UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

High-throughput genomic assays : applications and analysis of DSL technology and next-generation sequencing

Permalink

https://escholarship.org/uc/item/1wq123m1

Author

Hutt, Kasey Robert

Publication Date

2009

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

High-throughput genomic assays: Applications and analysis of DSL technology and next-generation sequencing

A Dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics

by

Kasey Robert Hutt

Committee in charge:

Professor Michael G. Rosenfeld, Chair

Professor Terry Gaasterland, Co-Chair

Professor Xiang-Dong Fu

Professor William McGinnis

Professor Pavel Pevzner

2009

Copyright

Kasey Robert Hutt, 2009

All rights reserved.

The dissertation of Kasey Robert Hutt is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2009

To my parents who have always urged me to better myself.

To my brothers who have always tried their best to be better than me.

To my dog who has never cared about doing anything better.

Whilst he sits on the cushion of advantages he goes to sleep. When he is pushed, tormented, defeated, he has a chance to learn something;

Ralph Waldo Emerson

TABLE OF CONTENTS

Signature Page ..... **iii**

Dedication ..... **iv**

Epigraph ..... **v**

Table of Contents ..... **vi**

List of Figures ..... **ix**

List of Tables ..... **xi**

Acknowledgements ..... **xiii**

Vita ..... **xvi**

Publications ..... **xvi**

Abstract of the Dissertation ..... **xviii**

1 Introduction **1**

2 DSL Applications and Analysis **4**

2.1 Background 4

2.2 Methods 7

2.2.1 Probe Design 7

2.2.2 Tiling Probe Design 8

2.2.3 ChIP, 3C, and 3D Descriptions 9

2.2.4 Microarray Analysis 11
2.2.5 3D Microarray Analysis 12
2.2.6 Motif Analysis 13
2.3 Results 13
2.3.1 ChIP-DSL of ERα on Human Promoter Array 13
2.3.2 ChIP-DSL of ERα Targets on Human Tiling Array 15
2.3.3 3D Analysis Using Human Tiling Arrays 17
2.3.4 ChIP-DSL of LSD1 on Human Promoter Array 19
2.4 Conclusions 20
3 High-throughput Sequencing and Analysis 37
3.1 Background 37
3.2 Methods 40
3.2.1 Genome Analyzer Description 40
3.2.2 Mapping Tags 42
3.2.3 Peak-Finding 43
3.2.4 Motif Analysis 46
3.2.5 Genome Ontology Analysis 46
3.3 Results 50

3.3.1 ChIP-seq of ERα 50

3.3.2 ChIP-seq of Androgen Receptor (AR) 52

3.3.3 ChIP-seq of FOXA1 55

3.3.4 ChIP-seq of REST/NRSF 56

3.3.5 Analysis of Histone Modifications in $CD4^+$ T-cells 59

3.4 Conclusions 62

4 Future Work 104

4.1 3C, DSL and High-throughput Sequencing Combined 104

4.2 Genome Ontology Improvements 106

References 117

LIST OF FIGURES

Figure 2.1 Workflow diagram of ChIP-on-chip 22

Figure 2.2 DSL probe generation 23

Figure 2.3 3C workflow schematic 24

Figure 2.4 ERα log-log intensity scatterplot 25

Figure 2.5 ChIP-DSL of positive marks surrounding *TFF1* 26

Figure 2.6 ChIP-DSL of positive marks surrounding *GREB1* 27

Figure 2.7 ChIP-DSL of positive marks surrounding *RARβ* 28

Figure 2.8 3D interaction map around *TFF1* 29

Figure 2.9 3D interaction map around *GREB1* 30

Figure 2.10 3D interaction map around *CASP7* 31

Figure 2.11 3D interaction map around *DIO1* 32

Figure 2.12 LSD1 log-log intensity scatterplot 33

Figure 2.13 Pol II log-log intensity scatterplot 34

Figure 3.1 Schematic of dye-terminating sequencing 67

Figure 3.2 Dye-terminating sequencing output 68

Figure 3.3 Genome Analyzer adapter ligation and amplification 69

Figure 3.4 Schematic of cluster generation 70

Figure 3.5 Schematic of GA sequencing reaction 71

Figure 3.6 Image capture of a single base within one sequencing cycle 72

Figure 3.7 TreeView output of Genome Ontology clustering..... 73

Figure 3.8 TreeView output of Genome Ontology clustering: activation marks. 74

Figure 3.9 TreeView output of Genome Ontology clustering: silencing marks 75

Figure 3.10 TreeView output of Genome Ontology clustering: FOXA1, ERα, and AR 76

Figure 3.11 TreeView output of Genome Ontology clustering: REST 77

Figure 3.12 TreeView output of Genome Ontology clustering: LSD1 78

Figure 3.13 TreeView output of Genome Ontology clustering: tri-methyl-H3K36..... 79

Figure 4.1 UCSC genome browser of PHF8 ChIP-seq over *ERα* locus 111

Figure 4.2 UCSC genome browser of PHF8 and tri-methyl-H4K20 ChIP-seq over *ERα* locus..... 112

Figure 4.3 UCSC genome browser of CLIM ChIP-seq over *BCAS1*..... 113

Figure 4.4 UCSC genome browser of CLIM ChIP-seq over an expanded region covering *BCAS1* 114

Figure 4.5 UCSC genome browser of CLIM ChIP-seq over *BCAS3*..... 115

LIST OF TABLES

Table 2.1 Human and HSV DSL coverage 35

Table 2.2 Mouse DSL coverage 36

Table 3.1 HOMER motif-finding of ERα ChIP-seq proximal peaks 80

Table 3.2 HOMER motif-finding of ERα ChIP-seq distal peaks 81

Table 3.3 HOMER motif-finding of ERα ChIP-on-chip data [50] 82

Table 3.4 HOMER motif finding of ERα ChIP-on-chip peaks not overlapping with ERα ChIP-seq peaks 83

Table 3.5 Genome Ontology of ERα ChIP-seq 84

Table 3.6 HOMER motif-finding of AR +DHT ChIP-seq distal peaks 85

Table 3.7 HOMER motif-finding of AR +DHT/si*FOXA1* ChIP-seq distal peaks 86

Table 3.8 Genome Ontology of AR +DHT ChIP-seq 87

Table 3.9 Genome Ontology of AR +DHT/si*FOXA1* ChIP-seq 88

Table 3.10 HOMER motif-finding of FOXA1 -E_2 ChIP-seq distal peaks 89

Table 3.11 HOMER motif-finding of FOXA1 +E_2 ChIP-seq distal peaks 90

Table 3.12 HOMER motif-finding of FOXA1 -DHT ChIP-seq distal peaks 91

Table 3.13 HOMER motif-finding of FOXA1 +DHT ChIP-seq distal peaks 92

Table 3.14 Genome Ontology of FOXA1 –E_2 ChIP-seq 93

Table 3.15 Genome Ontology of FOXA1 +E_2 ChIP-seq 94

Table 3.16 Genome Ontology of FOXA1 -DHT ChIP-seq 95

Table 3.17 Genome Ontology of FOXA1 +DHT ChIP-seq 96

Table 3.18 HOMER motif-finding of REST NPC ChIP-seq distal peaks 97

Table 3.19 HOMER motif-finding of REST dNPC ChIP-seq distal peaks 98

Table 3.20 HOMER motif-finding of REST 293T ChIP-seq distal peaks 99

Table 3.21 HOMER motif-finding of REST enriched in NPC versus dNPC. 100

Table 3.22 HOMER motif-finding of REST enriched in dNPC versus NPC. 101

Table 3.23 Genome Ontology of Pol II and gene-centered annotations 102

Table 3.24 Genome Ontology of tri-methyl-H3K36 and gene-centered annotations 103

Table 4.1 Results of aligning PHF8 sequencing data to vector library..... 116

ACKNOWLEDGEMENTS

The first person I would like to thank is my advisor, Dr. Geoff Rosenfeld. His ceaseless drive towards scientific discovery is truly astounding, and is only matched by his near-addiction for perfection. I doubt I will ever meet another person who works as hard for the pursuit of truth, and I know my life is all the better for being a part of his lab for the past five years.

I would like to thank my committee: Dr. Xiang-Dong Fu, Dr. Terry Gaasterland, Dr. William McGinnis, and Dr. Pavel Pevzner. They have certainly endured my many shortcomings as a scientist for the selfless purpose of teaching. Their contributions and criticisms have provided me with a great sense of humility and confidence, and I hope to someday pass on all they have taught me.

I am greatly indebted to several individuals with whom I have collaborated with. Dr. Xiang-Dong Fu is a supremely talented technologist, without whom I may never have even known about next-generation sequencing first-hand. Dr. Young-Soo Kwon, an incredible researcher in Dr. Fu's lab, has provided me with a constant supply of high quality data, without which my analyses would probably not have been successful. Above and beyond mere data, Dr. Kwon has always been a brilliant researcher, and every scientific conversation I have had with him has led to important insights. Dr. Victoria Lunyak, a former member of the Rosenfeld lab, was basically my second mentor, and I am still in

awe of her ingenious foresight when it came to future research directions. Her creativity could probably only be equaled by her ferocity at performing sound science, and I truly appreciate her demanding work ethic. Dr. Ivan Garcia-Bassets, a current Rosenfeld lab member, has been a tireless scientist, and has generated a large portion of the data not only analyzed here, but in my entire graduate student career in this lab.

I would like to thank the past and present members of the Rosenfeld lab. Though every person has provided me with helpful advice or criticisms, there are some important individuals I need to recognize. Marie Fisher, Dr. Rosenfeld's administrative assistant (actually, more like administrative project manager), has been the most helpful person I have ever met, and she is like this with everyone in lab. That is about forty people, and I cannot imagine how the lab could run without her. Equally important is Kenny Ohgi, the lab manager. He runs nearly every next-generation sequencing sample for everyone in lab (along with the help of Jane Zhang), and he is still the lab manager and must deal with the constant supply of reagents for forty different people. Not only is Kenny a major source of data for my analyses, but I count him as a good friend as well, and someday, I will beat him in a one-on-one basketball game. Esperanza Núñez is a fellow graduate student and very good friend. She is a constant source of happiness, and surely, without her, my graduate student life would have been drastically different and boring, and I may have never come to appreciate coffee as much as I do now.

I would like to thank all my friends I have met throughout graduate school, especially Chris Benner, Kristine Briedis, Michele Day, Sven Heinz, Eugene Ke, Coleman Mosley and Joshua Stender. Without fail, at least one of these people is good for a laugh, 24 hours a day, 7 days a week.

Finally, I would like to thank my family for their love and support throughout this arduous journey. My parents, despite not telling me that I would be in school for 25 years when they first sent me to kindergarten, are the best parents I could ever ask for, especially since I am their favorite son. I must also thank my brothers for putting up with this fact. But I must also thank them for always being there for me, no matter the circumstances. I hope they know I am always here for them.

VITA

2001 Bachelor of Science, Biochemistry
University of New Mexico, Albuquerque, NM

2004 Teaching Assistant, University of California, San Diego
"Applied Genomic Technologies"
Bioengineering 183

2006 Teaching Assistant, University of California, San Diego
"Functional Genomics",
Bioengineering 203

2009 Doctor of Philosophy, Bioinformatics
University of California, San Diego

PUBLICATIONS

- Garcia-Bassets, I., Y.S. Kwon, F. Telese, G.G. Prefontaine, **K.R. Hutt**, C.S. Cheng, B.G. Ju, K.A. Ohgi, J. Wang, L. Escoubet-Lozach, D.W. Rose, C.K. Glass, X.D. Fu, and M.G. Rosenfeld, *Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors.* Cell, 2007. **128**(3): p. 505-18.

- Kwon, Y.S., I. Garcia-Bassets, **K.R. Hutt**, C.S. Cheng, M. Jin, D. Liu, C. Benner, D. Wang, Z. Ye, M. Bibikova, J.B. Fan, L. Duan, C.K. Glass, M.G. Rosenfeld, and X.D. Fu, *Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters.* Proc Natl Acad Sci U S A, 2007. **104**(12): p. 4852-7.

- Lunyak, V.V., G.G. Prefontaine, E. Nunez, T. Cramer, B.G. Ju, K.A. Ohgi, **K. Hutt**, R. Roy, A. Garcia-Diaz, X. Zhu, Y. Yung, L. Montoliu, C.K. Glass, and M.G. Rosenfeld, *Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis.* Science, 2007. **317**(5835): p. 248-51.

- Hu, Q., Y.S. Kwon, E. Nunez, M.D. Cardamone, **K.R. Hutt**, K.A. Ohgi, I. Garcia-Bassets, D.W. Rose, C.K. Glass, M.G. Rosenfeld, and X.D. Fu, *Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules.* Proc Natl Acad Sci U S A, 2008.

ABSTRACT OF THE DISSERTATION

High-throughput genomic assays: Applications and analysis of DSL technology and next-generation sequencing

by

Kasey Robert Hutt

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2009

Professor Michael G. Rosenfeld, Chair

Professor Terry Gaasterland, Co-chair

Determining DNA sequence has been a principle tool for several methods in biology research. From whole genome sequencing to RNA expression assays to several types of immunoprecipitation experiments, sequencing DNA has been a staple detection technique. Recent advances in sequencing and detection of DNA has revealed many new possibilities, and problems, with regards to data analysis. Here, I present a study analyzing a novel detection technology and sequencing method. These are both important contributions for not only providing new insights for utilizing a more sensitive detection technique, but also creating a method which enables any researcher to quickly analyze the unheralded amount of sequence data now being produced, soon be available to everyone.

Chapter 2 focuses on the current uses and analysis of a novel DNA detection technique called DNA Selection and Ligation (DSL). By taking advantage of the more sensitive and specific DSL strategy, any assay that is dependent on DNA detection is improved. In this chapter, I show how DSL can be used to modify the standard chromatin immunoprecipitation (ChIP)-on-chip assay (termed ChIP-DSL), in both the promoter-specific and tiling cases. Additionally, I show how the ChIP-DSL method gives promising results for a high-throughput version of the chromatin conformation capture (3C) assay, which is used to measure if two regions of DNA are interacting.

Chapter 3 concentrates on next-generation sequencing, more specifically on the Illumina Genome Analyzer (GA). After describing the details of how the

sequencing is performed and analyzed, I discuss the current flaw in these datasets, and propose a solution to the problem, the Genome Ontology. I then give several examples where the Genome Ontology is helpful in extracting knowledge from these incredibly large datasets.

Chapter 4 describes numerous future directions of my own work as well as several of my observations that resulted from working with these next-generation sequencing datasets.

1 Introduction

DNA sequencing has been indispensable in modern biology for a multitude of assays. Ever since the Sanger chain-termination method [1, 2] was used in conjunction with fluorescently labeled ddNTPs [3, 4], automated sequencing of DNA has continued to advance our knowledge at an ever-increasing rate. Previously, sequencing 24 bases was worthy of publication [5]; however, today we are very close to sequencing an individual human genome for $1000 in the span of several days [6]. This jump in sequencing throughput did not happen overnight as several intermediate technologies had to first be invented.

It was quickly realized that despite the automation and speed of the Sanger-based sequencing method, one major drawback kept it from becoming a high-throughput technology: the inability to process multiple sequences at once. While the technology to accomplish this task was being researched, an indirect method was utilized that could do the same thing: DNA microarrays The first microarray technology was born by spotting cDNA sequences onto filter paper [7], and the race for more advanced microarrays began. Today, we have microarrays that can cover entire mammalian genomes across several arrays with tiling probes, or can cover an entire transcriptome in a single slide.

As useful as microarrays are, they are limited by their dependence on hybridization. DNA complementarity is not an exact science, and every step of the protocol that depends on inexact bonding introduces more and more error. The first example of this bias comes from the PCR amplification of the target DNA sample. Normally accomplished with ligation-mediated PCR, universal primers are ligated to the edges of the DNA fragments for a less-biased amplification. While the primers themselves do not contribute to the error, the processivity of the polymerase changes depending on the sequence content of what is being amplified. Any single-stranded sequence that contains a stable secondary structure can hinder the amplification efficiency, as can polyN tracts. These sequence-specific problems arise again during the hybridization to the microarray itself. Theoretically, these issues are addressed during the analysis; however, no amount of analysis will save a seemingly blank spot from being used as data. The recent technological advancement called DNA, selection and ligation (DSL) aims to solve this problem and the applications and analysis of DSL-modified assays constitutes part of my dissertation.

The latest innovations for sequencing have created a huge opportunity for incredible amounts of research. Termed “next-generation sequencing,” these methods solve the problem of serial DNA sequencing by changing it to parallel DNA sequencing. As with anything new, however, the solved problem is replaced by a new one: How do we analyze all of this data? It is perfectly fine for a single

lab to generate genome-wide datasets, but unless it is easily comparable to other datasets, a researcher could spend months looking for interesting discoveries. To this end, I present my work on the Genome Ontology, a simple statistical framework for assimilating and comparing multiple datasets simultaneously, with specific emphasis on next-generation sequencing data.

2 DSL Applications and Analysis

2.1 Background

A DNA microarray is a hybridization-based, high-throughput detection method for inferring the sequence content of a DNA sample. The fabrication of such arrays is a competitive and well-established field, but is only part of the complete experiment. Any multi-step protocol is only as good as the least-efficient step. This “weakest link” proverb also applies to techniques involving microarrays. Sample preparation is a key component, and is the step where the DSL strategy most heavily improves upon previous designs.

As an example and short introduction to DSL technology, a typical ChIP-on-chip experiment is considered (Figure 2.1). After the cross-linking, shearing, and immunoprecipitation steps at the beginning of a ChIP, an amplification step is required to increase signal from the relatively low abundance of protein-bound DNA. The usual method is ligation-mediated PCR (LM-PCR), which ligates universal primers to both ends of the blunted DNA fragments [8]. Though LM-PCR removes primer bias in the amplification, the high complexity of the fragment sequences, in terms of variable length and nucleotide content, introduces fragment-specific amplification bias. Repetitive sequence, poly-N tracts, and the

common helix-hairpin-helix structure all contribute to non-specific priming between fragments [9].

The DSL method overcomes these biases by giving the experimenter control over the sequence content that is being amplified. Rather than directly amplifying the fragments, a representative 40mer is chosen and is called the DSL probe [10]. However, the DSL probe does not exist as a 40mer initially, it is two, 20mer halves, which are also attached to the universal primers T3 or T7 (Figure 2.2). These halves are designed to hybridize next to each other on the target DNA sequence, which allows for efficient ligation, the purpose of which is to create the full 40mer DSL probe sequence. The T3 and T7 universal primers are then used to amplify only the successfully ligated probes. The importance of these details cannot be understated. The designed 40mer sequence entirely avoids the problem of an amplified region containing the aforementioned sequence biases. The initial separation of the probe followed by ligation also helps to avoid non-specific hybridization by requiring two compatible 20mers to hybridize directly next to each other. The remaining steps of the ChIP-on-chip protocol follow much the same path, except only the 40mer sequences need to be spotted or synthesized on a microarray slide. The DSL method provides increased sensitivity and specificity when applied to the interrogation of DNA sequence, as exemplified in ChIP microarray applications [10-12].

Any DNA sequence extracted from an experiment is a potential target of the DSL method, and a specific set of experiments concerning estrogen receptor alpha (ERα) and lysine specific demethylase 1 (LSD1) is presented in this chapter. ERα belongs to the nuclear receptor (NR) superfamily of transcription factors which, upon ligand binding, becomes activated and up- or down-regulates its target genes [13, 14]. One of the proposed mechanisms of this regulation is via recruitment of histone-modifying factors which change the chromatin landscape surrounding the target gene, either towards activation (euchromatin) or repression (heterochromatin) [15]. ERα is also a key factor in many diseases, especially in breast cancer where ERα is usually overexpressed. At the time of this study, the ERα binding program was extremely limited, and it was of great interest to learn how extensive the binding of ERα was in the genome. The investigation of ERα binding profile on a human promoter-wide DSL array, which contained over 20,000 known promoters, was highly successful in identifying novel gene targets [12]. Subsequent publications then pointed towards a significant non-promoter binding profile in which trans-activation of target genes via intrachromosomal interactions with distal sites was proposed [16]. A tiling DSL array was designed and implemented to investigate this phenomenon. Finally, a novel application of the tiling array was used to query the possibility of long-distance interactions in a higher-throughput manner by combining aspects of the chromosome conformation capture (3C) technique, described later, and DSL technology (termed 3D).

LSD1 was first identified as a component of the corepressor complex, CoREST [17-25], and was soon discovered to be the first example of a new class of enzymes, protein demethylases. Thus far, only histone targets of LSD1 have been identified, specifically H3K4me1/2 (in the case of CoREST) [26, 27], and H3K9me3 (in the case of androgen receptor, AR) [28], which are generally considered activation and repression marks, respectively [29, 30]. This finding radically changed the model that post-translational methylation was a relatively permanent mark responsible for epigenetic memory, which could only be changed upon passage through at least one cycle of mitosis [29, 31].

Though the initial findings suggested LSD1 could play a role in both activation and repression depending on the composition of its associated proteins, the extent of LSD1 functionality was relatively small (CoREST or AR targets are not widespread). In this chapter, the LSD1 promoter binding program was elucidated using the ChIP-DSL technique on the human 20k promoter array with the expectation of discovering and validating CoREST targets.

2.2 Methods

2.2.1 Probe Design

DSL probe design begins with the targeted sequence fragment. This fragment, usually 500-1000bp in length, is checked for uniqueness in the source organism's genome using BLAST [32]. A word size of 7 (the minimum) and an unlimited e-value maximum is chosen to detect as many matching sequences as possible. The target fragment is BLASTed against the repeat-masked genome sequence extracted from the UCSC genome browser [33, 34]. Any subsequence of the fragment larger than 23nts with higher than 95% identity found elsewhere in the genome is additionally masked, and the masked fragment is submitted to Primer3 to design a 40mer, internal oligo [35]. A slight modification had to be made to allow Primer3 to design internal oligos of 40nts, as the hard-coded maximum is 36 due to concerns of inaccurate melting temperature calculation. This inaccuracy is not a concern for the 40mer design, and the T_m limits were set to a wide range (50-90°C) to allow for greater design freedom. Up to ten 40mers were designed for a single fragment, and each oligo is checked for compatible 20mer paired T_m. For each pair of left and right 20mers with less than 10°C difference of T_m, the 40mer oligo is then measured for its predicted secondary structure using RNAfold with DNA parameters [36]. From this list of filtered 40mers, the oligo with the least stable secondary structure (free energy closest to zero) is chosen.

2.2.2 Tiling Probe Design

Designing tiling probes takes the previously described DSL probe design and queries a much larger sequence of DNA. This is done by fragmenting the sequence into smaller pieces to the size of the desired resolution. In practice, the minimum resolution over a large region is 500bp (due to prevalence of repetitive elements), though tiling resolution in non-repetitive regions can probably be improved to less than 200bp.

The targeted areas in the ERα studies for tiling include a larger scale region surrounding *TFF1* (~1.7mb on chromosome 21), as well as several smaller scale regions focused on promoters of *GREB1*, *DIO1*, *RARβ*, *CASP7*, *KLK3*, and *KAI1*, usually encompassing 50kb upstream to 10kb downstream of the transcription starts and ends of each gene, respectively (Table 2.1). Additional tiling regions for other projects are also listed, including the whole genome of *Human herpesvirus 1* (HSV1), and a larger sampling of many mouse loci, though no data for these is presented (Table 2.2).

2.2.3 ChIP, 3C, and 3D Descriptions

Chromatin Immunoprecipitation (ChIP) is a commonly used technique for detecting protein binding on a region of DNA. The basic premise is that intermolecular interactions, including protein-DNA interactions, are stabilized by formaldehyde cross-linking, and are then sonicated to shear off extraneous DNA (Figure 2.1). These DNA fragments are then immunoprecipitated by an antibody

specific to the protein of interest, and this enriched subset of DNA sequence can be identified through several methods, including region specific PCR, microarrays and high-throughput sequencing. The Brown lab previously described the ChIP procedure [37], while the Fu lab previously described the DSL and hybridization procedures [12].

Chromatin Conformation Capture (3C) is a method which detects DNA-DNA interactions [38]. Starting with formaldehyde cross-linking again, the DNA-DNA interactions are also stabilized. Instead of sonicating to obtain random DNA fragments, a restriction enzyme is used, which results in known DNA fragment ends (Figure 2.3). The overhanging-ends of any fragment can be ligated to any other fragment, but the efficiency of the localized, interacting DNA fragments is much greater than the non-interacting fragments. A subset of these ligations results in non-canonical sequences (such as sequences directly connecting a distal enhancer to a promoter), and can now be interrogated through several conventional methods.

3C combined with tiling DSL technology (3D) is a mixture of several methods. The 3C protocol is followed through ligation, but rather than interrogating the DNA directly at that point, an immunoprecipitation step is done to enrich for a specific set of interacting regions. Specifically, a biotinylated capture oligo is designed to strongly hybridize a target region of DNA, such as a promoter or enhancer. The digested 3C DNA fragments are then hybridized to the

capture oligo and immunoprecipitated against the biotin modification by streptavidin. By enriching for any ligation products that overlap the capture oligo, the 3C technique can now be scaled up to a semi-high-throughput method via DSL tiling arrays. The ligation products are now put through the DSL procedure, and the resultant DSL probes can be hybridized to tiling arrays. The Kleckner lab previously described the 3C procedure [38].

2.2.4 Microarray Analysis

Promoter-specific or tiling DSL arrays were analyzed using the single-array error model [8, 39, 40]. As a brief description, the Cy3 and Cy5 intensities are normalized through background subtraction and median-stabilization. The significance of each spot is then measured by calculating an intensity/noise-dependent statistic X, as defined as $X = \frac{(a_2 - a_1)}{\sqrt{\sigma_1^2 + \sigma_2^2 + f^2(a_1^2 + a_2^2)}}$, where $a_{1,2}$ are the intensities from each channel, $\sigma_{1,2}$ are the uncertainties of the background subtraction, and f is a fractional multiplicative error coming from hybridization errors, dye bias, etc. The distribution of this statistic should be approximately Gaussian if the Cy3 and Cy5 channels are similar. However, in an experiment where the ChIP sample is in one channel and a control genomic sample is in the other, this assumption is not true. To approximate a Gaussian distribution, the spots whose X statistic is enriched towards the genomic input (on the left half of

the proposed Gaussian curve, with values less than zero) are mirrored across the mean by duplicating all negative spots such that $X^+ = -X^-$. The mean and standard deviation of this mirrored distribution can now be used to calculate p-values according to $P = 1 - Erf(\frac{X-\mu}{\sigma})$, where μ is the average of X and σ is the standard deviation of X. $Erf(x)$ is the standard normal cumulative distribution function. Multiple experiments can be combined using a weighted average [41].

2.2.5 3D Microarray Analysis

Despite the proven utility of the single-array error model, data coming from the 3D technique proved to be too noisy to analyze using this method. As an alternative, I turned to using a rank statistic more robust towards noisy data, the median percentile rank (MPR) [42]. The percentile rank is defined as the percent of data points below the current data point when all of the data is sorted according to some parameter, such as the log ratio of the intensities. The MPR is found by taking the median value of a data point's percentile ranks across several replicate experiments. A cutoff can be determined by plotting a histogram of all the MPR values for a set of experiments, hopefully resulting in a bimodal distribution, and choosing a cutoff at the local minimum between the two peaks. In addition to this statistic, a smoothing window was applied to take advantage of the tiling aspect of the DSL probes, which has been a common strategy concerning tiling experiments [43-45]. Specifically, I take the median of all the MPR values in a 10kb window

as the test statistic, and empirically calculate p-values for different median MPR values by randomizing the MPR values 1000 times across the entire tiling region.

2.2.6 Motif Analysis

Motif analysis was done by using the novel Hypergeometric Optimization of Motif EnRichment (HOMER) algorithm [46]. HOMER uses a differential enrichment algorithm similar to those previously described [47]. For differential enrichment, an appropriate background must be selected for comparison. In the case of finding promoter motifs, all known promoters are chosen as background. The target set and background set of promoters are parsed in *n*-mer length oligos (default is $n = 10$), and a p-value for each oligo's enrichment of target versus background (allowing 2 mismatches) is calculated using the cumulative hypergeometric distribution (described more completely in Section 3.2.5). The top 200 oligos found are converted to probability matrices and optimized for maximal hypergeometric enrichment.

2.3 Results

2.3.1 ChIP-DSL of ERα on Human Promoter Array

To investigate the extent of the ERα binding program on human promoters, a ChIP-DSL experiment was done. ChIP was performed using an

antibody against ERα in the ERα-expressing breast cancer cell line MCF7, and the resulting DNA sample was interrogated with a set of DSL probes containing over 20,000 human promoters. The amplified DSL probes, along with genomic input, were then hybridized to a microarray of the same DSL probes. Analysis by the single-array error model showed significant enrichment on 1003 promoters with a p-value < 0.001 (Figure 2.4). When combined with two different analysis methods, SAM [48] and Chipper [49], the resulting overlap of similarly sized significance-sorted promoters was 578. A subset of these overlapping promoters was validated by conventional ChIP and the false positive and negative rates were found to be 3% (2/58 promoters) and 33% (9/27 promoters), respectively [12], highlighting the specificity of the DSL method. This finding represented a substantially larger list of ERα-bound promoters than previously known [16, 50, 51], and hopefully provides an additional resource for breast cancer researchers.

Motif-finding on this bound set of promoters revealed an unexpected result. The top motif found was not that of the canonical estrogen receptor element (ERE), but had the consensus sequence of ACTAnnnnnCCCR, an orphan motif with no known binding protein, and is a widely used promoter motif discovered in another study [52]. The ERE was found as the second motif, but surprisingly FOXA1, a known cofactor with ERα and a known cis-regulatory element with EREs [16, 51, 53, 54], was not found, suggesting that FOXA1 is not

directly required for all ERα-bound promoters. This data does not exclude the association of FOXA1 and ERα on distal sites, as will be discussed in Chapter 3.

To examine the functional significance of ERα binding, the ChIP-DSL results were compared to a gene-expression experiment done in MCF7 cells. RNA from several timepoints of estradiol (E_2) treatment (0, 3, 6, 9, and 12 hours) was converted to cDNA and hybridized to Illumina HumanWG-6 expression arrays. 879 genes were found to be differentially expressed at one or more time-points, however, only 54 overlapped with ERα binding at their promoter. Of these, 49 were up-regulated and 5 were down-regulated, indicating a larger role for indirect regulation via distal regulatory elements.

2.3.2 ChIP-DSL of ERα Targets on Human Tiling Array

In addition to the promoter array, ChIP-DSL was performed on several tiled loci (Section 2.2.2). Several antibodies were used, including specific transcription factors/cofactors such as the enhancer-binding oncogene, AP1; lysine-specific demethylase, LSD1 (aka BHC110); CREB binding protein, CBP; ERα; SP1; nuclear receptor coactivator, SRC1; and the histone acetyl-transferase, TIP60; along with the histone marks acetylated H3, AcH3, di-/tri-methyl H3K4 and H3K9, di-methyl H4K20, and mono-/tri-methyl H3K27.

The region surrounding the known ER-responsive gene *TFF1* is shown (Figure 2.5). As expected, several activation marks are overlapping the *TFF1* gene itself, including acetyl-H3, di-/tri-methyl-H3K4, and di-methyl-H3K79. Bound transcription factors/cofactors on the *TFF1* promoter include AP1, LSD1, CBP, and ERα. A known enhancer of *TFF1* found 10kb upstream also binds a large subset of these activation marks and regulatory factors, only excluding AP1 and di-methyl-H3K79. The downstream gene *TFF3* has a very similar promoter profile to *TFF1*, yet the interceding *TFF2* gene contains only two of the same marks, LSD1 and tri-methyl-H3K4, suggesting the *TFF2* gene may be active, but not controlled by the ERα program. Indeed, according to the gene expression profiling microarray, both *TFF1* and *TFF3* are induced by estradiol while *TFF2* is not regulated, but is expressed.

Another known ERα target is *GREB1*, and ChIP-DSL results are shown (Figure 2.6). The promoter marks are again very similar to *TFF1*, though SP1 is now present as well. Putative enhancer regions upstream of *GREB1*, marked by three separate peaks of ERα binding also demonstrate similar marks as the *TFF1* enhancer. An interesting difference can be seen in the di-methyl-H3K79 and tri-methyl-H3K9 marks. The somewhat activating di-methyl-H3K79 mark is found at a low but consistent level upstream of *GREB1* and continues to the end of the shorter isoforms (*GREB1b*, *GREB1c*). A seemingly anti-correlated presence of the silencing mark, tri-methyl-H3K9, has similarly low but consistent levels starting

from the end of the shorter isoforms and continues through to the end of the longest isoform (*GREB1a).* This evidence might suggest that the higher density heterochromatin that tri-methyl-H3K9 marks would prevent elongation of the covered isoform, in this case, *GREB1a*. However, the gene expression data supports the opposing hypothesis. Not only is the expression level of *GREB1a* much higher than the other two, *GREB1a* is the only estrogen-inducible isoform found. Based on this data alone, one might suggest that intragenic tri-methyl-H3K9 would affect elongation regulation in a novel, positive way, however much more detailed experiments should be attempted.

The tiling maps of *CASP7*, *KAI1*, and *DIO1* were unremarkable. *DIO1* had no marks found on its promoter, while both *CASP7* and *KAI1* had acetyl-H3 and tri-methyl-H3K4 on their promoters. However, only *CASP7* was detected as expressed. The tiling map of RARβ shows the ability of ChIP-DSL to detect repressed regions (Figure 2.7). The upstream region contains high levels of the known silencing mark, tri-methyl-H3K27, and undetectable levels of all activation marks. Interestingly, another known silencing mark, tri-methyl-H3K9, was not detected, suggesting that these silencing marks can be differentially recruited for what seems to be a similar purpose.

2.3.3 3D Analysis Using Human Tiling Arrays

One of the interesting findings from chromosome-wide location analysis of ERα was the discovery of long-distance interactions of putative enhancers with their target promoters using the 3C method [16]. One of the disadvantages of this method is the requirement of target regions *a priori*, as well as the ability to only test one pair of interactions at a time. Considering the number of possible pair-wise interactions in the genome, the 3C method as described is unable to scale up to this level. As a possible solution to this problem, the 3C protocol was adapted to be used in conjunction with ChIP-DSL, termed 3D. To test this novel technique, a 3C experiment was done in MCF7 cells treated with E_2. As described in Section 2.2.3, the 3D adaptation begins after the ligation step (Figure 2.3). Following capture-probe enrichment, tiling DSL probes are hybridized, ligated, and amplified, then detected by tiling arrays corresponding to these probes. The specific capture-probe used overlaps the *TFF1* enhancer located ~10kb upstream of the *TFF1* promoter. The arrays were then analyzed using the modified MPR method as described in Section 2.2.5. Looking at the 1.7mb surrounding *TFF1*, the resulting signal shows an interesting interaction profile. Excluding the expectedly highly positive region surrounding the capture-probe, several peaks are seen. Peaks can be found overlapping several gene annotations, such as *UBASH3A*, *SLC37A1*, and *PDE9A*. A non-coding RNA, NCRNA00111, also has an overlapping peak. The functional significance of this is unknown, as *UBASH3A* is not expressed in MCF7 cells, *SLC37A1* is expressed but not

regulated by E_2, and *PDE9A* is expressed and repressed by E_2. No expression information is known for the non-coding RNA.

Results coming from the other tiling loci proved to be more interesting. Out of all the other tiling loci, the only one that showed significant positive signal was *GREB1* (Figure 2.9). Intriguingly, the positive 3D signal overlapped not only the *GREB1* promoter, but also the putative enhancer region. This data represents the *TFF1* enhancer, found on chromosome 21, which is directly interacting with the *GREB1* promoter and putative enhancers, found on chromosome 2, and is recapitulated by two other methods, conventional 3C and FISH [55]. This finding is not entirely surprising, as will be discussed in Section 2.4. Other tiling loci were found to be negative (Figures 2.10, 2.11).

2.3.4 ChIP-DSL of LSD1 on Human Promoter Array

A ChIP-DSL promoter study was done to investigate the extent of LSD1 binding program on promoters. The purpose was to validate known targets of the neuron-restrictive silencing factor (NRSF/REST) as evidenced by the previously discovered requirement of LSD1 for these NRSF targets [26]. NRSF targets are required in non-neuronal tissues to suppress neuronal differentiation-specific genes. A ChIP-DSL experiment was done on the human 20k promoter array in MCF7 cells treated with E_2 with an antibody against LSD1 to validate known targets, as well as discover novel targets for this function.

Intriguingly, an entirely unexpected result showed a massive recruitment of LSD1 to 4,217 promoters. This large subset did include several known or predicted targets of NRSF, such as *SYN1*, *SCN3A*, *CX36*, *DNM1*, *PAX4*, and *SYT2* [17-19, 22], however, this represented only a slim minority of positive promoters. Overlapping these results with the gene expression analysis in the same cell type and conditions revealed that 74% of the LSD1 positive promoters were also expressed [11]. This data suggests LSD1 is primarily associated with transcriptionally active genes, and is supported by the data that 84% of the LSD1 positive promoters are also positive for RNA polymerase II (Pol II), as seen in the ChIP-DSL experiment using a Pol II antibody (Figure 2.13).

Motif analysis of the LSD1 positive promoters detected the binding site for nuclear respiratory factor 1 (NRF1). NRF1, though initially described as being important for metabolism [56] and respiratory function [57], has a binding motif that is actually quite widespread and conserved across human and mouse. In fact, it is one of the most prevalent, conserved motif found very proximal to promoters [52]. The widespread promoter recruitment and high motif prevalence strongly suggests a connection between the LSD1 and NRF1 proteins, which has not yet been elucidated.

2.4 Conclusions

DSL technology at its core simply represents a different method for interrogating a DNA sample. Rather than using LM-PCR or random priming and the inherent noise associated with these types of “anonymous” amplification, DSL introduces an optimized design step for interrogation. The results speak for themselves, from detecting a more widespread promoter binding program of the ERα protein and subsequent detection of overlapping cofactors [12], to discovery of LSD1 being a core transcriptional component [11], to the capability of identifying 3C interaction partners in a normally unfeasible experimental setup (high-throughput 3C methods are impossibly noisy). Despite the shortcomings of the DSL method, namely the cost, limited resolution, and the inability to scale up to genome-wide assays (with the exception of small genomes), DSL remains an important technology even in today’s era of high-throughput sequencing. In fact, 3C and 3D are not mutually exclusive, and can be combined for even more powerful, focused assays. These possibilities will be discussed in Chapter 4.

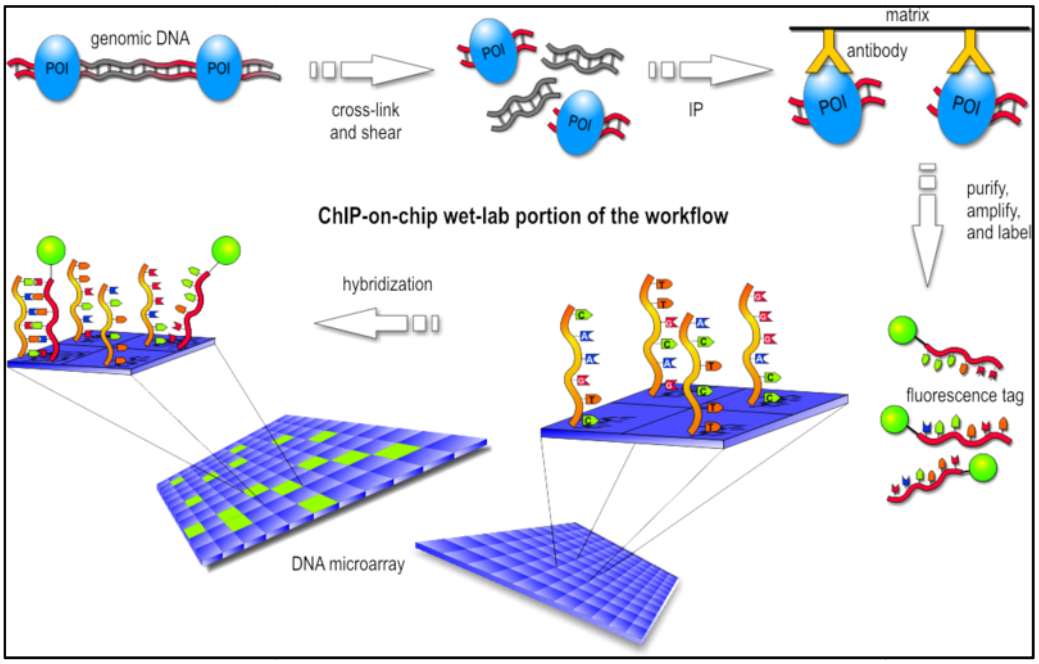

Figure 2.1 Workflow diagram of ChIP-on-chip

The general workflow of a ChIP-on-chip experiment is depicted. Extraction of genomic DNA is followed by formaldehyde cross-linking and shearing, which is then immunopreciptated against target protein. Purified, amplified, labeled DNA fragments that were formerly bound to target protein are then hybridized to a microarray. This picture was taken from http://en.wikipedia.org/wiki/ChIP-on-chip.

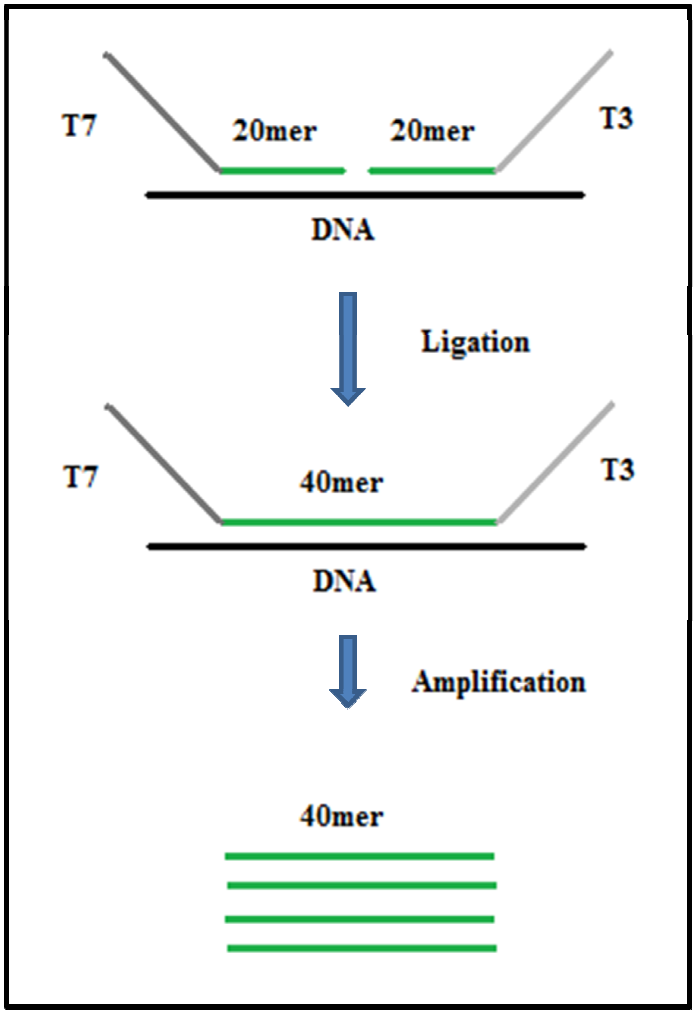

Figure 2.2 DSL probe generation

DSL probes start out as two 20mer sequences designed to hybridize next to each other on the target DNA sequence. After ligation, the 20mers now become the 40mer DSL probe, and are amplified by universal primers T3 and T7.

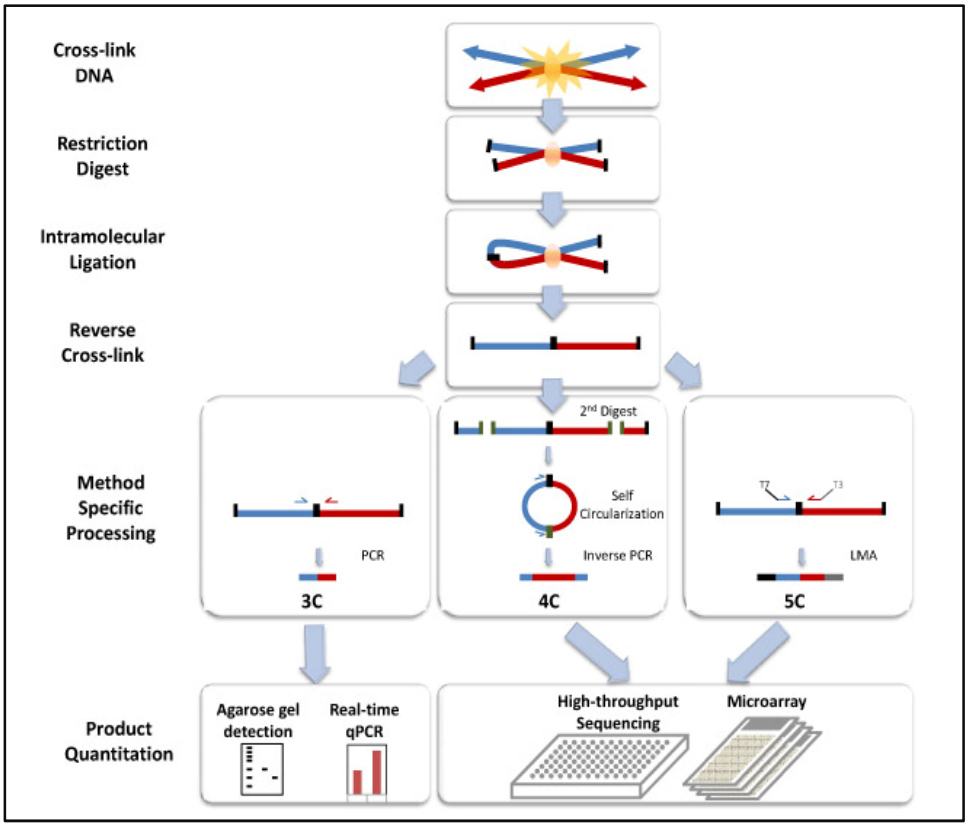

Figure 2.3 3C workflow schematic

3C takes cross-linked DNA and restriction digests so that DNA-DNA interactions now have sticky-ends of DNA fragments in close proximity to each other. Subsequent ligation should be highly efficient only for the interacting partners and detection across the restriction enzyme sites provides evidence of this interaction. This picture was taken from http://en.wikipedia.org/wiki/Chromosome_conformation_capture

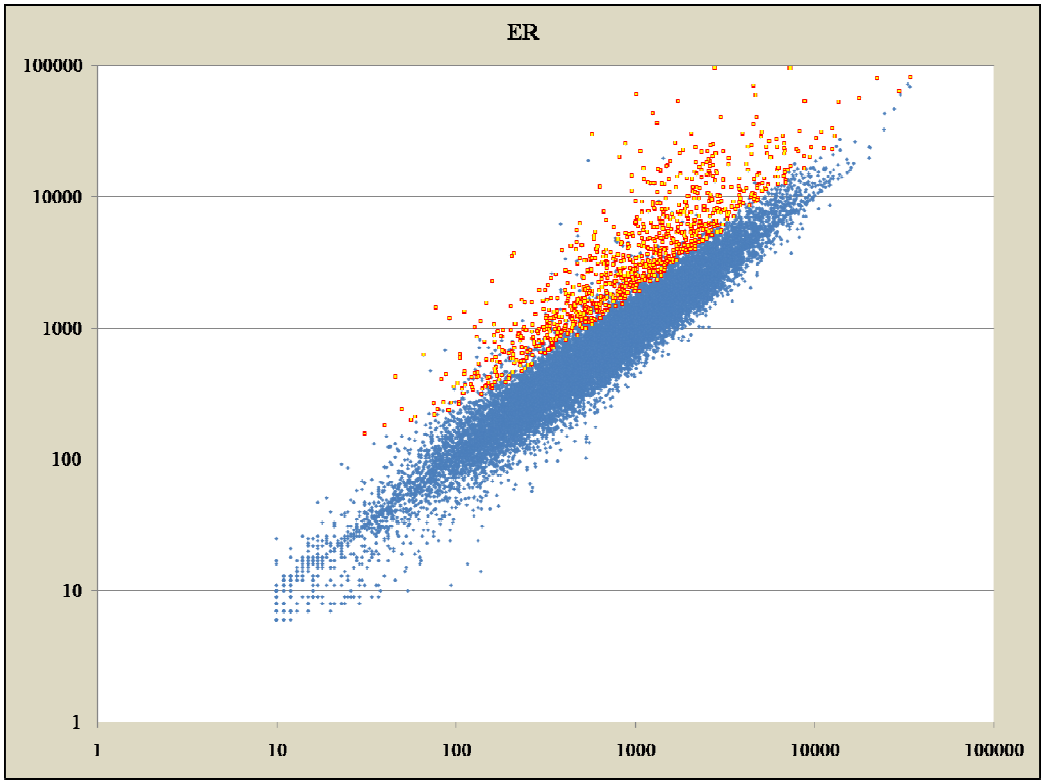

Figure 2.4 ERα log-log intensity scatterplot

ChIP-DSL results as represented by the log-log intensity scatterplot. The x-axis is the genomic input and the y-axis is the signal coming from the ERα IP. The red-yellow labeled points represent promoters with a p-value < 0.001.

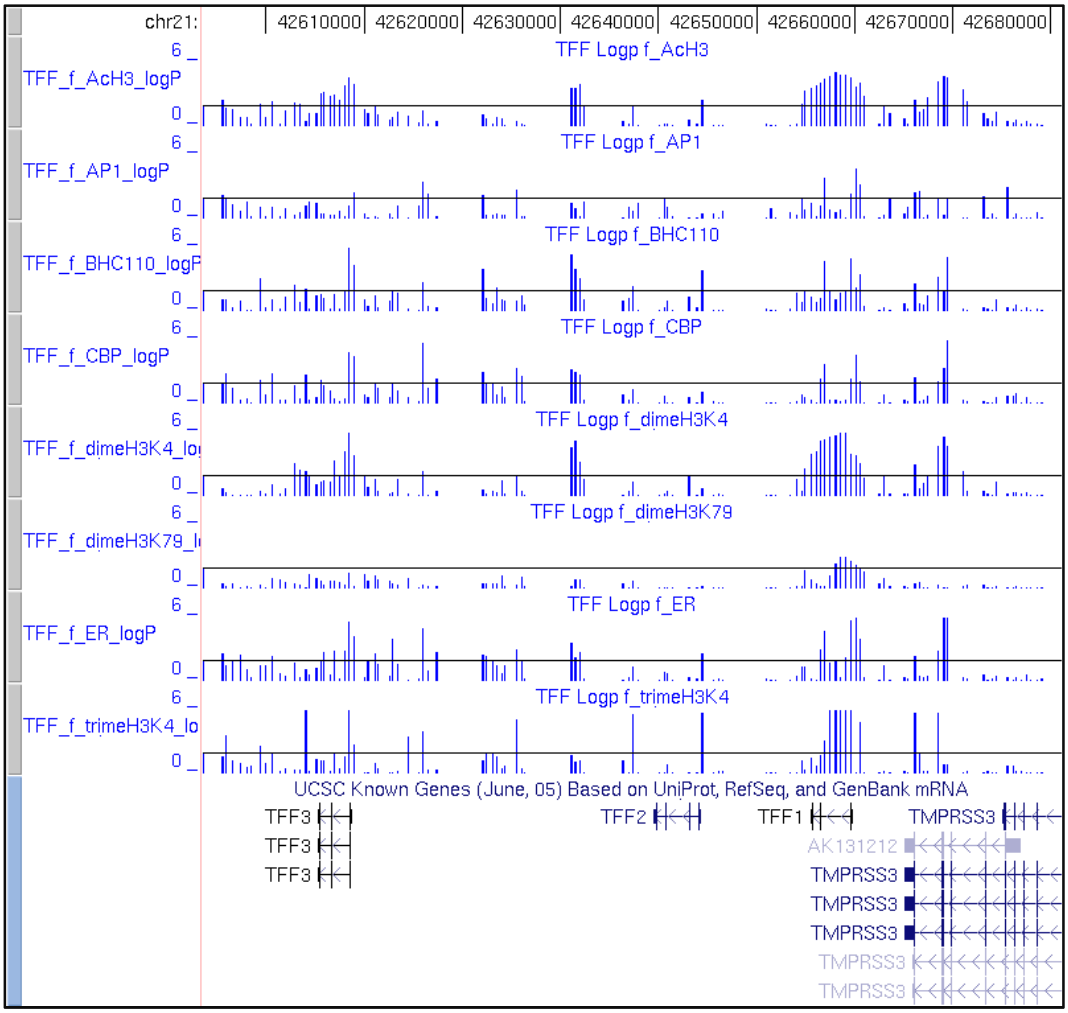

Figure 2.5 ChIP-DSL of positive marks surrounding *TFF1*

Visualization of ChIP-DSL results are shown on the UCSC genome browser for human genome version hg17. Quantitative results are shown as the negative logarithm of the p-value, as calculated by the single-array error model.

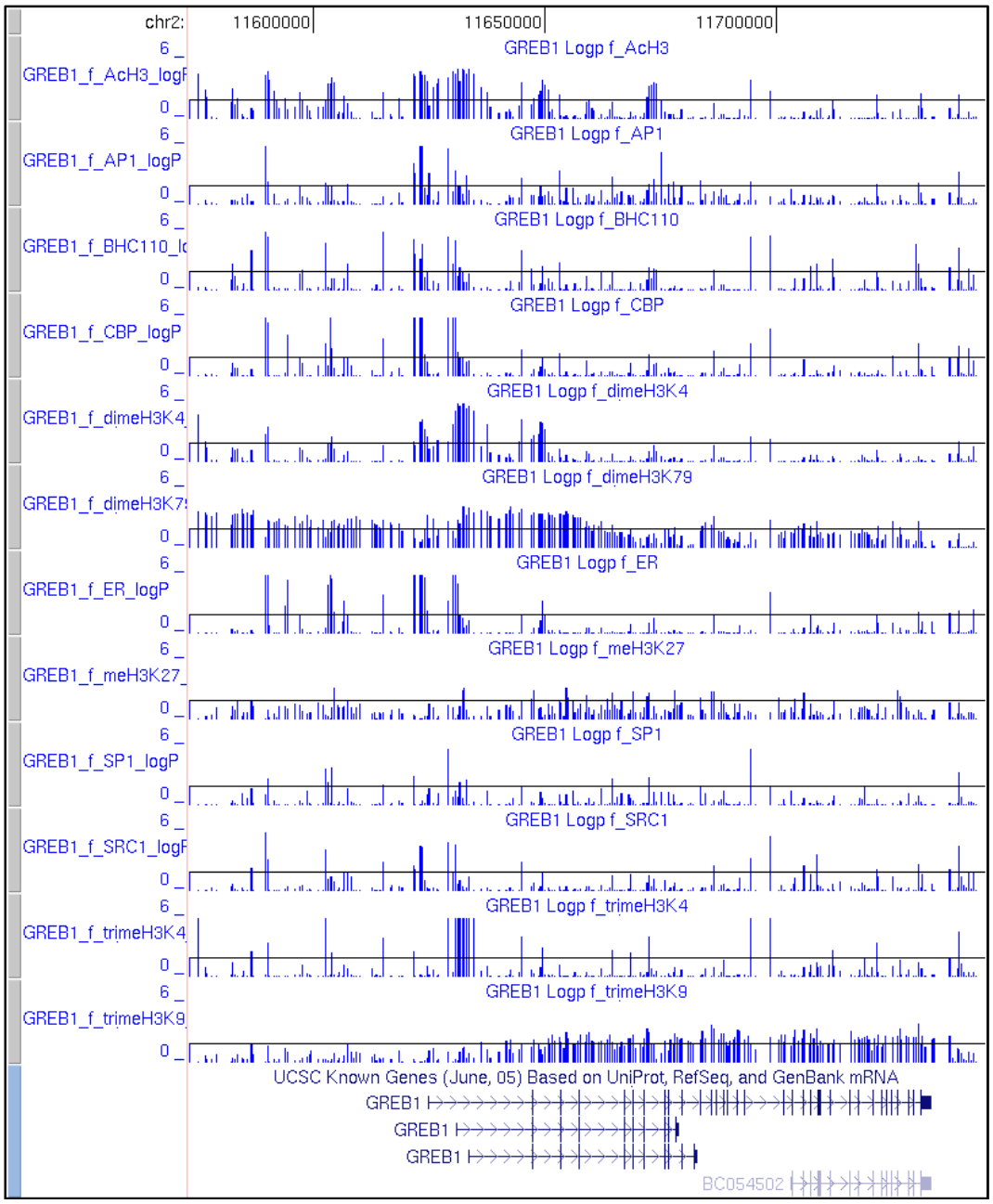

Figure 2.6 ChIP-DSL of positive marks surrounding *GREB1*

Visualization of ChIP-DSL results are shown on the UCSC genome browser for human genome version hg17. Quantitative results are shown as the negative logarithm of the p-value, as calculated by the single-array error model.

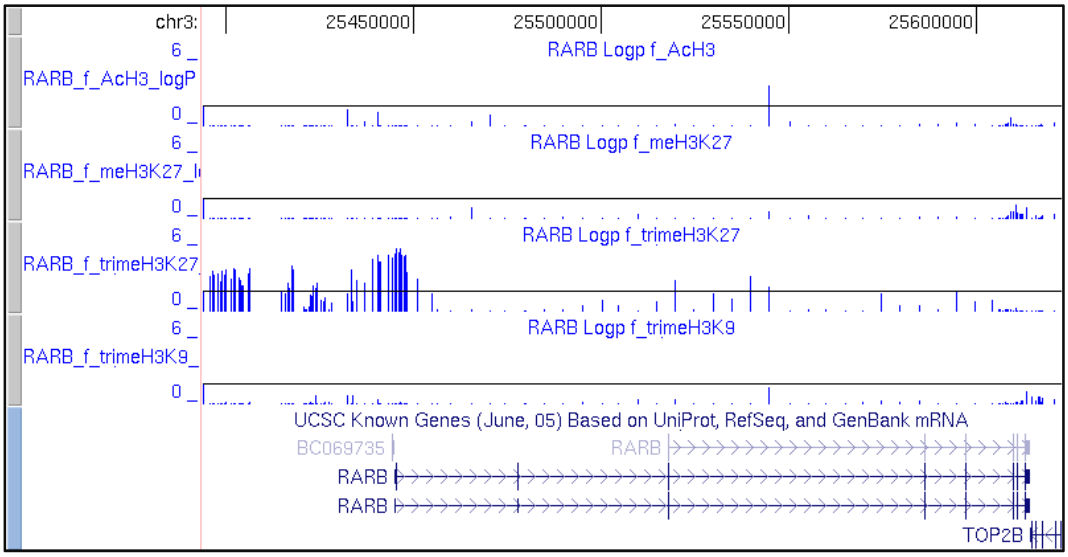

Figure 2.7 ChIP-DSL of positive marks surrounding *RARβ*

Visualization of ChIP-DSL results are shown on the UCSC genome browser for human genome version hg17. Quantitative results are shown as the negative logarithm of the p-value, as calculated by the single-array error model.

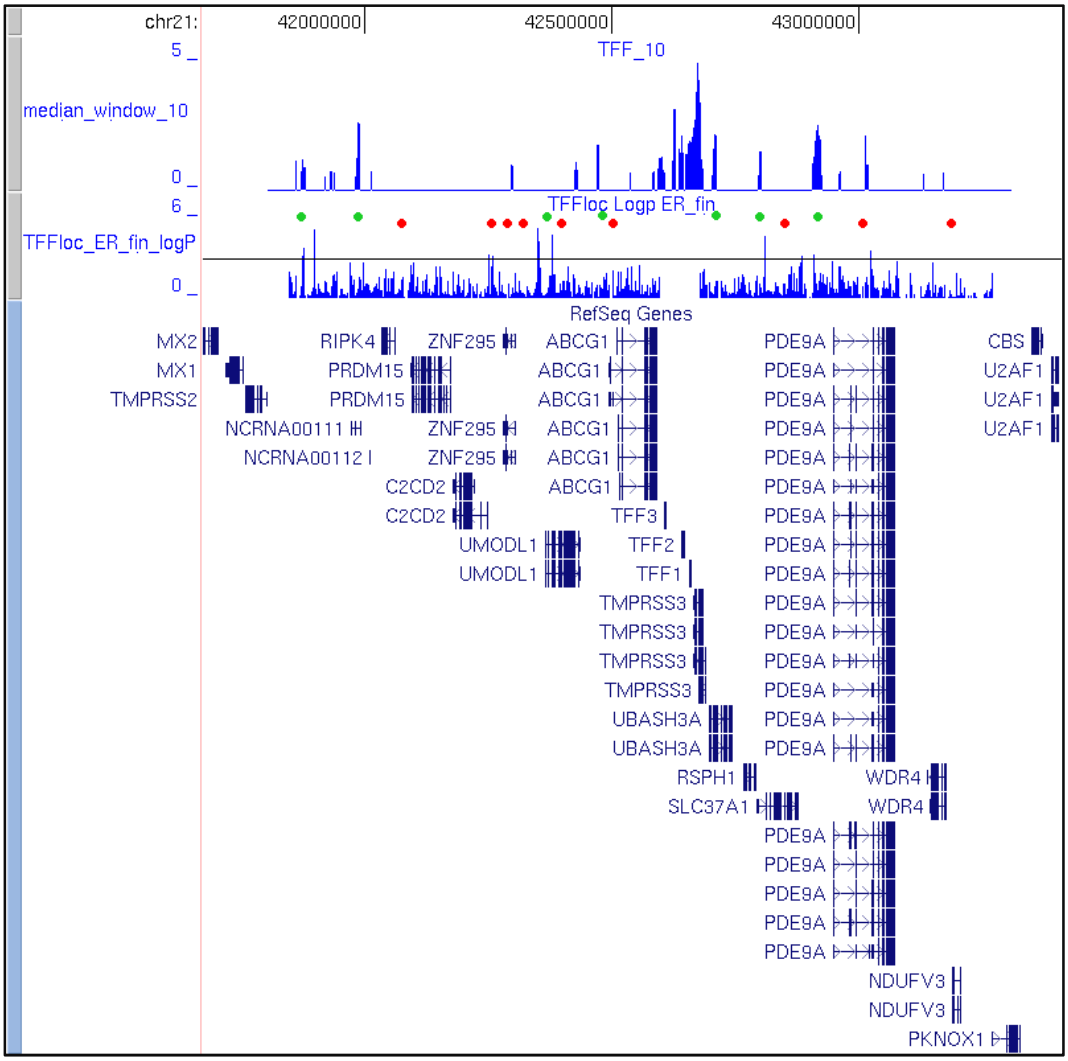

Figure 2.8 3D interaction map around *TFF1*

3D interaction peaks are visualized on the first custom track on the UCSC genome browser (top). ERα ChIP-DSL binding is also shown in the second custom track. Both tracks are quantified by the negative logarithm of the p-value. 3C or FISH validated regions are shown with green dots. Negative regions are shown with red dots.

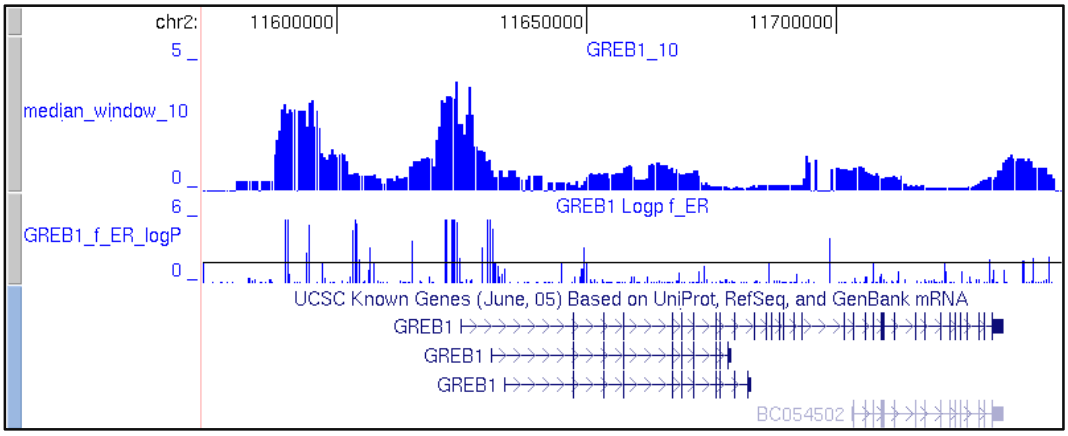

Figure 2.9 3D interaction map around *GREB1*

3D interaction peaks are visualized on the first custom track on the UCSC genome browser (top). ERα ChIP-DSL binding is also shown in the second custom track. Both tracks are quantified by the negative logarithm of the p-value.

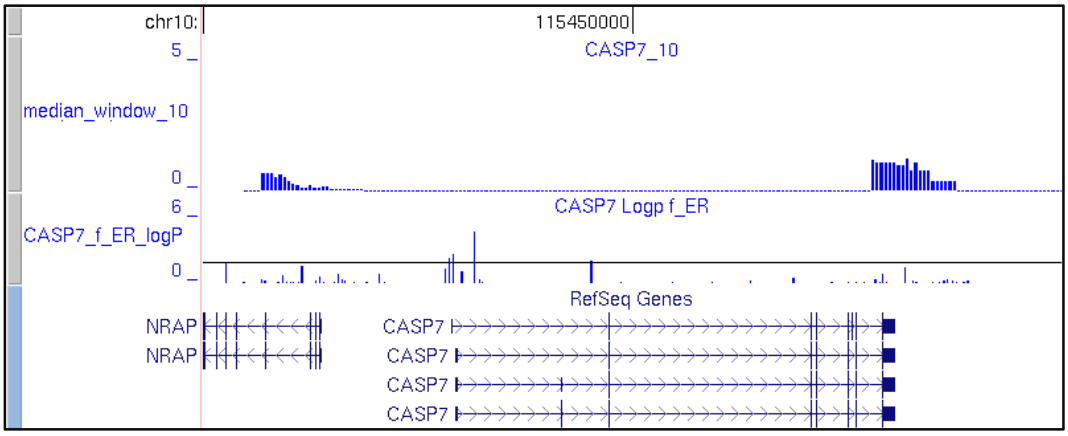

Figure 2.10 3D interaction map around *CASP7*

3D interaction peaks are visualized on the first custom track on the UCSC genome browser (top). ERα ChIP-DSL binding is also shown in the second custom track. Both tracks are quantified by the negative logarithm of the p-value.

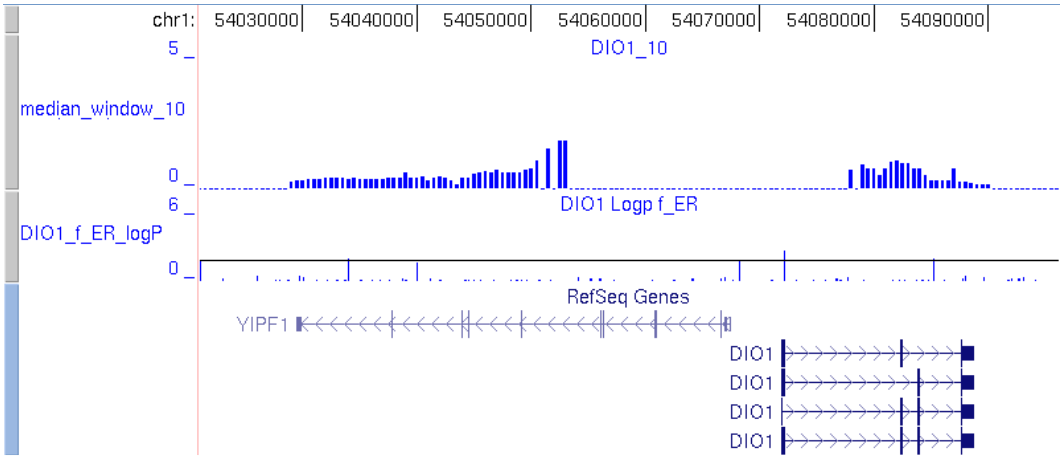

Figure 2.11 3D interaction map around *DIO1*

3D interaction peaks are visualized on the first custom track on the UCSC genome browser (top). ERα ChIP-DSL binding is also shown in the second custom track. Both tracks are quantified by the negative logarithm of the p-value.

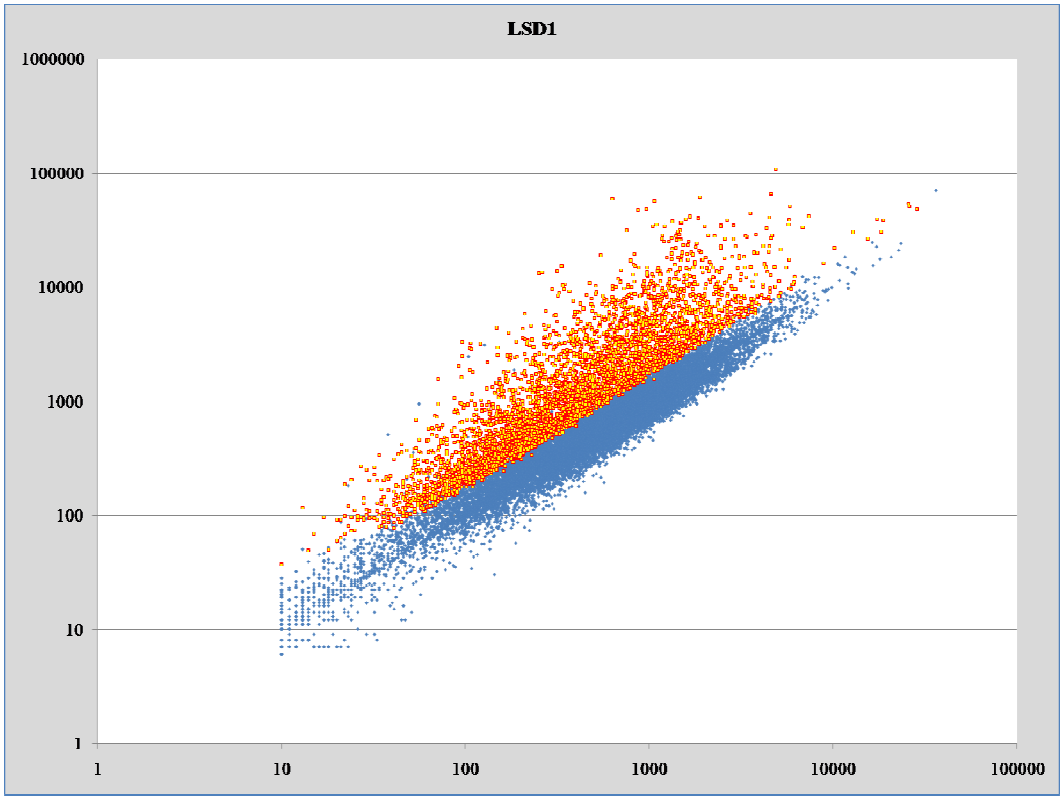

Figure 2.12 LSD1 log-log intensity scatterplot

ChIP-DSL results are represented by the log-log intensity scatterplot. The x-axis is the genomic input, while the y-axis is the signal coming from LSD1 IP. The red-yellow labeled points represent promoters with a p-value < 0.0001.

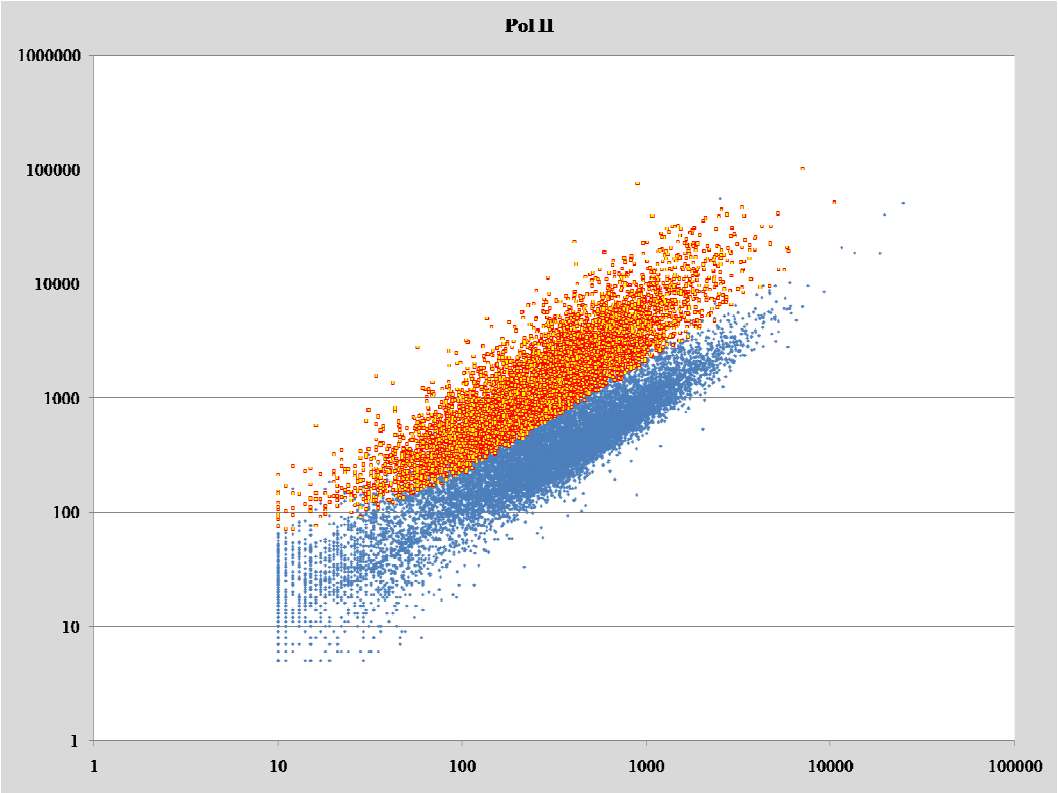

Figure 2.13 Pol II log-log intensity scatterplot

ChIP-DSL results are represented by the log-log intensity scatterplot. The x-axis is the genomic input, while the y-axis is the signal coming from Pol II IP. The red-yellow labeled points represent promoters with a p-value < 0.0001.

Table 2.1 Human and HSV DSL coverage

Gene Name	Chromosome	Start	Stop	Probes
CASP7	chr10	115404768	115489080	53
DIO1	chr1	54023024	54094176	68
GREB1	chr2	11575105	11743139	226
KAI1	chr11	44495446	44607080	101
KLK locus	chr19	56044961	56229537	69
RARB	chr3	25395689	25620914	105
TFF locus	chr21	41849029	43580005	2117
HSV genome	genome	1	144951	294

Table 2.2 Mouse DSL coverage

Gene Name	Chromosome	Start	Stop	Probes
Ascl1	chr10	86238258	87066699	2230
Bclx	chr2	152442023	152567877	148
Bdnf	chr2	109426764	109536854	142
betaTM	chr4	43526572	43593517	106
Bk	chr7	118163629	118283225	156
Camk2d	chr3	126539607	126843448	380
Cd45	chr1	139873974	140039707	203
cSrc	chr9	57416193	57492478	91
Cypher	chr14	33349101	33464644	172
Ets1	chr9	32396098	32512809	202
Fgfr2	chr7	129945376	130106777	256
Fn1	chr1	71509030	71635298	186
Gad1	chr2	69152486	70695182	2146
Gh	chr11	105938941	105992093	87
Hoxd locus	chr2	74463320	74570896	188
Itga3	chr11	94850863	94941759	140
Lef1	chr3	130949252	131319258	318
Lpl	chr8	67856884	67954301	104
Mmp	chr9	7198408	7460733	158
Neurod1	chr2	79144190	79503933	289
Nmdar1	chr2	25104355	25189590	139
Pit1	chr16	64993472	65510750	292
Pomc1	chr12	3905735	3969827	87
Prop1	chr11	50684739	50944481	184
Ptbp1	chr10	79207996	79277116	90
Ptbp2	chr3	119704691	119822066	107
Scn2a	chr2	65294308	66118295	918
Setdb1	chr3	95399047	95478716	56
Shh	chr5	26883435	26697111	65
Tbx19	chr1	166965682	167043886	108
Tshb	chr3	102896597	102951285	38
Zfhx1a	chr18	5542671	5782213	243

3 High-throughput Sequencing and Analysis

3.1 Background

As discussed in Section 2.1, a DNA microarray is a high-throughput method for detecting sequence content of a DNA sample. The typical microarray experiment is really an indirect method for DNA sequencing, as the hybridization of DNA strands is not always exact [9], and mismatched bases are quite common. Another disadvantage is that a microarray can only detect the same sequence that is on the array, making the typical microarray experiment close-ended. This means a researcher can only detect what is specifically being searched.

A more specific method for DNA sequencing is the universally-used adaptation of the classic chain-termination Sanger method [1, 2], termed dye-terminating sequencing. In this method, a DNA polymerase reaction is done on the target DNA sequence in the presence of a mix of dNTPs and labeled ddNTPs. If a ddNTP is incorporated, that particular fragment can no longer be elongated because of the lack of the required 3'-OH. Because of the relative concentrations of dNTP versus ddNTP, and the large number of the same DNA fragments being elongated, the incorporation of a labeled ddNTP is a stochastic process, and typically all nucleotide positions within the target DNA fragment are labeled with

a ddNTP. The specific nucleotide incorporation is detected via mass separation by capillary electrophoresis combined with optical detection of the nucleotide-specific label (Figure 3.1). Successive peak calls of the four different fluorescent labels produce an ordered, readable sequence (Figure 3.2). Despite over thirty years of improvements and optimizations, the dye-terminating sequencing method is incapable of high-throughput sequencing since sequences can only be assayed by the capillary electrophoresis/optical sensor assembly one at a time.

Not surprisingly, one solution to high-throughput, specific sequencing is a combination of the two aforementioned methods, as implemented by the Illumina Genome Analyzer. As will be detailed in the next section, the Genome Analyzer basically works by fabricating a random microarray derived from a DNA sample, which is then followed by a sequencing-by-synthesis (SBS) reaction [58-61]. The SBS reaction generally follows a similar approach to dye-terminating sequencing, except the terminator bases are reversible, thus enabling consecutive base-by-base incorporation of a fluorescently labeled nucleotide. The output of this method is several million short sequences which correspond to the sequence content of the DNA sample. This type of approach has been very successful for a large number of genome-wide assays, including SNP detection [62], RNA profiling [63], small RNA profiling [64], nucleosome positioning [65], histone modification [66-68], DNA methylation [69], and mapping translocation events [70]. Though none of these assays were necessarily impossible before high-throughput sequencing

became available, all were heavily dependent on the close-ended microarray technology. Even the genome-wide tiling arrays currently available are not completely open-ended. They are fabricated across several microarrays for a single genome, and usually require several replicate experiments to be meaningful.

The Illumina Genome Analyzer is not the only high-throughput sequencing method available, as several other competing technologies, namely ABI/SOLiD, 454/Roche and Helicos, all generally aim to sequence millions of DNA fragments in a massively parallel fashion. Regardless of technology, the resulting datasets are of unprecedented size, as the raw image data are on the order of hundreds of gigabytes, and the sequence data produced is on the order of gigabases. The first-pass analysis of such data is being constantly researched and optimized, as the raw data is processed and sequences are base-called by proprietary methods provided by the parent company or processed by third party methods [71], and the sequences are then mapped [72, 73]. For the ChIP sequencing type of experiment (ChIP-seq), an enrichment of tags covering a specific genomic area is defined as a peak through several different methods [74-77], and the resulting peak positions are equivalent to binding regions found in ChIP-on-chip experiments. Each ChIP-seq experiment represents the genome-wide binding profile for that antibody's target, and the biggest question left to answer is what this profile means. The usual data-mining of this profile usually

centers on promoter binding and the associated knowledge of these potential gene targets. However, the unprecedented amount of data being generated with several different antibodies in several different conditions necessitates a much faster and generalized method for data-mining, as will be introduced in the subsequent sections.

3.2 Methods

3.2.1 Genome Analyzer Description

It is vitally important to understand the details of the high-throughput sequencing protocol as downstream analysis parameters will depend on some of these details. The ChIP-seq method, done in conjunction with the Illumina Genome Analyzer (GA), will be used as an example. Following the ChIP protocol (Figure 2.1), the DNA fragments have the GA adapters (A and B) ligated to both sides of the fragment using a proprietary ligation method. This results in DNA fragments with only the A-adapter linked to the 5' side of both strands and the B-adapter linked to the 3' side of both strands (Figure 3.3, top). PCR amplification using the A and B sequences for priming results in two different products containing the same DNA fragment (Figure 3.3, bottom). The complementary A-adapter (A'), which is used later for priming the sequence-by-synthesis reaction, is found on the 3' end of both the original forward and reverse strands.

The amplified sample is now submitted to the Illumina Cluster Generator which fabricates the random microarray, called a “flowcell” (Figure 3.4). The flowcell is covered with a high density of chemically bound oligos corresponding to the A and B adapters, thus providing a solid phase hybridization platform for capturing DNA fragments from the sample. A low concentration of the sample DNA is washed onto the flowcell, ensuring that a hybridization event is rare enough such that a single molecule of DNA is sufficiently distant from a neighboring molecule. Subsequent polymerase elongation not only chemically attaches the nascent complementary sequence to the flowcell, but also allows double-stranded bridge products to form (Figure 3.4, step 4). Denaturation of these bridges doubles the amount of single-strand DNA molecules attached to the flowcell that are in close proximity to each other, which can then form new bridges for additional amplification. By the end of the cluster generation protocol, these polymerase-formed colonies, or clusters, represent several hundred copies of the same DNA sequence.

After cluster generation, the flowcell is submitted to the GA sequencer machine, which follows a proprietary version of sequence-by-synthesis (Figure 3.5). The sequencing reaction is primed using a primer corresponding to the A-adapter. This primer hybridizes to the 3’ attached A’ sequence, and fluorescently-labeled, terminating nucleotides are added one at a time, thus preventing additional nucleotides from being added. After each single nucleotide addition, an

image is captured of the fluorescing clusters. Once all four possible nucleotides are added, the terminating nucleotides are unlocked for another cycle of base addition and image capture. Thus, each cycle represents a nucleotide position, starting from the edge where the A' sequence was ligated. A typical captured image is shown in Figure 3.6.

Once all cycles are completed (generally 26-36 cycles), the company-provided analysis software, the Genome Analyzer Pipeline, starts image analysis by locating the clusters on the flowcell so that base-specific intensities can be measured across cycles for the same cluster. Intensity measurements for each base are then used for base-calling in a fashion similar to the dye-terminating sequencing output, where the base with the highest intensity should be the base that was incorporated at that position on that cluster. Each base-called cluster (tag) is basically the sequence found at the edges of the DNA fragments which comprise that cluster. The entire collection of tags is now ready for downstream analysis.

3.2.2 Mapping Tags

The GA Pipeline also includes a proprietary mapping program, called Efficient Large-Scale Analysis of Nucleotide Databases, or ELAND. Though the details are not available, this algorithm is capable of aligning several million short sequences (of up to 32nts) to a human genome-sized reference sequence while

allowing two mismatches. All tags are reported in the results file, defining if a tag mapped uniquely or in multiple locations, if a tag did not map, or if a tag was of too low quality to be used for mapping (according to the GA Pipeline quality controls). Alternative mapping methods are available [72, 73], but the ELAND algorithm was used in all cases considered here.

3.2.3 Peak-Finding

The desired result from any ChIP experiment is enrichment of DNA fragments overlapping the putative binding region of the target protein. This result manifests itself, in high-throughput sequencing data, as an abundance of tags mapping to a narrow region on the reference sequence, such as a genome. Determining a cutoff for this tag enrichment is as varied as the mapping techniques, but most methods usually use one of a few approaches, either by an estimation of significance assuming a Poisson distribution parameterized by a genome-wide tag frequency [68, 78], an empirical approach using control sequencing experiments as background [79], or both [80]. Assuming a Poisson distribution (or another appropriate background distribution, the negative binomial [57]) has the distinct advantage of not requiring a control experiment, thus making experiments as much as 50% faster and cheaper. However, any significant location-specific bias found in the experiment, such as telomeric/pericentromeric areas (discussed in Section 3.3), will be designated as

real peaks. A control sequencing experiment is the most direct method for quantifying such bias, but comes at the cost of fewer experiments being run.

At the time of my involvement with analyzing the GA data, no peak-finding methods were published or available, and I wrote my own Perl scripts based on the Poisson approach. The Poisson distribution describes the probability of a low frequency, discrete event happening, such as the number of car accidents per day at an intersection. This distribution assumes each event is independent, and the probability function is as follows: $f(k;\lambda) = \frac{\lambda^k e^{-k}}{k!}$, where k is a non-negative integer, and λ is the frequency of events, then $f(k;\lambda)$ calculates the probability of exactly k events occurring in the given interval (as defined by λ). When applied to peak-finding, we want to know the probability of a certain number of tags overlapping a region, using the genomic frequency of tags, which is calculated by summing all probabilities up to the total number of tags within the interval, or $p = \sum_{n=1}^{k} f(n;\lambda)$, where $f(n;\lambda)$ is the Poisson distribution defined earlier. In practice, tags are not the event that should be considered. As described in Section 3.2.1, tags are the ends of the DNA fragments being sequenced, and because both ends are definitively prepared for sequencing, these events are highly dependent. The simple adjustment to the equation is that tags are no longer the event being considered; instead, it is the fragment, which is easily done by dividing the frequency by 2. After determining the cutoff (i.e. number of

fragments within a region) for the desired p-value, the cutoff is adjusted back to "tag space" by multiplying by 2.

The general workflow goes as follows. Mapped tags are sorted by chromosome and position, and a total tag count is determined by summing all non-replicate tags. Non-replicate tags are used to remove any amplification bias by the multiple rounds of PCR in sample preparation. Tag density is calculated by dividing the sum of non-replicate tags by the size of the mappable genome. The mappable genome is defined as all gapped regions and highly repetitive regions subtracted from the total genome size. To obtain a sensible interval size for the usual ChIP experiments that are sonicated to 500bp, the tag density is further divided by 1000bp, thus giving an expected background frequency of tags over 1kb. The previously described p-value calculation is then used to determine the tag number cutoff to achieve a p-value of less than $1x10^{-5}$. Taking this cutoff, peaks can now be found in the sorted tag list simply by iterating through each chromosome. Anytime a region is found to be over the genomic-frequency-defined cutoff, an additional calculation is made to calculate a local cutoff based on the neighboring 10^5 tags using the same Poisson calculation. This local calculation is used to correct for genomic amplification bias as seen in many cancer cell lines, which will be shown in Section 3.3. All enriched regions passing the genomic and local cutoffs have their peak location calculated as the average

location of all overlapping tags, which are then output to UCSC-compatible WIG files. Downstream analyses presented in this chapter use these files as input.

3.2.4 Motif Analysis

As described in Section 2.2.6, motif analysis is performed using the novel HOMER algorithm [46]. One difference for genome-wide datasets is the background set is a non-trivial decision. In practice, 200nt sequences are extracted from within 100kb of a transcription start site (TSS) and are binned into CpG density intervals of 0.02%. 100k sequences are chosen to match the same CpG density profiles of the target sequences. For example, if the target set of 10 sequences has 9 sequences with less than 0.02% CpG, and 1 sequence between 0.02-0.04% CpG, then the background set will have correspondingly 90k sequences with less than 0.02% CpG, and 10k with 0.02-0.04% CpG.

3.2.5 Genome Ontology Analysis

To date, most secondary analyses of ChIP-seq data has followed a predictable path. As with ChIP-on-chip data, the first proof that an experiment worked is a motif analysis, which is expected to result in the discovery of the transcription factor's known consensus site. On the other hand, if the target is a histone mark, then the experiment is validated through a correlation analysis with expression data. Many of these analyses are simply correlations between binding

data and a regulatory outcome, such as expression or chromatin landscape [66, 67], especially near or within genes, resulting in a somewhat common thread amongst ChIP-seq papers. A set of ChIP-seq experiments is done, each of which is proven to work because of motif analysis. Then a breakdown of the genomic peak locations that is gene-centric is given (e.g. percentage of overlapping introns, exons, proximal promoters, and "other"). Finally, the particular protein of interest, which is already known to have several specific functions during some relevant biological process, is shown to behave as expected (regulated binding/dismissal upon some treatment or differentiation), and any abnormal behaviors are mentioned, but rarely investigated further. Nothing is wrong with this particular research "plotline," except for the fact that most of these papers are somewhat self-contained and incredibly specific to the singular interest of the author. As the experiments are open-ended, genome-wide studies, the difficulty in comparing data between papers is exceedingly frustrating. Each dataset should be opening up incredible amounts of insight, not just one or two known outcomes. So how do we find these insights?

One possible solution stems from the well-known Gene Ontology (GO) Consortium. The primary purpose of the GO project is to build and maintain a common vocabulary for all known genes so that many different sources of gene-centered knowledge can be quickly accessed [81]. Once an ontology containing all of the desired descriptions for genes is built, assaying the ontology is a

relatively simple process, as demonstrated by the fact that a GO analysis is a staple among almost all genome-wide studies, whether whether they are binding or expression profiles. The usual method for assaying the GO is over-representation analysis, wherein the subset of positive genes is checked for enrichment in any particular GO term using the hypergeometric distribution. The hypergeometric distribution is classically explained as the "urn" problem. Given an urn filled with red and green marbles, this distribution will calculate the probability of getting a specific number of red marbles when picking out a sample of marbles from the urn. The hypergeometric distribution requires that each marble chosen is not replaced, and the marble choice will result in either a success or a failure. In the GO context, this distribution is defined as $f(k; m, N, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$, where *N* is the total number of genes, *m* is the total number of genes in a particular GO term, *n* is the group of genes being queried (such as all regulated genes from an expression array), and *k* is the subset of queried genes that are in the GO term in question. The hypergeometric distribution is appropriate because the genes from any experiment cannot be chosen twice, so the non-replacement assumption is satisfied. The success-failure assumption is satisfied because a gene is either within a GO term or it is not.

Here, I propose to use a similar method for looking at ChIP-seq data, called the Genome Ontology. Instead of using user-defined definitions for genes, I

am using the ChIP-seq-defined binding regions as “terms,” for example, all the regions binding the target protein. When two datasets are compared, the significance of the overlap can be calculated using the same hypergeometric distribution described above. The same assumptions remain valid because peaks from one experiment cannot overlap with each other, and thus cannot be chosen twice, while peaks compared between two experiments can be defined as overlapping (success) or non-overlapping (failure). More specifically, when peaks are defined within a particular interval (1kb in this case), the same function is used, $f(k;m,N,n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$, where N is now the total number of 1kb intervals in the genome, m is the total number of peaks from experiment 1, n is the total number of peaks from experiment 2, and k is the number of overlapping peaks. N is simply calculated by dividing the mappable genome size (Section 3.2.3) by the interval size.

This method retains the advantage of the GO by keeping all data as annotations for later comparison, and future data can be instantly assayed against the Genome Ontology. Additionally, each comparison can be used to generate new, more detailed annotations. Lastly, the Genome Ontology is extensible to any genome-wide annotation, not just ChIP-seq data. Any genome-wide ChIP-on-chip data can also be used, as well as already defined annotations such as gene regions or transcription start sites.

3.3 Results

3.3.1 ChIP-seq of ERα

To further investigate the genome-wide functionality of ERα, ChIP-seq was performed using an ERα antibody in MCF7 cells treated with estrogen (E_2). The 13.8 million sequenced tags that uniquely mapped to the human genome resulted in 12,874 peaks after peak-finding. 245 of these peaks were found within -800 to +200bp of known TSS (defined here as proximal peaks), leaving 12,629 peaks outside of this region (defined here as distal peaks). Motif-finding results of the proximal and distal peaks are found in Tables 3.1 and 3.2, respectively. The proximal peaks show a weak enrichment for the ERα half-site, TGACC, while the distal peaks show a much stronger enrichment for the canonical half-site. Interestingly, the FOXA1 binding site that was predicted by other findings [16, 51, 53, 54] was not found using this ChIP-seq dataset (Table 3.3, motif 7). A possible reason for this finding is that previous studies have looked at E_2-responsive ERα binding, while the ChIP-seq dataset detected all ERα bound in MCF7 cells in the presence of E_2. If a particular peak were to be found in the absence of E_2, this ChIP-seq experiment would be unable to differentiate an E_2-dependent and an E_2-independent peak.

To validate that our ChIP-seq peaks did contain the previously discovered FOXA1-associated binding regions, our ERα ChIP-seq data was overlapped with the ERα ChIP-on-chip data [50]. Unsurprisingly, the majority of the ChIP-on-chip peaks overlapped with the ChIP-seq peaks (2,422 of 3,665). HOMER analysis revealed the FOXA1 motif ranked 14th in this overlapping set. The remaining 1,243 ChIP-on-chip peaks not found within 1kb of the ChIP-seq were instead all found within 10kb, and were highly enriched in the FOXA1 motif (Table 3.4, motifs 2-4). These findings suggest that FOXA1 is correlated with E_2-dependent ERα recruitment, and the FOXA1 binding site does not need to be proximal to the ERα binding region.

To find additional evidence that FOXA1 and ERα are correlated, a Genome Ontology analysis was done using the 12,874 ChIP-seq ERα peaks (Table 3.5). The top two overlapping datasets corresponded to ERα binding data from ChIP-on-chip experiments in MCF7 cells and ChIP-seq experiments in MDA-MB-231 cells. The next two significantly overlapping datasets were FOXA1 ChIP-seq data from MCF7 cells in the absence or presence of E_2. This finding demonstrates the utility of the Genome Ontology by discovering the ERα/FOXA1 correlation without requiring data separation. Motif analysis without separating the hormone-dependent and independent peaks did not find the FOXA1 correlation in the ChIP-seq data. In this case, it seems somewhat elementary to undertake this more specific analysis, but in an unknown dataset,

the experimenter does not know *a priori* what specific analyses to attempt. The Genome Ontology can quickly reveal these correlations, assuming the correct experiment was done.

Additionally, finding these correlations does not necessarily require the experiment be done in exactly the same background. The 6th and 7th top hits to the ERα ChIP-seq data corresponded to FOXA1 ChIP-seq binding in the prostate cancer cell line, LNCaP (Table 3.5). Because these experiments were done in different cell lines, the correlation found cannot equate biological significance. However, this finding does demonstrate the robustness of the Genome Ontology by providing hints at real correlations. In this case, the correlation of ERα binding in MCF7 cells with FOXA1 binding in LNCaP cells would direct the experimenter to attempt a FOXA1 binding experiment in MCF7 cells.

3.3.2 ChIP-seq of Androgen Receptor (AR)

AR, is a steroid nuclear receptor that, upon activation by androgenic steroids, such as testosterone or dihydrotestosterone (DHT), can regulate several biologically relevant processes, including direct gene transcription [82], signal transduction [83, 84], and indirect gene transcription [85]. Like the estrogen-responsive steroid nuclear receptor, ERα, AR seems to play a significant role in cancer progression (specifically prostate cancer in the case of AR), as evidenced by the elevated AR expression levels in advanced prostate tumors [86, 87]. To

investigate the genome-wide role of AR binding events, ChIP-seq was done using an AR antibody in LNCaP cells, an androgen-dependent prostate cancer cell line, in the absence or presence of DHT.

The AR –DHT ChIP-seq experiment resulted in 11.8 million mapped tags, with 344 peaks found, while the AR +DHT ChIP-seq experiment resulted in 16.5 million mapped tags, with 1,832 peaks found. This large increase in peaks found is not surprising since LNCaP cells are androgen responsive. The HOMER motif-finding results contained no significant motifs in the –DHT case, while the +DHT case contained the AR binding site as the top 2 motifs (Table 3.6). Interestingly, the FOXA1 binding site was found as the 3rd most significant motif, echoing previously published data from ChIP-on-chip experiments that found the FOXA1 motif associated with AR binding regions [88].

To further investigate the association of FOXA1 with AR, additional AR ChIP-seq experiments were performed. The same experimental setup was used (AR ChIP-seq in LNCaP cells in the absence or presence of DHT), however, an siRNA against *FOXA1* was introduced to knockdown the mRNA levels of the *FOXA1* gene. The –DHT/si*FOXA1* case contained 9.1 million mapped tags, resulting in 375 peaks, while the +DHT/si*FOXA1* case contained 7.4 million mapped tags, resulting in 6,276 peaks. These findings suggest a similar relationship between FOXA1 with AR and FOXA1 with ER in that FOXA1 is mostly found on E_2-induced binding of ERα rather than the constitutively-bound

ERα. The AR binding program is only modified by *FOXA1* knockdown in the case of hormone induction as evidenced by the 3.4 fold increase in peak numbers between the +DHT/si*FOXA1* and the +DHT alone experiments. No increase is found when comparing the –DHT/si*FOXA1* and –DHT alone experiments. Motif-finding shows that the increased number of peaks found in the +DHT/si*FOXA1* experiment are highly enriched in the AR binding site, however the FOXA1 binding site is no longer found (Table 3.7).

The Genome Ontology results for the AR +DHT and the AR +DHT/si*FOXA1* are shown (Tables 3.8 and 3.9, respectively). The unsurprising findings show that the top two overlapping terms with AR +DHT is the FOXA1 +DHT ChIP-seq and AR +DHT/si*FOXA1*, while the top hit for AR +DHT/si*FOXA* is not FOXA1, but the AR +DHT ChIP-seq experiment. Additionally, despite being unable to discover the FOXA1 binding site in the AR +DHT/si*FOXA1* regions, the Genome Ontology still finds a significant overlap with FOXA1 bound regions. Though this result could indicate a finding similar to the ERα data in that the FOXA1 function may not require proximal binding of FOXA1 to AR (and hence, difficulty in motif-finding), recently published work demonstrates that FOXA1 is a corepressor of the *AR* gene itself [89]. Because the AR +DHT/si*FOXA1* experiment would knockdown *FOXA1* mRNA expression, this would lead to increased expression of the *AR* gene, which might explain why there is an increased binding profile of AR. These additional peaks are not

strongly associated with the FOXA1 binding motif or with FOXA1 itself, though the AR binding site is even more enriched in the AR +DHT/si*FOXA1* versus the AR +DHT without any siRNA.

3.3.3 ChIP-seq of FOXA1

As described previously, FOXA1 is a central focus for several biological studies. Differential expression of *FOXA1* is associated with multiple types of cancer. Decreased *FOXA1* expression results in increased *AR* expression [89], and this decreased expression is also correlated with breast cancer [51, 90]. To further investigate its role in breast and prostate cancers, FOXA1 ChIP-seq experiments were performed in MCF7 and LNCaP cells, in the absence or presence of their target hormones (E_2 in MCF7 and DHT in LNCaP).

A FOXA1 ChIP-seq was performed in MCF7 cells in the absence or presence of E_2. The FOXA1 -E_2 experiment yielded 13.4 million mapped tags resulting in 15,662 peaks, while the FOXA1 +E_2 experiment yielded 14.9 million mapped tags resulting in 16,374 peaks. Motif-finding of the MCF7 experiments discovered the FOXA1 binding site as the top hit, while the AP-1 motif was also a high-ranking hit in both experiments (Tables 3.10 and 3.11). FOXA1 ChIP-seq was also performed in LNCaP cells in the absence or presence of DHT. The FOXA1 –DHT experiment yielded 7.5 million mapped tags, resulting in 26,904 peaks, while the FOXA1 +DHT experiment yielded 7.4 million mapped tags,

resulting in 18,306 peaks. Motif-finding of the LNCaP experiments found the same FOXA1 binding site as the top-ranking hits, but no AP-1 site was found (Tables 3.12 and 3.13). The prevalence of the AP-1 site in the MCF7 data strongly suggests a FOXA1/AP-1 association, specifically in this breast cancer cell line.

Genome Ontology analysis of the FOXA1 ChIP-seq binding regions in MCF7 cells in the $-E_2$ and $+E_2$ cases are shown in Tables 3.14 and 3.15, respectively, while the ChIP-seq binding regions in LNCaP cells in the –DHT and +DHT cases are shown in Tables 3.16 and 3.17, respectively. While the most significant overlapping experiments for all the FOXA1 binding regions was always the other FOXA1 experiments, an interesting result occurs just below the multiple FOXA1 datasets. In the FOXA1/MCF7 experiments, the next highest-ranking experiments are the ERα binding regions, while in the FOXA1/LNCaP experiments the next highest-ranking experiments are the AR binding regions. This result suggests cell-type specificity for a subset of the FOXA1 binding, although in LNCaP case, Genome Ontology analysis indicates ERα binding regions follows directly after the AR binding regions.

3.3.4 ChIP-seq of REST/NRSF

As introduced previously in Section 2.3.4, REST is an essential protein for the regulation of neuronal genes, specifically for the repression of these genes in

non-neuronal tissues and in undifferentiated neuronal precursors [91-94]. This repression consists of REST recruiting several silencing-related proteins and enzymes, such as the methylated DNA binding protein MeCP2, H3K9 methyltransferases G9a and SuVar39H1, and the H3K4 demethylase LSD1. Defining the complete REST binding program is an important step in understanding the extent of REST-related repression. This importance is evidenced by the fact that REST ChIP-seq experiments have already been published, although both were done in the non-neuronal Jurkat T-cell line [76, 79]. To investigate the REST binding program in a setting more relevant to how REST controls neuronal differentiation, a REST ChIP-seq experiment was performed in a human neuronal precursor cell (NPC) line (derived from fetal cortex in the Svendsen lab) in the absence or presence or of FGF, as well as in a non-neuronal cell line, 293T cells. The NPC cells should remain precursor cells in the presence of the growth factor, FGF, while the NPC cells should begin to differentiate down a neuronal path when FGF is taken away.

The REST ChIP-seq experiment in the presence of FGF (referred to as the REST NPC experiment) yielded 2.6 million mapped tags, resulting in 1,889 peaks. Motif analysis revealed the REST binding site as all of the most significant motifs found (Table 3.18). Without FGF, the NPC cells will spontaneously begin to differentiate down a neuronal path. The REST ChIP-seq experiment in the absence of FGF (referred to as the REST dNPC experiment) yielded 2.1 million

mapped tags, resulting in 754 peaks. Motif analysis revealed the REST binding site as all of the most significant motifs found (Table 3.19). The REST ChIP-seq experiment in 293T cells yielded 2.1 million mapped tags, resulting in 407 peaks. Motif analysis found the REST binding site as all of the most significant motifs found (Table 3.20).

The results found are as expected. Upon differentiation, protein levels of REST are supposed to decrease [93], which is validated by the fewer peaks found in the dNPC data versus the NPC data. Although peaks are still found, this can be explained by the incomplete differentiation of the NPC cells. If these were the only results found, the data would be otherwise unremarkable. However, an interesting result was observed when comparing the locations of the binding regions to promoters. Normally, REST binding is found distal to TSS, as evidenced by the TSS overlaps of REST in the Wold data [76] with 127 out of 1,874 peaks being proximal (6.8%), in the Sidow data [79] with 188 out of 4,438 peaks being proximal (4.2%), or in the 293T data with 32 out of 397 peaks being proximal (8.1%). The NPC data had 396 out of 1,889 peaks being proximal (21.0%), while the dNPC data had 303 out of 754 peaks being proximal (40.2%). This large difference between the non-neuronal tissues and the NPC or dNPC cells was investigated further by looking at the relative binding strength of overlapping peaks between the NPC and dNPC experiments. By defining enrichment of overlapping peaks as the logarithm of the ratio of NPC to dNPC tag

numbers, peaks that are more strongly associated with NPC cells or with dNPC cells can be found as positive or negative values, respectively. Motif analysis on the REST binding more enriched in NPC cells found the REST binding site as the top several hits (Table 3.21). However, motif analysis on the REST binding more enriched in dNPC cells found a weak ETS motif (Table 3.22). While the identity of the specific ETS factor cannot be known from just motif analysis, the surprising result is not the weak ETS site, but the lack of the REST site being found. The significance of the large subset of promoters bound by REST in the dNPC is unknown as the Genome Ontology analysis found no interesting associations besides overlapping with other REST experiments, and the data is still under investigation.

3.3.5 Analysis of Histone Modifications in $CD4^+$ T-cells

Histones are a specific class of proteins that make up the nucleosome particle. DNA is wrapped around the nucleosome to enable efficient packaging and regulated processing. The generally accepted model for this regulation is differential combinations of modifications on the N-terminal histone tails which protrude out of the nucleosome-DNA structure, a model commonly called the "histone code" [95]. Several recent papers have utilized the ChIP-seq assay to more efficiently explore this code, and here I look at two papers which demonstrate the utility of next-generation sequencing for this purpose [66, 68].

Between these two papers, 41 different profiles are assayed, including 20 histone methylations, 18 histone acetylations, as well as the boundary element protein CTCF, the histone variant H2A.Z, and Pol II. Additionally, DNase hypersensitive sites assayed in $CD4^+$ T-cells are also considered [96]. These particular datasets were chosen because it is the largest dataset arising from the same cell line ($CD4^+$ T-cells), making them the best group of data to investigate the utility of the Genome Ontology.

Although looking at each individual experiment's Genome Ontology results would eventually lead to a better understanding of how these histone marks overlap with each other, a much faster method is to cluster the results together. By using the Eisen lab's Cluster 3.0 program [97], the pairwise Genome Ontology comparisons can be clustered together, and similar ontology terms will fall near each other in the resulting tree. Using the Eisen lab's TreeView program allows us to visualize and browse this tree (Figure 3.7). The most noticeable trend is that all of the known activation marks cluster together (Figure 3.8). These marks include all of the histone acetylations, Pol II, H2A.Z, CTCF, LSD1, DNase hypersensitive sites, and eight different histone methylations. On the silencing side, four methylation marks clustered together. These were tri-methylations of H3K27, H3K79, H3K9, and H4K20 (Figure 3.9).

Previously discussed data also showed informative clustering. The ERα, AR, and FOXA1 experiments clustered together (Figure 3.10), reinforcing the

correlation found earlier in Section 3.3.3. The REST experiments also clustered together (Figure 3.11). Interestingly, an LSD1 ChIP-seq experiment did not cluster with ERα, despite being associated with some ERα bound genes [11]. This difference can be explained because despite a visible correlation between LSD1 and ERα, the vast majority of LSD1 sites overlap activation marks (Figure 3.12).

These findings provide an important framework for future analyses. Because the activation and repression marks clustered separately from each other, any future experiment that specifically clusters into either of these groups can be assigned as predominantly active or silent. Other clustering results are also useful, as the ERα/AR/FOXA1 and REST clusters would reveal any future associated histone marks or proteins. With all Genome Ontology analyses, the only biologically meaningful comparisons are between those experiments done in the same background. This caveat explains why LSD1 and REST did not overlap strongly, as the LSD1 experiment was done in MCF7 cells while REST was done in NPC cells.

The Genome Ontology approach is also useful for describing gene-centered trends in the histone mark data. By using known annotations for exons, introns, and UTRs, another set of genomic regions can be used in the Genome Ontology analysis. For example, the Pol II data was compared to five different annotations: TSS, 5’ UTRs, 3’UTRs, exons, and introns (Table 3.23). As expected, Pol II shows enrichment over all of these gene-centered annotations,

with the strongest enrichment occurring on 5' UTRs and TSS. Tri-methyl-H3K36 is a histone mark associated with transcriptional elongation [98]. According to the gene-centered annotation profile, this mark is most strongly enriched on introns and exons, but not enriched on TSS or 5' UTRs (Table 3.24). Intriguingly, although this mark clusters with the other activation marks, it does have a significant overlap with the cluster of four silencing marks (Figure 3.13). The ability to take other non-experimental annotations and run the same analysis makes the Genome Ontology a robust solution to genome-wide analysis.

3.4 Conclusions

High-throughput sequencing has had a profound effect on genomic studies, as whole-genome datasets can now be generated in the span of days instead of weeks or months. This increased flux of data enables a single lab to generate dozens of genomic profiles with only limitation being the quality of antibody used. However, an increase in data flow does nothing to improve analysis methods, as evidenced by many of the recently published papers each using different methods to declare significance [62, 66-70, 74, 76, 77, 79]. Restricting ourselves to these individualized analyses belies the utility of high-throughput sequencing data. When any lab can generate a genome-wide dataset in a week, the hundreds of datasets that are being published will be almost impenetrable as a whole.

As a putative solution to this problem, I have developed a simple statistical framework for performing this analysis, called the Genome Ontology. This method is simple to use because all it requires is a BED-formatted file (as commonly used by the UCSC genome browser). In terms of sequencing, this file would represent peak locations as calculated from a tag enrichment algorithm. However, only requiring a BED-formatted file enables the use of many different data sources. ERα ChIP-on-chip data has been successfully analyzed using this method, and was found to be highly similar to the ERα ChIP-seq experiment (Table 3.5). It was also seen that despite not finding the expected FOXA1 binding site from motif-finding in the ERα ChIP-seq, the Genome Ontology was able to discover the association. The results seemed to point towards a FOXA1 association only with the E_2-dependent ERα binding regions, which explains why the FOXA1 site was not found in the total ERα binding regions. More experiments need to be done to reveal the importance of the E_2-independent ERα binding regions. Predefined genomic annotations such as introns and exons have also been used to quickly verify the known associations of Pol II and tri-methyl-H3K36 with their respective genic locations. Allowing different input data types enable the user to annotate any desired subset of genomic regions to be used in the Genome Ontology.

Subsequent analysis of several AR ChIP-seq datasets revealed an interesting phenomenon. In both the AR –DHT and AR –DHT/+si*FOXA1*

experiments, very few peaks were found, compared to the AR +DHT and AR +DHT/+si*FOXA1* datasets. This means AR is not binding to many regions in the genome, which is probably caused by a lower expression of the *AR* gene. This lower expression is supported by the fact that the AR +DHT/+si*FOXA1* experiment has many more peaks than the AR +DHT. Because FOXA1 is a known corepressor of the *AR* gene [89], the AR +DHT/+si*FOXA1* experiment probably overexpressed *AR*, enabling it to bind several additional locations. Motif-finding of these AR ChIP-seq experiments supports this explanation, as the AR +DHT dataset contained the FOXA1 binding site as the third ranked motif (Table 3.6) while the AR +DHT/+si*FOXA1* did not contain the FOXA1 binding site (Table 3.7). These results suggest FOXA1 is required for proper regulation of *AR* expression and cooperative binding to appropriate AR binding regions. It would be interesting to see how FOXA1 affects AR in normal prostate cells, as well as other prostate cancer cell lines. LNCaP cells are an AR-expressing cancer line, and their proliferation is androgen-dependent. Two other prostate cancer cell lines, PC-3 and DU-145, express little AR, and their proliferation is androgen-independent. Repeating the AR and FOXA1 ChIP-seq experiments in normal prostate and the androgen-independent cancer cell lines would be very informative for investigating cancer cell progression from normal prostate to either LNCaP cells or PC-3 cells.

The REST ChIP-seq experiments in human NPC, dNPC, and 293T cells also revealed a novel finding. On the surface, the data looked much like other REST ChIP-seq data performed in non-neuronal cell lines in that the REST motif was found as the top motifs in all cases (Tables 3.18-3.20). The Genome Ontology discovered that an unusually high percentage of REST sites in the NPC and dNPC datasets were overlapping TSS. Further analysis of the REST binding regions found in both the NPC and dNPC datasets showed that the peaks more strongly present in the dNPC experiment made up the majority of these proximal peaks. This result suggests that upon differentiation a subset of promoters may actually recruit REST. Although this might indicate that these target genes are repressed, recent data suggests REST can have an activation role in mature neurons [99, 100]. To further investigate what role the increased binding of REST on TSS has, an RNA expression analysis needs to be done to describe how these targets are being regulated during differentiation. The expression analysis would also help detect which REST splice variant is being expressed, which is an essential distinction in the recently published activation roles of REST.

The Genome Ontology is most useful when using multiple datasets from the same background. Recent publications contain ChIP-seq experiments from 41 different antibodies in the same $CD4^+$ T-cell background [66, 68]. Rather than exploring each experiment's Genome Ontology results, hierarchical clustering was performed to help visualize groups of related marks (Figure 3.7). The

clustering results showed a clear association of many known activation histone marks (Figure 3.8), while only a few known repression histone marks clustered together (Figure 3.9). By overlapping experiments with gene-centered annotations, some histone marks could be additionally validated, as Pol II overlapped very well with TSS and 5' UTRs (Table 3.23), while the transcriptional elongation mark, tri-methyl-H3K36 was most highly associated with introns and exons (Table 3.24). The same clustering diagram can be used to easily view the associations among the ERα, AR and FOXA1 ChIP-seq experiments (Figure 3.10), as well as the associations among the different REST ChIP-seq experiments (Figure 3.11). Interestingly, although the LSD1 ChIP-seq data unsurprisingly clustered with the known activation histone marks, the association with a subset of ERα binding regions can also be seen (Figure 3.12).

Overall, the Genome Ontology has been a very useful tool for incorporating multiple data sources into a single framework. Data from any genome-wide assay can be used, whether from ChIP-on-chip or any next-generation sequencing method, as the program only requires a list of regions of interest. In fact, any genomic annotation can be used, as shown with the overlapping of peak data with introns, exons, UTRs, and TSS. Additional uses for the Genome Ontology will be discussed in Chapter 4.

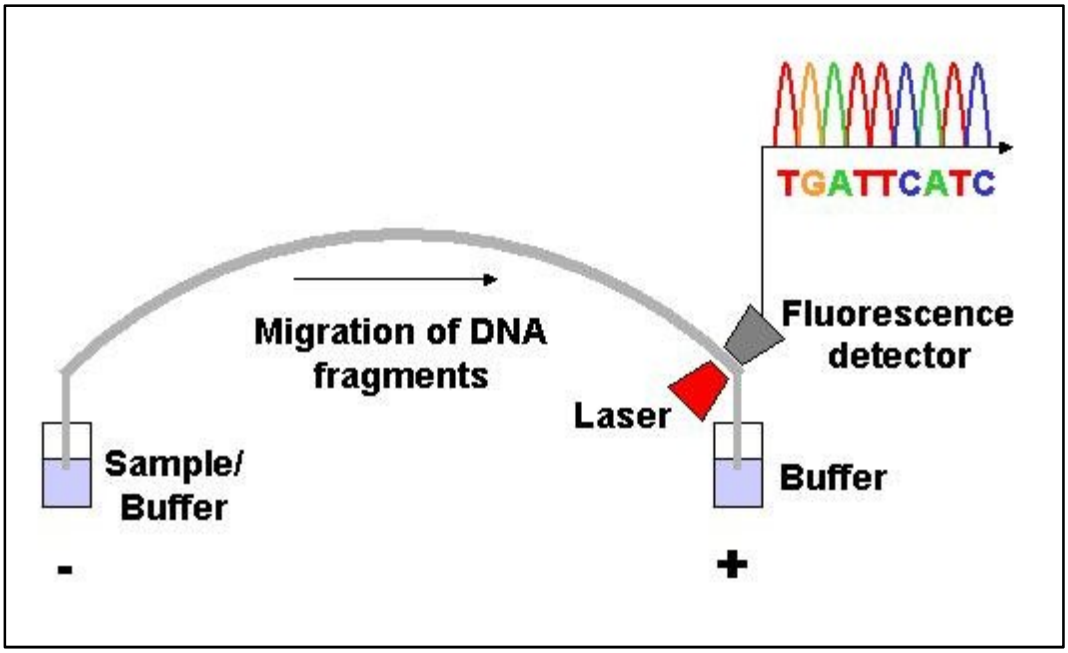

Figure 3.1 Schematic of dye-terminating sequencing

Amplified sample of specific DNA sequence is primed and elongated using the dye-terminating sequencing method. A mix of dNTPs and fluorescently labeled ddNTPs are used, resulting in multiple elongation products of all possible lengths ending in fluorescent label. The now-labeled sample is separated through capillary electrophoresis, with the shortest fragments traveling fastest towards the optical sensor. Peak-detection of one of the four specific fluorescent labels (one for each of the ddNTPs) indicates the sequence at that position. This picture was taken from http://en.wikipedia.org/wiki/Dna_sequencing.

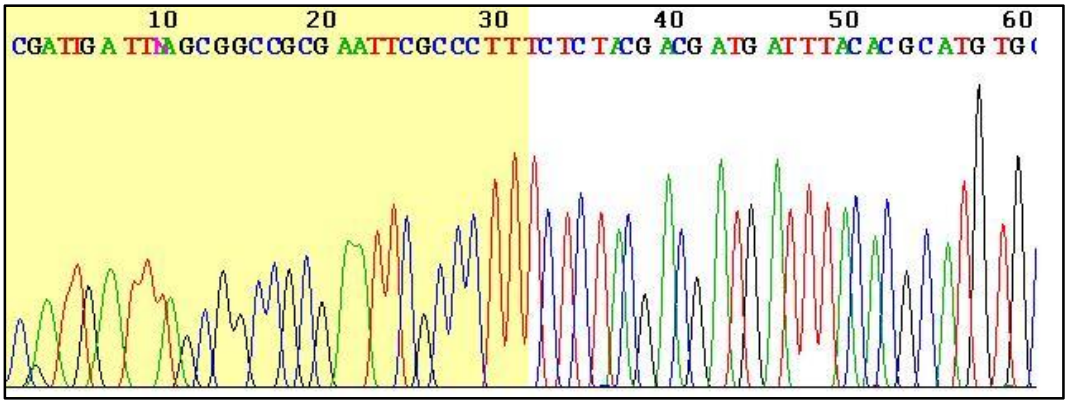

Figure 3.2 Dye-terminating sequencing output

The optical detection of fluorescent labels from size-sorted DNA fragments results in ordered base calls based on peak color. This picture was taken from http://en.wikipedia.org/wiki/Dna_sequencing.

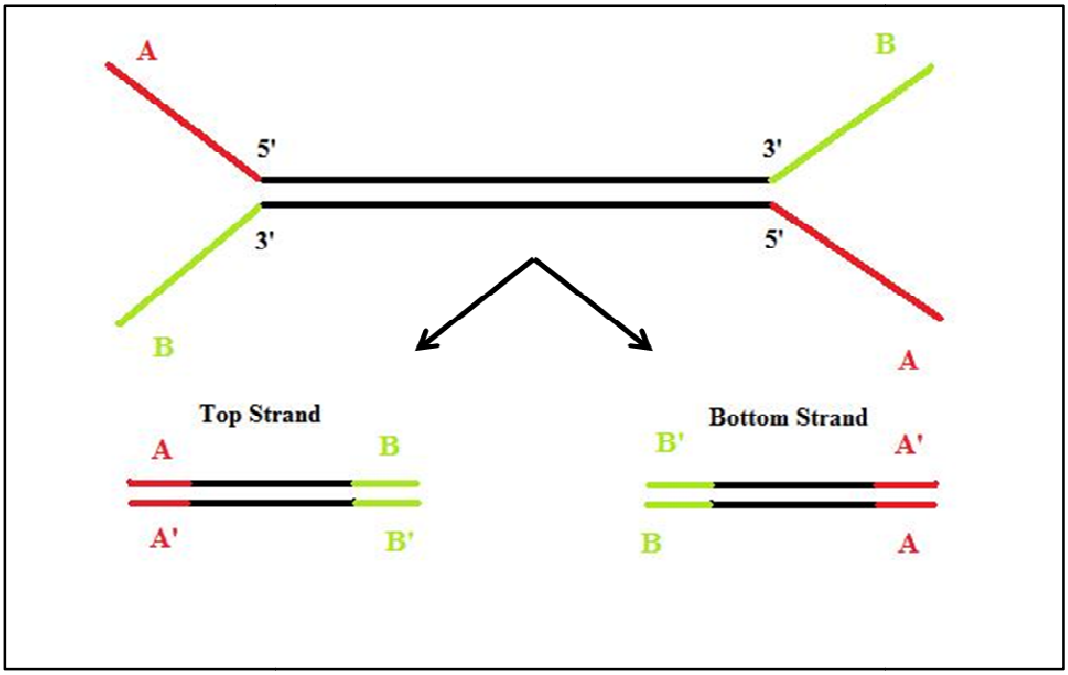

Figure 3.3 Genome Analyzer adapter ligation and amplification

Proprietary ligation of GA adapters, A and B, ensures specific ligation of the A-adapter on the 5' end of each strand and the B-adapter on the 3' end of each strand (top). Subsequent PCR amplification results in two slightly different products of the same DNA fragment such that the sequencing priming sequence (A') is found on the 3' side of both strands.

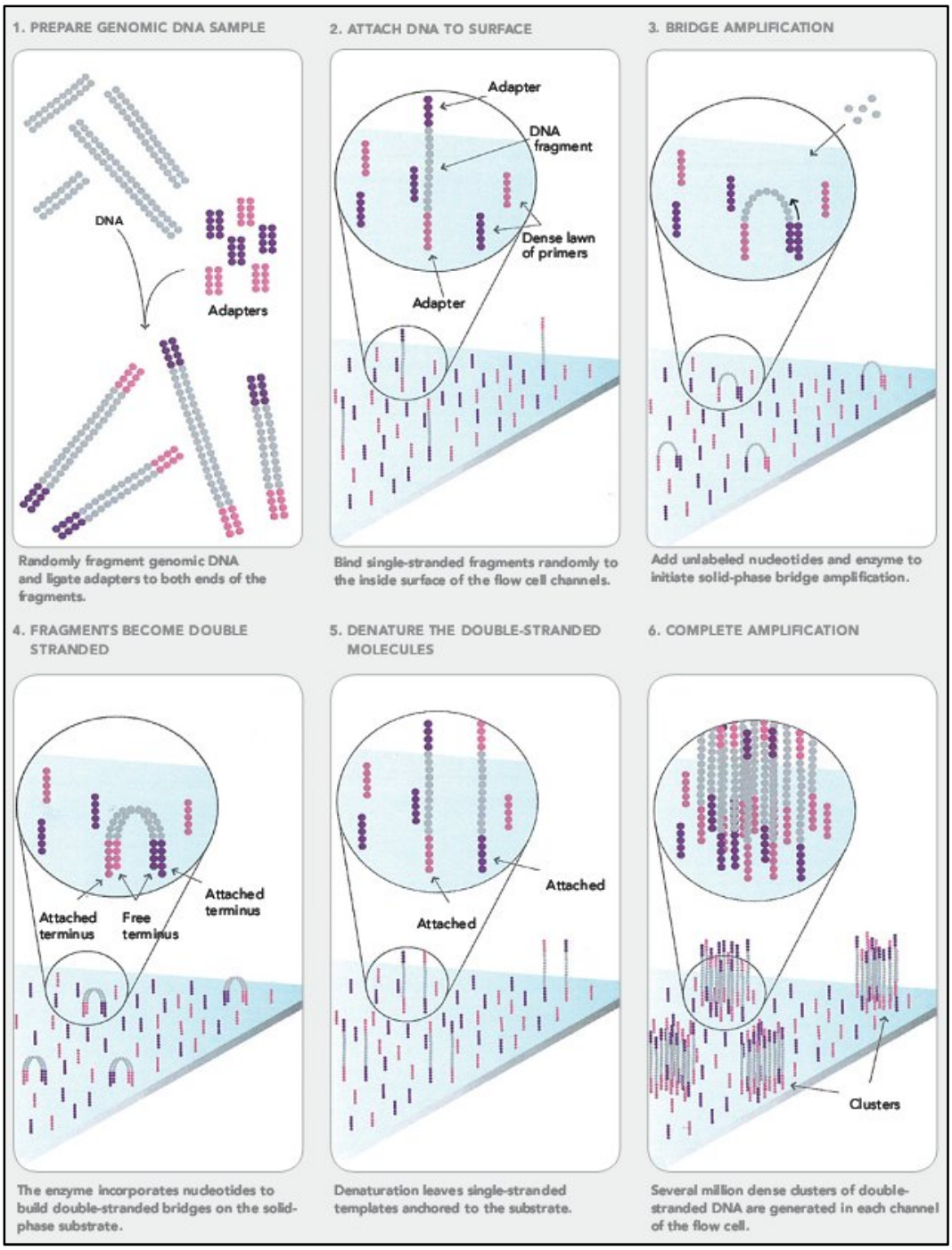

Figure 3.4 Schematic of cluster generation

This picture was taken from http://seqanswers.com/forums/showthread.php?t=21.

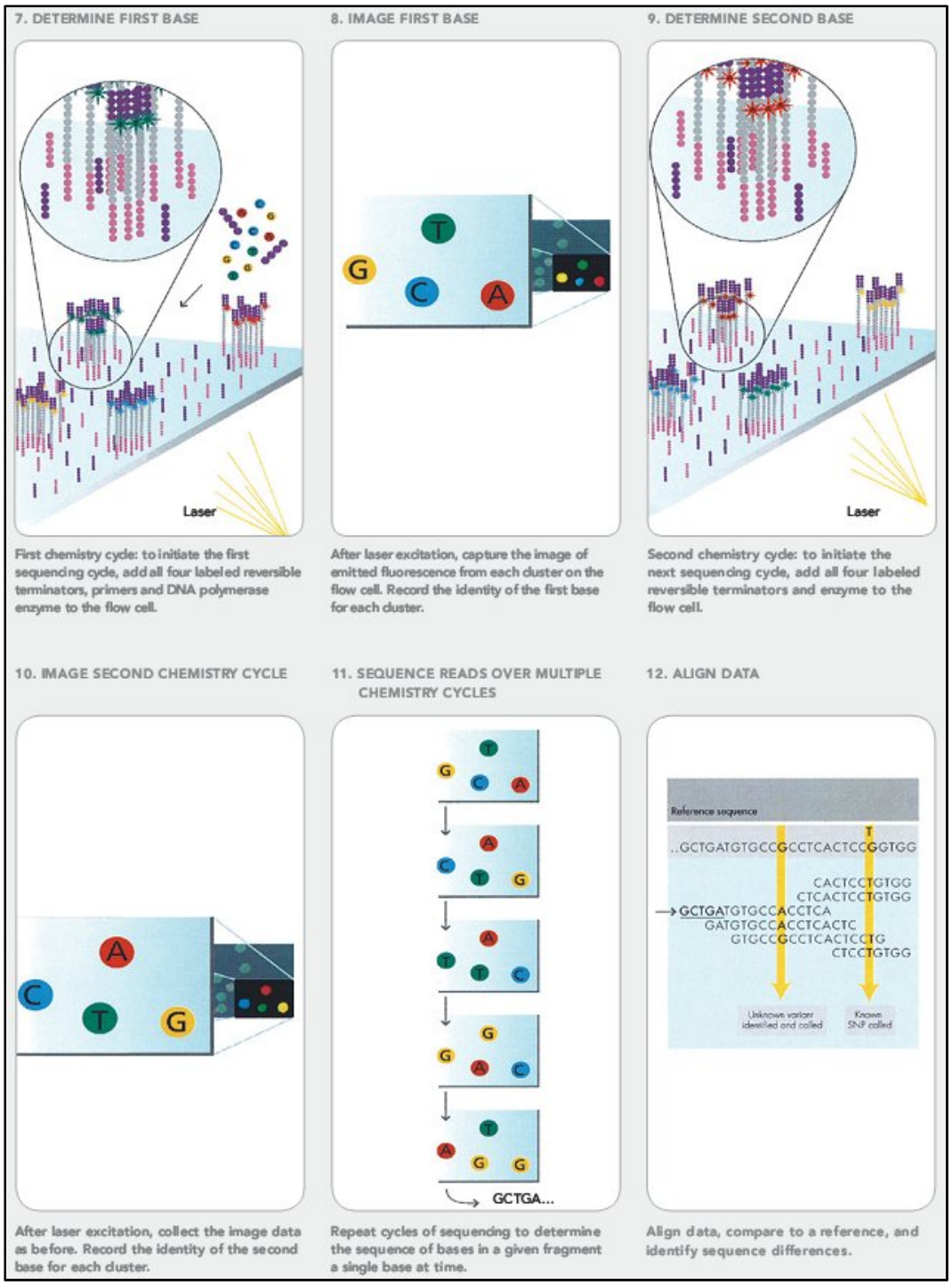

Figure 3.5 Schematic of GA sequencing reaction

This picture was taken from http://seqanswers.com/forums/showthread.php?t=21.

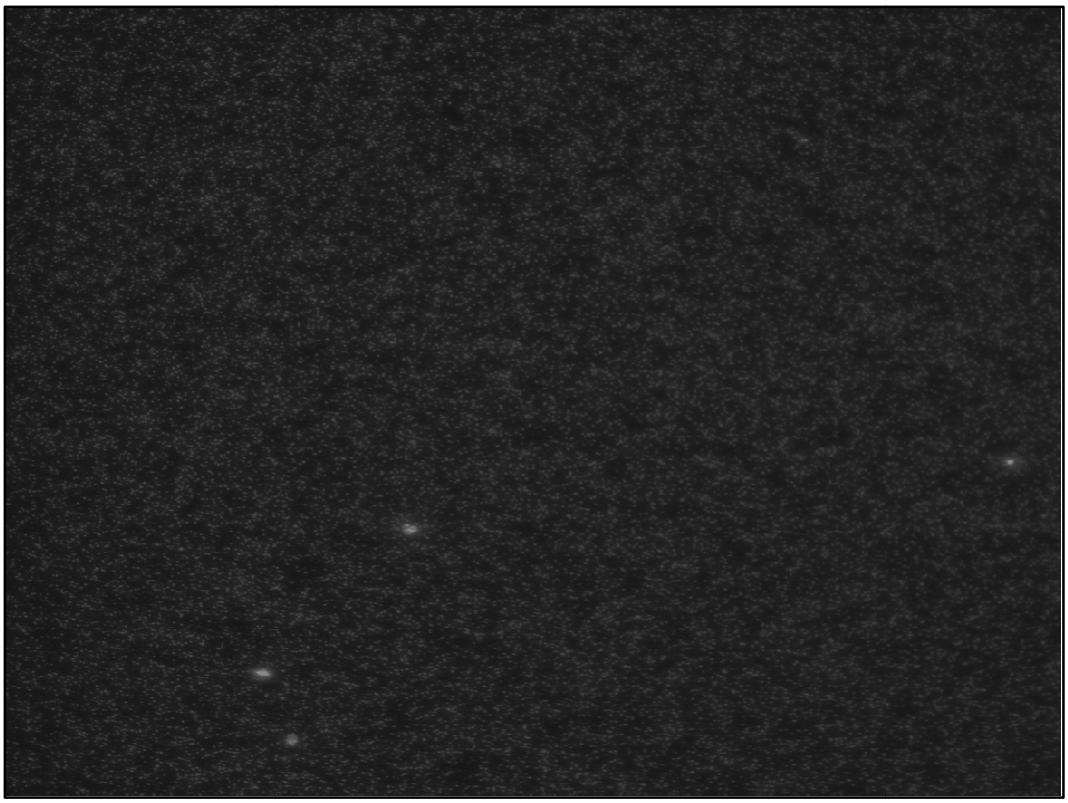

Figure 3.6 Image capture of a single base within one sequencing cycle

Each spot is a cluster representing several hundred copies of a single DNA fragment. Because only one of the four possible nucleotides is recorded at one time, approximately three times more spots are not seen. A typical flowcell will contain over 100 million clusters.

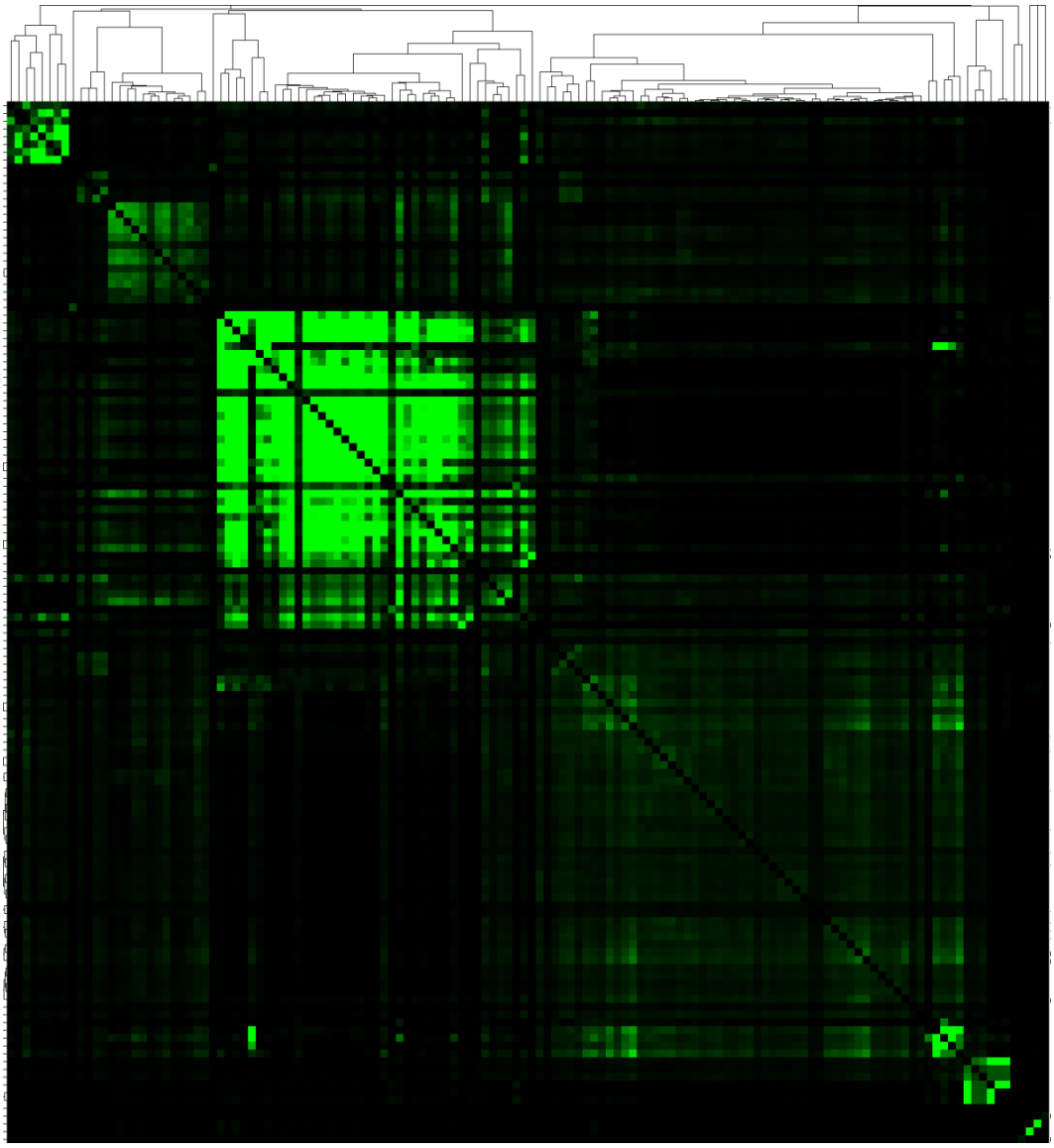

Figure 3.7 TreeView output of Genome Ontology clustering

A different experiment is represented in each column, which is mirrored across the diagonal. A row/column combination has the value of the ln(p-value) as calculated by the Genome Ontology overlap analysis of those two experiments. The larger the value, the more green it is represented.

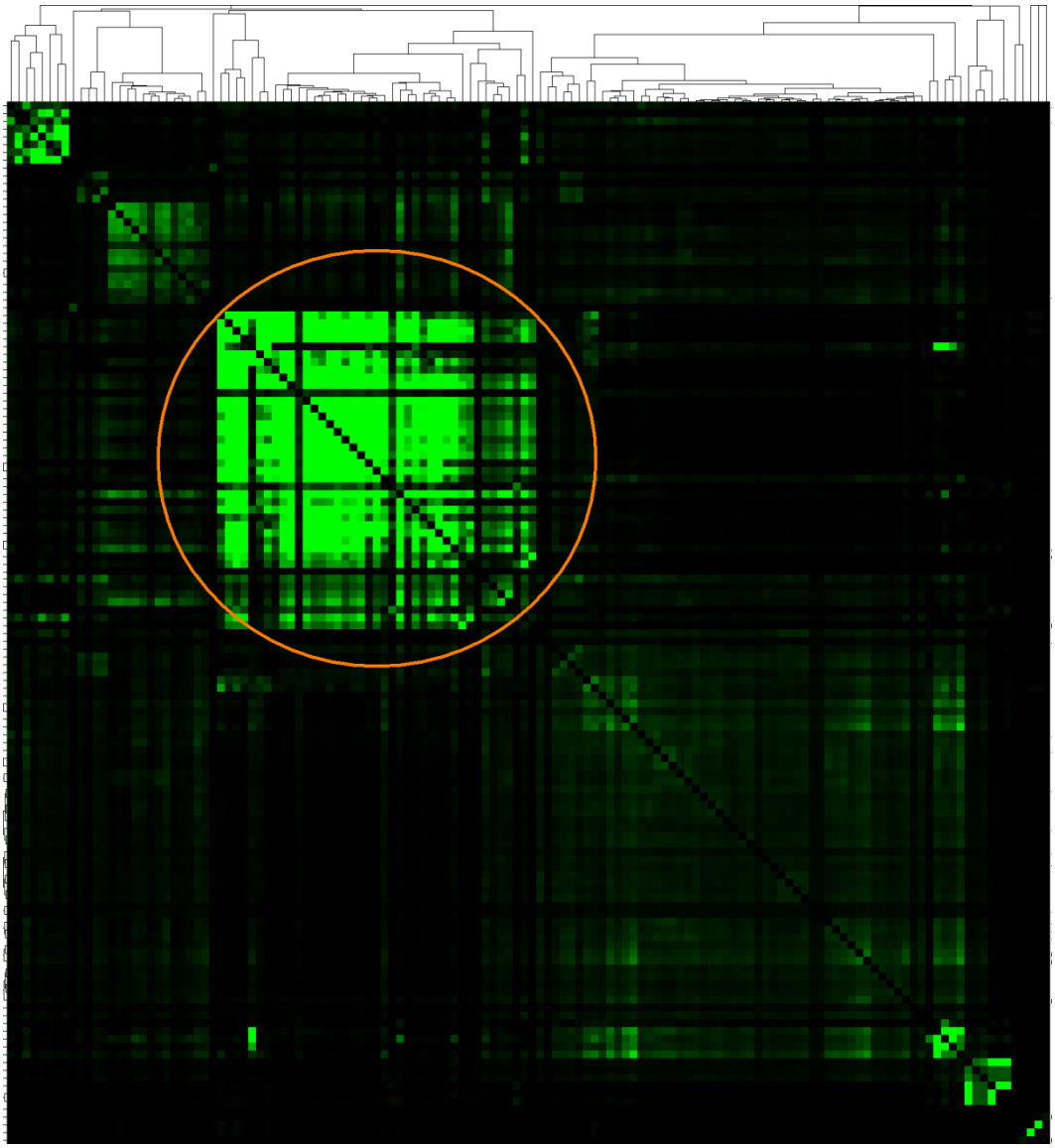

Figure 3.8 TreeView output of Genome Ontology clustering: activation marks

Known activation histone marks are circled in orange.

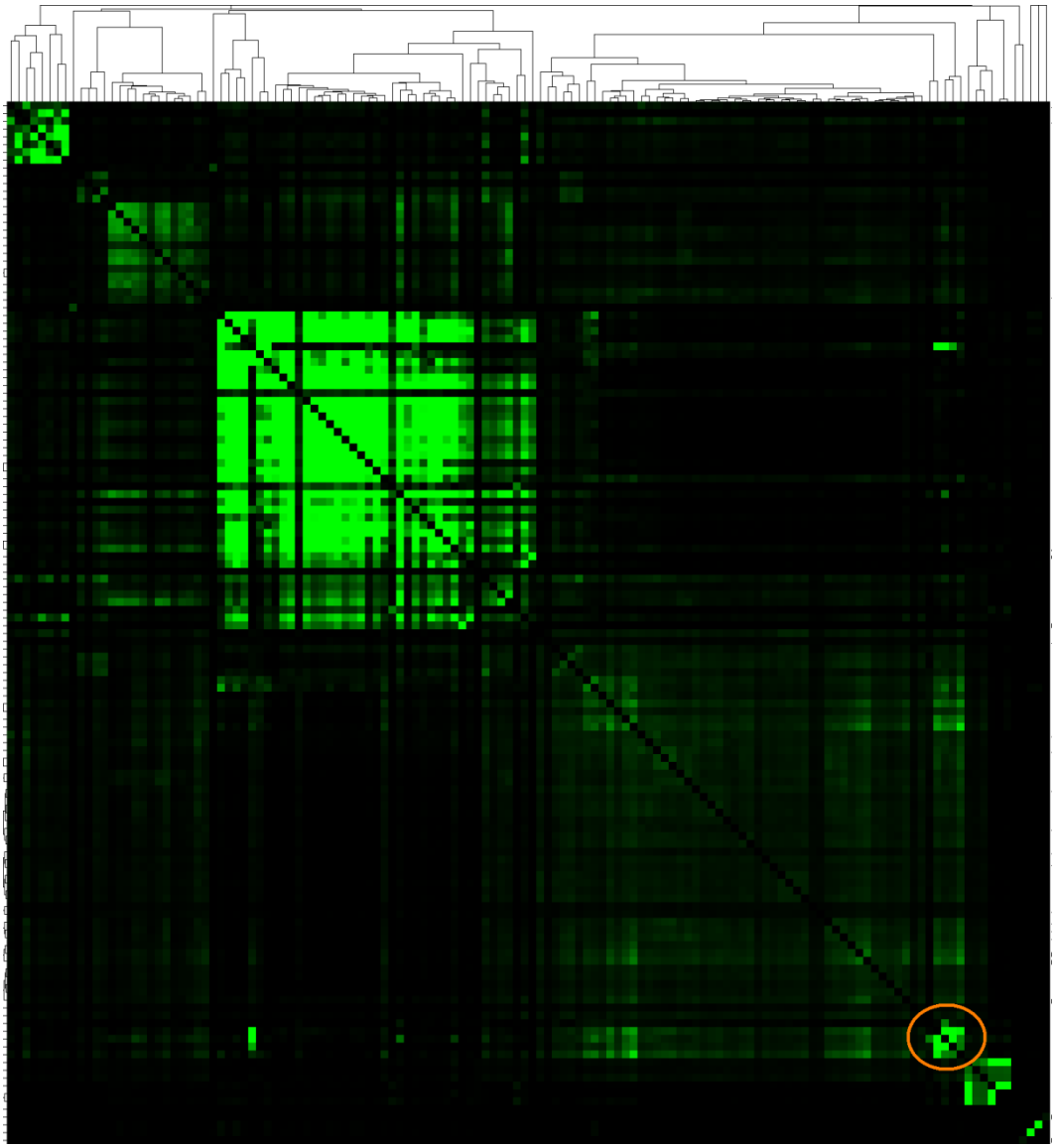

Figure 3.9 TreeView output of Genome Ontology clustering: silencing marks

Known silencing marks are circled in orange.

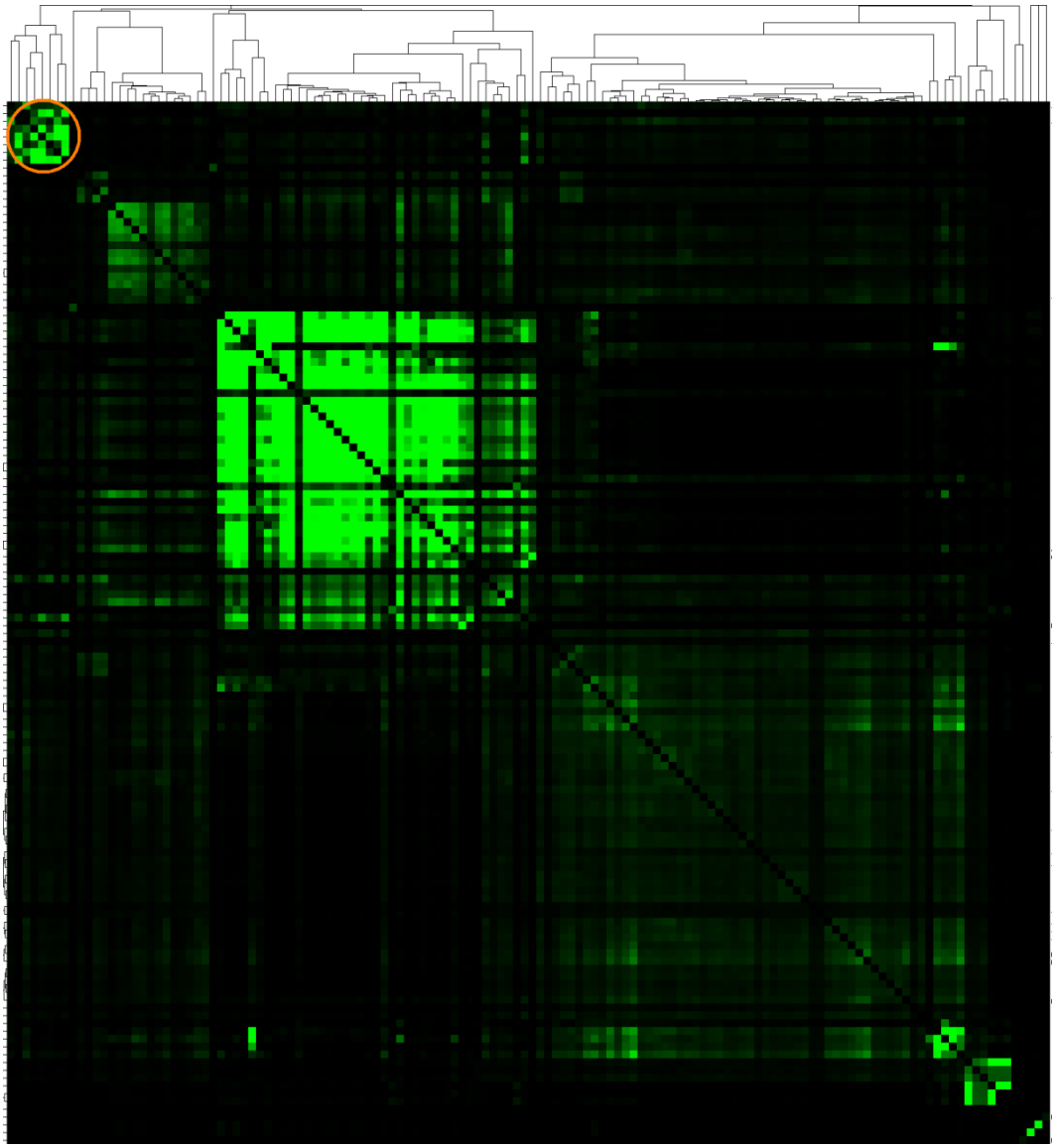

Figure 3.10 TreeView output of Genome Ontology clustering: FOXA1, ERα, and AR

The overlap of ERα, AR, and FOXA1 data with themselves is circled in orange.

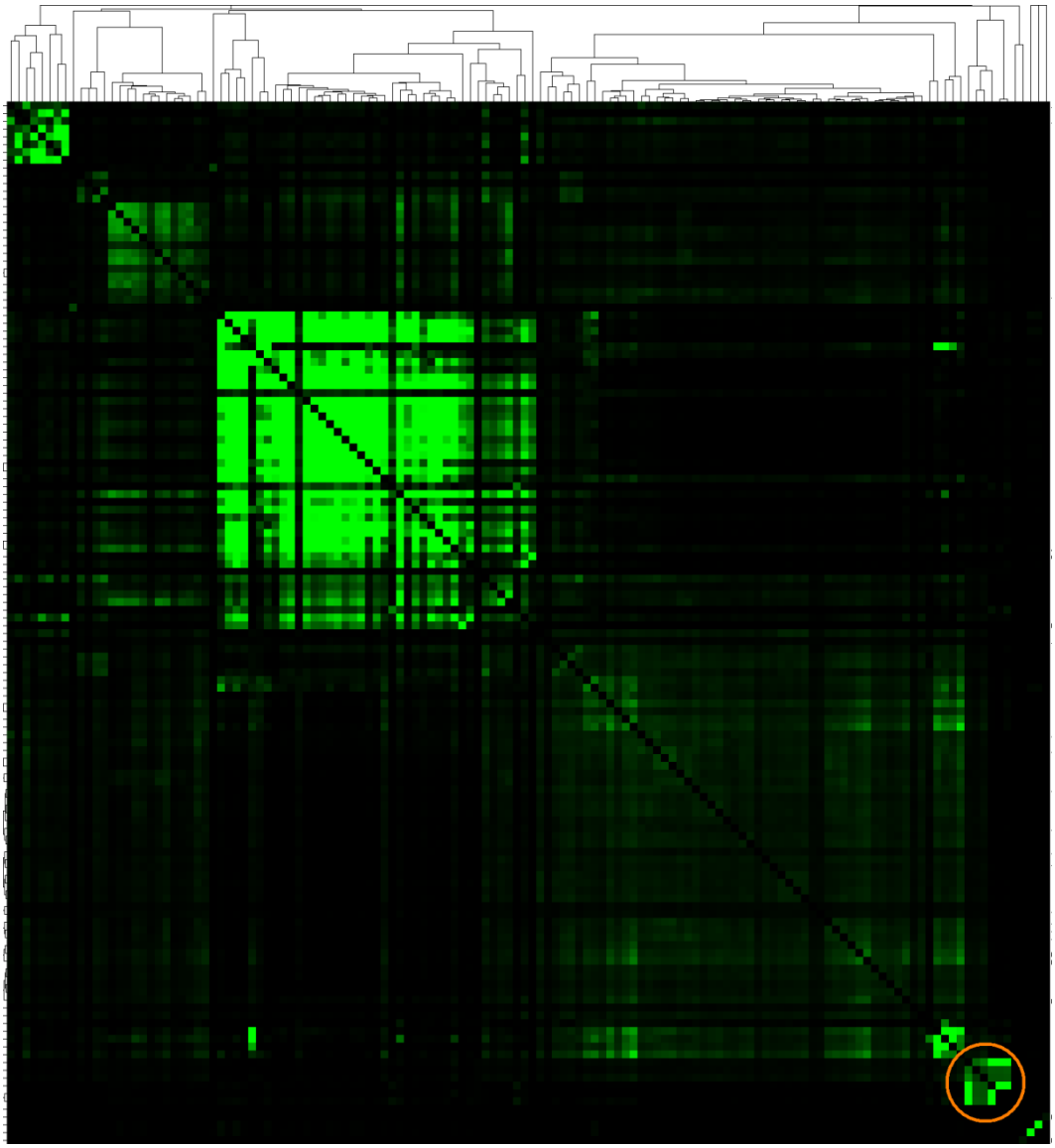

Figure 3.11 TreeView output of Genome Ontology clustering: REST

The overlap of REST/NRSF data with itself is circled in orange.

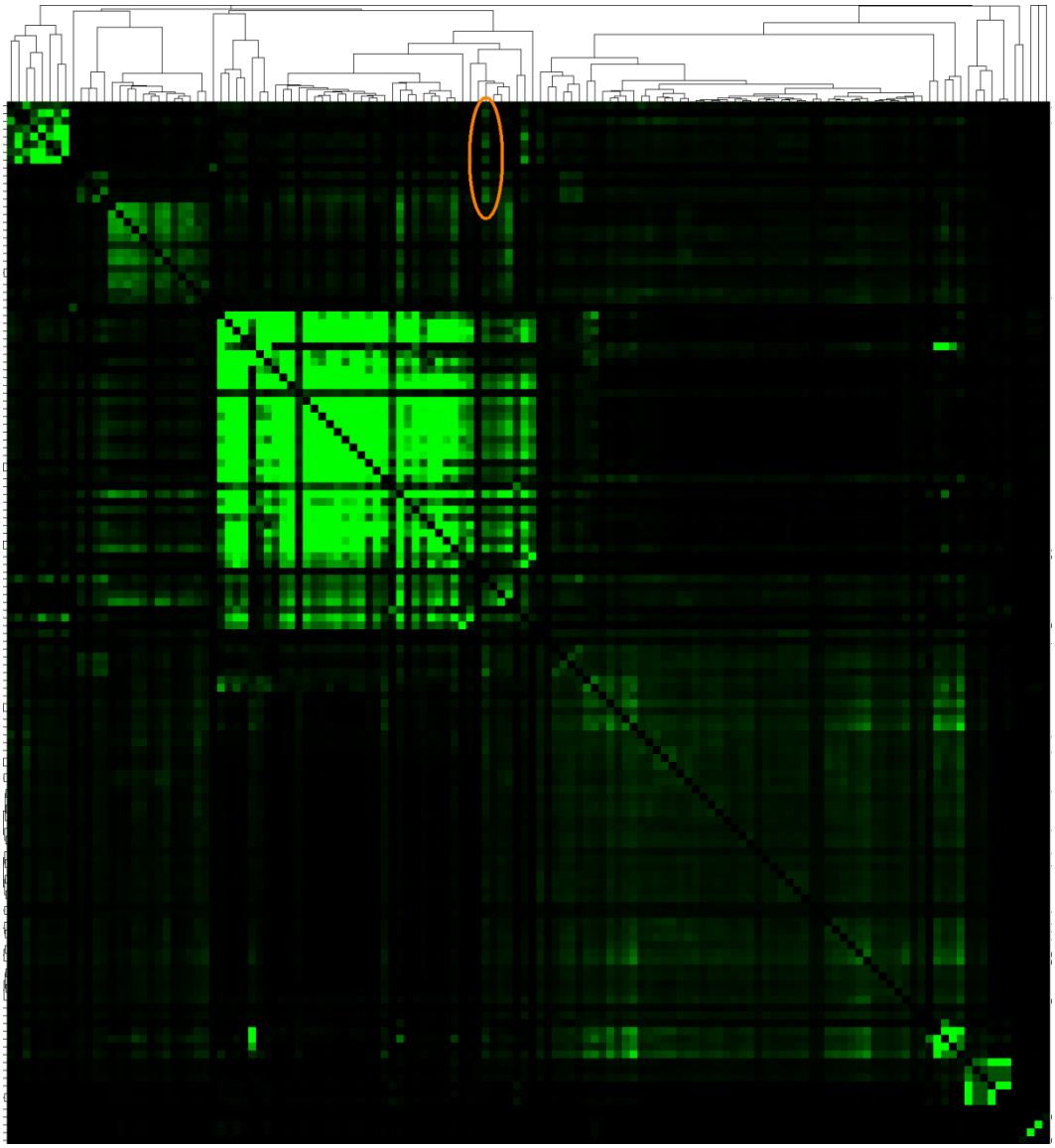

Figure 3.12 TreeView output of Genome Ontology clustering: LSD1

The LSD1 overlap with ERα, AR, and FOXA1 is circled in orange.

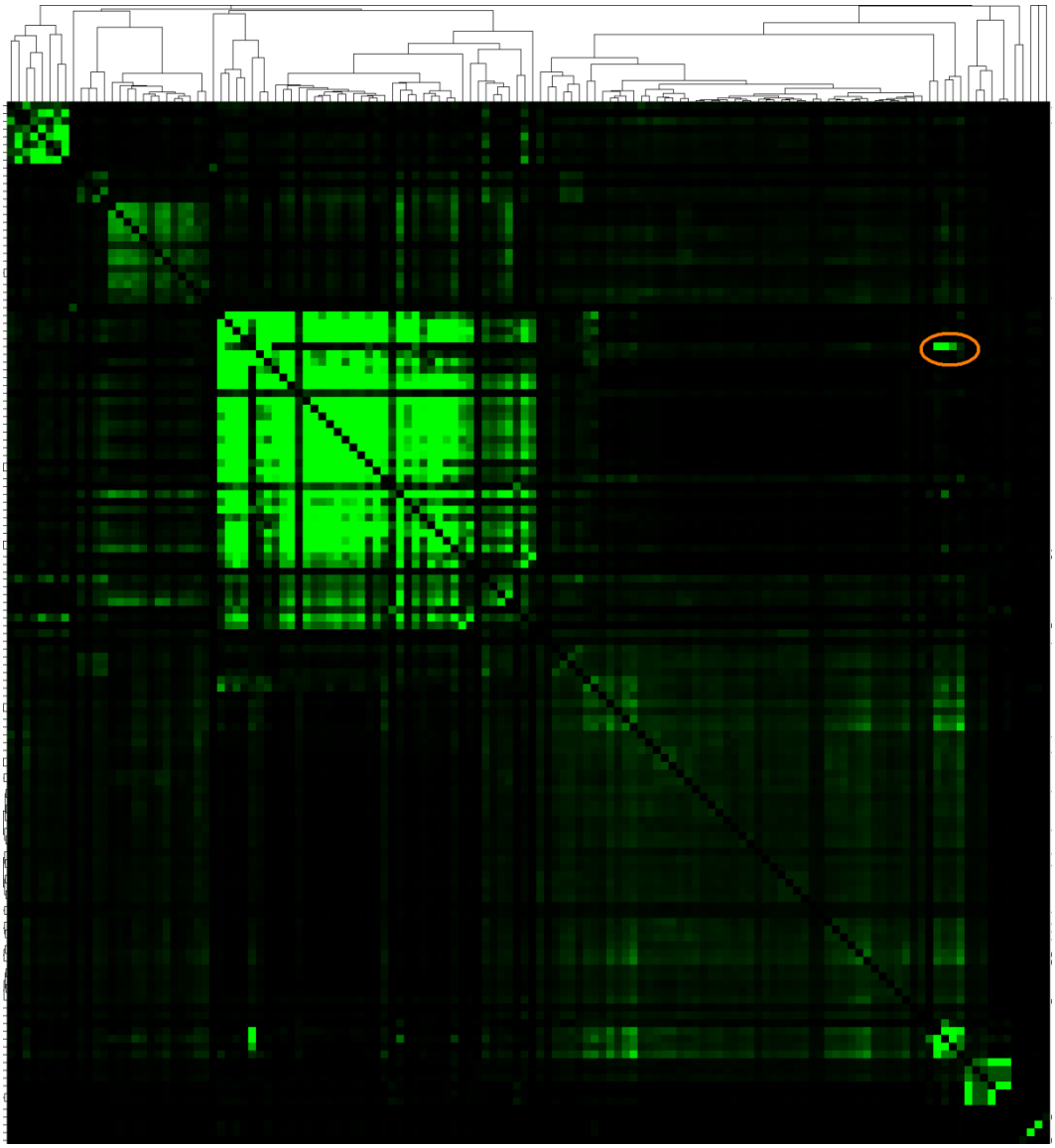

Figure 3.13 TreeView output of Genome Ontology clustering: tri-methyl-H3K36

The tri-methyl-H3K36 data overlap with four silencing methylation marks is circled in orange.

Table 3.1 HOMER motif-finding of ERα ChIP-seq proximal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		4.778e-17	-3.758e+01	CF1 / USP
2		2.290e-15	-3.371e+01	ER
3		9.669e-14	-2.997e+01	ER

Table 3.2 HOMER motif-finding of ERα ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-4.452e+03	PPARG
2		0.000e+00	-4.110e+03	PPARG
3		0.000e+00	-3.867e+03	RORA_1
4		0.000e+00	-3.811e+03	v-ErbA
5		0.000e+00	-1.582e+03	ER

Table 3.3 HOMER motif-finding of ERα ChIP-on-chip data [50]

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-9.332e+02	PPARG
2		0.000e+00	-9.004e+02	ER
3		0.000e+00	-8.709e+02	NR1H2-RXRA
4		0.000e+00	-8.322e+02	v-ErbA
5		2.790e-173	-3.973e+02	Bach1
6		1.371e-136	-3.128e+02	Bach1
7		3.777e-120	-2.750e+02	Freac-4

Table 3.4 HOMER motif finding of ERα ChIP-on-chip peaks not overlapping with ERα ChIP-seq peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		3.032e-136	-3.120e+02	Bach1
2		6.608e-99	-2.261e+02	Freac-4
3		1.038e-82	-1.888e+02	Freac-4
4		1.098e-82	-1.887e+02	Freac-4
5		3.351e-71	-1.623e+02	NR1H2-RXRA
6		7.031e-68	-1.546e+02	ER

Table 3.5 Genome Ontology of ERα ChIP-seq

ERα ChIP-seq peaks were compared with several other genome-wide location analysis experiments and the top 7 results are listed in decreasing order of significance. The top hit corresponds to previously published ERα ChIP-on-chip results [50], while the remaining hits are ChIP-seq experiments.

ln(pvalue)	Partner Description	Overlap Size
-10807.72	ER ChIP on chip	2413
-9289.31	ER MDA MB 231	2563
-4786.28	FOXA1_MCF7_minusE2	1960
-4026.46	FOXA1_MCF7_plusE2	1766
-1717.28	LSD1_MCF7_plusE2	613
-1209.80	FOXA1_LNCAP_plusDHT	853
-1142.79	FOXA1_LNCAP_minusDHT	965

Table 3.6 HOMER motif-finding of AR +DHT ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		6.963e-198	-4.540e+02	AR
2		3.220e-193	-4.432e+02	AR
3		7.572e-141	-3.226e+02	Freac-4

Table 3.7 HOMER motif-finding of AR +DHT/si*FOXA1* ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-2.466e+03	AR
2		0.000e+00	-2.125e+03	AR
3		0.000e+00	-1.381e+03	AR

Table 3.8 Genome Ontology of AR +DHT ChIP-seq

ln(pvalue)	Partner Description	Overlap Size
-3375.10	FOXA1_LNCAP_plusDHT	929
-3326.46	AR_LNCAP_plusDHT_siFOXA1	746
-2127.12	FOXA1_LNCAP_minusDHT	728
-1333.60	FOXA1_MCF7_plusE2	462
-945.50	AR_LNCAP_minusDHT	161
-944.37	AR_LNCAP_minusDHT_siFOXA1	163
-759.10	FOXA1_MCF7_minusE2	306

Table 3.9 Genome Ontology of AR +DHT/si*FOXA1* ChIP-seq

ln(pvalue)	Partner Description	Overlap Size
-3326.46	AR_LNCAP_plusDHT	746
-1316.28	FOXA1_LNCAP_plusDHT	690
-891.641	AR_LNCAP_minusDHT_siFOXA1	188
-877.888	ER_MCF7_plusE2_hg18	469
-744.645	FOXA1_MCF7_plusE2	459
-741.037	FOXA1_LNCAP_minusDHT	555

Table 3.10 HOMER motif-finding of FOXA1 -E_2 ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-5.811e+03	Freac-4
2		0.000e+00	-2.326e+03	Freac-4
3		0.000e+00	-1.701e+03	Bach1
4		0.000e+00	-1.000e+03	HFH-4
5		4.791e-291	-6.685e+02	Freac-4

Table 3.11 HOMER motif-finding of FOXA1 +E_2 ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-6.447e+03	Freac-4
2		0.000e+00	-1.921e+03	
3		0.000e+00	-1.459e+03	HFH-4
4		0.000e+00	-1.078e+03	Bach1
5		0.000e+00	-9.977e+02	Freac-4

Table 3.12 HOMER motif-finding of FOXA1 -DHT ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-1.162e+04	Freac-4
2		0.000e+00	-5.064e+03	FOXF2
3		0.000e+00	-2.979e+03	Freac-4
4		0.000e+00	-2.802e+03	XFD-3
5		0.000e+00	-2.746e+03	HFH-8

Table 3.13 HOMER motif-finding of FOXA1 +DHT ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-7.869e+03	Freac-4
2		0.000e+00	-3.285e+03	Freac-4
3		0.000e+00	-2.078e+03	HFH-4
4		0.000e+00	-1.503e+03	HFH-4
5		0.000e+00	-1.473e+03	Freac-4

Table 3.14 Genome Ontology of FOXA1 –E_2 ChIP-seq

ln(pvalue)	Partner Description	Overlap Size
-57619.65	FOXA1_MCF7_plusE2	11780
-20418.91	FOXA1_LNCAP_minusDHT	6549
-17435.33	FOXA1_LNCAP_plusDHT	5331
-5298.58	ER ChIP-on-chip	1493
-4786.28	ER_MCF7_plusE2	1960

Table 3.15 Genome Ontology of FOXA1 +E_2 ChIP-seq

ln(pvalue)	Partner Description	Overlap Size
-57619.65	FOXA1_MCF7_minusE2	11780
-27667.39	FOXA1_LNCAP_minusDHT	8168
-25331.55	FOXA1_LNCAP_plusDHT	7007
-4605.94	ER ChIP-on-chip	1362
-4026.46	ER_MCF7_plusE2	1766

Table 3.16 Genome Ontology of FOXA1 -DHT ChIP-seq

ln(pvalue)	Partner Description	Overlap Size
-62605.64	FOXA1_LNCAP_plusDHT	14577
-27667.39	FOXA1_MCF7_plusE2	8168
-20418.91	FOXA1_MCF7_minusE2	6549
-2127.12	AR_LNCAP_plusDHT	728
-1261.47	ER ChIP-on-chip	633
-1142.79	ER_MCF7_plusE2	965

Table 3.17 Genome Ontology of FOXA1 +DHT ChIP-seq

ln(pvalue)	Partner Description	Overlap Size
-62605.64	FOXA1_LNCAP_minusDHT	14577
-25331.55	FOXA1_MCF7_plusE2	7007
-17435.33	FOXA1_MCF7_minusE2	5331
-3375.10	AR_LNCAP_plusDHT	929
-1316.28	AR_LNCAP_plusDHT_siFOXA1	690
-1302.29	ER ChIP-on-chip	575
-1209.80	ER_MCF7_plusE2	853

Table 3.18 HOMER motif-finding of REST NPC ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		0.000e+00	-1.648e+03	NRSF
2		0.000e+00	-1.513e+03	NRSF
3		0.000e+00	-1.143e+03	REST/NRSF
4		1.907e-84	-1.928e+02	NRSF
5		1.273e-82	-1.886e+02	NRSF

Table 3.19 HOMER motif-finding of REST dNPC ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		7.854e-201	-4.608e+02	NRSF
2		7.060e-183	-4.194e+02	NRSF
3		1.556e-123	-2.828e+02	REST/
4		2.913e-76	-1.739e+02	NRSF
5		2.602e-20	-4.510e+01	REST

Table 3.20 HOMER motif-finding of REST 293T ChIP-seq distal peaks

Rank	Motif	P-value	ln(pvalue)	Best Match
1		2.347e-203	-4.666e+02	REST
2		6.140e-196	-4.495e+02	NRSF
3		6.631e-119	-2.721e+02	NRSF
4		9.765e-33	-7.371e+01	REST
5		2.505e-30	-6.816e+01	NRSF

Table 3.21 HOMER motif-finding of REST enriched in NPC versus dNPC

Rank	Motif	P-value	log pvalue	Best Match
1		8.843e-114	-2.603e+02	NRSF
2		8.640e-110	-2.511e+02	NRSF
3		1.865e-103	-2.365e+02	NRSF
4		3.412e-87	-1.991e+02	REST
5		4.874e-29	-6.519e+01	NRSF

Table 3.22 HOMER motif-finding of REST enriched in dNPC versus NPC

Rank	Motif	P-value	ln(pvalue)	Best Match
1		4.997e-23	-5.135e+01	FLI_01
2		2.242e-14	-3.143e+01	MOK2_01
3		2.772e-14	-3.122e+01	ETS

Table 3.23 Genome Ontology of Pol II and gene-centered annotations

The Genome Ontology can overlap not just experimental data, but any genome-wide information, including predefined annotations such as exons and introns. The "% in Query" represents how many of the peaks in the query data overlaps with the target annotation, while the "% in Genome" represents how much of the mappable genome is found in the target annotation. Here, 10.81% of the Pol II data overlaps with Refseq defined exons, while exons make up 1.31% of the mappable genome.

Query	Target	% in Query	% in Genome
PolII	exons	10.81	1.31
PolII	introns	47.26	34.59
PolII	UTR3	3.79	0.91
PolII	UTR5	20.23	4.65
PolII	TSS	28.08	0.63

Table 3.24 Genome Ontology of tri-methyl-H3K36 and gene-centered annotations

Query	Target	% in Query	% in Genome
H3K36me3	exons	11.07	1.31
H3K36me3	introns	76.08	34.59
H3K36me3	UTR3	5.27	0.91
H3K36me3	UTR5	2.94	4.65
H3K36me3	TSS	0.26	0.63

4 Future Work

In the previous chapters, I describe two recent technologies which advance the method of DNA sequencing. DSL technology allows a more specific and sensitive detection of DNA, and has been successful in identifying a wider binding profile of ERα [12] and a novel binding profile of LSD1 [11]. Perhaps most interestingly, the DSL method was able to convert the 3C assay into a high-throughput, array-based assay, despite the array not being designed for such a purpose [55]. The second technology I describe is the Illumina Genome Analyzer (GA), capable of high-throughput sequencing. This method opens up several possibilities, as any lab can now quickly generate genome-wide datasets, whether the purpose is genome sequencing, RNA expression profiles, SNP detection, ChIP-seq, or any other sequencing-based assay.

4.1 3C, DSL and High-throughput Sequencing Combined

The DSL and next-generation sequencing technologies are not mutually exclusive, and very powerful assays are now being developed which take advantage of both. For example, the conventional 3C method relies on designing primers surrounding restriction enzyme sites (Figure 2.3), and an immediate improvement being tested is the use of DSL probes to detect the 3C ligation

product. This is easily accomplished by designing the DSL probe (Figure 2.2) across the restriction site such that the two-20mer oligos exist on either side of the restriction site. If the universal primers, T3 and T7, are replaced with the Illumina adapter sequences, the DSL probes can now be sequenced directly rather than hybridized to an array. However, the number of possible interactions is too many to count, let alone design probes for. By taking an idea from the previously discussed 3D method (Section 2.2.3), we can set a target region by assigning the Illumina B-adapter to a single DSL 20mer oligo, such as the *TFF1* promoter. All other DSL 20mer oligos will have the sequencing A-adapter attached. This means that the GA assay will sequence all of the partner 20mer oligos that have been ligated to the target region. Work on this method is currently underway, although several drawbacks exist. First, ligation efficiency of the DSL 20mer oligos is a key step in providing specificity, but the fact that the DSL 20mer pairs all have the same six nt center (arising from the common restriction site) makes this ligation more non-specific. Second, the proposed method is still not open-ended and we are not able to take full advantage of the sequencing technology. An alternative solution is to forego the DSL probes, and just amplify the 3C ligation products by extending from a common target (for example, the *TFF1* promoter), and then use a random primer for the second strand synthesis. The random primer will contain the Illumina A-adapter, such that the GA assay will sequence all of the randomly primed, 3C ligation products from the unknown side. However, variations of this method have been tried and are shown to be extremely noisy

[101-103]. Regardless of method, a genome-wide 3C assay would introduce an incredibly novel dataset, which can be easily integrated into the Genome Ontology.

4.2 Genome Ontology Improvements

The Genome Ontology itself has several optimizations to be implemented. Currently, only the first level of pairwise comparisons is done, meaning that all results come from two experiments. In addition to a p-value assessment, the Genome Ontology also generates three new lists corresponding to which regions overlap and which regions are exclusive to each experiment. These files can be informative, and ideally would become new ontology terms to be used for subsequent analysis. An example of this analysis would be the comparison of the Pol II data to tri-methyl-H3K4. The comparison would give a very significant p-value in its output (as this histone mark is a strong activation mark), but the output would also include three files which correspond to the overlapping peaks, the peaks of Pol II that do not contain the histone mark, and the peaks of tri-methyl-H3K4 that do not contain Pol II. In this simple example, the Pol II exclusive peaks would probably represent promoters that are poised for activation, while the histone mark-exclusive peaks might represent active transcription units being transcribed by a different polymerase (such as tRNAs). However, the vast number of pairwise comparisons makes the implementation of this type of analysis non-

trivial. While the number of pairwise comparisons follows a polynomial growth, recursive pairwise comparisons on the output files lead to an exponential growth, which is unacceptably slow. Some possible solutions to this problem are to only exhaustively search a very small subset of experiments, or set a minimum threshold on the number of regions to be considered (an experiment with only a few peaks is unlikely to be interesting in this type of analysis). Setting a limit on how deep the algorithm should search would work equally well.

In a related optimization, the calculated overlap of two experiments can be further dissected in the same way as the analysis on the REST NPC/dNPC ChIP-seq data. Because the overlap is performed with any significantly enriched peaks, any quantitative information the data holds is disregarded. A pair of overlapping peaks probably should not be defined as overlapping if a peak in one experiment has a drastically higher tag count than the other. This case would normally be interpreted as a significant change in binding affinity, and should be assigned to the appropriate exclusion list. This improvement would definitely help the experimenter interpret their data.

The fact that this phenomenon was not only seen, but was informative in the REST case prompts the experimenter to consider the sensitivity of next-generation sequencing. As evidence of this sensitivity, I present several examples. One of the biggest problems with the sequencing data is not knowing from where a sequence actually came. One particular experiment I analyzed was a PHF8

ChIP-seq in Hela cells. After peak-calling and overlapping with TSS, I found one of the highest-ranking promoters was *ERα* itself, which was an exciting discovery because the demethylase PHF8 was only binding during M phase. Especially with stronger peaks, I prefer to examine the data in the UCSC genome browser, and in this example, I was quite surprised (Figure 4.1). While the overlap analysis correctly indicated a strong peak overlapping the TSS of *ERα*, several absurdly strong peaks were occurring over several of *ERα*'s exons. This finding was definitely a contamination of some sort, but the source was unclear. The fact that this happened over the *ERα* gene gave me the first clue. This gene is extensively studied in our lab, and many constructs have been made which contain the coding sequence of *ERα*. I contended that an *ERα* construct may have contaminated that sample To test my hypothesis, I remapped the sequence data from the entire flowcell to a library of common vector sequences used in lab. The results validated my hypothesis (Table 4.1). For the PHF8 lane 3, almost 40k tags overlapped the vector sequences, while the tri-methyl-H4K20 lane 7 had almost 8k vector tags. The contamination was obviously *ERα* (Figure 4.2). Although the tag counts mapping to vectors seem insignificant when compared to the total tag counts (< 0.5% of the total), the analysis of such contaminations can be very misleading, as well as difficult to detect. In this case, I was fortunate that this contamination happened to be a widely used construct in our lab. Mapping to a vector library is only useful for detecting a contamination, but not in identifying its exact source. A more sophisticated approach involving mapping to an mRNA

library would be essential to identify a contamination. In the case of using the sequencer for RNA profiling, it may not be possible to confidently detect this type of error.

Another intriguing case of sensitivity comes from ChIP-seq experiments performed in a cancer cell line. It is well known that cancers display a significant amount of genomic instability, leading to a variety of chromosomal aberrations such as translocations, duplications, and deletions. However, sequencing data that is mapped to a reference genome does not reflect these changes, and the results can be somewhat confusing. A CLIM ChIP-seq experiment done in the breast cancer cell line, MCF7, displays the problem quite clearly (Figure 4.3). As seen in the peak-calling track using only a global cutoff, a large region surrounding the *BCAS1* gene seems to be covered in CLIM binding regions. However, upon zooming out on this region, other loci are also seen that are covered in CLIM binding, such as *EYA2* and *BMP7* (Figure 4.4), as well as the *BCAS3* gene (Figure 4.5). The regions encompassing these genes are known to be amplified in breast cancer [104, 105], and it is assumed that the widespread signal is coming from higher background sequencing of these amplified regions as compared with non-amplified regions. Even though this analysis was not designed to detect this type of phenomenon, the results visually demonstrate the capability of detecting the amplification aberration. As an initial step, I attempted to extract the boundaries of these amplified regions and run motif-finding, but those results were

inconclusive, probably due the few numbers of amplified regions being identified (only 3). Future steps should be taken to improve this analysis because extracting any information regarding the reason for the amplification of these regions is incredibly valuable.

Related to the amplification aberration is the spreading of certain histone marks. The analysis of such regions would be easy because the analysis would be similar to that of chromosomal amplifications, and the results are no less important. Edges of such chromatin regions are termed “boundary elements,” and currently very few details are available to explain what constitutes a boundary element. Thus far, very few of these elements have been identified, but would provide a necessary component to our annotation of the genome.

The current state of sequencing technology provides science with an unprecedented amount of usable data, and this technology will only improve. Our current analysis methods need to greatly improve to take advantage of these datasets, and I hope the Genome Ontology is one such improvement.

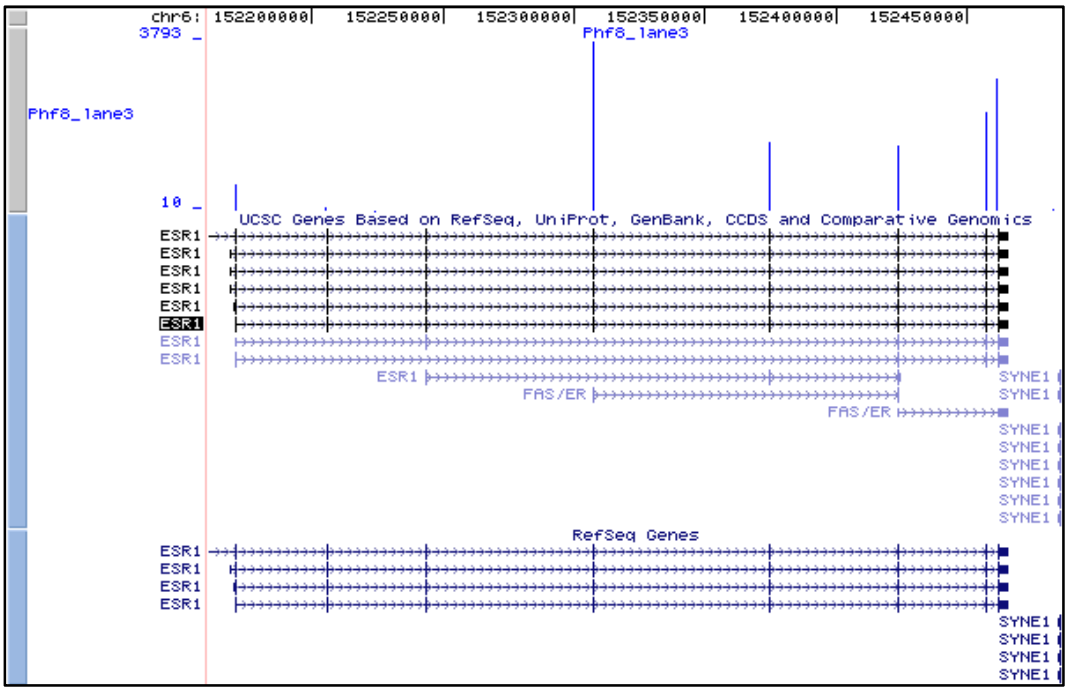

Figure 4.1 UCSC genome browser of PHF8 ChIP-seq over *ERα* locus

Figure 4.2 UCSC genome browser of PHF8 and tri-methyl-H4K20 ChIP-seq over *ERα* locus

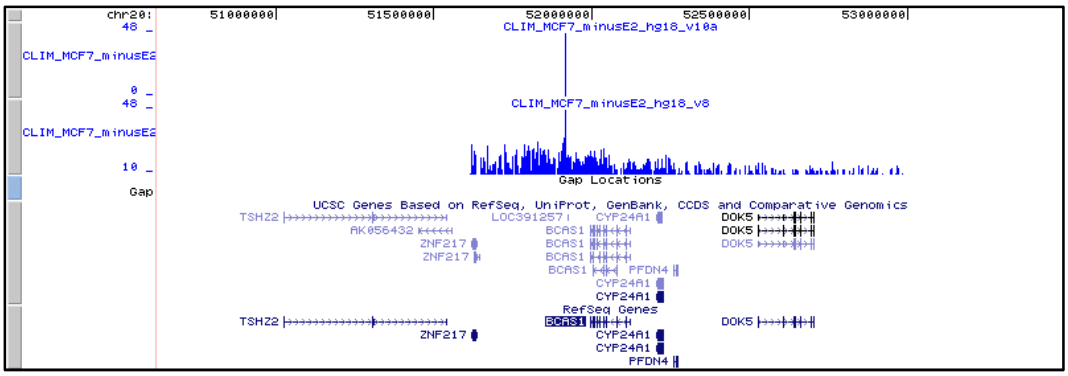

Figure 4.3 UCSC genome browser of CLIM ChIP-seq over *BCAS1*

The top track corresponds to significant peaks after applying an additional local cutoff. The bottom track corresponds to significant peaks after applying only a global cutoff.

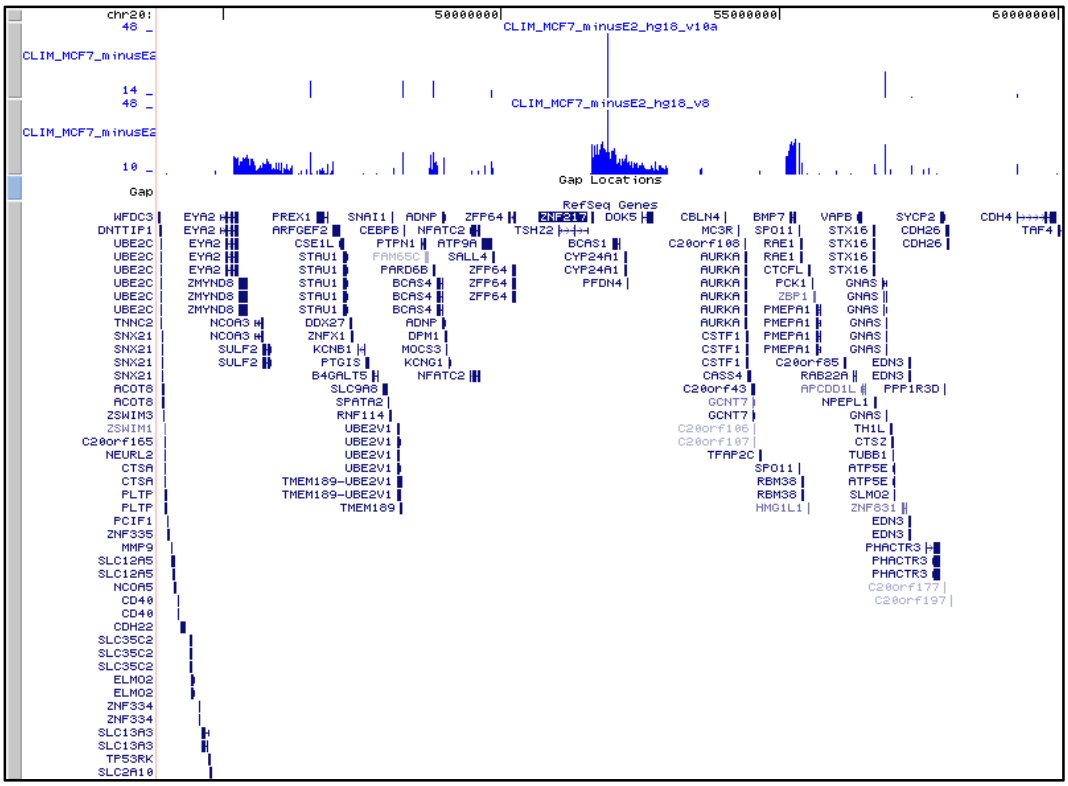

Figure 4.4 UCSC genome browser of CLIM ChIP-seq over an expanded region covering *BCAS1*

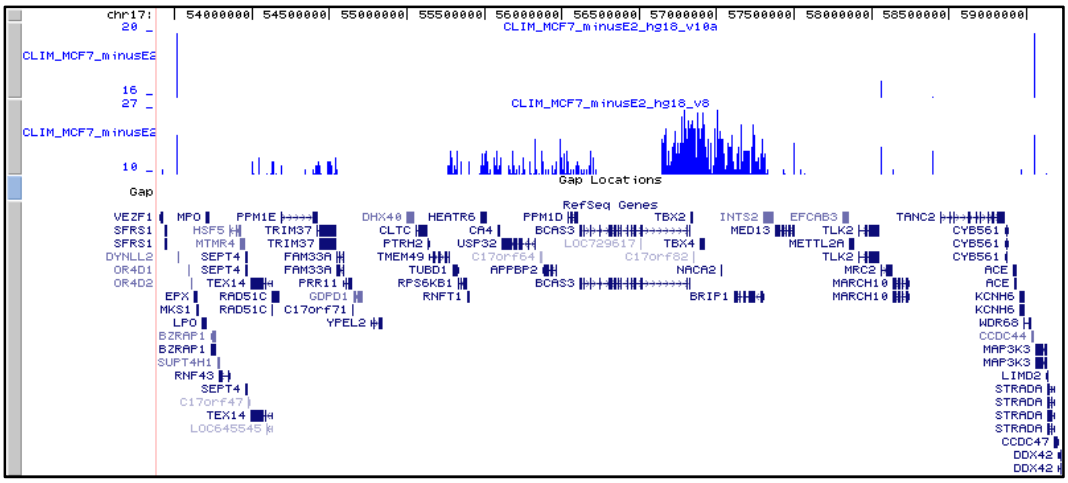

Figure 4.5 UCSC genome browser of CLIM ChIP-seq over *BCAS3*

Table 4.1 Results of aligning PHF8 sequencing data to vector library

Lane	Contamination Tags	Total Tags
1	172	10643693
2	158	6498513
3	38306	8780261
4	86	3707281
5	120	3632144
6	7845	8318134
7	69	1899971
8	121	5024559

References

1. Sanger, F. and A.R. Coulson, *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J Mol Biol, 1975. **94**(3): p. 441-8.

2. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.

3. Smith, L.M., S. Fung, M.W. Hunkapiller, T.J. Hunkapiller, and L.E. Hood, *The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis.* Nucleic Acids Res, 1985. **13**(7): p. 2399-412.

4. Smith, L.M., J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, S.B. Kent, and L.E. Hood, *Fluorescence detection in automated DNA sequence analysis.* Nature, 1986. **321**(6071): p. 674-9.

5. Gilbert, W. and A. Maxam, *The nucleotide sequence of the lac operator.* Proc Natl Acad Sci U S A, 1973. **70**(12): p. 3581-4.

6. Service, R.F., *Gene sequencing. The race for the $1000 genome.* Science, 2006. **311**(5767): p. 1544-6.

7. Kulesh, D.A., D.R. Clive, D.S. Zarlenga, and J.J. Greene, *Identification of interferon-modulated proliferation-related cDNA sequences.* Proc Natl Acad Sci U S A, 1987. **84**(23): p. 8453-7.

8. Ren, B., F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, and R.A. Young, *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.

9. Pavlov, A.R., G.I. Belova, S.A. Kozyavkin, and A.I. Slesarev, *Helix-hairpin-helix motifs confer salt resistance and processivity on chimeric DNA polymerases.* Proc Natl Acad Sci U S A, 2002. **99**(21): p. 13510-5.

10. Fan, J.B., J.M. Yeakley, M. Bibikova, E. Chudin, E. Wickham, J. Chen, D. Doucet, P. Rigault, B. Zhang, R. Shen, C. McBride, H.R. Li, X.D. Fu, A. Oliphant, D.L. Barker, and M.S. Chee, *A versatile assay for high-throughput gene expression profiling on universal array matrices.* Genome Res, 2004. **14**(5): p. 878-85.

11. Garcia-Bassets, I., Y.S. Kwon, F. Telese, G.G. Prefontaine, K.R. Hutt, C.S. Cheng, B.G. Ju, K.A. Ohgi, J. Wang, L. Escoubet-Lozach, D.W. Rose, C.K. Glass, X.D. Fu, and M.G. Rosenfeld, *Histone methylation-dependent mechanisms impose ligand dependency for gene activation by nuclear receptors.* Cell, 2007. **128**(3): p. 505-18.

12. Kwon, Y.S., I. Garcia-Bassets, K.R. Hutt, C.S. Cheng, M. Jin, D. Liu, C. Benner, D. Wang, Z. Ye, M. Bibikova, J.B. Fan, L. Duan, C.K. Glass, M.G. Rosenfeld, and X.D. Fu, *Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters.* Proc Natl Acad Sci U S A, 2007. **104**(12): p. 4852-7.

13. Ascenzi, P., A. Bocedi, and M. Marino, *Structure-function relationship of estrogen receptor alpha and beta: impact on human health.* Mol Aspects Med, 2006. **27**(4): p. 299-402.

14. Bourguet, W., P. Germain, and H. Gronemeyer, *Nuclear receptor ligand-binding domains: three-dimensional structures, molecular interactions and pharmacological implications.* Trends Pharmacol Sci, 2000. **21**(10): p. 381-8.

15. Rosenfeld, M.G., V.V. Lunyak, and C.K. Glass, *Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response.* Genes Dev, 2006. **20**(11): p. 1405-28.

16. Carroll, J.S., X.S. Liu, A.S. Brodsky, W. Li, C.A. Meyer, A.J. Szary, J. Eeckhoute, W. Shao, E.V. Hestermann, T.R. Geistlinger, E.A. Fox, P.A. Silver, and M. Brown, *Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1*. Cell, 2005. **122**(1): p. 33-43.

17. Andres, M.E., C. Burger, M.J. Peral-Rubio, E. Battaglioli, M.E. Anderson, J. Grimes, J. Dallman, N. Ballas, and G. Mandel, *CoREST: a functional corepressor required for regulation of neural-specific gene expression.* Proc Natl Acad Sci U S A, 1999. **96**(17): p. 9873-8.

18. Ballas, N., E. Battaglioli, F. Atouf, M.E. Andres, J. Chenoweth, M.E. Anderson, C. Burger, M. Moniwa, J.R. Davie, W.J. Bowers, H.J. Federoff, D.W. Rose, M.G. Rosenfeld, P. Brehm, and G. Mandel, *Regulation of neuronal traits by a novel transcriptional complex.* Neuron, 2001. **31**(3): p. 353-65.

19. Hakimi, M.A., D.A. Bochar, J. Chenoweth, W.S. Lane, G. Mandel, and R. Shiekhattar, *A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes.* Proc Natl Acad Sci U S A, 2002. **99**(11): p. 7420-5.

20. Hakimi, M.A., Y. Dong, W.S. Lane, D.W. Speicher, and R. Shiekhattar, *A candidate X-linked mental retardation gene is a component of a new family of histone deacetylase-containing complexes.* J Biol Chem, 2003. **278**(9): p. 7234-9.

21. Humphrey, G.W., Y. Wang, V.R. Russanova, T. Hirai, J. Qin, Y. Nakatani, and B.H. Howard, *Stable histone deacetylase complexes distinguished by the presence of SANT domain proteins CoREST/kiaa0071 and Mta-L1*. J Biol Chem, 2001. **276**(9): p. 6817-24.

22. Lunyak, V.V., R. Burgess, G.G. Prefontaine, C. Nelson, S.H. Sze, J. Chenoweth, P. Schwartz, P.A. Pevzner, C. Glass, G. Mandel, and M.G. Rosenfeld, *Corepressor-dependent silencing of chromosomal regions encoding neuronal genes.* Science, 2002. **298**(5599): p. 1747-52.

23. Shi, Y., J. Sawada, G. Sui, B. Affar el, J.R. Whetstine, F. Lan, H. Ogawa, M.P. Luke, Y. Nakatani, and Y. Shi, *Coordinated histone modifications mediated by a CtBP co-repressor complex.* Nature, 2003. **422**(6933): p. 735-8.

24. Tong, J.K., C.A. Hassig, G.R. Schnitzler, R.E. Kingston, and S.L. Schreiber, *Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex.* Nature, 1998. **395**(6705): p. 917-21.

25. You, A., J.K. Tong, C.M. Grozinger, and S.L. Schreiber, *CoREST is an integral component of the CoREST- human histone deacetylase complex.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1454-8.

26. Shi, Y., F. Lan, C. Matson, P. Mulligan, J.R. Whetstine, P.A. Cole, R.A. Casero, and Y. Shi, *Histone demethylation mediated by the nuclear amine oxidase homolog LSD1.* Cell, 2004. **119**(7): p. 941-53.

27. Shi, Y.J., C. Matson, F. Lan, S. Iwase, T. Baba, and Y. Shi, *Regulation of LSD1 histone demethylase activity by its associated factors.* Mol Cell, 2005. **19**(6): p. 857-64.

28. Metzger, E., M. Wissmann, N. Yin, J.M. Muller, R. Schneider, A.H. Peters, T. Gunther, R. Buettner, and R. Schule, *LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription.* Nature, 2005. **437**(7057): p. 436-9.

29. Lachner, M. and T. Jenuwein, *The many faces of histone lysine methylation.* Curr Opin Cell Biol, 2002. **14**(3): p. 286-98.

30. Martin, C. and Y. Zhang, *The diverse functions of histone lysine methylation.* Nat Rev Mol Cell Biol, 2005. **6**(11): p. 838-49.

31. Kouzarides, T., *Histone methylation in transcriptional control.* Curr Opin Genet Dev, 2002. **12**(2): p. 198-209.

32. Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

33. Jurka, J., *Repbase update: a database and an electronic journal of repetitive elements.* Trends Genet, 2000. **16**(9): p. 418-20.

34. Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas, R.J. Weber, D. Haussler, and W.J. Kent, *The UCSC Genome Browser Database.* Nucleic Acids Res, 2003. **31**(1): p. 51-4.

35. Rozen, S. and H. Skaletsky, *Primer3 on the WWW for general users and for biologist programmers.* Methods Mol Biol, 2000. **132**: p. 365-86.

36. Hofacker, I.L., *Vienna RNA secondary structure server.* Nucleic Acids Res, 2003. **31**(13): p. 3429-31.

37. Shang, Y., X. Hu, J. DiRenzo, M.A. Lazar, and M. Brown, *Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription.* Cell, 2000. **103**(6): p. 843-52.

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

39. Li, Z., S. Van Calcar, C. Qu, W.K. Cavenee, M.Q. Zhang, and B. Ren, *A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8164-9.

40. Ren, B. and B.D. Dynlacht, *Use of chromatin immunoprecipitation assays in genome-wide location analysis of mammalian transcription factors.* Methods Enzymol, 2004. **376**: p. 304-15.

41. Roberts, C.J., B. Nelson, M.J. Marton, R. Stoughton, M.R. Meyer, H.A. Bennett, Y.D. He, H. Dai, W.L. Walker, T.R. Hughes, M. Tyers, C. Boone, and S.H. Friend, *Signaling and circuitry of multiple MAPK*

pathways revealed by a matrix of global gene expression profiles. Science, 2000. **287**(5454): p. 873-80.

42. Buck, M.J. and J.D. Lieb, *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.* Genomics, 2004. **83**(3): p. 349-60.

43. Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D.K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D.S. Gerhard, and T.R. Gingeras, *Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.* Science, 2005. **308**(5725): p. 1149-54.

44. Kampa, D., J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T.R. Gingeras, *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.* Genome Res, 2004. **14**(3): p. 331-42.

45. Kapranov, P., J. Cheng, S. Dike, D.A. Nix, R. Duttagupta, A.T. Willingham, P.F. Stadler, J. Hertel, J. Hackermuller, I.L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T.R. Gingeras, *RNA maps reveal new RNA classes and a possible function for pervasive transcription.* Science, 2007. **316**(5830): p. 1484-8.

46. Benner, C., Heinz S., Hutt K., Rosenfeld MG., Subramaniam S., Glass CK., *Novel approach to sequence analysis of ChIP-Seq data reveals detailed cis-regulatory code underlying embryonic stem cell transcription factors*. 2009, UCSD: San Diego.

47. Barash, Y., Bejerano, G., Friedman, N. *A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites*. in *Proceedings of the First International Workshop on Algorithms in Bioinformatics*. 2001. Aarhus, Denmark: Springer-Verlag.

48. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.

49. Gibbons, F.D., M. Proft, K. Struhl, and F.P. Roth, *Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization.* Genome Biol, 2005. **6**(11): p. R96.

50. Carroll, J.S., C.A. Meyer, J. Song, W. Li, T.R. Geistlinger, J. Eeckhoute, A.S. Brodsky, E.K. Keeton, K.C. Fertuck, G.F. Hall, Q. Wang, S. Bekiranov, V. Sementchenko, E.A. Fox, P.A. Silver, T.R. Gingeras, X.S. Liu, and M. Brown, *Genome-wide analysis of estrogen receptor binding sites.* Nat Genet, 2006. **38**(11): p. 1289-97.

51. Laganiere, J., G. Deblois, C. Lefebvre, A.R. Bataille, F. Robert, and V. Giguere, *From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response.* Proc Natl Acad Sci U S A, 2005. **102**(33): p. 11651-6.

52. Xie, X., J. Lu, E.J. Kulbokas, T.R. Golub, V. Mootha, K. Lindblad-Toh, E.S. Lander, and M. Kellis, *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.

53. Beck, S., P. Sommer, E. dos Santos Silva, N. Blin, and P. Gott, *Hepatocyte nuclear factor 3 (winged helix domain) activates trefoil factor gene TFF1 through a binding motif adjacent to the TATAA box.* DNA Cell Biol, 1999. **18**(2): p. 157-64.

54. Robyr, D., A. Gegonne, A.P. Wolffe, and W. Wahli, *Determinants of vitellogenin B1 promoter architecture. HNF3 and estrogen responsive transcription within chromatin.* J Biol Chem, 2000. **275**(36): p. 28291-300.

55. Hu, Q., Y.S. Kwon, E. Nunez, M.D. Cardamone, K.R. Hutt, K.A. Ohgi, I. Garcia-Bassets, D.W. Rose, C.K. Glass, M.G. Rosenfeld, and X.D. Fu,

Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules. Proc Natl Acad Sci U S A, 2008.

56. Patti, M.E., A.J. Butte, S. Crunkhorn, K. Cusi, R. Berria, S. Kashyap, Y. Miyazaki, I. Kohane, M. Costello, R. Saccone, E.J. Landaker, A.B. Goldfine, E. Mun, R. DeFronzo, J. Finlayson, C.R. Kahn, and L.J. Mandarino, *Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: Potential role of PGC1 and NRF1.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8466-71.

57. He, Z., Y. Hu, L. Feng, Y. Li, G. Liu, Y. Xi, L. Wen, and A. Lucia, *NRF-1 genotypes and endurance exercise capacity in young Chinese men.* Br J Sports Med, 2008. **42**(5): p. 361-6.

58. Braslavsky, I., B. Hebert, E. Kartalov, and S.R. Quake, *Sequence information can be obtained from single DNA molecules.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3960-4.

59. Mitra, R.D., J. Shendure, J. Olejnik, O. Edyta Krzymanska, and G.M. Church, *Fluorescent in situ sequencing on polymerase colonies.* Anal Biochem, 2003. **320**(1): p. 55-65.

60. Ronaghi, M., M. Uhlen, and P. Nyren, *A sequencing method based on real-time pyrophosphate.* Science, 1998. **281**(5375): p. 363, 365.

61. Seo, T.S., X. Bai, D.H. Kim, Q. Meng, S. Shi, H. Ruparel, Z. Li, N.J. Turro, and J. Ju, *Four-color DNA sequencing by synthesis on a chip using photocleavable fluorescent nucleotides.* Proc Natl Acad Sci U S A, 2005. **102**(17): p. 5926-31.

62. Hillier, L.W., G.T. Marth, A.R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, J.I. Glasscock, M. Hickenbotham, W. Huang, V.J. Magrini, R.J. Richt, S.N. Sander, D.A. Stewart, M. Stromberg, E.F. Tsung, T. Wylie, T. Schedl, R.K. Wilson, and E.R. Mardis, *Whole-genome sequencing and variant discovery in C. elegans.* Nat Methods, 2008. **5**(2): p. 183-8.

63. Marioni, J.C., C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad, *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome Res, 2008. **18**(9): p. 1509-17.

64. Hafner, M., P. Landgraf, J. Ludwig, A. Rice, T. Ojo, C. Lin, D. Holoch, C. Lim, and T. Tuschl, *Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing.* Methods, 2008. **44**(1): p. 3-12.

65. Schones, D.E., K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao, *Dynamic regulation of nucleosome positioning in the human genome.* Cell, 2008. **132**(5): p. 887-98.

66. Barski, A., S. Cuddapah, K. Cui, T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

67. Mikkelsen, T.S., M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.K. Kim, R.P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E.S. Lander, and B.E. Bernstein, *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.

68. Wang, Z., C. Zang, J.A. Rosenfeld, D.E. Schones, A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, W. Peng, M.Q. Zhang, and K. Zhao, *Combinatorial patterns of histone acetylations and methylations in the human genome.* Nat Genet, 2008. **40**(7): p. 897-903.

69. Cokus, S.J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C.D. Haudenschild, S. Pradhan, S.F. Nelson, M. Pellegrini, and S.E. Jacobsen, *Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning.* Nature, 2008. **452**(7184): p. 215-9.

70. Chen, W., V. Kalscheuer, A. Tzschach, C. Menzel, R. Ullmann, M.H. Schulz, F. Erdogan, N. Li, Z. Kijas, G. Arkesteijn, I.L. Pajares, M. Goetz-Sothmann, U. Heinrich, I. Rost, A. Dufke, U. Grasshoff, B. Glaeser, M.

Vingron, and H.H. Ropers, *Mapping translocation breakpoints by next-generation sequencing.* Genome Res, 2008. **18**(7): p. 1143-9.

71. Dolan, P.C. and D.R. Denver, *TileQC: a system for tile-based quality control of Solexa data.* BMC Bioinformatics, 2008. **9**: p. 250.

72. Li, R., Y. Li, K. Kristiansen, and J. Wang, *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-4.

73. Lin, H., Z. Zhang, M.Q. Zhang, B. Ma, and M. Li, *ZOOM! Zillions of oligos mapped.* Bioinformatics, 2008. **24**(21): p. 2431-7.

74. Fejes, A.P., G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S.J. Jones, *FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology.* Bioinformatics, 2008. **24**(15): p. 1729-30.

75. Ji, H., H. Jiang, W. Ma, D.S. Johnson, R.M. Myers, and W.H. Wong, *An integrated software system for analyzing ChIP-chip and ChIP-seq data.* Nat Biotechnol, 2008.

76. Johnson, D.S., A. Mortazavi, R.M. Myers, and B. Wold, *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.

77. Jothi, R., S. Cuddapah, A. Barski, K. Cui, and K. Zhao, *Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.* Nucleic Acids Res, 2008. **36**(16): p. 5221-31.

78. Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

79. Valouev, A., D.S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R.M. Myers, and A. Sidow, *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nat Methods, 2008.

80. Zhang, Y., T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nussbaum, R.M. Myers, M. Brown, W. Li, and X.S. Liu, *Model-based Analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

81. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

82. Mooradian, A.D., J.E. Morley, and S.G. Korenman, *Biological actions of androgens.* Endocr Rev, 1987. **8**(1): p. 1-28.

83. Fix, C., C. Jordan, P. Cano, and W.H. Walker, *Testosterone activates mitogen-activated protein kinase and the cAMP response element binding protein transcription factor in Sertoli cells.* Proc Natl Acad Sci U S A, 2004. **101**(30): p. 10919-24.

84. Heinlein, C.A. and C. Chang, *The roles of androgen receptors and androgen-binding proteins in nongenomic androgen actions.* Mol Endocrinol, 2002. **16**(10): p. 2181-7.

85. Vlahopoulos, S., W.E. Zimmer, G. Jenster, N.S. Belaguli, S.P. Balk, A.O. Brinkmann, R.B. Lanz, V.C. Zoumpourlis, and R.J. Schwartz, *Recruitment of the androgen receptor via serum response factor facilitates expression of a myogenic gene.* J Biol Chem, 2005. **280**(9): p. 7786-92.

86. Ford, O.H., 3rd, C.W. Gregory, D. Kim, A.B. Smitherman, and J.L. Mohler, *Androgen receptor gene amplification and protein expression in recurrent prostate cancer.* J Urol, 2003. **170**(5): p. 1817-21.

87. Linja, M.J., K.J. Savinainen, O.R. Saramaki, T.L. Tammela, R.L. Vessella, and T. Visakorpi, *Amplification and overexpression of androgen receptor gene in hormone-refractory prostate cancer.* Cancer Res, 2001. **61**(9): p. 3550-5.

88. Wang, Q., W. Li, X.S. Liu, J.S. Carroll, O.A. Janne, E.K. Keeton, A.M. Chinnaiyan, K.J. Pienta, and M. Brown, *A hierarchical network of transcription factors governs androgen receptor-dependent prostate cancer growth.* Mol Cell, 2007. **27**(3): p. 380-92.

89. Lee, H.J., M. Hwang, S. Chattopadhyay, H.S. Choi, and K. Lee, *Hepatocyte nuclear factor-3 alpha (HNF-3alpha) negatively regulates androgen receptor transactivation in prostate cancer cells.* Biochem Biophys Res Commun, 2008. **367**(2): p. 481-6.

90. Badve, S., D. Turbin, M.A. Thorat, A. Morimiya, T.O. Nielsen, C.M. Perou, S. Dunn, D.G. Huntsman, and H. Nakshatri, *FOXA1 expression in breast cancer--correlation with luminal subtype A and survival.* Clin Cancer Res, 2007. **13**(15 Pt 1): p. 4415-21.

91. Ballas, N., C. Grunseich, D.D. Lu, J.C. Speh, and G. Mandel, *REST and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis.* Cell, 2005. **121**(4): p. 645-57.

92. Battaglioli, E., M.E. Andres, D.W. Rose, J.G. Chenoweth, M.G. Rosenfeld, M.E. Anderson, and G. Mandel, *REST repression of neuronal genes requires components of the hSWI.SNF complex.* J Biol Chem, 2002. **277**(43): p. 41038-45.

93. Lunyak, V.V. and M.G. Rosenfeld, *No rest for REST: REST/NRSF regulation of neurogenesis.* Cell, 2005. **121**(4): p. 499-501.

94. Yeo, M., S.K. Lee, B. Lee, E.C. Ruiz, S.L. Pfaff, and G.N. Gill, *Small CTD phosphatases function in silencing neuronal gene expression.* Science, 2005. **307**(5709): p. 596-600.

95. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.

96. Crawford, G.E., I.E. Holt, J. Whittle, B.D. Webb, D. Tai, S. Davis, E.H. Margulies, Y. Chen, J.A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T.J. Vasicek, M.J. Daly, T.G. Wolfsberg, and F.S. Collins, *Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).* Genome Res, 2006. **16**(1): p. 123-31.

97. Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

98. Bannister, A.J., R. Schneider, F.A. Myers, A.W. Thorne, C. Crane-Robinson, and T. Kouzarides, *Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes.* J Biol Chem, 2005. **280**(18): p. 17732-6.

99. Abramovitz, L., T. Shapira, I. Ben-Dror, V. Dror, L. Granot, T. Rousso, E. Landoy, L. Blau, G. Thiel, and L. Vardimon, *Dual role of NRSF/REST in activation and repression of the glucocorticoid response.* J Biol Chem, 2008. **283**(1): p. 110-9.

100. Kim, C.S., C.K. Hwang, K.Y. Song, H.S. Choi, K. Kim do, P.Y. Law, L.N. Wei, and H.H. Loh, *Novel function of neuron-restrictive silencer factor (NRSF) for posttranscriptional regulation.* Biochim Biophys Acta, 2008. **1783**(10): p. 1835-46.

101. Tiwari, V.K., L. Cope, K.M. McGarvey, J.E. Ohm, and S.B. Baylin, *A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations.* Genome Res, 2008. **18**(7): p. 1171-9.

102. Wurtele, H. and P. Chartrand, *Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology.* Chromosome Res, 2006. **14**(5): p. 477-95.

103. Zhao, Z., G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K.S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson, *Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.* Nat Genet, 2006. **38**(11): p. 1341-7.

104. Bieche, I., C. Tomasetto, C.H. Regnier, C. Moog-Lutz, M.C. Rio, and R. Lidereau, *Two distinct amplified regions at 17q11-q21 involved in human primary breast cancer.* Cancer Res, 1996. **56**(17): p. 3886-90.

105. Tanner, M.M., M. Tirkkonen, A. Kallioniemi, K. Holli, C. Collins, D. Kowbel, J.W. Gray, O.P. Kallioniemi, and J. Isola, *Amplification of chromosomal region 20q13 in invasive breast cancer: prognostic implications.* Clin Cancer Res, 1995. **1**(12): p. 1455-61.