Green).

The distance between the two closest ends of the two motifs was 7-8 bp (i.e., started 11-12 bp upstream of the “A” at WGATAA midpoint) (Figure 3-8, upper panel). This enrichment was not observed in no/low TAL1 binding locations (Figure 3-8, lower panel).
Figure 3-8. Distance of TG (half E-box) to WGATAA sites in GATA1, TAL1 co-bound locations.
Frequency histogram of distances of TG relative to WGATAA sites in GATA1 binding sites with high TAL1 occupancy (Left) and in GATA1 binding sites with no/low TAL1 occupancy (Right).
One interpretation of this TG dinucleotide is that it comprises the fifth and sixth nucleotides of an E-box (CANNTG) that has been previously linked to WGATAA motifs at TAL1-bound sites in vivo (Chen, Yang et al. 1990, Kassouf, Hughes et al. 2010, Soler, Andrieu-Soler et al. 2010), and to gene activation [19]. We investigated this possibility by inspecting the remaining sequences of a putative E-box configuration, where the conserved TG represented the most WGATAA-proximal side of the E-box. Indeed, matches to an E-box consensus were the most abundant (Figure 3-9), but nevertheless represented a minority of all possible configurations. Those with a CANNTG motif had the central two nucleotides being largely KC dinucleotides (IUPAC nomenclature). Thus, TAL1-GATA1 genomic binding sites are predominantly TGn7-8WGATAA, but having a bias towards CAKCTGn7-8WGATAA. The tight positional linkage of the TG motif with WGATAA, and the degeneracy of the remaining E-box sequence, suggest that the DNA interactions at TG on the WGATAA proximal side are predominantly contributing towards specificity (in addition to WGATAA via GATA factor binding).

Figure 3-9. Pie chart for nucleotide composition of E-box portion of composite motif [TG(N7-
Structural relationship between points of TAL1/GATA1 crosslinking and DNA sequence

We next attempted to orientate and position the TAL1/E2A crystal structure (El Omari, Hoosdally et al. 2013) to the $\text{NNNNTG}_{7,8}$WGATAA motif (position 1 and 2 are underlined), as illustrated in Figure 3-6. The two primary TAL1 crosslinking points are located 21 and 13 bp upstream of the “A” at the WGATAA midpoint. Those locations flank position 1 and 2 of the degenerate E-box, based on TG starting 11-12 bp upstream of the WGATAA midpoint. Positions 1 and 2 are located 15-16 bp upstream of the WGATAA midpoint. Making the reasonable assumption that the TAL1 DNA-binding alpha helix is what is crosslinking to DNA, the most likely placement of the TAL1 alpha helix is midway between its two points of crosslinking, and thus within 1 bp of positions 1 and 2. If this interpretation is correct, then in the context of GATA1, TAL1 may not be binding to DNA with much sequence-specificity. Rather, we suggest it is its E2A partner that is binding to the TG motif [30], and thus positionally closer to GATA1 along the DNA. This interpretation would place the TAL1/E2A heterodimer in one single predominant orientation (with TAL1 being distal) with respect to GATA1/WGATAA, in addition to being located primarily upstream of WGATAA (i.e., essentially as oriented in the Figure 3-6 structural model).

A second, alternative interpretation of the double peak-pairs generated by TAL1 is that one peak-pair is from TAL1 and the other is from its partner, in which the heterodimer is directionally oriented relative to the motif. We deem this to be a more complex and less likely
scenario since crosslinking detection would require that TAL1 crosslink to its partner, and its partner crosslink to DNA. Due to the inherent inefficiency of crosslinking, this is expected to be a low-frequency situation. A third possibility is that TAL1 and its partner can bind in both the forward and reverse orientation. If the heterodimer were binding in both orientations, then in a population of molecules, the crosslinking points should be equidistant from the E-box midpoint. However, that was not observed. Thus, we favor a model in which TAL1 and its partner have a directional orientation upstream of GATA1 binding, where TAL1 is the more distal partner and thus its binding is not being specified directly by the TG motif.

**ChIP-exo versus ChIP-seq in location detection**

We compared ChIP-exo to ChIP-seq with the intention of understanding why some locations were called by one method and not by the other. For ChIP-exo, we used the set of locations defined by the intersection of peak-pairing and MultiGPS (filtered to be > 2-fold over background, q-value < 0.05). The sets of peaks showed substantial overlaps between ChIP-exo and ChIP-seq (Venn diagrams in Figure 3-10). The vast majority of ChIP-exo peaks overlap with ChIP-seq peaks, as is expected since ChIP-exo is essentially a refined version of ChIP-seq. However, a substantial fraction of ChIP-seq peaks did not overlap with ChIP-exo peaks. One explanation may be that the outersects reflect putative binding events detectable in only one assay. Alternatively, thresholding of the data may result in a set of locations being marginally above a threshold in one assay and marginally below a threshold in the other assay.
We tested this by directly comparing occupancy levels between the two assays. A scatter plot comparison of occupancy levels from ChIP-seq versus ChIP-exo showed that they were well correlated $R \sim 0.65$, Figure 3-11) when both assays identified GATA1 locations. The correlation dropped to only 0.4 when locations were identified only by ChIP-seq. These locations were generally low occupancy in both cases. Therefore, the lack of a call in the ChIP-exo assay is
more likely due differences in occupancy thresholding between the two assays, or imposition of peak-pairing distance limits, rather than a qualitative distinction between the ChIP-seq and ChIP-exo assays. Locations called by ChIP-exo only, produced a weak occupancy correlation (Pearson R ~0.18). They were few in number (i.e., 5% of the calls made by ChIP-seq only), and low in occupancy in both assays. These therefore likely represent false positives.

Figure 3-11. Scatter plot with Spearman’s correlation coefficient of log10 transformed ChIP-exo and
ChIP-seq tag counts (window size = 400bp) around all binding locations of ChIP-exo and ChIP-seq at GATA1 G1E-ER4 3 h. Red to blue color indicates the high to low density of data points in the scatter plot. Similar results were obtained for the other time point datasets (not shown).

To further address the ChIP-exo/ChIP-seq correspondence of the outersects, we examined the ChIP-exo GATA1 tag distributions around the midpoints of “ChIP-seq-only” bound occupied segments. On average, the ChIP-exo peak pairs were centered on the ChIP-seq-only midpoint locations (Figure 3-12), demonstrating that they were indeed reporting on similar locations.

Taken together, many of the locations called in only one of the two assays likely reflect a substantial amount of real but low occupancy binding events. They also having a greater frequency of false positives. Since both ChIP-seq and ChIP-exo use essentially the same method (formaldehyde crosslinking) to trap in vivo interactions, and that ChIP-exo is a refinement of ChIP-seq, the primary the difference between called locations in the two assays lies in the level of false discovery (sum total of false positives and negatives), defined by assay-specific location-calling thresholds. While the use of thresholds adds high confidence to location calling by reducing false positives, it comes at the price of removing false negatives. The analysis in Figure 4 indicates that ChIP-seq has a greater false discovery rate than ChIP-exo, but in principle can capture a greater overall number of lower-confidence locations.
Dynamics of GATA1 and TAL1 binding during erythroid development

We next examined the changes in GATA1 binding during erythroid development. The set of ~10,000 GATA1 binding locations were classified by their differential occupancy between various induced differentiation time points (3 vs 0 hr, 24 vs 0 hr, and 24 vs 3 hr) (Figure 3-13).

Figure 3-12. Composite distribution of G1E-ER4 3 h ChIP-exo reads (smoothing = 20 bp) around occupied segments detected only by ChIP-seq (n= 14,071). Similar results were obtained for other time points.
Six kinetic classes were produced by k-means clustering, which were grouped into increased (clusters 1 and 2), unchanged (clusters 3 and 4), and decreased (clusters 5 and 6) GATA1 occupancy. Since GATA1 was ectopically induced, we assume that its total level of genome-wide binding would either increase or approximately stay the same. We normalized the total of all occupancy levels at each time point to be constant across time points. Therefore, an apparent relative “decrease” in GATA1 occupancy upon GATA1 induction may actually reflect less of an increase, compared to other locations. In general, the kinetic patterns of TAL1 occupancy were similar to those of GATA1, as expected of them binding as a complex (plots on the right portion of Figure 3-13). We note that prior to induction (0 hr), GATA1 is in both the cytoplasm and the nucleus (Jing, Vakoc et al. 2008), and thus has a significant level of binding genome-wide. This pre-activation state is insufficient to promote differentiation, and it may be in an inactive until released by estradiol.

Figure 3-13. Heatmap of GATA1, TAL1 total tag count normalized tags (strand merged) plotted around GATA1 binding locations which were clustered by differential occupancy of GATA1 between time points (left). Differential occupancies between time points (G1E-ER4 3h/0h, 24h/0h,
and 24h/3h) were determined as log2 fold change of ratio of GATA1 tags over input control between time points using EdgeR and MultiGPS. For each binding location, a heatmap of Log2 expression fold change of the closest gene between G1E-ER4 3h/0h is shown (right) where increased, unchanged, decreased expression are depicted as yellow, black, and blue, respectively.

The closest annotated mouse gene was assigned to each GATA1 binding location to examine the effect of GATA1 and TAL1 binding on gene expression. These locations were largely associated with genes involved in blood cell development and maintenance, as previously determined (Figure S3-6) [26,30]. Binding locations with the most increase in GATA1 occupancy displayed the most increase in gene expression of the nearest gene in response to GATA1 activation (Jain, Mishra et al. 2015) (Clusters 1 and 2, Figure 3-14). GATA1 was closer to their TSS on average, when compared to the average of all other genes linked to a GATA1 location that did not experience the same relative increase in GATA1 occupancy (i.e., clusters 3-6).

Figure 3-14. Box plot of expression change (G1E-ER4 30h/0h) (left), and distance to the closest transcription start site (right).
Conclusions

In this study we determined the genome-wide positional organization of GATA1 and TAL1 at near single bp resolution using ChIP-exo. Comparison of location-calling using peak-pairing versus MultiGPS revealed that they largely call the same locations. Peak-pairing picks up low-complexity binding locations that are missed by MultiGPS, whereas MultiGPS picks up locations where tags are missing on one strand (likely due to molecule-specific biases arising during sample preparation and library construction). These are missed by peak-pairing. Often the biggest differences in called locations are due to difference in data thresholding (tag counts and patterning). A large fraction of locations typically fall near the threshold, and thus small differences in thresholding can create an appearance of incongruence between location-calling methods. We similarly compared ChIP-exo to ChIP-seq and found them to be highly similar, with the main differences attributable to differences in false discovery rates and thresholding.

We identified about 10,000 GATA1-bound WGATAA sites in GATA1-ectopically expressed mouse G1E cells, and about 15,000 TAL1 locations. About 3,000 of these locations correspond to TAL1/GATA1 complexes bound to a WGATAA motif having a half E-box located 7-8 bp upstream. TAL1 and presumably its E2A partner bind stereospecifically to GATA1-bound DNA, and TAL1 appears to contact DNA more distally upstream of WGATAA compared to E2A. We also find evidence of TAL1 being in close proximity to DNA on the downstream side of GATA1.

A small portion of both GATA1 and TAL1 bind in clusters, while remainder bind alone (non-clusters). The significance of this is unclear, although it likely reflects constraints imposed by higher order structures that involve multiple copies of GATA1 and TAL1. Our analyses did not identify major functional distinctions between clustered and non-clustered locations. Nonetheless, G1E differentiation along the hematopoietic lineage upon ectopic activation of
GATA1 is linked to a substantial increase in GATA1/TAL1 binding to about 3,000 sites genome-wide.
Methods

Cell culture

Cells were cultured as described previously (Welch, Watts et al. 2004). G1E and G1E–ER4 cells were grown in Iscove’s modified Dulbecco’s media (IMDM) with 15% fetal calf serum, 2 U/mL erythropoietin, and 50 ng/mL kit ligand. To activate the conditional GATA1–ER, cells were cultured in the presence of $10^{-7}$ mol/L beta-estradiol for 24 h (Cheng, King et al. 2008). Sonicated chromatin materials of G1E, G1E-ER4 0 hr, 3 hr and 24 hr cells were prepared by standard methods.

ChIP-exo

With prepared sonicated chromatin, chromatin immunoprecipitation (ChIP) was performed on GATA1 (antibody sc265 L1609) and TAL1 (sc12984). Standard ChIP methods were used, followed by lambda exonuclease treatment and library construction, as described (Rhee and Pugh 2012). Sequencing was performed using AB 5500xl Genetic Analyzer, Illumina HiSeq2000 and Illumina NextSeq. For single end reads from HiSeq, basecalls were performed using CASAVA version 1.7 and for paired end reads from NextSeq, basecalls were performed using Bcl2fq version 2.15. ChIP-exo reads were aligned to the mm9 genome assembly using Bowtie 1.00 for SOLiD, and BWA (Version: 0.6.2 for HiSeq single end reads and Version: 0.7.9a for NextSeq paired end reads) with default options. Non-uniquely mapped reads were filtered out in order to remove the reads with low mapping quality. Sequencing statistics are reported in Table 3-1.
Table 3-1. Sequencing statistics. Sequencing statistics including total reads, uniquely mapped read and percentage yield of all biological replicates and time points of GATA1 and TAL1.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Time point</th>
<th>Cell Line</th>
<th>Total read #</th>
<th>Unique #</th>
<th>Yield %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G1E</td>
<td>Rep1</td>
<td>4,190,070</td>
<td>820,435</td>
<td>19.6%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 0h</td>
<td>Rep1</td>
<td>19,702,763</td>
<td>4,360,209</td>
<td>22.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep2</td>
<td>29,613,874</td>
<td>16,554,432</td>
<td>55.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>29,698,806</td>
<td>15,651,119</td>
<td>52.7%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 3h</td>
<td>Rep1</td>
<td>24,024,479</td>
<td>5,500,092</td>
<td>22.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep2</td>
<td>32,044,346</td>
<td>19,169,197</td>
<td>59.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>33,464,704</td>
<td>18,581,787</td>
<td>55.5%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 24h</td>
<td>Rep1</td>
<td>65,889,106</td>
<td>21,844,285</td>
<td>33.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep2</td>
<td>23,877,797</td>
<td>5,310,437</td>
<td>22.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>15,912,120</td>
<td>4,542,129</td>
<td>28.5%</td>
</tr>
<tr>
<td></td>
<td>G1E</td>
<td>Rep1</td>
<td>46,589,482</td>
<td>5,342,547</td>
<td>11.5%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 0h</td>
<td>Rep2</td>
<td>7,529,177</td>
<td>4,897,307</td>
<td>65.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>8,657,300</td>
<td>6,162,922</td>
<td>71.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep4</td>
<td>8,805,668</td>
<td>6,081,456</td>
<td>69.1%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 3h</td>
<td>Rep1</td>
<td>7,474,239</td>
<td>3,307,412</td>
<td>44.3%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 24h</td>
<td>Rep2</td>
<td>5,277,712</td>
<td>2,530,215</td>
<td>47.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>2,567,457</td>
<td>1,614,888</td>
<td>62.9%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep4</td>
<td>1,025,038</td>
<td>660,045</td>
<td>64.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep5</td>
<td>22,538,102</td>
<td>14,363,908</td>
<td>63.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor</th>
<th>Time point</th>
<th>Cell Line</th>
<th>Total read #</th>
<th>Unique #</th>
<th>Yield %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>G1E</td>
<td>Rep1</td>
<td>8,623,055</td>
<td>6,116,733</td>
<td>70.9%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 0h</td>
<td>Rep2</td>
<td>5,069,351</td>
<td>3,650,832</td>
<td>72.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>5,098,704</td>
<td>3,283,999</td>
<td>64.4%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 3h</td>
<td>Rep1</td>
<td>5,092,403</td>
<td>3,602,146</td>
<td>70.7%</td>
</tr>
<tr>
<td></td>
<td>G1E-ER4 24h</td>
<td>Rep2</td>
<td>12,996,249</td>
<td>8,713,782</td>
<td>67.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rep3</td>
<td>5,530,386</td>
<td>3,960,887</td>
<td>71.6%</td>
</tr>
</tbody>
</table>
Determination of binding locations using intersection of peak-pairing and MultiGPS, Related to Figure 1

Prior to peak-pairing, tags from biological replicates were merged after demonstrating their reproducibility. During peak-pairing, sequence read distributions was used to identify peaks using the strand-separate peak-calling algorithm in GeneTrack (parameters: sigma = 5, exclusion zone = 10) [32]. After peaks within the blacklist regions were removed [33, 34], peaks were paired if the “+” strand peak was within upstream 5 bp or downstream 25 bp of “-” strand peak. Peak pairs that were enriched >2 fold over input control with q-value < 0.05 were selected (q-values are adjusted p-values from binomial test for multiple hypothesis testing). To collect binding locations that are present in one or more time points (“union” of binding locations), binding locations within 40 bp were determined as bound in multiple time points, while further than 40 bp were determined as time point-specific binding.

MultiGPS is designed to detect binding locations across multiple conditions while characterizing differential binding between conditions [35]. MultiGPS detected binding locations across multiple conditions with reads enriched >2 fold over input control and q-value < 0.05. MultiGPS command: --geninfo mm9.info --threads 4 --q 0.05 --d reb1_chipexo.distrib.txt --exclude blacklist.bed --design design_gata1 --verbose --probshared 0.99 --poisssongausspb --medianscale --prlogconf -5 --memepath usr/bin --mememinw 6 --mememaxw 16 --seq mm9.

Homotypic clustering of transcription factor binding analysis, Related to Figure 1

Occurrences of the distance between adjacent binding location midpoints were calculated. Binding location midpoints that were closer than 500 bp from the adjacent binding locations were clustered as one binding unit, homotypic cluster of transcription factor binding, while the median genomic coordinate was taken as the binding location. When a binding location
was further than 500 bp away from nearby binding location, it was considered as a non-cluster. Closest transcription start site (TSS) of mm9 RefSeq gene was assigned to the binding location as their target gene.

**Binding around WGATAA motifs, Related to Figure 2 and 3**

MEME was performed on 80 bp sequences surrounding the top 500 highly occupied binding locations of GATA1 and TAL1. Occurrences of top motifs were scanned across the binding locations using FIMO (p-value < 10^-3), to classify the binding locations by the presence of cognate motifs. Binding locations were centered on the most significant motifs if a motif was present within 40 bp distance. When a motif was present multiple times, the most significant motif (more significant p-value) was chosen. Figures displaying strand-specific sequencing tags of merged time points represent the merged raw data without normalization.

When sorting the binding locations by occupancy around WGATAA motif, GATA1 occupancy was measured as tags within 25 bp upstream to 25 bp downstream, while TAL1 occupancy was calculated as tags within 40 bp upstream to 30 bp downstream.

When 7,929 GATA1 binding sites with a WGATAA motif were sorted by TAL1 occupancy, approximately one third (2,997 locations) showed co-occupancy of GATA1 and TAL1, while remainder (4,930 locations) showed only GATA1 occupancy. Then frequencies of dinucleotide TG distance from WGATAA were calculated for both GATA1 and TAL1 co-occupied sites and GATA1 only bound sites. Among the top 2,997 GATA1 and TAL1 co-occupied WGATAA sites, TG was present at 1,650 sites between 7-9 bp distance from WGATAA. To examine the sequence composition of a full E-box (NNNNTG), the 4 upstream sequences of 1650 TG sites were examined.
TAL1 only binding locations around E-box, Related to Figure 3-5

Co-occupancy of GATA1 and TAL1 binding locations were determined if binding location midpoints were within 40 bp. When TAL1 only binding locations with WGATAA motif (FIMO p-value < 10^{-3}) were removed, the remainder sequences were most enriched with the E-box, CAGMTG motif. Then these binding locations were further classified by the presence and absence of an E-box motif within 40-bp of the binding location midpoint.

ChIP-exo and ChIP-seq comparison, Related to Figure 4

Overlap between binding locations of ChIP-exo and ChIP-seq for the same factor in identical cell types was examined. Binding location midpoints lying within 40 bp distance were defined as overlap. As a measure of consistency, Spearman correlation of log10 transformed read counts of ChIP-exo and ChIP-seq within 400 bp window surrounding binding locations midpoints of all ChIP-seq (n=22,729) and ChIP-exo (n=10,168) were calculated. A heat-map scatter plot was drawn using LSD package of R. Lastly, distribution of GATA1 G1E-ER4 3h ChIP-exo reads (total tag count normalized) around ChIP-seq only bound locations (n=14,071) were plotted.

Differential occupancy of GATA1 across time points, Related to Figure 5

EdgeR was run internally along with MultiGPS to call differential binding events between time points. K-means clustering was performed on log2 fold change of GATA1 occupancy between G1E-ER4 3h/0h, 24h/3h and 24h/0h. For analyses for occupancy levels, data were normalized such that the total tags counts are equal across all time points of a given factor and mapped around the binding location midpoints. Binding locations were assigned to the
closest annotated RefSeq gene to calculate their gene expression fold change between G1E-ER4 30 h over 0 h, and their distances to transcription start site.

**Abbreviations**

ChIP-exo: Chromatin immunoprecipitation followed by 3’-5’ exonuclease treatment then massively parallel DNA sequencing;  
ChIP-seq: Chromatin immunoprecipitation followed by massively parallel DNA sequencing;  
IUPAC: International Union of Pure and Applied Chemistry
Supplementary Figure 3-1. Workflow of ChIP-exo data analysis. Confident binding locations were determined by obtaining the intersection of union of significant peak pairs from multiple time points (Left), and binding events of MultiGPS (Right). Motif discovery analysis using MEME and MEME-ChIP showed enrichment of WGATAA, E-box, and composite half E-box/WGATAA motifs on these stringent binding locations validated by both methods. To classify binding locations by presence of motifs, further motif occurrences and downstream analyses were conducted.
Supplementary Figure 3-2. Binding locations detected by only one method, only peak pairing or MultiGPS. Browser shot of 5’-end of TAL1 ChIP-exo tags surrounding the binding locations detected by only peak pairing (Top left panel) or MultiGPS (Top Right panel). Binding locations called by union of peak pairing and MultiGPS are shown in the middle panel, while binding locations in intersect (Blue) and outersect (Red) of two methods are highlighted. WGATAA and E-box motifs are shown in black.
Supplementary Figure 3-3. TAL1 and GATA1 co-binding around WGATAA sites

Distribution of 5’ ends of GATA1, TAL1 ChIP-exo sequencing tags around 7927 WGATAA sites in GATA1 binding locations. Rows are linked across all panels and sorted by TAL1 occupancy of merged data of all time points. TAL1 occupancy is calculated by the sum of tags from 40 bp upstream and 30 bp downstream from motif reference point (WGATAA). The far-left panel reports the nucleotide composition of 25 bp upstream and downstream regions from GATA motif reference point.
Supplementary Figure 3-4. Tag distribution of GATA1 and TAL1 sorted by occupancy of individual peak of 4 peak patterns. GATA1 and TAL1 tag distribution around WGATAA sites sorted by the occupancy of individual peak of GATA1 four-peak locations. Tags in the same strand of motif (Blue) and opposite strand (Red) and sequence composition plot were sorted by occupancy of sense peak 1 (defined as 0 to 25 bp upstream of WGATAA reference point, Top left), sense peak 2 (0 to 25 bp downstream, Top right), antisense peak 1 (25 bp upstream and 5 bp downstream, Bottom left), and antisense peak 2 (5 to 25 bp downstream, Bottom right).
Supplementary Figure 3-5. Sequence composition around crosslinking sites when sorted by GATA1 crosslinking level in individual peak of 4 peak patterns. Relationship of nucleotide composition in -18 bp to +3p from WGATAA reference point and GATA1 crosslinking level in
individual peak of GATA1. For each position, the nucleotide compositions in low to high (Left to right) crosslinking of individual peak are shown through 4 adjacent bars.

Supplementary Figure 3-6. Mouse Genome Informatics Phenotype Ontology results from GREAT for the binding locations in all clusters in Figure 5. Binding locations were assigned to their single nearest gene for gene ontology analysis.
Chapter 4

ChIP-exo binding locations using peak shape

The coupling of chromatin immunoprecipitation (ChIP) and next generation sequencing (NGS) has revolutionized the genome wide mapping of protein-DNA interaction. Many algorithms and softwares have been developed to call “peaks” or binding locations, which are determined by identifying statistically significant reads enriched over background. Using ChIP-seq data, several studies have shown that transcription factor binding detection can be improved by utilizing the peak shape information (Hower, Evans et al. 2011, Mendoza-Parra, Nowicka et al. 2013, Wu and Ji 2014). Furthermore, changes of peak shape allowed detecting differential bound peaks (Schweikert, Cseke et al. 2013), and clustering by different peak shape speculated that there maybe biological meaning (Cremona et al., 2014). However, progresses in classifying the binding locations by peak shape and further extracting biological function have been limited by the ~200-300 bp resolution of ChIP-seq.

With high resolution ChIP-exo, complex patterns of binding and crosslinking have become more pronounced. Different peak patterns may reveal various mode of binding, such as recruitment of various cofactors. Sharpness of peak shape around a motif may reflect the strength or specificity of the motif, and the symmetry of peak may reflect the symmetry of protein binding. Furthermore, ChIP-exo peaks may reflect the pattern of crosslinking. Chapter 4 describes various exploratory analyses regarding peak shape and patterns of ChIP-exo data.

Interpreting the multiple peaks of ChIP-exo

The essence of ChIP-exo analysis is to identify the peaks on both strands to define a single binding location. During the conventional data analysis, a peak pair comprised of a single
peak on the positive strand and another on the negative strand defines the point of crosslinking. The midpoints of final remaining binding locations indicate the center locations of the binding events. Short WGATAA motif was identified from these GATA1 binding locations using MEME (Multiple Em for Motif Elicitation) (Bailey and Elkan, 1994, Crooks et al., 2004) as expected. When a composite plot was aligned to the WGATAA motif sites, surprisingly, multiple peaks were observed rather than a single peak pair (Figure 4-1).

![Figure 4-1: Summed tag distribution around GATA1 motif. Multiple peak patterns are observed around motif midpoints.](image)

This was contrasting to the single peak pair pattern that has been observed in several transcription factors of *S. cerevisiae*, including Reb1 (Rhee and Pugh 2011) (Figure 4-2A). However, it confirmed the multiple peak patterns, more specifically four peaks around the motifs, that had been reported from CTCF ChIP-exo of human (Rhee and Pugh 2011). (Figure 4-2B, C). In the case of p53, six peak pattern was found, where 4 peaks from each half site is overlapping, and two internal peaks are adjoining with each other (Chang, Chen et al. 2014) (Figure 4-2D).
Figure 4-2: (A) ChIP-exo tags around Reb1 binding sites centered by the motif midpoint. Figure 2A from (Rhee and Pugh 2011). (B) ChIP-exo sequencing tags around CTCF-bound locations centered by the motif midpoint. Figure 6B from (Rhee and Pugh 2011). (C) Model for presence
We hypothesize that these multiple peaks around the binding site may result from multiple cross-linking sites during formaldehyde fixation. Furthermore, the consistent distances between the peaks may reflect the effect of the header room space of λ-exonuclease during digestion (Figure 4-2C, 4-3).

Figure 4-3. Diagram explaining the multiple crosslinking and exonuclease digestion from GATA1 pattern

When various peak pair combinations are possible from a population of cells being cross-linked on one or two sites, the composite tag distribution may appear as multiple peaks (Figure 4-4).
We hypothesize that a more complex pattern of multiple small peaks may define a binding event. This raised the need for a fine-grain analysis method that distinguishes the small peaks and considers the various combinations of these peaks. This method may enable us to obtain a more comprehensive list of factor binding locations from the ChIP-exo data (Figure 4-5).

Moreover, high resolution of the precise multiple peak patterns and crosslinking points not only provides the transcription factor binding locations, but also the DNA sequence base that...
the crosslinking and binding prefers. Furthermore, the different peak pattern may reveal a possible different binding mechanism through other cofactors.

![Diagram of ChIP-seq and ChIP-exo methods](image)

**Figure 4-5.** Goal of examining multiple peak pattern of ChIP-exo. ChIP-exo provides higher precision compared to ChIP-seq. More comprehensive list of binding locations maybe detected using multiple peak pattern compared to conventional peak pairing.

The goal of Chapter 4 is to develop a bioinformatics method to determine a comprehensive list of transcription factor binding locations using the multiple peak patterns of ChIP-exo. Furthermore, we set out to classify the binding locations by the combinations of multiple peaks, to examine whether difference in peak patterns shed light on different chromatin functions and transcriptional mechanisms. To achieve this, two methods have been explored: quad method and peak pair clustering.
Quad Method

Quad refers to the four intervals that are determined from the summed tag distribution around the motif (Figure 4-5, Left). Quad method uses these quads as a “template” to examine the peak pattern and to determine the number of peaks that are present among the four peaks.
Algorithm & Assumptions

There are several assumptions of the quad method. 1) Quad method is a motif-independent approach, where the template is overlaid on the peak with the higher read counts (“reference peak”), and giving them higher priority to be called as binding locations. 2) The relative distances between four quads and size of quads are determined by the summed tag distribution. 3) The quad template can be overlaid on the reference peak in two possible ways, where the reference peak can be the upstream reference peak (URP) or the downstream (DRP). 4) By comparing the two possible quad sets with statistical measures, such as sum of read counts within the quad or correlation of tags to the quad template, one quad set is chosen and the binding is defined. 5)
Previously called binding locations cannot be called twice, thus, the next highest peak will be searched and previous steps will be repeated.

**Result**

Quad method was applied to a ChIP-exo data set that resulted in 871,351 GeneTrack peaks (parameters sigma = 5, exclusion zone = 5). Among these peaks, 62,899 binding locations (7.2%) with no reads in three adjacent quads were filtered out. When quad size was set to 14 bp, the quad method called 24,664 binding locations. When the overlap with ChIP-seq data, 48.9% of ChIP-exo overlapped with ChIP-seq (Figure 4-5).

![Diagram](image)

Figure 4-8. Overlap of ChIP-seq peaks with ChIP-exo peaks from quad method

**Limitations**

Although ChIP-exo peaks from quad method may determine binding locations with
multiple peak pattern, it suffers from the stringent fixed distances between quads. Furthermore, quad sizes and relative distances among quads can be subjective. To address these limitations, peak pair clustering (PPC), a method that allows flexible distances among quads was employed.

Peak Pair Clustering (PPC)

The goal of peak pair clustering (PPC) method is to obtain a comprehensive and accurate ChIP-exo binding locations, while allowing flexible distance range among multiple peaks. Furthermore, we hypothesize that different peak patterns may relate to different chromatin functions.

Algorithm

We hypothesize that multiple peak patterns of ChIP-exo arise from the lambda exonuclease digestion after the multiple crosslinking points between protein and DNA. First round of peak pairing in a narrow window detects a pair of peak around a single crosslinking point, while an orphan peak is often discarded during conventional peak pairing method. When there are two crosslinking points between protein and DNA, 2~4 peaks may arise. To detect these, another iteration of pairing, but pairing of crosslinking points, is performed. Four peaks may arise from two sets of peak pairs (PP2), while three-peak pattern is comprised of a peak pair and an orphan peak (PPO). Lastly, a two-peak pattern may be composed of a pair of orphan peaks (O2) or a peak pair around only one crosslinking point (PP1).
Figure 4-9: Possible peak patterns from two crosslinking sites of protein binding
Figure 4-10: Algorithm of Peak Pair cluster (PPC) method. **Step 1** - Solid red and blue peaks indicate the peaks that were paired after C-W peak pair distance calculation. Dashed C peaks indicate the possible peaks that were not chosen as peak pairs. **Step 2** – The distance between cross-linking sites are calculated using each as a reference cross-linking site (green circle in bold). In this diagram, the arrow was calculated as the mode of doublet pair distance. **Final Outcome** – Gray box indicates the region with the paired crosslinking points that forms a four-peak pattern (PP2). The final reference coordinate of the binding location is indicated as the inverted black triangle.
Peak pair cluster method takes the multiple cross-linking sites and ~12bp distance from λ-exonuclease header room into account. More specifically, during the peak pairing step (Figure 4-7, Step 1), by starting with the W peak with the highest number of reads, C peaks that are approximately ~12 bp ± 6 bp away are searched. After calculating the peak pair distances genome-wide, the most frequent peak pair distance (which is peak pair distance mode) is obtained. Then for each W peak, the C peak that has distance most closer to this mode is chosen as a match, and these W peak and C peak form a peak pair, indicating a single crosslinking point. Then, the distances between these crosslinking points are calculated throughout the genome, to calculate the mode of doublet pair distance (Figure 4-7, Step 2). Then the two cross-linking sites that have the distance closest to the mode (D) are grouped to form a pair of crosslinking points (or pair of peak pairs). The midpoint of the two crosslinking points are assigned as the coordinate for the binding location. Through this peak pair clustering method, the multiple peaks are grouped while allowing more flexibility between the distances between the two crosslinking points, which is improved from the quad method.

Results & Limitations

Peak pair clustering (PPC) was performed on GATA1 RefSet, which is a merged data set of all time points of G1E-ER4 cells, including 0 hour, 3 hour and 24 hours. We hypothesized that different peak patterning may relate to different biological features, such as different sequence motifs.

However, regardless of the peak pattern, all types of peak patterns resulted in the main WGATAA motif. Furthermore, when examining individual binding locations that are determined as 4 peaks (PP2), 3 peaks (PPO), 2 peaks (O2, PP1), the expected shapes were not observed in individual loci.
Figure 4-11: Tag distribution around GATA1 four peak (PP2) and three peak (PPO) reference points. Motif indicates the MEME logo from all peak pair clusters, including all possible 2~4 peaks. Also, depending on the orphan peak location, there are 4 possible PPO configurations. Therefore, when PPO are combined, the peak pattern seems like a four-peak pattern, when it is a combination of various orphan peak orientation.
Sorting with individual peak among four peaks

The uniqueness of ChIP-exo peak shape is that the peak patterns indicate the combination of crosslinking points. Through the various combinations of four-peak pattern, we can infer the TF’s preference of crosslinking point and the sequence nucleotides that are involved.

We sought out to examine whether the four-peak pattern from the composite plot is consistent throughout the genome, and whether the individual peaks among the four are linked to one another. To address this, GATA1 and TAL1 tags were sorted by GATA1 occupancy level of each of the four peaks separately (Figure 4-8). GATA peaks surrounding the crosslinking point on the upstream of WGATAA motif were mostly stronger than the downstream crosslinking point. Although the positioning of crosslinking point did not change, the level of occupancy varied among binding sites.
When the nucleotide sequences in the vicinity of crosslinking points were examined, the deviation was minimal compared to the overall average, when comparing high versus low GATA1 occupied sites.

Figure 4-12: Tag distribution of GATA1 and TAL1 sorted by strength of individual peak of four peak patterns
Multiple peak patterns of ChIP-exo provide insights into the precise crosslinking points of the protein and the nucleotide compositions in the vicinity, which was not possible with previous technologies. However, the binding location determination using multiple peak patterns did not show difference in sequence motif or possible cofactors.

**Summary**

Figure 4-13: Sequence composition around crosslinking sites when ordered by occupancy level of GATA1 in individual peak among four peaks. (Left) Sorted by occupancy of sense 1 peak, (Right) sorted by occupancy of sense 2 peak.
**Chapter 5**

**Conclusion**

Gene regulation, the process from genetic material, DNA, to the final product, protein, is regulated in various levels and layers. These regulations are essential, as misregulation and misreading are lethal, and the first step of gene regulation, transcription, is the most regulated. Transcription in eukaryotes is performed by three polymerases, which splits their labor. These polymerases are ignorant of the DNA templates, but are made aware by the binding of transcription factors. Therefore, transcription factor binding has a critical role in gene expression, thus, knowing their precise binding locations is important. The advance from single gene studies to high-throughput next generation sequencing (NGS) technology was a revolution in the biomedical field. This allowed the technical advance of genome-wide survey of transcription factors binding, transforming from ChIP-chip to ChIP-seq, then to ChIP-exo, which is a mega base pair to base pair clarity of protein binding boundaries. ChIP-exo is part of this endeavor for achieving higher resolution of binding information, which will allow us to achieve the closer dissection of the mechanism of transcription.

**Significance and Contributions**

Significant findings in this dissertation and their contributions to the scientific field:

1. Generation of high resolution ChIP-exo binding data for erythroid transcription factor (GATA1, TAL1) and general transcription factor (TFIIB, PolII) in mouse genome during erythroid development.

2. The ChIP-exo analysis method employed on an erythroid development model in the mammalian genome serves as a guideline for general ChIP-exo analysis.
Comparative analysis of peak pairing and MultiGPS methods, not only allows examining pros and cons of various DNA-protein binding analysis platforms, but also provides insights into ways of improving and developing novel computational methods.

3. Scripts and tools developed can be applied to studies of transcription factors involved in myriad of other biological questions. Furthermore, the concepts and methodologies that have been used for developing ChIP-exo analysis tools, such as parallel processing, can be applied to tackling bioinformatics problems in various biomedical fields, such as stem cell and cell therapies, cancer genomics and personalized medicine.

4. Conceptual considerations for ChIP-exo of erythroid factors in mammalian genome can be applied to the general analyses pipeline for ChIP-exo. For example, observation of prevalent black list regions in the mammalian genome compared to yeast genomes have allowed us to consider and implement the blacklist removal into ChIP-exo analyses pipeline.

5. High resolution of ChIP-exo dissects the precise positioning of GATA1 and TAL1, members of same multiprotein complex. The precise ~8bp distance between the proteins is evident through the ChIP-exo data, which was not possible through ChIP-seq.

6. Presence of half E-box (TG) on upstream of WGATAA was shown as sequence determinants for co-occupancy of GATA1 and TAL1. This was achievable by determining of driving motif in a narrower ChIP-exo binding location window.

7. High order of homotypic clusters of transcription factor binding sites (HCT), clusters containing multiple sites for one particular transcription factor, was
highlighted through ChIP-exo. Many previously identified broad ChIP-seq binding locations were dissected into numerous distinct binding locations forming a HCT.

8. Multiple types of -omic data were incorporated additional to ChIP-exo of GATA1 and TAL1 to study the transcriptional regulation of erythropoiesis. It is challenging to comprehend the behavior of cell with a single type of analysis, when result can be highly context-dependent. Thus, integrative analyses across multiple layers of data encompassing genomics, epigenomics, transcriptomics, proteomics, metabolome and phenome allow powerful genotype-phenotype connections (Ritchie, Holzinger et al. 2015).

Figure 5-1: Integration of multi-omic data: Multiple types of –omic data allows deeper understanding of biological mechanisms, such as transcriptional regulation. Study in Chapter 3 incorporated various types of data (italicized in figure), including ChIP-exo, RNA-seq, histone modification, and DNase-seq (generated by ENCODE).
Future Direction

High resolution profiling of transcription factor binding in erythroid development can serve as a pilot study for a more comprehensive list of proteins. Precise positioning of all protein members in the multiprotein complex can be explored by expanding the study to more number of members, such as LDB1, LMO2, KLF1, and FOG1. The scope of the study can be expanded from erythropoiesis to hematopoiesis, the cell differentiation of blood stem cells to various blood cellular components, including erythroid cells, lymphocytes and myelocytes. Furthermore, high throughput enhancer screening assays can used to validate which the binding locations and/or motif sites are functional.

Optimal methods on bioinformatics analysis of ChIP-exo data can be explored by expanding from the current comparative analysis of peak pairing and MultiGPS to numerous softwares. Currently available ChIP-exo analyses softwares include MACE (model based analysis of ChIP-exo) (Wang, Chen et al. 2014), GEM (Genome wide Event finding and Motif discovery) (Guo, Mahony et al. 2012), CexoR (Strand specific peak-pairing calling in ChIP-exo replicates) and more.
References


E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J.
MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul and J. A. Stamatoyannopoulos


Garam (Celine) Han

465 North Frear Building
University Park, PA 16802, USA

gzh105@psu.edu

EDUCATION

Ph. D. in Integrative Biosciences (IBIOS) 2008 - Present
Bioinformatics and Genomics Option
The Pennsylvania State University
Faculty Advisor: Dr. B. Franklin Pugh

B.S. in Science - Life Science Option 2004 - 2008
Eberly College of Science, The Pennsylvania State University

PUBLICATION & PRESENTATIONS AT MEETINGS


Talk at 3rd Bioinformatics and Genomics Retreat. Insights into Transcriptional Regulation of Erythropoiesis via High Resolution ChIP-exo. September 2013, University Park, PA.

Lecture at National Human Genome Research Institute (NHGRI)-funded National DNA Day event, "Exploring DNA with Penn State". Gave lecture and coordinated workshops on molecular basics of DNA, and human genome project to high school students from 5 school districts of Pennsylvania. Han GC and Rossi M. May 2012, University Park, PA.

Lecture at Introduction to Genomics Course for graduate students. Analyzing Next Generation Sequencing (NGS) data through Web-based Tools: GALAXY, BLAST, ClustalW. Han GC and Park BS. October 2010, University Park, PA.

Han G.C., Hardison R.C., Pugh B.F. A Comprehensive and High Resolution Genome-wide binding of GATA1 and TAL1 during Erythroid Development. Poster Presentation at Cold Spring Harbor Lab (CSHL) Conference on The Biology of Genomes, May 2013, Cold Spring Harbor, NY. (Poster presentation)

Han G.C., Hardison R.C., Pugh B.F. Transcriptional Regulation of Erythropoiesis via High Resolution GATA1 and TAL1 ChIP-exo. Poster at 2nd Bioinformatics and Genomics Retreat. August 2012, University Park, PA. (Poster presentation)

TEACHING EXPERIENCES

Certificate in College Teaching  Fall 2014

Graduate Teaching Assistant – The Pennsylvania State University

BMB 252H – Introduction to Cell & Molecular Biology for Honors students  Spring 2012
Supervision: Dr. B. Frank Pugh (Biochemistry and Molecular Biology)

IBIOS 496 – Introduction to Genomics for Graduate Students  Fall 2010
Supervision: Dr. Ross C. Hardison (Biochemistry and Molecular Biology)

EDITORIAL EXPERIENCE

Editorial Advisory Member for Book “Big Data Analytics in Bioinformatics and Healthcare” by Wang, Li and Perrizzo

ORGANIZATIONS & AFFILIATIONS

Sigma Delta Epsilon Graduate Women in Science (GWIS) Chapter – Officer and Webmaster 2010 - 2011

HONORS, AWARDS & FELLOWSHIPS & GRANTS

Golden Key International Honour Society – Invited member 2009
Dean’s List – The Pennsylvania State University 2004 - 2006