Chromatin profiles of human cells in health and disease using FAIRE

by
Paul G. Giresi

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology.

Chapel Hill
2011

Approved by:

Jason D. Lieb, PhD

Brian D. Strahl, PhD

Kerry S. Bloom, PhD

Frank L. Conlon, PhD

Wei Sun, PhD

Abstract

PAUL G. GIRESI: Chromatin profiles of human cells in health and disease using FAIRE
(Under the direction of Jason D. Lieb)

Breast cancer is a heterogenous disease comprised of molecularly distinct subtypes with diverse clinical outcomes. Understanding the molecular composition of each subtype will aid in the effective diagnosis and treatment of breast cancer. The composition and activity of subtype-selective regulatory pathways operate, in part, through binding of proteins at distinct sites throughout the genome, often referred to as regulatory elements, to govern levels of gene expression. One of the characteristics of these binding events is the displacement of nucleosomes. Here we have developed a technique called FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements), which is capable of the genome-wide identification of active regulatory elements in human cells based on the nucleosome-depleted nature of these sites.

Using FAIRE we have identified the genome-wide set of active regulatory elements in the luminal and basal-like tumor subtypes. Here most of the active regulatory elements were distinct to each subtype. Many of these unique sites also reflected the activity of the regulatory mechanisms present in a given subtype. For example, in the hormone-responsive luminal cells we detected strong FAIRE signals at estrogen-receptor alpha binding sites, whereas the signals are diminished or absent in the hormone non-responsive basal-like cells. These subtype-selective regulatory elements tended to be clustered around the set of expressed genes in the respective subtype, regardless of whether the gene was differentially expressed between the subtypes. The subtype-selective regulatory elements were also enriched with sequence motifs for DNA-binding

proteins, which included factors known to be active in the respective subtypes.

We also used FAIRE to investigate the set of active regulatory elements associated with the transformation of a mammary epithelial cell line to a cancerous phenotype. Here transformation occurred through the differential expression of members of transcription factors families, which recycled the set of existing regulatory elements to effect global changes in gene expression.

Together, these findings indicate that FAIRE will be a powerful tool for discovery of the molecular characteristics underlying cancer and that FAIRE holds promise as a clinical diagnostic tool.

Acknowledgments

The work presented in this dissertation would not have been possible without the contributions of many people whom I wish to acknowledge.

First and foremost I would like to thank my mentor and advisor Jason Lieb for both his patience and confidence in my abilities. His many talents have provided me with an example of what it takes to be a successful scientist, which extends well beyond the bench. The success of the laboratory is largely due to his scientific acumen, which is derived from an exceptional combination of reason, intuition and effective communication. I continue to strive to achieve the confluence of abilities his example provides.

There are several individuals both at UNC and other intuitions who made this work possible, which I would like to acknowledge. The expertise and assistance of Vishy Iyer and Jonghwan Kim at the University of Texas at Austin were critical for the early development of FAIRE for use in human cells. The foray into breast cancer was initiated by Chuck Perou at University of North Carolina at Chapel Hill. I would like to acknowledge Kevin Struhl at Harvard Medical School who collaborated with to investigate how active regulatory elements are utilized in the transformation mammary epithelial cells. Much of the analysis of the FAIRE-seq data in the breast cancer samples would not have been possible without the work of Naim Rashid and Wei Sun in Biostatistics at University of North Carolina who were the main architects of ZINBA.

All of the members of the Lieb lab were instrumental in the development of my

scientific abilities as a graduate student and offered a tremendous amount of support and insights for the work in this dissertation. There are too many to mention here, however there are a few of particular note. Cheol-Koo Lee and Greg Hogan, who had been characterizing FAIRE in yeast, taught me the FAIRE technique and provide many insights for transitions into human cells. Sean Hanlon and Sevinc Ercan always made themselves available to serve as critical ear and provide many important insights into this work.

I am especially grateful to Joachim Theilhaber and Nathalie van Bockstaele for providing me the opportunity to pursue a career in science, without them none of this would have been possible.

I am especially grateful for the support of my family, my parents Bob and Kathy, and my brothers Todd, Brad and Kyle.

Most importantly I would like thank wife, Tiffany, and son, Reece, for putting up with me and my schedule as I conducted this research. I have certainly asked a lot and I am indebted for your for patience. Thank you.

Table of Contents

List of Figures

List of Abbreviations

BIC	Bayesian Information Criteria
CGH	Comparative Genomic Hybridization
ChIP	Chromatin ImmunoPrecipitation
ChIPOTle	Chromatin ImmunoPrecipitation On Tiling arrays
CNV	Copy Number Variation
DHS	DNase Hypersensitive Site
ENCODE	ENCyclopedia Of Dna Elements
ER	Estrogen-receptor 1
FAIRE	Formaldehyde Assisted Isolation of Regulatory Elements
GAII	Illumina Genome Analyzer II
H3K36me3	Histone H3 lysine 36 trimethylation
HER2	Human Epidermal growth factor Receptor 2 (ErbB-2)
KB	Kilobase
NGS	Next generation sequencing
PBS	Phosphate Buffered Saline
PR	progesterone receptor
qPCR	Quantitative PCR
SBPC	Single Basepair Overlap Count
TSS	Transcription Start Site
UCSC	University of California at Santa Cruz
ZINBA	Zero Inflated Negative Binomial Algorithm

Chapter 1

Introduction

Cancer is a neoplastic disease characterized by alterations of the molecular components within normal cells that disrupts the regulation of cell division and apoptosis leading to abnormal tissue growth and the formation of a tumor. The molecular events that lead to cancer are typically accumulated over time and result in changes to both cellular morphology and growth characteristics. Early on these changes are reversible and characterized by increases in the overall number of cells resulting from an external stimulus, termed hyperplasia. Dysplasia occurs as the irreversible alteration of cells, which is characterized by the loss of cell structure and disorganization of cells within tissues. The final and most severe stage is an invasive malignancy, where the cancerous cells are no longer confined to the originating tissue and can form tumors in distant tissues. There are several biological hallmarks common amongst cells in the development of malignant tumors, including sustaining proliferative signals, evading growth suppressors, resisting apoptosis, loss of senescence, inducing angiogenesis and activating invasion and metastases [59, 60].

Acquisition of these hallmarks in the progression of tumorigenesis often requires alteration of both tumor suppressors and oncogenes [57, 87]. Oncogenes encode proteins involved in the regulation of proliferation and can be classified into six broad categories:

transcription factors, chromatin remodelers, growth factors, growth factor receptors, signal transducers and regulators of apoptosis [33]. Typically, alterations to oncogenes results in them becoming inappropriately expressed at high levels or altered to possess new functions. While tumor suppressor genes, which play a role in inhibiting cellular proliferation and promoting apoptosis, are often disabled during tumorigenesis [166]. There are a variety of mechanisms by which tumor suppressors limit proliferation and promote apoptosis, which range from transducing extracellular signals to regulating gene expression through DNA binding.

The function of oncogenes and tumor suppressors can be transformed through both genetic and epigenetic alterations [57, 87]. Genetic alterations includes changes to the nucleotide composition and/or rearrangement of the genome. Mutations of individual nucleotides can operate at many different levels to alter gene function, including changes to the genomic sequence involved in the regulation of the gene to non-synonomous mutations in the coding regions effecting the production and function of the resultant protein [33]. Nucleotide composition can also be altered through large-scale amplification or deletion of genomic segments, resulting in the overproduction or elimination of genes altogether. Genomic rearrangements entail fusion of separate chromosomal segments, which can produce either novel regulatory mechanisms or gene fusions. While epigenetic alterations involve multiple processes, including noncoding RNAs, covalent modifications of chromatin, chromatin remodeling and DNA methylation [9, 44, 77].

Together the set of alterations that result in cancer redefine the molecular identity of normal cells. Identification of the molecular components associated with the formation, progression and clinical behavior of cancer has been the focus of research for several decades [13, 117, 135]. Advances in technology, including DNA microarrays and high-throughput sequencing, have provided researchers with the capacity to identify and investigate these components genome-wide. These technologies have been employed

to measure cellular abundance of RNAs [121, 153], estimate genomic copy number variations [3], characterize genotypic variation [34, 43], determine DNA methylation profiles [27, 95] and map the genome-wide binding of proteins [25]. The information obtained from these technologies has provide valuable insights into the diagnosis and treatment of cancers. However our understanding of the molecular underpinnings of cancer is still incomplete.

Here we examine for the first time the genome-wide regulatory status of chromatin in breast cancer and determine the extent to which we can observe both known and novel regulatory activity. We have used a technique termed FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) to identify the genome-wide set of active regulatory elements. The work presented in this dissertation first describes the development and characterization of FAIRE for human cell culture and tissues. The data derived by FAIRE also necessitate the development of novel analysis tools. We then performed FAIRE throughout a timecourse of transformation of mammary epithelial cells to a cancerous phenotype, which included creation of cancer stem cells. Finally we identified the set of active regulatory elements from tumors capable of distinguishing the luminal and basal-like breast cancer subtypes using FAIRE.

1.1 Breast cancer

One in every eight women will be diagnosed with breast cancer over the course of their lifetime. Among women, breast cancer is the leading diagnosed cancer (Figure 1.1A) and the second leading cause of cancer-related deaths (Figure 1.1B) [159]. Effective diagnosis and treatment is complicated by the heterogeneous nature of the disease. Breast cancer can arise from several different cell types throughout the mammary epithelial hierarchy [163]. Additionally, genetic instability can give rise to clonal variation during the growth and progression of each tumor [11]. Therefore a positive prognosis

is dependent on the accurate diagnosis and effective treatment of the specific cellular composition of each tumor.

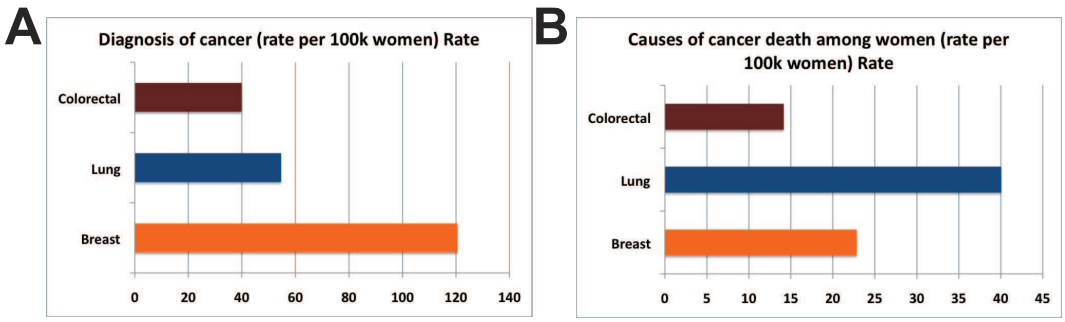

Figure 1.1: Cancer-related diagnoses (A) and mortality (B) rates among women (1999-2007) were obtained the CDC (Centers for Disease Control and Prevention) website

1.1.1 Pathophysiology of breast cancer subtypes

The normal breast is composed of a series of branching ducts radiating from the nipple which terminate in lobules (Figure 1.2A). The lobules are the sites of milk-production, which is then carried to the nipple through the ducts. The stroma is composed of fatty and connective tissues, blood vessels and lymphatic vessels. The majority of breast cancer arises from cells in the terminal ductal lobular unit (TDLU, Figure 1.2B). The TDLU is composed of an inner luminal epithelial and basal myoepithelial cell layers (Figure 1.2C). The myoepithelial cell layer is contains a heterogeneous mix of cells adjacent to the basement membrane. These cells have a similar morphology to smooth muscle cells, including possessing contractile function, but exhibit features of epithelial cells [152]. While the inner cell layer is composed of luminal epithelial cells, which are polarized glandular cells. At present it is thought that all breast cancer is derived from cells involved in the formation of the inner luminal epithelial cell layer. However there is still not a consensus for the precise cell type of origin for some forms of breast cancer [163].

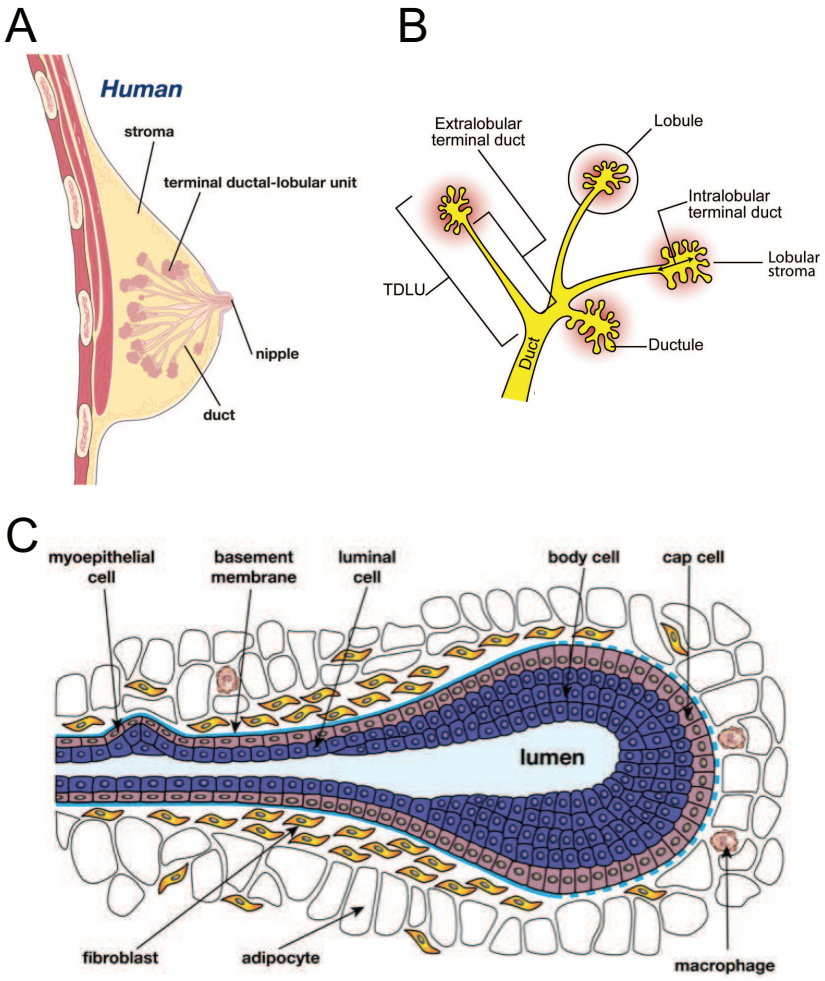

Figure 1.2: The images in this figure were obtained from a review of the mammary hierarchy and breast cancer [163] (A) The breast is composed of a branching network of ducts that terminate at the lobules. The loubules are the site of milk production, which are delivered to the nipple through the ducts. (B) The majority of breast cancer originates within the terminal ductal lobular untis (TDLU). (C) The ducts and lobules are composed of the luminal epithelial and myoepithelial layers.

To date six different subtypes of breast cancer have been identified and are distinguished based on the presence/absence of the estrogen and progesterone hormone receptors and overexpression of the HER2 amplicon [121, 123, 153]. Each of these subtypes are thought to be derived from cell at various stages of commitment in the formation of the mammary gland (Figure 1.3). The luminal subtypes are the most

common form of breast cancer (60%) and are characterized by the presence of the estrogen and progesterone receptor and are further divided in to an A and B group based on being either HER2 negative or positive, respectively. Luminal tumors are thought to originated from the terminally differentiated cells of the inner luminal epithelial cell hierarchy and in the case of luminal A has the best patient outcomes. The basal-like subtype is the next most common type of breast cancer ($\sim$20%) and is characterized as negative for both the estrogen and progesterone hormone receptors and HER2. These are thought to originate from a luminal progenitor cell and typically give rise to aggressive tumors that have high rates of metastasis and recurrence [127, 132]. The HER2 subtype is relatively rare ($\sim$12%) and is HER2 positive, but is negative for both hormone receptors. More recently the claudin-low subtype has been identified [123] and is characterized by the absence of luminal differentiation markers and enrichment of epithelial-to-mesenchymal transition markers, immune response genes and cancer stem cell-like features. The remaining set of tumors that do not fit these criteria are classified as normal-like.

Although all breast cancer are treated by surgically removing the tumor, the specificity and efficacy of secondary (adjuvant) therapies are critical to long term survival. The luminal subtypes are often treated with hormone therapies, such as Tamoxifen, which interferes with the activity of estrogen receptor. While the HER2 subtype is treated with a monoclonal antibody against the receptor (Trastuzumab) along with chemotherapy. However for the basal-like subtype treatment typically entails some combination of radiation and chemotherapy, since no targetable markers have been identified. Even for the subtypes where targeted treatments exist the ability to predict response to treatments, risk of reoccurrence and long term survival is not complete [119, 136]. Therefore the better we understand the molecular characteristic that predict clinical outcomes the more effectively breast cancer can be treated.

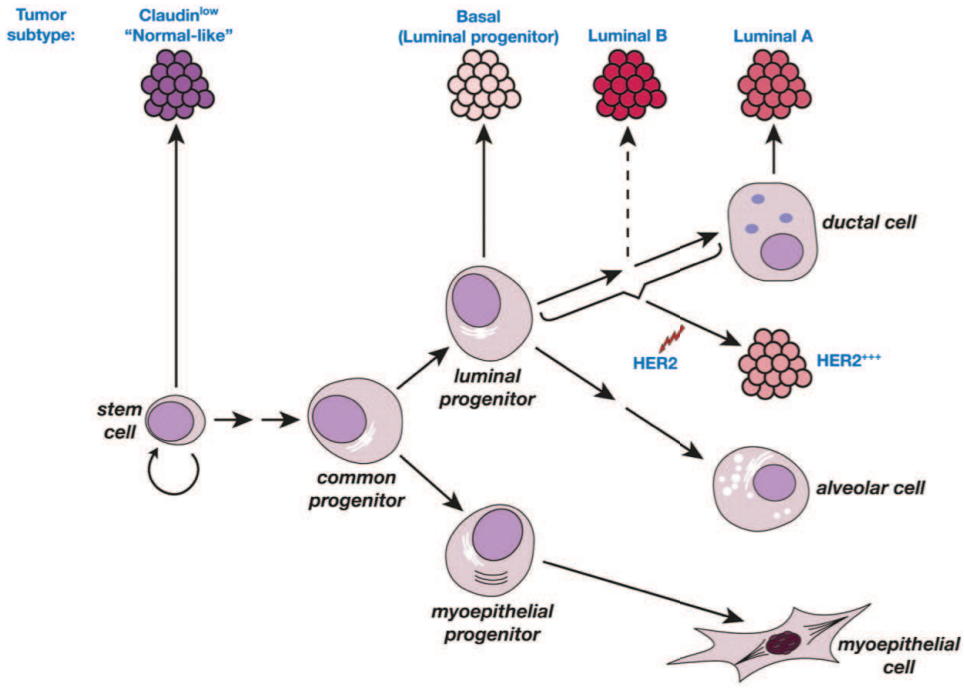

Figure 1.3: The image in this figure were obtained from a review of the mammary hierarchy and breast cancer [163]. The mammary epithelial hierarchy originates from a single stem cell that gives rise to the cells of the myoepithelial and luminal epithelial cell layers. Each of the subtypes has been related back to a cell type in this hierarchy based on similarity of gene expression profiles.

1.1.2 Molecular characterization of breast cancer subtypes

Several different approaches have been applied to better understand the molecular composition of breast cancer subtypes and its relationship to clinical outcomes, including identification of markers using immunohistochemistry [13, 80, 117], alteration in DNA methylation [27], aberrant microRNAs [72, 84], recurrent genomic copy number variations [160] and gene expression signatures [12]. Immunohistochemistry detects the presence and abundance of proteins in tissue samples and has long been used for the identification of markers capable distinguishing clinical behaviors, including the application of tissue microarrays for high-throughput screening of potential candidates [13, 112, 117, 135]. Copy number variations are typically identified using comparative

genomic hybridization (CGH) on DNA microarrays in which the genomic content of the tumor is compared against a diploid or matched normal counterpart. Hundreds of regions containing amplifications and deletions of oncogenes and tumor suppressors have been identified in breast tumors to date [26, 80, 91, 156]. While analysis of gene expression data from DNA microarrays has identified differences in the abundance of mRNAs that can be used to classify samples based on subtype [121, 153]. Gene expression can also be compared between tumors and matched normal samples to identify the genetic pathways altered in cancer and possibly serve to identify targets for treatment. Together this information provides a comprehensive picture regarding the transcriptional output, proteomic content and epigenetic state of breast cancer. However, several questions remain unanswered. Such as, which of the transcriptional regulatory proteins that have been identified as differentially expressed are actually functionally active? What are the genomic sites are bound by regulatory factors? What is the set of regulatory proteins governing the expression of each gene?

1.2 Nucleosome depletion is a hallmark of active regulatory elements

Cellular identity is established through the coordination of DNA-templated processes to utilize the information encoded in the genome. Coordination of these events is accomplished through interactions between regulatory proteins, enzymes, and DNA. In eukaryotes, these interactions are controlled largely through the regulation of chromatin composition . The most basic organizational unit of chromatin is the nucleosome, which is composed of DNA and an octamer of histone proteins. Given the central role of nucleosome formation in regulating genome function, many mechanisms have evolved to control nucleosome stability at loci across the genome. These include regulation

of nucleosome deposition following DNA replication [55], the activity of nucleosome-remodeling complexes like SWI/SNF and RSC [107], binding of regulatory proteins to DNA [110], transcriptional initiation and elongation by RNA polymerase II [140], incorporation of histone variants, post-translational modification of histones [161] and inherent properties of DNA sequence [144]. Access to the DNA template therefore requires the loss or destabilization nucleosomes at the cognate binding sites of regulatory proteins (Figure 1.4).

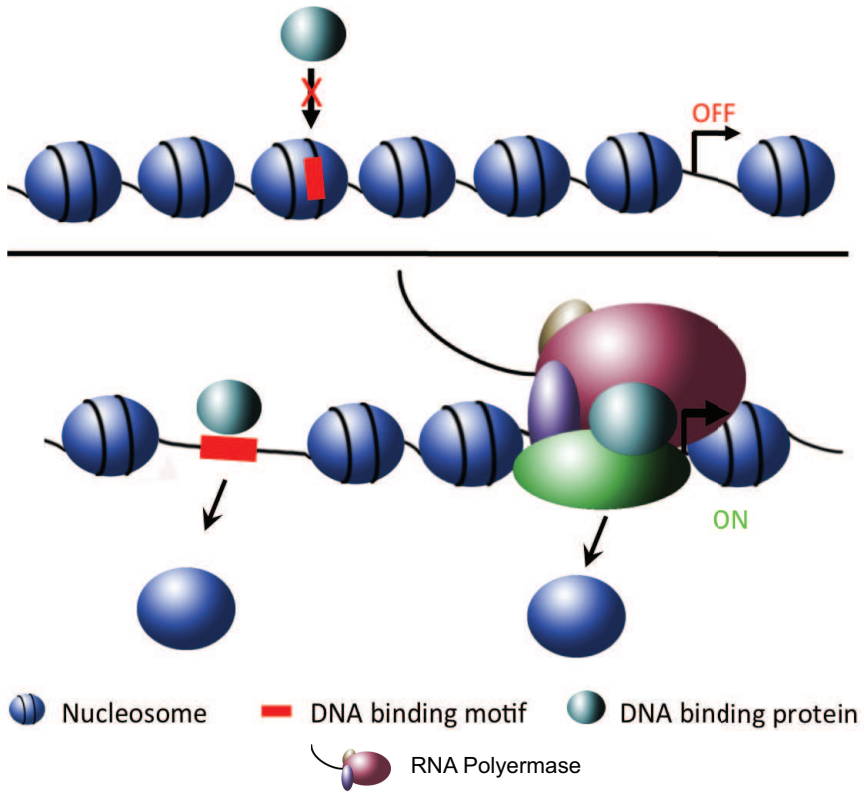

Figure 1.4: In the top panel, DNA is packaged into nucleosomes (dark blue spheres) preventing the DNA-binding protein (light blue sphere) from accessing its cognate binding site (red box). In the bottom panel, nucleosome loss is an indicator of regulatory activity. This is the case at both transcription start sites and at distal regulatory elements.

One may define nucleosome stability broadly to refer to the population of intact nucleosome-DNA complexes at a given locus, relative to the population of those in an

unfolded or disrupted state. This definition encompasses histone-histone, histone-DNA, and nucleosome-nucleosome interactions. Nucleosome stability is regulated through the concerted actions of both intrinsic and extrinsic mechanisms (Figure 1.5). Intrinsic mechanisms are those that affect the stability of the nucleosome itself. While extrinsic mechanisms consist of external factors that act upon the nucleosome.

Intrinsic mechanisms include variation in the flexibility of DNA sequences, incorporation of histone variants and post-translational modification of histones (Figure 1.5A). The inherent flexibility of a given DNA sequence can either promote or disfavor nucleosome formation [143]. Sequences containing AT dinucleotides at $\sim$10 bp periodicity appear to favor nucleosome occupancy. Whereas stretches of polyA appear to disfavor nucleosome occupancy. Incorporation of histone variants can alter nucleosome stability by altering the structural properties of the histone octamer [79]. While post-translational modification of histones, such as acetylation can reduce the energetic cost of removing a nucleosome.

Extrinsic mechanisms include the activity of nucleosome remodeling complexes, assembly and progression of proteins involved in DNA metabolism and binding of regulatory factors. Nucleosome remodeling complex facility the ATP-dependent movement or displacement nucleosomes to assist DNA-binding proteins in gaining access to their cognate sites [107]. The enzymes involved in DNA metabolism, such as transcription and replication, are recruited to their initiation sites by the stepwise assembly of proteins complex. The sites of assembly and initiation may also facilitate nucleosome clearance. Nucleosome "breathing" is the periodic unwrapping of DNA from the nucleosome, which allow regulatory proteins to bind DNA and lock in a disrupted state. While the progression DNA-templated processes, such as transcriptional elongation, must disrupt the histone octamer to gain access to DNA.

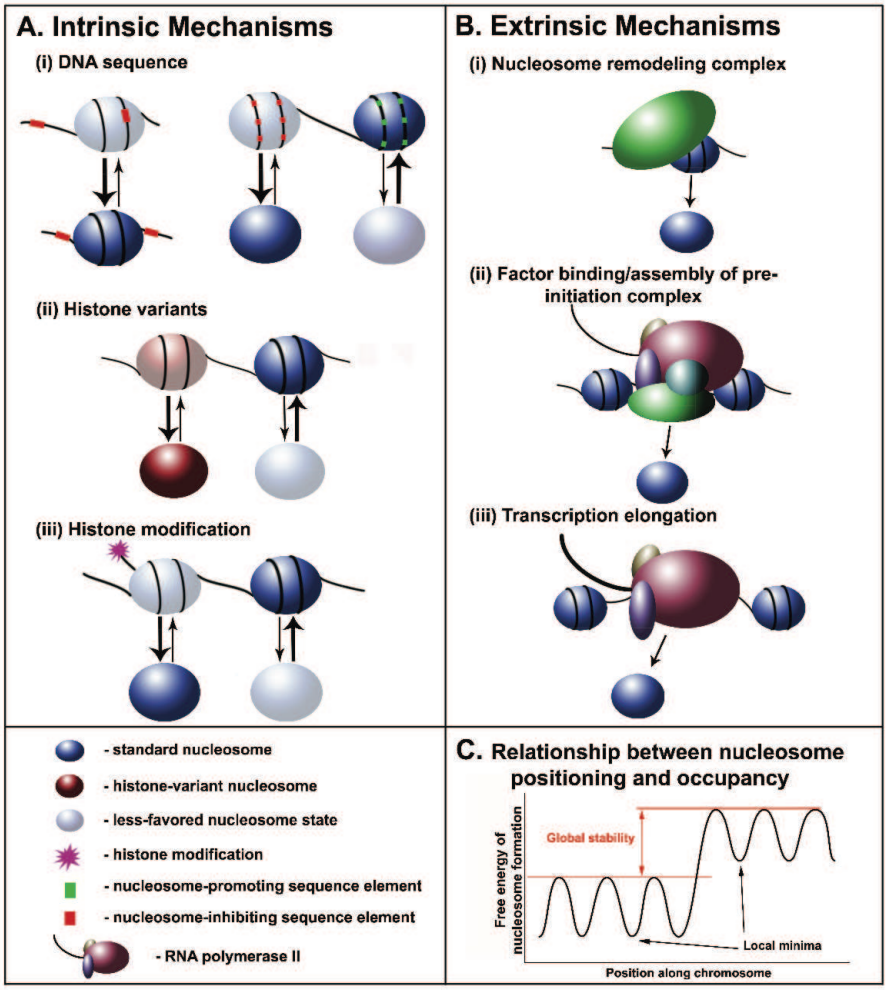

Figure 1.5: Processes involved in modulating nucleosome stability can be divided into intrinsic and extrinsic mechanisms. (A) Intrinsic mechanisms affect the stability of the nucleosome itself. (i) The inherent flexibility of a given DNA fragment can either promote (green dots) or disfavor the formation of a nucleosome. (ii) Incorporation of histone variants alters the interactions between histones, possibly stabilizing or destabilizing nucleosomes for specific tasks. (iii) Some post-translational modifications like histone acetylation reduce the energetic cost of removing a nucleosome. (B) Extrinsic mechanisms involve proteins that act upon nucleosomes to alter their stability. (i) Nucleosome remodeling complexes can move or displace nucleosomes. (ii) Assembly or initiation of the transcriptional machinery and binding of regulatory proteins can facilitate nucleosome clearance at promoters. (iii) Transcriptional elongation promotes the removal of nucleosomes. (C) Nucleosome occupancy versus positioning. A schematic representation of hypothesized nucleosome positioning via local energy minima. The concept of "occupancy" can be thought of as positioning at a larger scale.

In any given cell type the composition and activity of regulatory pathways coordinate these mechanisms to establish an overall chromatin structure, which includes the set of open chromatin sites or active regulatory elements. Genome-wide maps of active regulatory elements allow for a better understanding of how the availability of sequence-based regulatory elements are coordinated with the regulation of factors that utilize them in a given cellular environment.

1.3 Detection of nucleosome-depleted regions of genome

Sites of nucleosome loss or destabilization have traditionally been detected by virtue of their increased sensitivity to digestion by nucleases (Figure 1.6), which include deoxyribonuclease (DNase I) and micrococcal nuclease (MNase). DNase I is an endonuclease that cleaves the phosphodiester linkages in the DNA backbone. Initial studies noticed that the $5'$ ends of heat shock genes in *Drosophila* became hypersensitive to cleavage upon activation [82, 108, 169, 168]. MNase is an endo-exonuclease that preferentially digest the linker DNA in the intervening regions between nucleosomes [81] and allows nucleosome positions to mapped. Originally detection of the cleavage pattern for both nucleases was carried out at an individual locus using Southern blots. However both have since been adapted for genome-wide detection of chromatin structure [32, 138, 172] using either DNA microarrays or high-throughput sequencing.

1.3.1 A novel method for detection of nucleosome-depletion

FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements) is a simple procedure for the genome-wide isolation of nucleosome-depleted DNA from chromatin [51, 68, 113]. To perform FAIRE, chromatin is crosslinked with formaldehyde, sheared

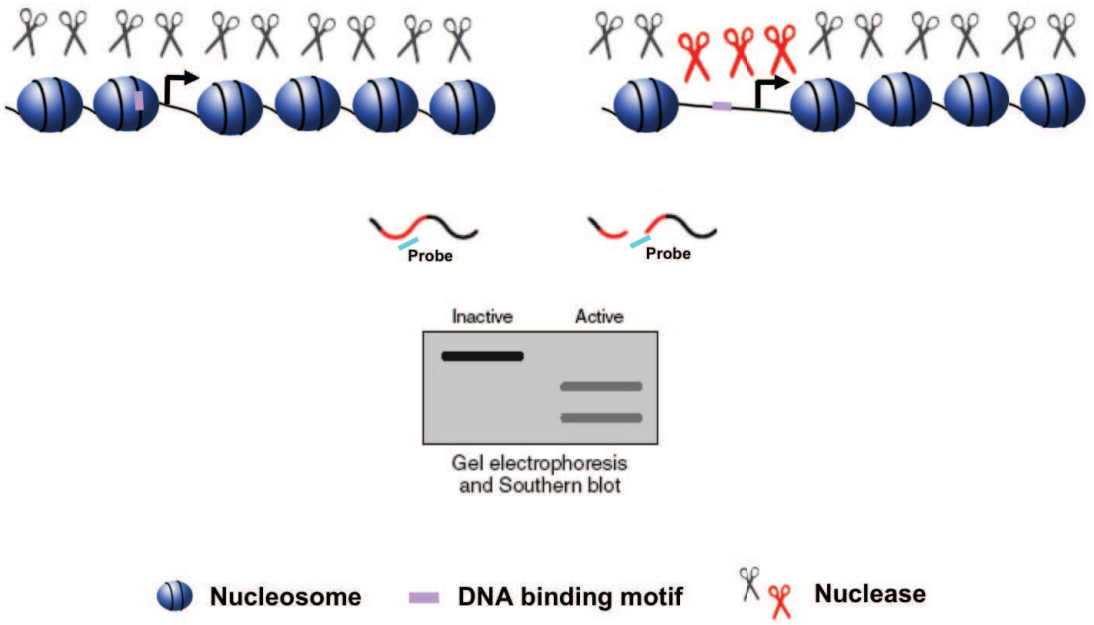

Figure 1.6: Nucleosomes assembled onto DNA are displayed as blue spheres and nucleases are shown as scissors. Upon loss or destabilization of nucleosomes nucleases are able to cleave DNA (red scissors). Traditionally nuclease sensitivity was probed at an individual locus using Southern blots to compare the cleavage patterns.

by sonication, and phenol-chloroform extracted. The crosslinking profile is likely dominated by nucleosomes, which are by far the most abundant protein-DNA interaction throughout the genome and are much more efficiently crosslinked [21, 122, 150]. Therefore, active regulatory elements preferentially segregate to the aqueous phase due to these DNA fragments being less efficiently crosslinked to proteins. The genomic regions preferentially segregated into the aqueous phase are then mapped back to the genome by either hybridization to tiling microarrays or are read directly using next-generation DNA sequencing.

This difference in crosslinking efficiency is likely due in part to the short crosslinking distance of formaldehyde. Formaldehyde is a small molecule (HCHO) and crosslinks are only formed between proteins and DNA in direct contact. There are approximately 10 to 15 histone-DNA interactions within a nucleosome that serve as potential crosslinking

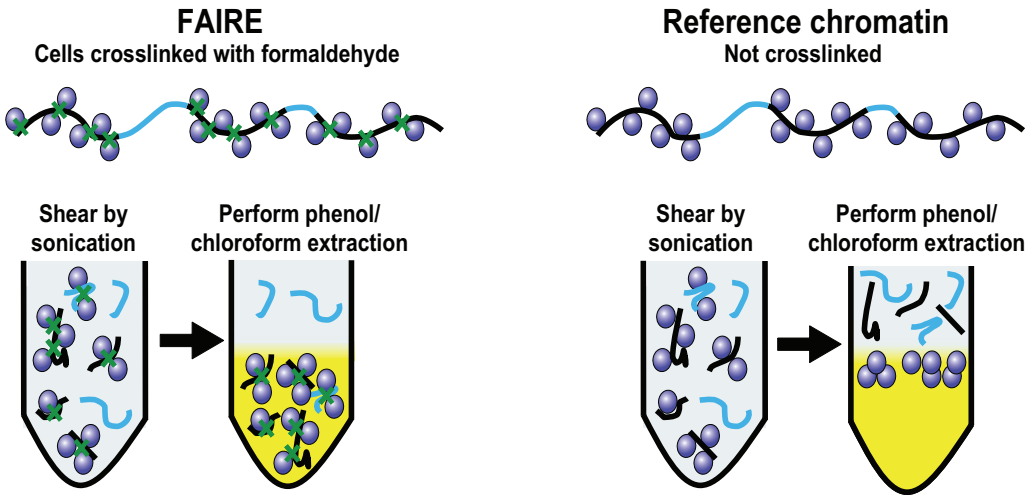

Figure 1.7: During a standard phenol/chloroform extraction proteins are sequestered to the interface due their hydrophilic and hydrophobic characteristics. While hydrophilic DNA remains in the aqueous phase (right panel). However, when DNA fragments are crosslinked to proteins they are also sequestered to the interface (left panel).

sites [103]. However, for most DNA-binding proteins there are far fewer potential crosslinking sites. The average binding sites are 5 to 15 bp [24], with only a few of the bases close enough to the protein contacts to be crosslinked [46]. In addition, formaldehyde requires a ϵ-amino group such as occurs on lysine, to form a crosslink [21, 150]. Approximately 10% of the amino-acid composition of histones are lysine, a much higher proportion than a typical protein. Due to both of these factors nucleosomes are much more readily crosslinkable to DNA, and are likely to dominate the crosslinking profile (Figure 1.8).

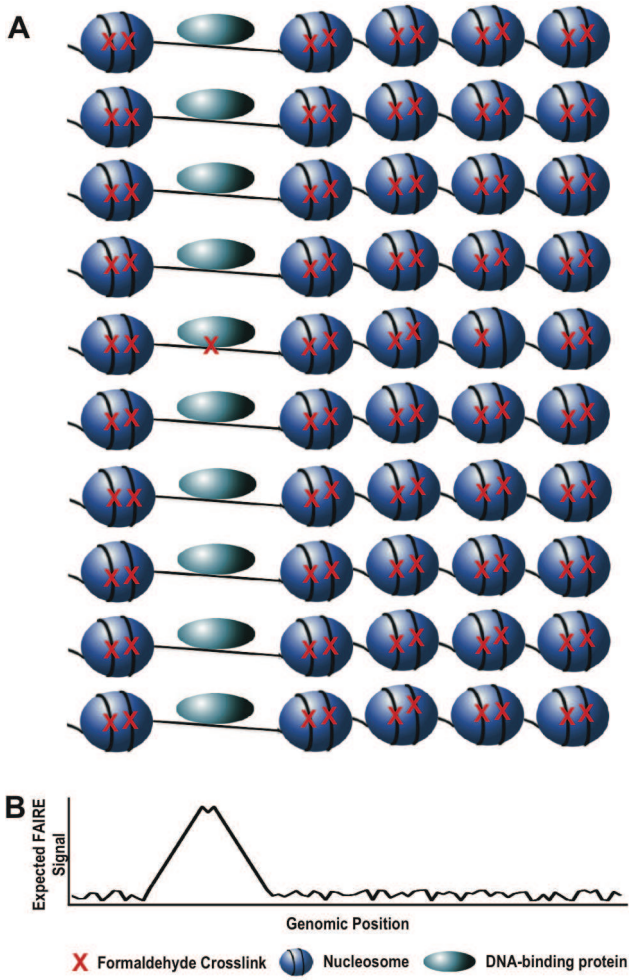

Figure 1.8: Crosslinking between histones and DNA (or between histones) likely dominate the chromatin crosslinking profile. (A) Here are a representative set of features from eukaryotic chromatin, including nucleosomes (blue spheres), a DNA-binding protein (light blue oval), and DNA (black line). Crosslinking with formaldehyde (red X) for most genomics applications only captures a portion of the potential interactions. Given that histone-DNA interactions constitute the majority of crosslinkable interactions in the genome, in a population of cells (ten rows) all of these interactions are likely to be captured. Whereas only a small proportion of the interactions between other DNA-binding proteins and DNA is actually captured by formaldehyde crosslinking. (B) The plot represents the expected FAIRE signal, which is inversely correlated with the occurrence of crosslinkable protein-DNA interactions.

Chapter 2

Methodology

The following section describes the steps for performing FAIRE in human tissues and cell culture that I developed during my graduate studies and were used throughout the subsequent chapters.

2.1 Performing FAIRE

To perform FAIRE, chromatin is crosslinked with formaldehyde, sheared by sonication, and phenol-chloroform extracted Figure 1.7.The following provides a general framework for performing FAIRE, which specifically emphasizes performing FAIRE on cells grown in culture. Modifications to the FAIRE protocol for tissue samples are noted in Section 2.1.7.

2.1.1 Formaldehyde crosslinking

For cells grown in culture, add 37% formaldehyde directly to the growth media to a final concentration of 1% and incubate at room temperature on an orbital shaker at 80 rpm for 5 minutes. To quench the fixation, add 2.5 M glycine to a final concentration of 125 mM and incubate for 5 min at room temperature while continuing to shake. Cells grown in suspension should be collected by centrifugation at 700 x g for 5 min at 4^{o}C.

For adherent cells, first remove the media containing formaldehyde and glycine, add ice-cold PBS to cover the cell layer, scrape, and transfer the cells to a conical tube. For both adherent cells and cells in suspension, wash two more times with ice-cold PBS to ensure all residual media is removed.

2.1.2 Cell lysis

Resuspend cells in 1 ml of lysis buffer (2% Triton X-100, 1% SDS, 100 mM NaCl, 10 mM Tris-Cl pH 8.0, 1 mM EDTA) per 10^7 (or 0.4g) of cells. Transfer 1 ml of lysis solution to 2 ml screw-capped tube with rubber seal and add 1 ml of 500 μM glass beads. Cell disruption is performed in a mini bead-beater (Mini-BeadBeater-8, BioSpec Inc.) set to homogenize for five 1-minute sessions with 2-minute incubations on ice between sessions (see the alternative protocol if a Beadbeater is not available). To recover the lysate, puncture the bottom of the 2 ml tube with a 25G syringe and drain into 15 ml tube on ice. Once the lysate has drained, add an additional 500 μl lysis buffer to clear any remaining sample from the beads. Filtered air can be used to push the liquid through the hole in the bottom of the tube. Proceed directly to sonication.

2.1.3 Cell lysis (alternative)

If a bead-beater is not available, the following procedure is suitable for human or similar cell types, but not yeast [93]. This procedure often requires additional rounds of sonication. Add 10 ml of Lysis Buffer 1 (50 mM HEPES-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) per 10^8 cells and rock at 4°C for 10 minutes. Spin at 1,300 x g for 5 minutes at 4°C and remove supernatant. Add 10 ml of Lysis Buffer 2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) per 10^8 cells and rock at room temperature for 10 minutes. Spin at 1,300 x g for 5 minutes at 4°C and remove supernatant, at this point the pellet should appear

white and fluffy. Add 3.5 ml of Lysis Buffer 3 (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-Deoxycholate, 0.5% N-lauroylsarcosine) per 10^8 cells. Proceed directly to sonication.

2.1.4 Sonication

Transfer the lysate to 1.5 ml tubes in 300 μl aliquots and sonicate for 15 minutes using a Bioruptor UCD-200 (Diagenode) set to pulse on high for 30 seconds followed by 30 seconds of rest. The water bath should be maintained at a constant temperature of 4^{o}C using a recirculator. Alternatively lysate can be transferred to 15 ml conical tubes for use with a microtip sonicator (Branson Sonifier 450) set at 15% amplitude for five sessions of sixty pulses (1 second on/1 second off), incubating the sample on ice for two minutes between sessions. Clear the lysate of cellular debris by spinning at 15,000 x g for 5 minutes at 4^{o}C, transfer supernatant to a new tube. Run an aliquot, equivalent to 100 ng total genomic DNA, on a 1% agarose gel to ensure fragment sizes range between 100-1000 bp.

2.1.5 Phenol/Chloroform extraction

Add an equal volume of phenol/chloroform (Sigma #P3803 phenol, chloroform, and isoamyl alcohol 25:24:1 saturated with 10mM Tris, pH 8.0, 1 mM EDTA) to the lysate, vortex well, spin at 12,000 x g for 5 minutes, and transfer the aqueous fraction to a fresh 1.5 ml tube. If there is very little aqueous phase due to an exceptionally large interface, remove aqueous phase, add 500 μl TE to old interface, vortex, and spin again. To ensure all protein has been removed, perform an additional extraction by adding an equal volume of phenol/chloroform to the isolated aqueous fraction. Finally, add an equal volume of chloroform (Fluka BioChemika 25666, chloroform, isoamyl alcohol 24:1) to the aqueous fraction, spin, and transfer aqueous phase to a new tube.

2.1.6 DNA precipitation

Add 3M sodium acetate (pH 5.2) to a final concentration of 0.3 M, and add 1 μl of 20 mg/ml glycogen. Mix by inverting. Add two volumes of 95% ethanol mix by inverting and incubate at -20°C overnight. Although overnight incubations are routinely performed, incubation as short as one hour should be sufficient. Pellet the precipitate by spinning at 15,000 x g for 30 minutes at 4°C, wash the pellet with 500 μl ice cold 70% ethanol, spin at 15,000 x g for 5 minutes at room temperature, remove the supernatant, and dry pellet in a speed-vac. Resuspend the dried pellet in 50 μl of 10 mM Tris-HCl pH 7.5. Add 1 μl of 10 mg/ml RNase A and incubate for 1 hour at 37°C. Earlier versions of the protocol included a step that incubates DNA from crosslinked samples at 65°C overnight to ensure that any DNA-DNA crosslinks do not interfere with downstream enzymatic steps. However, we have found that skipping this step results in no detectable difference in the efficiency of downstream enzymatic reactions.

Clean up the sample using either a spin column capable of recovering small DNA fragments (75-200 bp) or perform an additional phenol/chloroform extraction and ethanol precipitation. We have found that this is necessary to achieve accurate spectrophotometric measurements of our samples for subsequent reactions. Depending on the number of cells used for FAIRE and the final concentration, it may be possible to see the size distribution of FAIRE DNA fragments on a 1% agarose gel, which typically ranges between 75-200 bp. However, gel verification is not necessary and is often omitted.

2.1.7 Modifications for tissue samples

The following modifications for performing FAIRE in tissues include steps to prepare the tissue sample for crosslinking, disassociating the cells, and cell lysis. These modifications have been successfully used on tissue samples as small as 20 mg. Other

considerations for working with tissue samples include whether it is fresh or frozen, and how fibrous the tissue is. For fresh soft tissues, such as brain, simply mince the tissue into small pieces using a scalpel, transfer to a dounce with 1 ml of PBS containing 37% formaldehyde at a final concentration of 1%, and incubate for 5 minutes at room temperature (22-25°C) with swirling. Add 2.5 M glycine to a final concentration of 125 mM glycine and incubate for an additional 5 minutes. Disassociate the cells with a dounce homogenizer, wash two times with ice cold PBS, and proceed with cell lysis and all remaining steps for FAIRE as described above.

For previously frozen tissues or fresh fibrous tissues, samples should be disassociated in a tissue pulverizer, precooled in a liquid nitrogen bath. Tissue should then be resuspended in 1.5 ml room temperature PBS per 10 mg of tissue and transferred to 15 ml conical tissue grinder (VWR #47732-446). Formaldehyde should be added to a final concentration of 1% and incubated 5-10 minutes. Following quenching with 125 mM glycine and washing with ice cold PBS, pelleted tissue samples should be frozen by submerging 15 ml tube in liquid nitrogen bath. Tissue samples should be ground to roughly the consistency of sand. For most tissue types you can proceed with the protocol described above, but for especially tough tissue types use larger 2.8 mm ceramic or metal beads (Precellys CK28 or MK28) and perform additional cycles in the mini bead-beater for an efficient lysis before sonication.

2.1.8 Optimization of the FAIRE procedure

The two critical steps in FAIRE that should be optimized when first starting out or working with a new tissue or cell type are crosslinking time and sonication. A 5 minute incubation time has been sufficient for all cell types tested so far, however tissue often require longer incubations. If ChIP has been performed in the tissue type it is best to consider this the maximum incubation time and typically shorter incubation times are

optimal. For sonication, the exact parameters will vary based on the actual machine used. However, factors that will effect sonication efficiency include density of cells in solution, total volume of solution to be sonicated and the power setting. Sonication conditions should be optimized to deliver the proper size range of DNA fragments (100-1000 bp) with the fewest number of cycles while avoid using power settings that are so high that they excessively heat the sample or cause foaming. Optimized conditions based on the recommended equipment are provided above.

2.2 Detection of FAIRE DNA

2.2.1 Quantitative PCR

Quantitative PCR (qPCR) is used both as a method for detecting open chromatin sites and as a means to validate sites identified using either DNA microarray or high-throughput sequencing data. There are several considerations when designing qPCR experiments, including selection of an appropriate set of reference regions, exact primer localization, and methods for quantitation of the results. It is important to select an appropriate set of reference regions since these will be used to calculate relative enrichment for all other sites tested. This can be difficult due to the limited knowledge of gold standard sites of closed chromatin available for most species. Even for cells in which sites of closed chromatin have been mapped, these may be limited to a specific growth condition. Therefore we often use a tiling approach Figure 2.1 for detection of open chromatin sites using qPCR. Here, primer pairs are designed such that the products are either overlapping or closely spaced across the genomic regions being interrogated. The reference regions are those primer sets flanking the regions isolated by FAIRE. This strategy is also useful for validating results from microarray and sequencing data, which requires a set of positive and negative sites to determine both

sensitivity and specificity. Primer design is also critical for obtaining accurate results from qPCR, since primer pairs spanning or near the edges of open chromatin sites may be able to only detect a subset of the DNA fragments isolated in the aqueous phase. Optimally, primer pairs should be designed to amplify 60-100 bp products within the central portion of the identified regions. We typically calculate the relative enrichment for each amplicon using the comparative cT method [99]. Here, a ratio is calculated using the signal from the FAIRE sample relative to the signal from DNA prepared from an uncrosslinked sample. All ratios are then normalized to the amplicon with the lowest ratio, which is typically from the reference regions. Relative quantitation is used in part because FAIRE enriches for mitochondrial DNA, and since the mitochondrial content can vary considerably between cells it is difficult to get an accurate measurement of the proportion of genomic DNA enriched in each of the FAIRE samples.

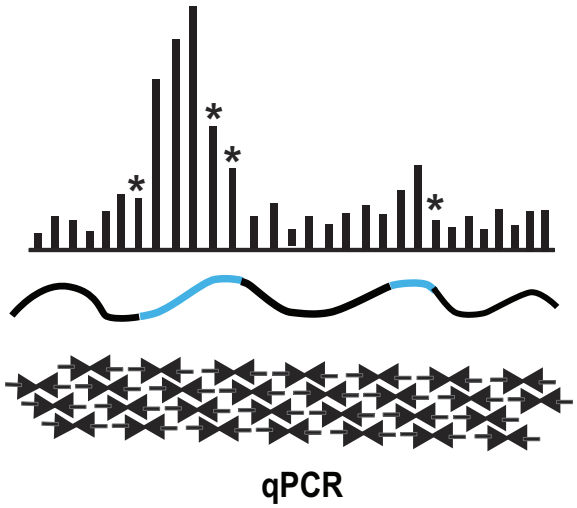

Figure 2.1: For qPCR, a series of primers, depicted as convergent arrows, are designed to span a genomic region of interest. Sites of open chromatin are highlighted in blue, with qPCR results depicted above. Amplicons that span or are near the boundaries of open chromatin often result in lower relative enrichment due to shearing of DNA fragments, as shown by asterisks.

2.2.2 DNA microarrays

High quality FAIRE data has been obtained from several microarray platforms, including Agilent, NimbleGen (Roche), and PCR-based arrays. Any microarray platform will suffice, but there are several factors to consider, such as the type of probe, the genomic regions covered, and the resolution [23]. One of the most important for FAIRE is selecting a microarray design with sufficient resolution. For oligonucleotide (50-75 bp) tiling microarrays, probe-to-probe spacing should not exceed 100 bp if possible. Doing so reduces the number of probes per FAIRE site to just one or two.

For dual-channel microarray platforms, DNA derived from uncrosslinked cells (right side of Figure 1.7), processed in parallel to the crosslinked cells, is hybridized as the reference or input sample. If it is not possible to obtain uncrosslinked cells, which is often the case when cells are limited or with tissues, crosslinks from a portion of the sample can be reversed and used as a reference. Remove an aliquot from the cleared lysate following sonication. Reverse crosslinks by incubating at 65°C overnight, and perform a phenol/chloroform extraction, ethanol precipitation, and RNase A treatment.

Typically, we amplify the DNA using ligation-mediated (LM) PCR [133]. The DNA fragments are made blunt using T4 DNA polymerase, asymmetric linkers (5′GCGGT GACCCGGGAGATCTGAATTC′3 and 5′GAATTCAGATC′3) are ligated to the blunt ends using T4 DNA ligase, and then amplified by PCR with a primer complementary to the linker. To avoid potential jackpot effects introduced during PCR, two amplification reactions are carried out in parallel for each sample, so 2 amplifications for FAIRE DNA and 2 amplifications for input DNA sample. Sample labeling and hybridization procedures follow the manufacturers recommended protocols

For tiling microarrays, raw data extraction is specific to the particular platform selected and entails image acquisition and feature quantitation. Data can be expressed

as a raw intensity for single-channel platforms or as a $\log_2$ ratio for dual-channel platforms. For data preprocessing, we typically normalize each dataset by calculating the z-score for each $\log_2$ ratio. The z-score is calculated by subtracting the mean $\log_2$ ratio and dividing by the standard deviation, which centers every dataset on the mean and standardizes the variance. In this way, every dataset has a mean of 0 and standard deviation of 1. This methodology is only applicable to dual channel platforms, although alternative strategies are available for single channel platforms [15, 76].

Identification of regions enriched by FAIRE can be accomplished using most existing peak-finding algorithms used for ChIP-chip experiments [49, 74, 86, 174]. For microarray data we typically use ChIPOTle [22], which uses a sliding window to identify statistically significant signals that comprise a peak. The significance of each region is determined by reflecting the negative portion of the data about zero, and then assuming a Gaussian distribution to estimate the null distribution.

The three main user-adjustable parameters in ChIPOTle are window size, step size, and threshold. Briefly we have found the following parameters to be optimal for analyzing FAIRE data from oligonucleotide tiling microarrays. For microarrays with probes spaced every 38 bp we use a window size of 300 bp. Whereas for probe spacing of every 60 to 100 bp we use a 500 bp window size. The larger window size is necessary to ensure a sufficient number of probes are included in each window. We use a step size that is the average probe spacing, which is measured as the start of one probe to the start of the next. We often try a range of thresholds and look at how the overlap changes between replicates and genomic features.

2.2.3 High-throughput sequencing

Each of the high-throughput sequencing platforms utilizes a different sample preparation procedure. We are most familiar with library preparation of FAIRE DNA for the Illumina Genome Analyzer II (GAII). We use 100 to 200 ng of DNA for starting material. This procedure involves blunting the ends of the DNA fragments (Epicentre #ER0720), adding an A overhang (Epicentre #KL06041K), and ligating double-stranded adapters containing a T-overhang to the DNA fragments (Illumina #1000521). Illumina adapters should be diluted for samples less than 500 ng. Adapters should be diluted 1:10 with ddH2O. For 500 ng use only 1 μl of undiluted adapters. Ligated samples are then amplified using PfuUltraII (Stratagene #600670) and primers complimentary to the adapters (Illumina #1000537 and 1000538). To avoid potential jackpot effects introduced during PCR, two amplification reactions are carried out in parallel, so 2 amplifications for FAIRE DNA. Amplified products are loaded into a 2% agarose gel using sample loading buffer (50 mM Tris pH 8.0, 40 mM EDTA, 40% w/v sucrose) and run at 120V for 1 hour. The brightest portion of the smear is excised, which typically corresponds to 150 bp (+/- 75 bp). If using a UV tray for the gel excision step, exposure should be limited to avoid crosslinking DNA products (alternatively a transillumination tray can be used). If using the Qiagen gel extraction kit for recovery of the DNA fragments, the incubation in QG buffer should be carried out at room temperature, not 55°C as per the manufacturers recommendation, since this induces a GC-bias. Subsequent sequencing of samples is carried out per the manufacturer's procedures.

Raw data acquisition for the GAII entails image acquisition and base calling. Approximately 25 million mapped 36 bp reads are typically required for robust detection of FAIRE peaks in a mammalian sample. Several algorithms are available for mapping the reads back to the genome, each utilizing different computational and alignment

strategies [90, 97, 148, 96]. Typically we use Bowtie [90], which incorporates information about read quality into the alignment. Since only the first 36 bp from either end of each 100 bp double-stranded DNA molecule is sequenced, we computationally extend each aligned read to produce 100 bp extended reads. For visualization, we count the number of extended reads overlapping every basepair in the genome, which we call a single basepair overlap count (SBPC). The baseoverlap count data for each basepair can be loaded into genome browsers, such as the UCSC genome browser [65].

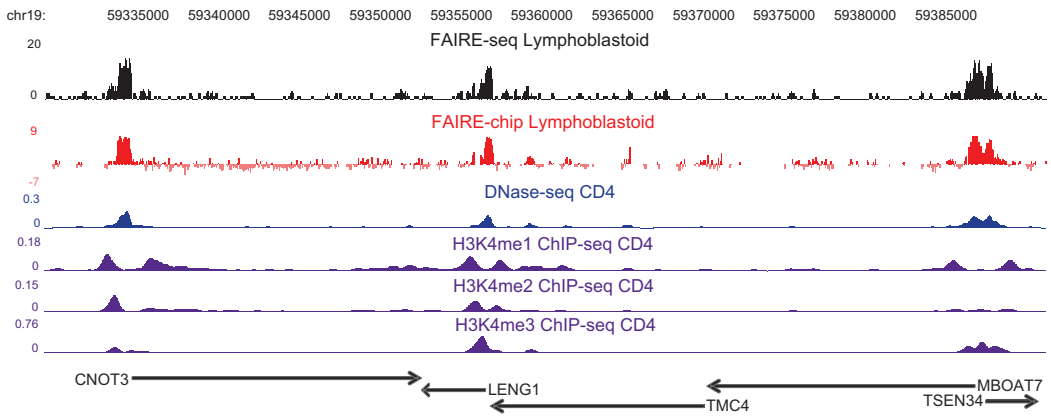

Figure 2.2: Here the exact same FAIRE DNA from lymphoblastoid cells were hybridized to a high density tiling DNA microarray (red) and sequenced using the Illumina GAII high-throughput sequencing technology. In general, sites identified by microarrays are highly concordant with those found by sequenceing. For comparison DNase-seq (blue) [19] and histone H3 lysine3 mono-, di- and tri-methylation ChIP-seq (purple) [7] from CD4 cells is shown as an indicator of active chromatin in a comparable cell type.

Chapter 3

Flexible and robust detection of enriched regions from DNA-seq experiments using ZINBA

The development of ZINBA was carried out in collaboration with Naim Rashid, a graduate student in Biostatistics at UNC. Naim was the main architect of the statistical components of the algorithm, including the development and implementation of the mixture regression model, model selection and peak refinement procedures. While I provided insights into the biological and computational challenges of data analysis, performed evaluation and feedback in the development of the algorithm and implemented many aspects of the data handling components. This included development of the local background estimate, as described below. The subsequent sections describe the principles of the algorithm and the performance of ZINBA, which is a portion of the work that has recently been accepted for publication [129].

3.1 Introduction

Next generation sequencing (NGS) technologies are now routinely utilized for genome-wide detection of DNA fragments isolated by a diverse set of assays interrogating genomic processes [118]. We refer to these collectively as DNA-seq experiments, which include Chromatin Immunoprecipitation (ChIP-seq), DNase hypersensitive site mapping (DNase-seq) [19], and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) [50], among others. Currently there are several algorithms available for the identification of genomic regions enriched by a given experiment [120]. Although each is well suited for the analysis of a particular intended data type, the underlying assumptions are not always suitable for the multitude of possible enrichment patterns found in DNA-seq datasets Figure 3.1. In particular, none of the existing algorithms performed adequately for the analysis FAIRE-seq data, which was due in part to the complex nature of the background signal and the relatively low signal-to-noise.

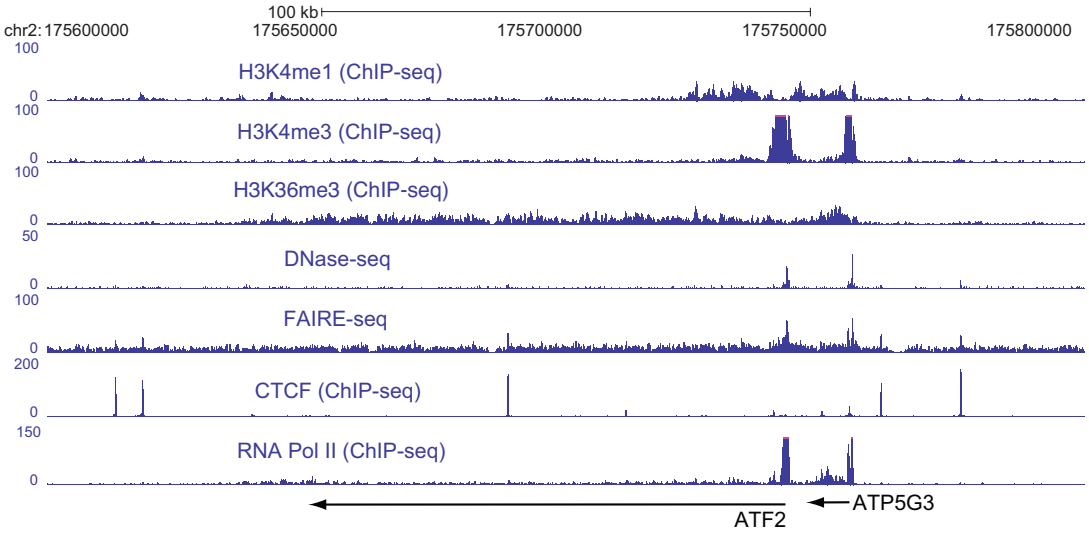

Figure 3.1: A 100 kb region of chromosome 2 at the ATF2 gene locus illustrating the diversity of enrichment patterns found in DNA-seq data. The y-axis represents the number of aligned reads overlapping a given basepair. The genes are represented by arrows showing the direction of transcription.

To this end, we developed ZINBA (Zero Inflated Negative Binomial Algorithm) for the robust detection of enrichment across a multitude of enrichment patterns in a variety of ChIP-seq and related DNA-seq experiments. ZINBA is capable of calling both broad and narrow modes of enrichment across a range of signal-to-noise ratios, with performance comparable to the existing set of algorithms specific to each data type. In addition, ZINBA models and accounts for factors that co-vary with background or experimental signal, such as G/C-content, and identifies enrichment in genomes with complex local copy number variations (CNV). Therefore ZINBA not only provides an interpretable analysis of FAIRE-seq data, but also serves as a unified framework for the analysis of a multitude of data types.

3.2 Overview of ZINBA

ZINBA implements a mixture regression approach, which probabilistically classifies genomic regions into three general components: background, enriched, and an artificial zero count. The regression framework allows each of the components to be modeled separately using a set of covariates, which leads to better characterization of each component and subsequent classification outcomes. In addition, the mixture-modeling approach affords ZINBA the flexibility to determine the set of genomic regions comprising background without relying on any prior assumptions of the proportion of the genome that is enriched.

ZINBA consists of three steps: data preprocessing, determination of significantly enriched regions, and an optional boundary refinement for more narrow sites Figure 3.2. The first step involves tabulating the number of reads falling into contiguous non-overlapping windows tiled across each chromosome and scoring corresponding covariate information. Covariates can consist of any quantity that may co-vary with signal in a

given region, including, for example, G/C-content, a smoothed average of local background, read counts for an input control sample, and the proportion of mappable [137] bases which we define as the mappability score. Additional sets of contiguous windows with offset starting positions can be tabulated as an option for increased resolution. Each set of offset windows is analyzed independently in the next step.

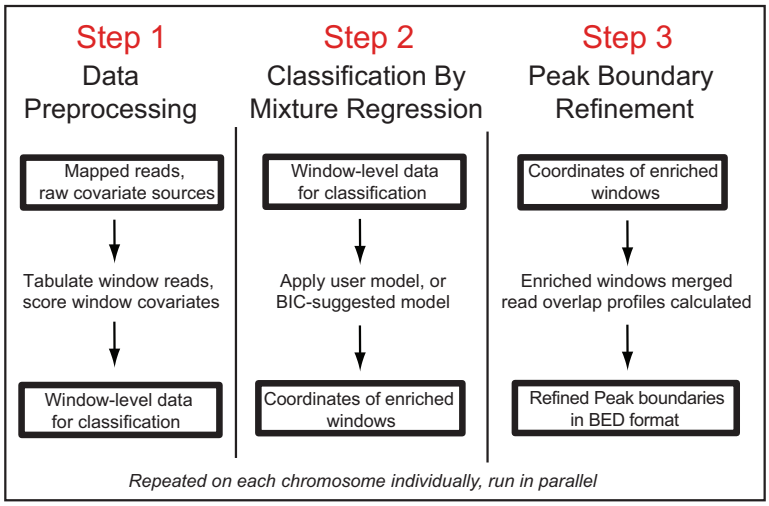

Figure 3.2: ZINBA is comprised of three steps that can each operate as an independent module. In Step 1, the set of aligned reads from the experiment along with a set of covariate measures are collated for each contiguous non-overlapping window spanning the genome. In Step 2, the component-specific model formulations of covariates are employed by the mixture regression framework to compute the posterior probability of each window belonging to either the zero-inflated, background or enrichment components. The component-specific model formulations of covariates can be generated using an automated model selection procedure or specified by the user. In Step 3, the set of windows exceeding the user-specified probability threshold (default 0.95) are merged to form broad regions of enrichment and a shape detection algorithm is employed on the read overlap representation of the data to refine the boundary estimates of distinct punctate peaks

In the second step, a novel mixture regression model is used to probabilistically classify each window into one of three components: background, enriched, or zero-inflated. Where enrichment refers to genomic segments captured specifically as the result of the biological experiment under consideration and background includes those

DNA sequences appearing due to noise from experimental and/or computational procedures. Zero-inflated regions are those genomic locations at which we might expect genomic DNA sequence from either the background or enrichment signal components, but which are not represented in the data. Zero-inflation typically occurs due to a lack of sequencing depth and is common in many NGS datasets. ZINBA utilizes an iterative approach [37] to determine for each window the relative likelihood of belonging to each component, in addition to estimating the relationship between average signal in each component and a set of covariates.

In the third step, all overlapping or adjacent windows classified as enriched are merged. For the detection of broader elements, especially helpful for histone modifications demarcating broad genomic regions (such as H3K36me3), an additional broad setting is available that merges enriched windows within a fixed distance An optional shape-detection algorithm may then be applied to identify sharp enrichment signals within broader enriched regions.

3.3 Data representation and calculation of covariates

The vast majority of algorithms for the analysis of DNA-seq data derive a null distribution by fitting a statistical distribution to a subset of the data deemed as background and then impose a "hard" cutoff to identify the enriched genomic segments. The effectiveness of this strategy relies on the extent to which the background data conforms to the assumptions regarding the estimation of the null distribution. This approach can also suffer diminished sensitivity and specificity as the signal at enriched sites approaches the background. The principle of the mixture model approach is not to derive a "hard" cutoff for distinguishing the two populations, but instead attempts to make

an informed decision about group membership using additional criteria. The benefit to sensitivity and specificity is most evident when the background and enriched signal is similar. The mixture model approach was particularly advantageous for the analysis of FAIRE-seq data due to the signal at enriched sites being less distinct from background and the violations of the assumptions regarding background used with other algorithms.

The following is a description of the set of additional criteria used for distinguishing background and enriched genomic regions, which are referred to as covariates.

3.3.1 Experimental and input DNA-seq data

Raw NGS data is comprised of millions of relatively short (25-75 bp) reads aligned to a reference genome sequence. A sequence read often does not represent the entire DNA fragment recovered with a given assay, but instead only a portion of the ends of each fragment. Therefore the set of aligned reads for each DNA fragment results in two distributions in the genome that correspond to the ends of fragment. The average distance between these distribution approximates the actual fragment length of the original fragment. The fragment length can often be approximated based on the fragment sizes generated during the preparation of material for sequencing, however it can also be computed using the cross-correlation coefficient [83]. ZINBA extends each aligned read in the $3'$ direction by this approximate fragment length and records the position of the central base. For FAIRE-seq data we extend the aligned reads by 100 bp, which is approximately the length of the DNA fragments recovered following phenol/chloroform extraction and following sample preparation for sequencing. The genome-wide set of positions for DNA fragments are summarized by counting the number falling within a set of contiguous non-overlapping windows, which were typically 300 to 500 bp in length. To avoid bisecting potentially significant regions, sets of contiguous

non-overlapping windows were also tabulated using offset starting positions.

3.3.2 GC-content

GC-content was selected because there is likely sequence composition dependent differences in the biological activity of factors studied with DNA-seq experiments, which can aid in the determining whether a DNA fragment belongs to the enriched or background regions. For each of the contiguous non-overlapping windows from above the G/C-content is calculated as the proportion of G and C bases in a given window.

3.3.3 Mappability

The mappability score is calculated as the proportion of all bases within a window that met the criteria for uniqueness imposed during alignment of the raw reads. Typically raw reads will only be aligned to positions that are unique throughout the genome. However, in some instances a more relaxed criteria may be used, such as with the FAIRE-seq data, raw reads could be aligned to a position that occurred four or less times throughout the genome. ZINBA implements the mappability software provided by PeakSeq [137] to calculate for each base pair the number of times a given k-mer (36 bp) starting at that base occurs throughout the genome. If a base pair receives a score of 1 then only one occurrence of the given k-mer exists throughout the genome and would be a mappable position under the absolute uniqueness criteria. Whereas if a base pair received a score of 5 then it would not be considered mappable for either the uniqueness or relaxed criteria described above. Before the mappability scores are summarized into the windows, those bases that meet the specified criteria are assigned a score of 1, while those that do not are assigned a score of 0. Finally, since the central position of each extended read is used for window assignment, the mappability data is shifted in the same way, where for each base the score of 0 or 1 is shifted both +/- one

half the average fragment length. As a result, each base in the genome has a score of 0, 1 or 2 depending on whether neither, one or both of the up- and downstream base pairs were mappable, respectively. The sum of mappability scores are tabulated and divided by two times the window size to derive the proportion of mappable bases in the window.

3.3.4 Local background

The local background estimate aims to roughly approximate large-scale fluctuations in background signal resulting from local variations in genomic copy number. It is calculated using a sliding window approach where, by default, 100 kb windows are stepped every 2.5 kb across each chromosome. The size of these large windows were selected to be sufficiently large to prevent sites of enrichment from influencing the estimate, but small enough to preserve enough resolution to capture local fluctuation in background signal. The number of reads per mappable base pair is calculated for each window. Windows that span the boundaries of CNVs are problematic, resulting in artificially inflated and deflated estimate of local background. Therefore an additional step is employed to identify these boundaries as change points, remove windows straddling these boundaries and calculate scores for new windows immediately adjacent to the boundaries. For each ZINBA window, which is considerably smaller than 100kb, the local background estimate is computed as the average of all overlapping 100kb windows, multiplied by the length of the ZINBA window.

3.4 Relationship between covariates and component signal can vary between experiments

The mixture regression framework implemented by ZINBA requires the user to specify the relationship between covariates and the set of components. Since the relevant covariates are not always known a priori ZINBA employs the BIC [141] to select the best formulation of covariates for each component. BIC balances model fit and model complexity and has long been employed as a statistical assessment of model performance. Covariates with weak relationships with the mean signal in a component will have little effect on classification and contributes to model complexity. The BIC criterion helps to remove such covariates to balance model fit and model size.

Evaluation of the relationships between the set of component-specific covariates selected using the automated model selection procedure for the RNA Pol II ChIP-seq and FAIRE-seq datasets shown in Figure 3.1 [42, 134], revealed covariates vary both between components and assays. As shown in Figure 3.3, the mappability score and input/local background were positively related with signal in the background component for both the RNA Pol II ChIP-seq and FAIRE-seq datasets, which is consistent with previous reports [137, 173]. For the RNA Pol II data, model estimates reveal that G/C-content had a positive relationship in background regions, similar to previous reports on G/C-content bias [38, 64, 125] (Figure 3.3A). However, in FAIRE-seq data, G/C-content was negatively associated with the background component (Figure 3.3B). These differences can easily be observed from scatter plots of the raw read counts from windows classified as background versus the corresponding G/C-content for the RNA Pol II and FAIRE-seq datasets (Figure 3.3C,D). The exact cause of the differences in the relationship between G/C-content and background signal between datasets, and whether it could be technical or biological, is not known.

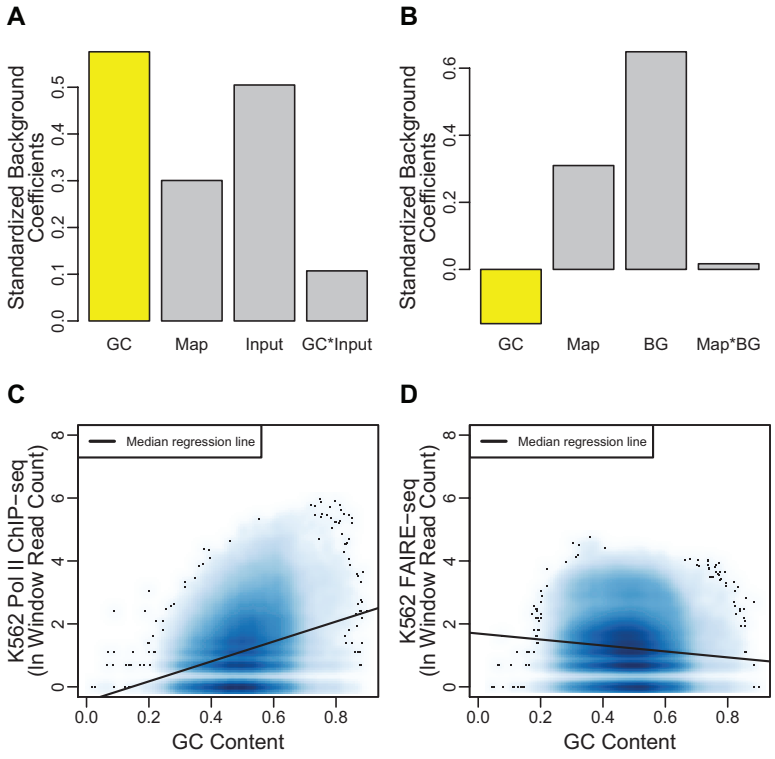

Figure 3.3: Estimates for the set of BIC selected covariates for the background components of the (A) RNA Pol II ChIP-seq and (B) FAIRE-seq data from K562 cells, chromosome 22. The set of covariates were standardized to a mean of 0 and variance of 1, which included G/C-content (GC), mappability score (Map), the local background estimate (BG), and input control (Input). The G/C-content covariate had a positive and negative effect on the background component of the RNA Pol II and FAIRE data, respectively (yellow shaded bar). Density plots of G/C-content (x-axis) versus the natural log of window read count (y-axis) in non-enriched windows (enrichment posterior probability <0.50) from the (C) RNA Pol II and (D) FAIRE data. Median regression lines fit to the set of background windows from each dataset (windows with enrichment posterior probability <0.50) mirror the ZINBA-estimated relationships between G/C-content and signal in background regions.

The relationship for each covariate also differed in magnitude and direction across components of the same dataset. For FAIRE-seq, although there was a negative relationship with G/C-content in background regions there was a positive relationship in the enrichment regions. A similar difference between the relationship of G/C-content in the background and enrichment regions was found for the RNA Pol II ChIP-seq data.

These results suggest that the relationships of covariates may not be consistent with background signal across different data types, or may different in their relationships between signal in background and enrichment regions.

3.4.1 Performance using FAIRE-seq data

ZINBA was compared with MACS [173] and F-Seq [18] which represent two classes of peak calling algorithms that also do not require an input control sample to call regions of enrichment. MACS represents a class of algorithms that uses a sliding window approach for the detection of enriched regions compared to a matching input control sample or local background estimate. F-Seq represents a class of algorithms that use kernel density estimation to estimate local read density and identifies enriched regions as those density estimates that exceed a user-defined threshold, which is estimated using simulations assuming random assortment of sample reads.

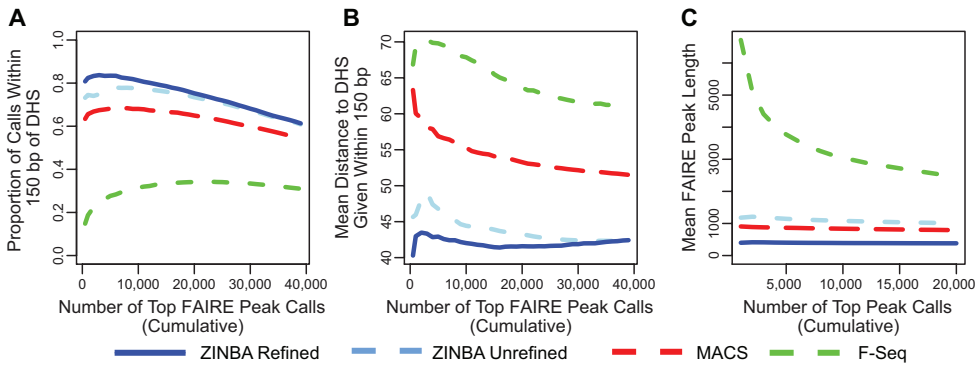

Figure 3.4: For FAIRE-seq data, the top N ranked peaks from MACS (red dashed line), F-Seq (green dashed line), ZINBA unrefined regions (light blue dashed line), and ZINBA refined region (blue solid line) were compared based on the proportion overlapping a DNase hypersensitive site (A), the average distance to the DNase hypersensitive site (B) and average length of peaks called by each algorithm (C).

Peak calls from ZINBA, MACS and F-Seq using FAIRE-seq data from K562 cells,

which lacked a matching input control, were compared to a set of DNase I hypersensitivity sites (DHS) [19, 31] isolated from the exact same set of cells. The DHS were called by F-seq, and were selected as a standard because of the longstanding use of DNase as a method for identification of open chromatin sites. Both ZINBA and MACS called a high proportion of FAIRE sites that were overlapping a DHS, but only a very low proportion of FAIRE sites called by F-seq were localized to a DHS (Figure 3.4A). The set of sites called by both MACS and F-Seq tended to be longer and more errant in K562 CNV regions, where approximately 50% of ZINBA peaks were localized to a DHS compared to only 37% and 27% of MACS and F-seq peaks. Overlap between called peak sets from ZINBA, MACS, and F-seq for FAIRE were more disparate than those found in high signal-to noise data.

3.4.2 Detection of regulatory elements within CNVs

One challenge for the analysis of DNA-seq data is fluctuations in background signal resulting from copy number variation (CNVs). If not properly accounted for, such changes in background can result in significant false-positives. This is especially true if there is no input control for comparison or the input control is not sequenced to a sufficient depth. To account for this, we constructed a new covariate to measure local background, and included this covariate in our mixture regression framework to account for local copy number changes. Changes in background signal levels due to CNVs were estimated locally using the DNA-seq sample itself, supplemented by a change-point detection method to determine boundaries of likely CNVs. Application of this approach provided an accurate estimation of signal changes due to local CNVs in a FAIRE-seq MCF-7 breast cancer dataset, which is aneuploid and has extensive CNVs [58] (Figure 3.5A).

Using a BIC-selected model considering the local background estimate, G/C-content,

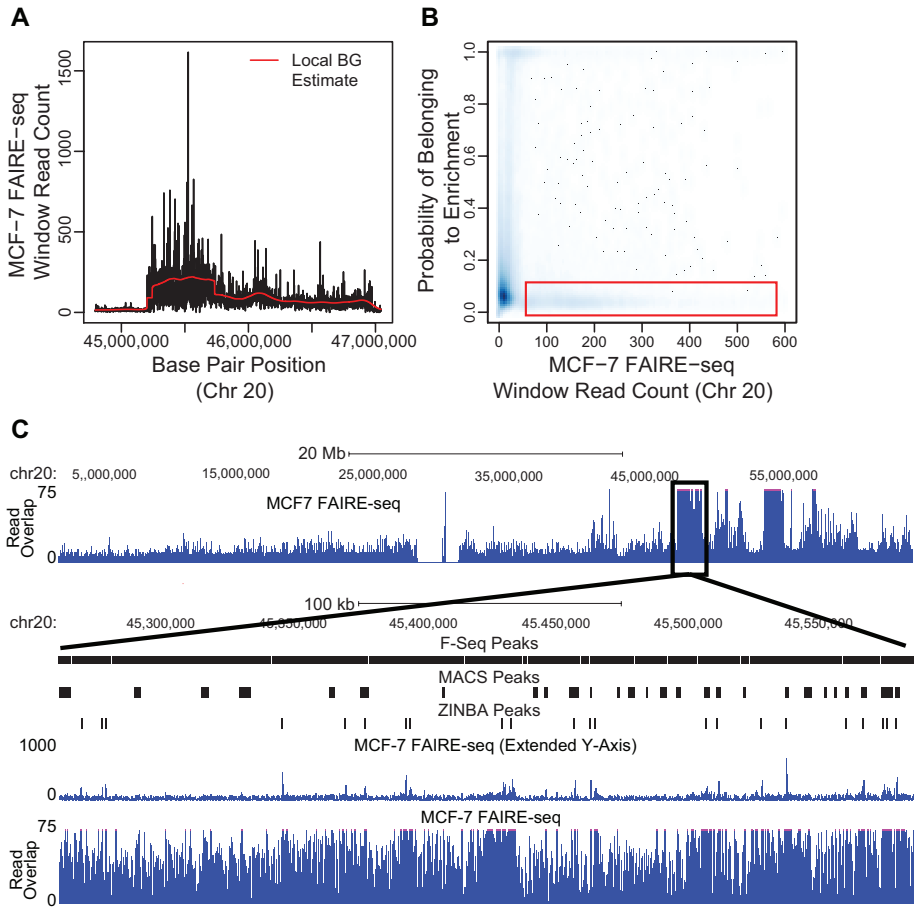

Figure 3.5: (A) The local background estimate (red line) approximates a CNV detected by FAIRE-seq (black line) within a 2 Mbp region of chromosome 20 in MCF-7 cells. (B) Density plot of the window read counts for FAIRE-seq data in MCF-7 (chromosome 20) versus the posterior probability of a given window belonging to enrichment, which included the local background estimate as a starting covariate in the ZINBA model formulation. The red box highlights a set of windows with high read counts (CNV background) being assigned a low posterior probability of being enriched. (C) The read overlap representation of MCF-7 FARE-seq data for all of chromosome 20 (top row) is displayed in the UCSC Genome Browser. The bottom panels zoom in on the black box outlining a CNV (same as panel A). Here a set of peak calls by F-Seq, MACS and ZINBA are shown as black boxes along with the FAIRE-seq data displayed using either an extended (top) or standard y-axis.

and mappability score as starting covariates, ZINBA was able to correctly classify background regions within CNVs (Figure 3.5B) and called 8 and 11 times fewer peaks (1,258) in MCF-7 CNV regions on chromosome 20 relative to MACS and F-seq (Figure 3.5C).

However estimation of local background from the experimental data is only effective when local background is sampled from a sufficiently large window size, where these large windows (default 100 kb) are not dominated by enriched signal. This is the case with the majority of data types, as most contain features that span no more than several kb. Nevertheless, other estimates of CNV boundaries can be used, because the flexibility of ZINBA allows for any CNV estimate to be included into the model selection procedure and determination of enrichment.

3.5 Conclusions

Although ZINBA was initially developed to aid in the detection of enriched regions for FAIRE-seq data, the strategies employed in the algorithm also provide a general and unified framework for the detection of enriched regions for a variety of DNA-seq data types. This framework will certainly be valuable as the production of NGS data continues to increase and diversity of data types included in a given experiment expands. Allowing researchers the opportunity to rely on a single algorithm instead integrating results from a variety of algorithms or using a single algorithm ill-suited for all the data types.

Chapter 4

Isolation of active regulatory elements from human chromatin using FAIRE

The development of FAIRE for human cells was carried our primarily in collaboration with Vishy Iyer at the University of Texas at Austin. All of the tissue culture and fixation of the fibroblast cells was carried out by Jonghwan Kim, then a graduate student in Vishy's lab. The ENCODE microarrays used for the detection of FAIRE samples was kindly provided by Roland Green at Nimblegen and hybridization was carried out under the supervision of Michael Singer.

4.1 Abstract

DNA segments that actively regulate transcription in vivo are typically characterized by eviction of nucleosomes from chromatin, and are experimentally identified by their hypersensitivity to nucleases. Here we demonstrate a simple procedure for the isolation of nucleosome-depleted DNA from human chromatin, termed FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). To perform FAIRE, chromatin is crosslinked

with formaldehyde in vivo, sheared by sonication, and phenol-chloroform extracted. The DNA recovered in the aqueous phase can be detected by hybridization to a DNA microarray or directly read using high-throughput sequencing. FAIRE performed in human cells strongly enriches DNA coincident with the location of DNaseI hypersensitive sites, transcriptional start sites, enhancers, insulators and active promoters. The set of active regulatory elements enriched by FAIRE also vary between cell-types. FAIRE has utility as a positive selection for genomic regions associated with regulatory activity, including regions traditionally detected by nuclease hypersensitivity assays.

4.2 Introduction

Chromatin at genomic loci that actively regulate transcription is distinguished from other chromatin types. The observation that the 5′ regions of genes became hypersensitive to both DNaseI and micrococcal nuclease upon gene activation in *Drosophila* was among the earliest demonstrations of this phenomena [82, 94, 169, 168]. The appearance of these hypersensitive sites reflects a loss or destabilization of nucleosomes at the promoters of active genes [14]. Several mechanisms act in concert to achieve this result. Loss of nucleosomes can be caused directly by a protein bound to its cognate site on DNA [171], facilitated in part by increased acetylation of the nucleosomes just prior to the activation of transcription [131], or can be mediated by the well-characterized SWI/SNF family of ATP-dependent nucleosome remodeling complexes [155, 158, 162]. Regardless of the specific mechanisms employed at any individual promoter, achieving nucleosome clearance at active regulatory regions is a conserved mechanism among eukaryotes [165].

Because nucleosome disruption is a conserved hallmark of active regulatory chromatin throughout the eukaryotic lineage, a simple, high-throughput procedure to isolate

and map chromatin depleted of nucleosomes would allow regulatory regions to be identified in a broad range of organisms and cell types. One such procedure, which we term FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements), was first demonstrated in the budding yeast *Saccharomyces cerevisiae* [113]. In yeast, the genomic regions immediately upstream of genes were preferentially segregated into the aqueous phase, in a manner that was strongly negatively correlated with nucleosome occupancy [10, 92, 172].

Human chromatin poses new challenges to FAIRE. Compared to the twelve million base-pair genome of yeast, the three billion base-pairs of the human genome makes it nearly 300 times as large. Only 1.5% of human DNA is coding, with perhaps 30% of the genome transcribed (introns plus exons), relative to 50% coding for yeast, with 85% of the genome being transcribed under a single growth condition [35, 69, 128, 167]. In addition, mammalian chromatin is inherently more complex than that of yeast. The majority of mammalian genes contain introns, regulation can occur at much greater distances from the initiation of transcription, there are more repetitive and heterochromatic regions, and the baseline state of chromatin is more compact and repressive [1]. Therefore, it is reasonable to expect that a much smaller fraction of the genome will be in the "open" conformation representing regions of active chromatin. Moreover, it is not clear *a priori* whether the same physical properties of yeast chromatin that allow isolation of open regions by FAIRE can be successfully exploited for isolation of regulatory regions in human chromatin.

The work presented in this chapter outlines the initial set of experiments characterizing FAIRE in a human foreskin fibroblast cell line using DNA microarrays tiling the 1% of the genome defined by the ENCODE (ENCyclopedia Of Dna Elements) consortium [41]. Subsequent studies expanded upon this work by employing high-throughput sequencing for genome-wide detection and included additional cell lines.

The goal of the initial experiments [51] was to ascertain whether the active regulatory elements recovered by FAIRE in human cells had similar characteristics to what had been observed in yeast. Subsequently we sought to characterize the set active regulatory elements in human chromatin and determine the extent to which these varied between cell-types [151]. The results indicate that FAIRE is a simple genomic method for the isolation and identification of human functional regulatory elements, with broad utility for mammalian genomes.

4.3 Materials and methods

4.3.1 Cell lines

Four independent cultures (biological replicates) of Human foreskin fibroblast (ATCC CRL 2091) cells were grown in 245 x 245 cm plates to 90% confluence.

4.3.2 Sample amplification, labeling, hybridization, and quantitation

Samples were amplified using ligation-mediated (LM) PCR [133]. Briefly, DNA fragments in a sample from each time-point were made blunt using T4 DNA polymerase. Asymmetric linkers (5′GCGGTGACCCGGGAGATCTGAATTC′3 and 5′GAATTCA GATC′3) were ligated to the blunt ends, and the samples were amplified by PCR with a primer complementary to the linker.

Sample labeling and hybridization were performed at NimbleGen. Samples were labeled by incorporation of cyanine dyes by polymerization with Klenow fragment primed by random nonomers. FAIRE samples were labeled with Cy5, and genomic DNA (to be used as a reference) was labeled with Cy3. The labeled samples were mixed and hybridized to high-density oligonucleotide microarrays tiling the ENCODE

regions (NimbleGen Systems, Inc.). The microarray contains approximately 385,000 50-mer probes, sharing 6 bp with the adjacent probes, allowing measurements at 38 bp resolution across the non-repetitive sequence in the ENCODE regions. Hybridizations were performed in a MAUI hybridization station for 16 hours at 42o C. Arrays were washed and scanned with an Axon Scanner 4000B. Spot intensities were quantitated using GenePix software and normalized by NimbleGens in-house software. Data from all four crosslinking times, which were prepared from four independent biological samples, were averaged for all analyses.

4.3.3 qPCR validation

Portions of three ENCODE regions were selected for validation: chr8:119189349-119195557, chr21:32813792-32820968, and chr7:26978053-26987656. 96 primer pairs were designed for qPCR and divided between the three regions, spaced as evenly apart as possible. DNA used in the qPCR validation was obtained independently using an identical protocol and cell line as for the microarray analysis. PCR was performed using SYBR green chemistry on an ABI 7900 instrument. Relative enrichment of each amplicon in the FAIRE-treated DNA was calculated using the comparative cT method [99]. DNA from untreated fibroblast cells served as the control for the calculations.

4.3.4 Data analysis

The signal generated by FAIRE is similar to that generated by a conventional ChIP-chip experiment. Therefore we used the peak-finding algorithm ChIPOTle [22] to identify regions isolated with FAIRE. Briefly, ChIPOTle uses a sliding window (300 bp) to identify statistically significant signals that comprise a peak. The null distribution is determined by reflecting the negative data from the region of interest about zero and fitting a Gaussian distribution. For the analysis presented, values calculated from the

average of four FAIRE experiments were input to ChIPOTle. Displayed peaks correspond to a p-value of less than 10-25, after using the Benjamini-Hochberg correction to adjust for multiple tests [8]. All of the feature sets used to compare with FAIRE peaks were downloaded from the UCSC genome browser. For the DNase-chip data, we excluded peaks found in only 1 of the 3 DNase concentrations reported by [32].

For visualization, data was loaded to the UCSC genome browser [65]. Genomic annotations including TSSs were produced by the GENCODE project [4, 61], whose goal is to provide high-quality annotation of all protein-coding DNA sequences that have been experimentally verified. All coordinates reported are based on human genome sequence release hg17 (NCBI build 35). Each annotation track presented is available for download, along with the raw FAIRE data for each microarray. The FAIRE data is also available from GEO (GSM109841, GSM109842, GSM109843, GSM109844 and series GSE4886).

4.4 Results

4.4.1 DNA isolated by FAIRE in human cells corresponds to regions of active chromatin

Fibroblasts were grown in culture, and formaldehyde was added directly to actively dividing cells to a final concentration of 1%. The cells were then disrupted with glass beads. The resulting extract was sonicated to yield 0.5 to 1 kb chromatin fragments, and subjected to phenol-chloroform extraction. The DNA fragments recovered in the aqueous phase were fluorescently labeled and hybridized to high-density oligonucleotide microarrays tiling the ENCODE regions at 38-bp resolution. The ENCODE regions represent 1% of the human genome (30 Mb), consisting of manually-selected regions

of particular interest and randomly-selected regions of varying gene density and evolutionary conservation [41]. As a reference, DNA prepared in parallel from uncrosslinked cells was labeled with a different fluor and simultaneously hybridized to the arrays.

We compared the genomic regions enriched by FAIRE to hallmarks of active chromatin, including localization of the general transcriptional machinery [85, 86], histone H3 and H4 acetylation and methylation [88], DNaseI hypersensitivity [32, 139], and direct assays of promoter activity [29, 157]. Genomic regions enriched by FAIRE correspond well with each of these indicators of active regulatory elements (Figure 4.1).

4.4.2 Active promoters are enriched by FAIRE

Earlier experiments performed in yeast had revealed that the regulatory regions of highly transcribed genes are preferentially isolated by FAIRE [113]. To determine if this relationship holds in human cells, we compared FAIRE signal to measurements of promoter strength. Predicted promoters in the ENCODE regions have been analyzed for regulatory activity by cloning them upstream of reporters and measuring the resulting activity of the reporter gene in different cell types [29, 157]. We assigned each probe on the microarray that mapped to a predicted promoter to one of four classes, based on the average activity of the corresponding promoter. Analysis revealed that probes mapping to the most active promoters have a higher FAIRE signal than those that do not map to a promoter, or that map to a promoter of lower activity (Figure 4.2A, p-value $<$ 10-100). Therefore, more active promoters are more strongly enriched by FAIRE in human cells.

4.4.3 FAIRE isolates DNA encompassing transcriptional start sites

Yeast experiments had also revealed that FAIRE isolated the nucleosome-free region located at yeast transcription start sites (TSSs) [68, 113, 172]. Alignment of DNase-chip signal [32], FAIRE signal, and gene annotations suggested that a similar feature was enriched by FAIRE in human cells (Figure 4.1). To assess the extent to which this was generally true, we aligned all TSSs for all annotated genes within the ENCODE regions, and calculated the average FAIRE signal over a region spanning 1.5 kb upstream to 1.5 kb downstream of the TSS (solid line, Figure 4.2B). This analysis revealed that on average, the peak of enrichment by FAIRE occurs at the TSS. DNase-hypersensitive sites are an indicator of DNA accessibility and a well-established characteristic of TSSs and regulatory DNA. We performed the same analysis using DNase-chip data [32] and found that the pattern of DNA enrichment at TSSs was very similar to that generated by FAIRE (dashed line, Figure 4.2B).

4.4.4 Global comparison of FAIRE peaks to other annotated features

We also analyzed the overall concordance between the genomic regions enriched by FAIRE and other selected hallmarks of active chromatin (Figure 4.2C). The overlap of FAIRE peaks and these marks (TSS [4, 61], DNaseI hypersensitivity [32, 139], 75th percentile of promoter activity [29, 157], RNA polymerase II (RNAP) ChIP-chip, or TAF1 ChIP-chip [85, 86]) is very strong, in most cases over 10 times the frequency observed with permuted data. Many of the FAIRE peaks overlap multiple marks of active chromatin (60% of peaks shown, 21% of all peaks) (Figure 4.2C). In addition,

there are a number of FAIRE peaks, which do not correspond to any of the annotations selected for comparison. These likely arise to due to a number of factors, most significantly the difference in cell types used among the experiments being compared, the sparse state of current human genome annotations and yet uncharacterized distal regulatory elements.

4.4.5 FAIRE isolates regulatory elements specific to individual cell types

Although all somatic cells in an organism contain the same genomic DNA, different cell types express different genes. These differences reflect differential utilization of regulatory information encoded in the genome. To determine if FAIRE could detect regulatory elements specific to a certain cell type, we compared FAIRE data derived from fibroblasts and HeLaS3 cells (Figure 4.3A). The conservative set of peak calls correspond to an extremely stringent cutoff for detection and represent very high quality FAIRE sites. While the majority of liberal peak calls are still *bona fide* regulatory elements with lower amplitude signal. Comparisons along the diagonal indicate that regardless of the stringency ∼25% of the FAIRE sites are held in common between cell types. Comparison of the conservative versus liberal peak calls (top right and bottom left corners) indicate that these differences are not simply a matter of the same set of regulatory elements with different amplitudes, but instead that these differences reflect completely independent sets of regulatory elements between cell types.

Figure 4.3B displays high-throughput sequencing data from FAIRE performed in an expanded set of cell lines. Across all four cell lines there are various patterns of regulatory occurrence, which likely reflects differences and similarities in the lineage and activity of regulatory pathways.

4.5 Discussion

Several aspects of FAIRE make it a powerful genome-wide approach for detecting functional *in vivo* regulatory elements in mammalian cells. First, FAIRE requires no treatment of the cells prior to the addition of formaldehyde. Formaldehyde is applied directly to the growing cells and enters quickly because of its small size (HCHO), which is comparable to that of water. In yeast, 1% formaldehyde immediately stops cell growth and results in 50% lethality in just 100 seconds, with 99% lethality achieved in 360 seconds (data not shown). Therefore, the state of chromatin just prior to the addition of the formaldehyde is likely to be captured. In contrast, nuclease sensitivity assays often require that cells be permeabilized, or that nuclei be prepared, both of which allow time for artifacts based on these preparations to occur.

Second, each time a nuclease-sensitivity assay is performed, the appropriate enzyme concentration and incubation time must be determined, due to lot-to-lot variations in commercial DNase activity and variations in individual nuclei preparations. With FAIRE, a wide range of incubation times (1, 2, 4, and 7 minutes) at a single formaldehyde concentration (1%) appear to be equally effective. FAIRE involves few steps, few variables, and takes less than an hour, making the method easy to control and develop. Few reagents other than formaldehyde, phenol, and chloroform are required. These properties make FAIRE amenable to high-throughput. Third, in contrast to ChIP, there is no dependence on antibodies, supplies of which may be limited, or upon tagged proteins, which may be difficult to construct, impaired in function, or expressed at inappropriate levels. FAIRE can analyze any cells; wild-type, mutant, or those that contain transgenes that would make histone ChIPs technically difficult (for example, those containing Protein-A based tags).

Another important advantage of FAIRE is that it positively selects genomic regions

at which nucleosomes are disrupted. These same regions would be degraded in nuclease sensitivity assays, and require identification by their absence, or by cloning and identification of flanking DNA [31]. In contrast, DNA isolated by FAIRE is the DNA of interest, allowing the use of direct detection methods like DNA microarrays and high-throughput sequencing.

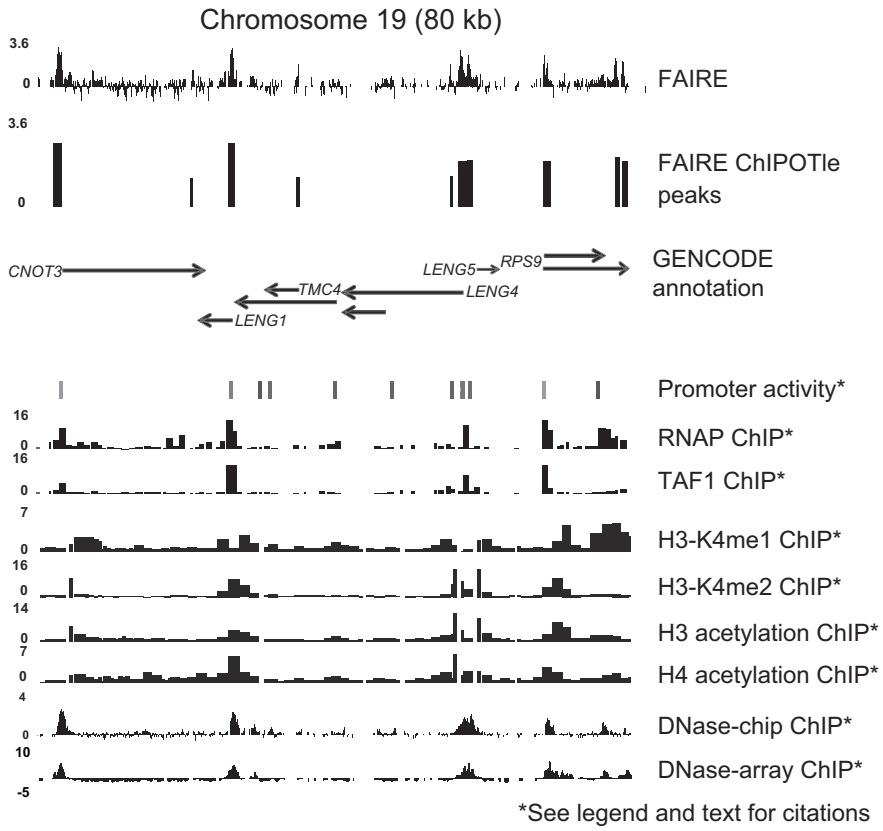

Figure 4.1: FAIRE data was loaded into the UCSC genome browser along with datasets generated by other ENCODE consortium members (labeled on the right-hand side). The top track represents the average $\log_2$ ratios for the FAIRE data from four independent cultures (biological replicates), each of which were crosslinked separately (for 1, 2, 4, and 7 minutes). The second track shows FAIRE peaks (cutoff = p ¡ 10-25) as determined by ChIPOTle [22]. The GENCODE annotations represent experimentally verified transcribed segments [4, 61]. Promoter activity represents the average activity of a reporter construct driven by each of the indicated regions and measured across 16 cell lines, where grey bars indicate high activity and black bars no activity [29, 157]. ChIP-chip data for RNAP and TAF1 from lung fibroblast cells (IMR90) is displayed as the $\log_1 0$ of the p-value for each probe scaled to 0-16 [85, 86]. ChIP-chip data for histone H3 and H4 acetylation and H3K4 mono-, di-, and trimethylation in embryonic lung fibroblast cells (HFL-1) is shown as the ratio of IP-signal over background [88]. Finally, data on DNaseI hypersensitivity is shown for two different techniques, DNase-chip and DNase-array, both techniques isolate DNA fragments flanking DNaseI cleavage sites and map them back to the genome using microarrays [32, 139]. The data shown for DNase-chip is the average $\log_2$ ratio for nine replicates (3 biological at 3 different enzyme concentrations), whereas the DNase-array data is the $\log_2$ ratios scaled so that a $\log_2$ ratio of 0 represents the 99% confidence bound on the experimental noise. The region shown corresponds to Chromosome 19 coordinates 59,330,000 to 59,409,000.

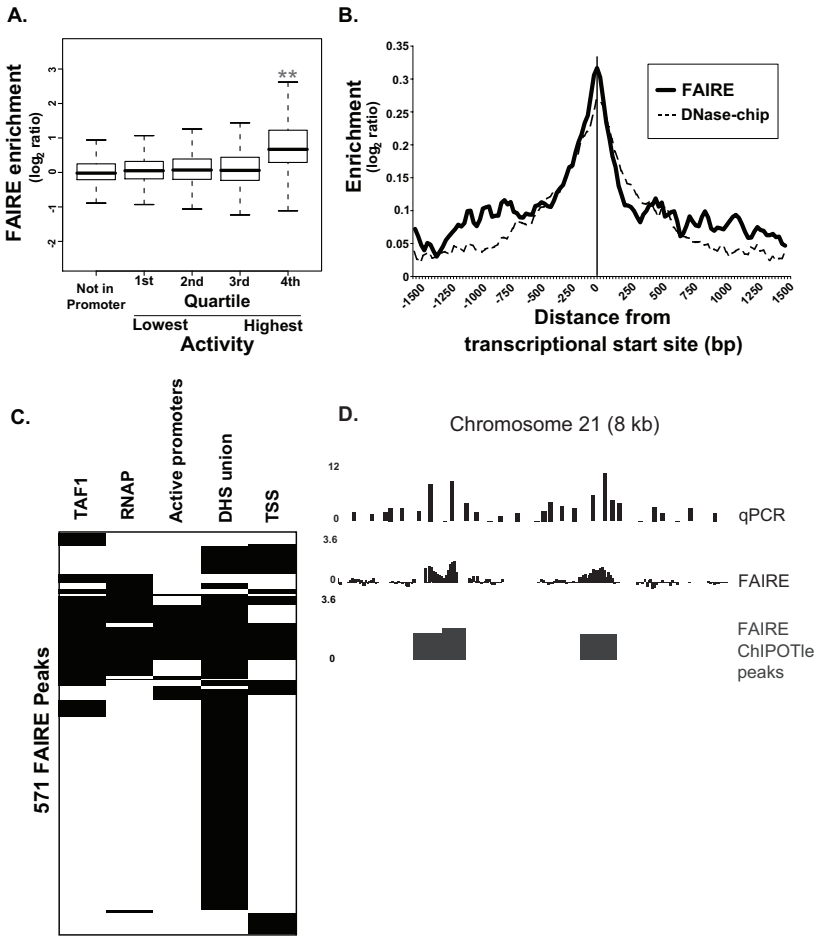

Figure 4.2: (A) Probes that mapped to predicted promoters were divided into quartiles based on the level of activity for each promoter, which was measured by driving a reporter construct [29, 157]. The black line in the center of box is the median value, while the bounds of the box represents the inter-quartile range of the FAIRE data. (B) Probes within +/- 1.5 kb of a GENCODE annotated transcriptional start sites [4, 61] were analyzed using a 50 bp sliding window (1 bp step) to calculate the average FAIRE enrichment at TSS (solid line). For comparison, the same analysis was performed using the DNase-chip dataset (dashed line). (C) A representation of the relationship between FAIRE peaks and other annotated features. Each row corresponds to one of the 571 FAIRE peaks that overlap with at least one of the following: a TSS [4, 61], DHS [32, 139], 75th percentile of promoter activity [29, 157], RNAP ChIP-chip, or TAF1 ChIP-chip [85, 86]. A black bar represents overlap with the FAIRE signal, while white represents no overlap (413 FAIRE peaks did not overlap with any of these marks). Data was clustered for display [40]. (D) qPCR validation of the microarray data was performed over three 8 kb regions. The height of the bars from the qPCR analysis represents the enrichment of the FAIRE samples relative to the uncrosslinked reference; the FAIRE data and peaks are the same as described in Figure 4.1.

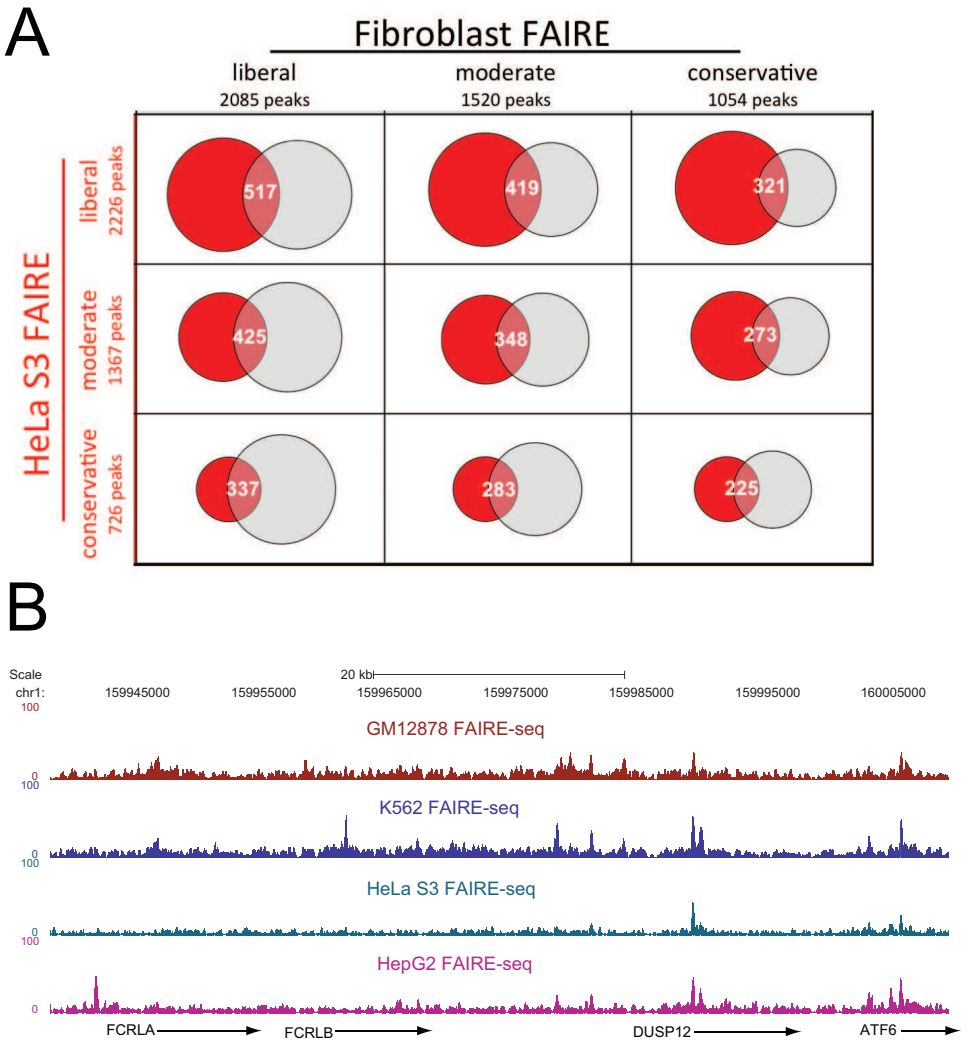

Figure 4.3: (A) FAIRE from HeLaS3 (rows) and fibroblast (columns) cells was hybridized to DNA microarrays covering the 1% ENCODE regions and analyzed using ChIPOTle at three levels of stringency. The thresholds were set to generate roughly equivalent number of peaks for each category and the numbers of peaks are listed on the corresponding rows and columns. The comparison is shown as a venn diagram, where the degree of overlap is proportional to the extent to which the peaks were the same. (B) Here high-throughput sequencing data for FAIRE performed in four cell lines GM12878, K562, HeLaS3 and HepG2 is shown in a representative 60 kb of chromosome 1. The y-axis is the count of overlapping sequence fragments after being aligned and extended.

Chapter 5

Classification of breast cancer subtypes using FAIRE

This work was carried out in collaboration with Charles Perou, professor in the Department of Genetics at UNC. All of the tissue culture of the MCF7 and SUM102 cells were carried out by Olga Karginova. While the tumor samples were provided by Xiaping He. All expression data, including RNA-seq and tumor expression microarrays, was generated by various members of the Perou lab.

5.1 Abstract

Breast cancer is a heterogenous disease comprised of molecularly distinct subtypes characterized by differential activity of regulatory pathways. Using FAIRE we were able to identify the genome-wide set of subtype-selective active regulatory elements, infer the function of regulatory pathways and classify tumor samples. The set of active regulatory elements allowed us to identify sets of both known and novel factors involved in transcriptional regulation. The set of subtype-selective sites were found to be colocalized throughout the genome with the set of genes up-regulated in the respective subtypes. Unexpectedly we found that separate regulatory mechanisms can

be employed between subtypes to achieve a comparable level of gene expression, suggesting that gene expression alone is not sufficient to explain the molecular complexity of breast cancer. FAIRE is also capable of identifying copy number variations (CNVs), which provides an additional molecular characteristic for the classification of samples based on clinical outcomes. FAIRE performed in tumors was capable of identifying subtype-selective regulatory elements indicative of the status of transcription factors, such as the presence of the estrogen responsive element (ERE) in the ER-positive luminal tumors. Together these findings suggest that FAIRE will be a powerful tool in the study of breast cancer and can be easily adapted for clinical research.

5.2 Introduction

One of the major challenges in the effective diagnosis and treatment of breast cancer is understanding how the molecular composition of tumors relates to the diversity of clinical outcomes. Breast cancer occurrence, progression and treatment outcomes have been characterized molecularly based on immunohistochemical markers, genetic mutations and gene expression signatures to reveal distinct subtypes [101]. These subtypes have distinct morphological features and clinical behaviors, which are broadly categorized by the originating cell type, the status of the estrogen and progesterone hormone receptors and the presence and activity of the HER2 amplicon. Using microarray gene expression data researchers initially characterized five intrinsic subtypes of breast cancer, which included luminal A, luminal B, HER2-enriched, basal-like and normal breast-like [121]. Together the luminal and basal-like subtypes account for ∼80% of all incidences of breast cancer. The vast majority being of the luminal subtypes (∼60%), which originate from the inner luminal epithelial cell layer, are positive for both hormone receptors and are separated into an A and B group based on the status of HER2. Generally tumors of the luminal subtypes are associated with better clinical outcomes, due in part

to the effectiveness of hormone therapies. Whereas for the ∼20% of basal-like tumors, which are negative for both hormone receptors and HER2, do not respond to hormone therapies and are generally associated with poor outcomes. It is thought that these tumors are derived from a luminal progenitor cell population [163, 164]. Currently there is not a clinical consensus for defining or diagnosing the basal-like subtype. Instead it has principally been defined based on the triple negative histology and gene expression data.

The composition and activity of subtype-selective regulatory pathways operate, in part, through binding of transcription factors to sets of regulatory elements throughout the genome to govern levels of gene expression. One of the characteristics of these binding events is the displacement of nucleosomes, resulting in an open chromatin region. Identification of open chromatin regions has been one of the most accurate and robust methods to identify functional promoters, enhancers, silencers, insulators, and locus control regions in mammalian cells. Here we have applied FAIRE to luminal and basal-like breast cancer cells and tumors to identify the genome-wide set of active regulatory elements. FAIRE offers several advantages for advancing of our understanding of the molecular composition of breast cancer subtypes. First, FAIRE is amenable to clinical applications because it is relatively easy to perform, requires only the administration of formaldehyde and works with a limited amount of tissue samples (∼20 mg). Second, identification of the set of regulatory elements provides a means for understanding the mechanisms driving expression of genes in a given subtype, which will aid the refinement of subtype classification and evaluation of clinical outcomes. Finally, identification of the set of regulatory motifs that differentiate subtypes provides a functional means for the identification of the relevant transcription factors and the set of putative targets.

The primary goals of the work presented here were to determine the extent to which the application of FAIRE to breast cancer samples was capable of distinguishing breast

cancer subtypes, recapitulated known regulatory mechanisms and could work effectively in clinical tumor samples. We also sought to characterize the regulatory information encoded at FAIRE sites and assess whether FAIRE provided any additional information regarding the molecular characteristics that distinguish breast cancer subtypes.

5.3 Materials and methods

5.3.1 Cell culture

All cells were grown at 37°C and 5% CO_2. The MCF7 cells were maintained in RPMI-1640 plus 10% FBS and SUM102 cells were cultured in HuMEC media with supplements (Gibco).

5.3.2 Expression analysis

For all expression data mRNA was collected using the Micro-FastTrack2.0 mRNA Isolation Kit (Invitrogen). Expresion data from the cell lines were measured using RNA-seq, while the tumor expression data was measured using Agilent microarrays where all samples were hybridized over a common reference.

For the RNA-seq data sample preparation was carried out as per the recommended protocol from Illumina. The set of raw sequencing reads were aligned to UCSC hg18 build of the human genome [45] using bowtie [90], where each aligned position was required to be completely unique. RPKM (Reads Per Kilobase of exon model per Million mapped reads) [111] values were then calculated for all isoforms of the set of genes from RefSeq [124]. The isoform with the maximal RPKM value was then recorded as the expression value for that gene. For the cells the set of differentially expressed genes were identified by calculating the standardized $\log_2$ ratio of RPKM values between samples and selecting those genes with a score greater than 1.5.

The expression data for the tumors was normalized by calculating the standardized $\log_2$ ratio between the tumor mRNA versus a common reference samples. For each gene in the tumor samples the mean and standard deviation of expression values was calculated within each subtype grouping. For each gene a difference score was calculated by subtracting the mean expression values for each subtype and dividing by the sum of the standard deviations. The set of differentially expressed genes were identified as those that met a false discovery rate of 0.05 based on an iterative permutation of samples amongst the subtypes to calculate a null distribution [130].

5.3.3 Analysis of sequencing data

The raw reads for each FAIRE-seq sample were aligned to the UCSC hg18 build of the human genome [45] using bowtie [90]. Each aligned position was allowed to occur up to four times throughout the genome, where one was selected at random for each those positions that were not unique. The set of enriched regions were then identified by ZINBA [129], using 500 bp windows with 125 bp offsets. The background and enriched components were modeled using G/C content and an interaction term between mappability and local background estimate. No peak refinement was included in this analysis. Overlap between datasets and with genomic features were carried using a suite of tools called BEDTools [126].

5.3.4 Genomic clustering analysis

The set of active regulatory elements were divided into three groups based on whether they only occurred in the luminal, basal-like or both subtypes. Each gene was categorized as either up-regulated in luminal, up-regulated in basal-like and for the RNA-seq data from the cell lines as expressed in both subtypes. Those genes in the cell lines with a RPKM value less 1 in both subtypes were excluded from the analysis as not

expressed. For each active regulatory element in a given FAIRE group the distance was calculated to nearest gene in each of the expression group. To determine what would be expected by chance the set same analysis was carried out by iteratively selecting the same number of genes for each expression group at random throughout the genome. Both the observed and the set random distributions were ranked in ascending order. For each distance the degree of enrichment was calculated as the number of standard deviation the observed distance was from the mean iterative background measure.

5.3.5 Copy number variation analysis

Copy number variation was estimated using large sliding windows (50 kb) with 10 kb steps across the genome. The value of each 50 kb window was calculated as the median value for all 500 bp windows falling within the larger window. The value of each 500 bp window was the number of reads per mappable base within the window, excluding windows where <25% of the bases were mappable. The set of amplified and deleted regions were determined using cnv-seq [170], where the mappability of the 50 kb windows were used as the reference.

5.3.6 Identification of subtype-selective sites in tumor samples

To identify the set of subtype-selective active regulatory elements from the tumor samples the union set of all peaks called from the FAIRE-seq samples was collapsed. The count of the number of reads for each tumor samples was calculated and normalized based on the total number sequencing sequencing tags from the given sample. The mean and standard deviation of normalized read counts for each subtype were calculated. The set of subtype-selective active regulatory elements were identified using a difference score (as described for expression data).

5.3.7 Enrichment of transcription factor binding motifs

The set of sequence motifs enriched for each subtype were identified using HOMER [63]. Here the set of active regulatory elements unique to subtype for either the cell lines or tumor samples were compared and the set of motifs common to both were considered the set of enriched motifs.

5.4 Results

5.4.1 Breast cancer subtypes can be distinguished based on the genome-wide set of regulatory elements identified by FAIRE

Initially the MCF7 and SUM102 cell lines, which are representative of the luminal A and basal-like breast cancer subtypes respectively, were used to evaluate the suitability of FAIRE for the detection of active regulatory elements between breast cancer subtypes. For both cell lines FAIRE was performed on three replicates followed by high-throughput sequencing using the Illumina GAII platform. There was a fairly high degree of concordance ($r >= 0.87$) between replicates for each cell line (Figure 5.1).

Comparison of the set of regulatory elements identified by ZINBA between MCF7 and SUM102 revealed that ∼60% of the sites identified by FAIRE were unique to each subtype (Figure 5.2A), which is slightly lower than what has been observed for two cell lines from independent lineages (see Figure 4.3A). The set of regulatory elements held in common between the subtypes were highly enriched within promoter regions (+/- 2kb of TSS) (Figure 5.2B). Whereas those that were unique to each subtype largely localized to introns and intergenic regions of the genome (Figure 5.2B). Together this suggests that the status of chromatin at promoters is less of an indicator of the degree of

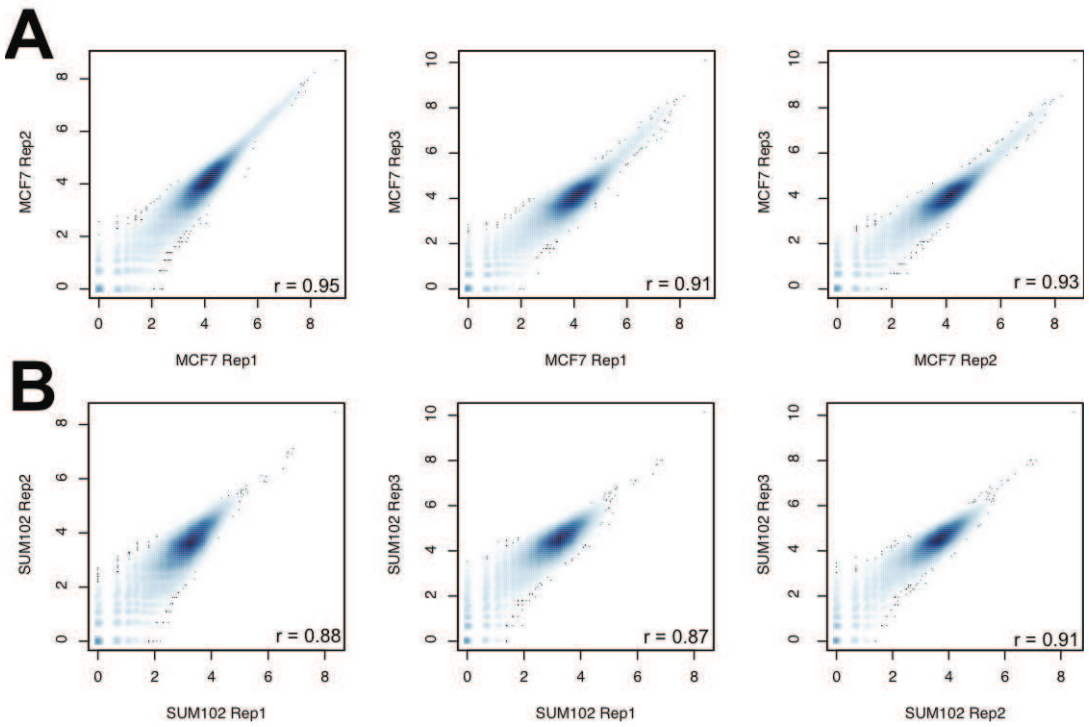

Figure 5.1: Correlation between replicates for FAIRE performed in the MCF7 (A) and SUM102 (B) breast cancer cell lines was carried out by counting the number of aligned sequencing tags falling within 10 kb windows spanning the genome. The values in each window were scaled by taking the natural logarithm of counts and the pearson correlation coefficient was calculated. Overall there was very high correlation between (r >= 0.87) indicating that the data is reproducible.

gene expression, instead it is largely driven through the binding of transcription factors to regulatory element at sites distal to the promoter.

5.4.2 Regulatory elements unique to each subtype cluster around both differentially and comparably expressed genes

To determine the extent to which the set of FAIRE sites unique to each subtype co-localize throughout the genome with the respective set of differentially expressed genes we constructed a clustering metric. Here the set of FAIRE sites were divided into groups based on being identified in MCF7-only, SUM102-only or common to both

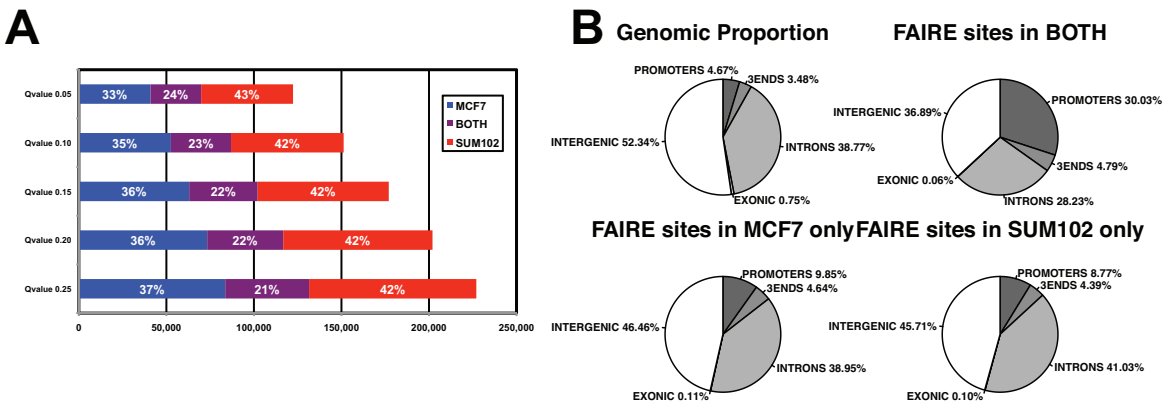

Figure 5.2: The set of enriched FAIRE sites in MCF7 and SUM102 cells were identified using ZINBA. (A) A series of peaks calls were performed using a range of thresholds (0.05-0.25) in each of the two cell types. Overlap between the set of calls for each cell type were compared at each threshold. In general the degree of overlap was consistent across the range of thresholds. (B) The set of peak calls at each threshold for those found only in MCF7, only in SUM102 and in both cell types were compared to a set of genomic features. The following of set of genomic features included promoters (+/- 2 kb of TSS), 3′ ends (+/- 2 kb of transcription stop), introns, exons and intergenic regions. The genomic background is shown in the first panel for reference.

cell-types. Genes were divided into groups based on expression patterns determined using RNA-seq data, which included up-regulated in MCF7, up-regulated in SUM102 and comparably expressed in both cell-types (genes not expressed in both subtypes were excluded). Then the distance between each FAIRE site and the closest gene in each group was recorded. The set of distances for each gene-FAIRE group comparison was evaluated with respect to an iterative random sampling of all genes to calculate enrichment. Here the set of FAIRE sites unique to each subtype were found to be enriched around the set of genes up-regulated in the respective subtype, while the set of FAIRE sites unique to the opposing cell type were found to be depleted (Figure 5.2, left and right panel). Surprisingly we also found that the set of FAIRE sites unique to each subtype were clustered around the set of comparably expressed genes (Figure 5.2, middle panel). To rule out the possibility that this was the result of using thresholds that were too conservative in calling genes differentially expression and/or correctly

identifying FAIRE sites as being present in both cell types we repeated the analysis using a more liberal set of calls for each. However, even when using a more liberal set of calls the relationship we originally observed remained (Figure 5.4). We also hypothesized that this could be the result of the comparably expressed genes being in close proximity with the set of differentially expressed genes and the resulting FAIRE sites would appear to be clustered. However, 80% of the FAIRE sites unique to each subtype were only within 50 kb of a comparably expressed gene and were not nearby a differentially expressed gene. In addition evaluation of the clustering of genes within the aforementioned groups throughout the genome were found to be more closely associated within- than between groups (Figure 5.5)

Examples for each of these gene-FAIRE groups can be found throughout the genome. The XBP1 gene locus, which is more highly expressed in MCF7 cells, there are a set of MCF7-only FAIRE sites ∼30 kb upstream of the promoter (Figure 5.6). These sites coincide with ER and FOXA1 binding, which is consistent with the estrogen responsive nature of this gene [145]. Whereas the FAIRE sites from SUM102 cells, where ER is absent, were localized to the promoter region. However for the KRT5 and KRT6 genes, which are highly expressed in SUM102 cells, there were SUM102-only FAIRE sites up to ∼5 kb upstream of the promoters (Figure 5.7). These regions have previously been reported to serve as the sites of epithelial-specific regulation of these genes, which includes motifs for the Sp1, AP-2 and AP-1 transcription factors [116]. Finally, at the SLC22A5 gene, which is comparably expressed in both cell types there are clear SUM102-only FAIRE sites ∼10 kb upstream of the gene and MCF7-only FAIRE sites in the introns of the gene that coincide with ER binding (Figure 5.8). At the 5′-end of the gene there is a modest common FAIRE site, which likely serves as the promoter element. Together this information suggests that differential expression alone does

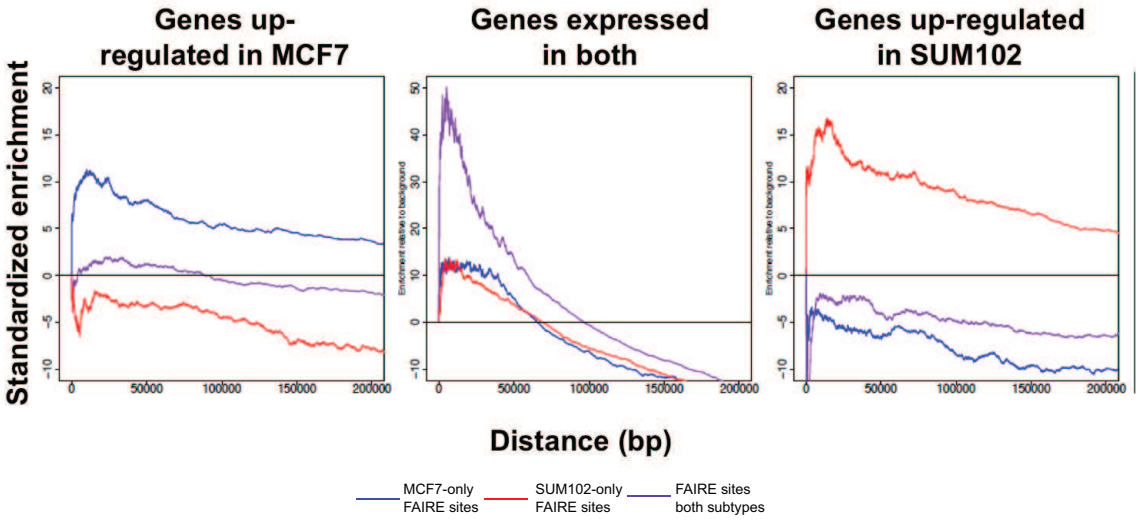

Figure 5.3: Here we constructed a metric to asses the extent to which the set of FAIRE sites found in each of the cell types were localized around the set of expressed genes. For each FAIRE site we calculated the distance to the nearest gene in each of the following groups, including up-regulated in MCF7, up-regulated in SUM102 and expressed in both cell types (excluding genes not expressed). For each of the gene groups the set of distances for the FAIRE sites found only in MCF7 cells, only in SUM102 or found in both cell types the set of distances were ranked in descending order. Enrichment was calculated by repeating the calculation using an iterative random resampling of all genes throughout the genome for each gene group and is represented as the number of standard deviations from the mean. The set of FAIRE sites unique to each cell type were enriched around the set of genes up-regulated in the respective cell types and depleted around those up-regulated in the opposing cell type. For the set of genes expressed in both cell types the set of FAIRE sites found in both were also highly enriched, but so were FAIRE sites unique to each cell type.

comprise the full molecular complexity of each cell type. Therefore accurate delineation of the molecular identity of samples will also rely understanding the regulatory mechanisms driving gene expression.

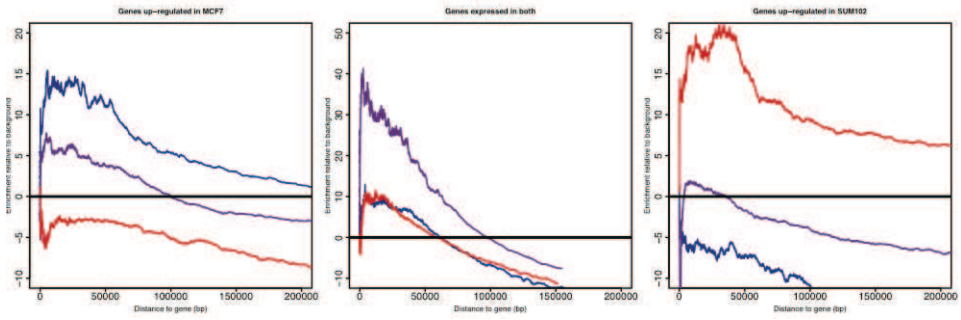

Figure 5.4: To determine whether the observed relationship between gene groups and FAIRE sites at comparably expressed genes was the result of thresholds that were too conservative the analysis was repeated using more lenient thresholds for differential expression and FAIRE sites. For the set of genes the threshold for differential expression was decreased from 1.5 fold to 1 fold (standardized $\log_2$ ratios). FAIRE sites called as unique to each cell type were reassigned to the both group if the site in the opposing cell type met a more lenient threshold (qvalue < 0.25). Together these liberal threshold leave only those genes with very similar levels gene expression and those FAIRE sites distinct to each cell type.

5.4.3 FAIRE is capable of detecting genomic copy number variations

In addition to detection of active regulatory elements, FAIRE is able to detect large-scale changes in the genomic content, including amplifications and deletions. This is due to the fact that although there is enrichment at active regulatory elements, signal is also generated within the intervening background regions. Where large-scale increases and decreases in the background signal correspond to amplifications and deletions, respectively. For example there are several amplifications on chr17 in MCF7 cells (Figure 5.9A and B). The FAIRE signal (Figure 5.9A) roughly approximates what is seen from high-throughput sequencing of genomic DNA from the same cells (Figure 5.9B). In general, the signal from FAIRE closely approximates that which is seen with genomic DNA genome-wide (Figure 5.10). However, we still retain the ability to detect regulatory elements within amplified regions using FAIRE (Figure 5.9C). Therefore FAIRE

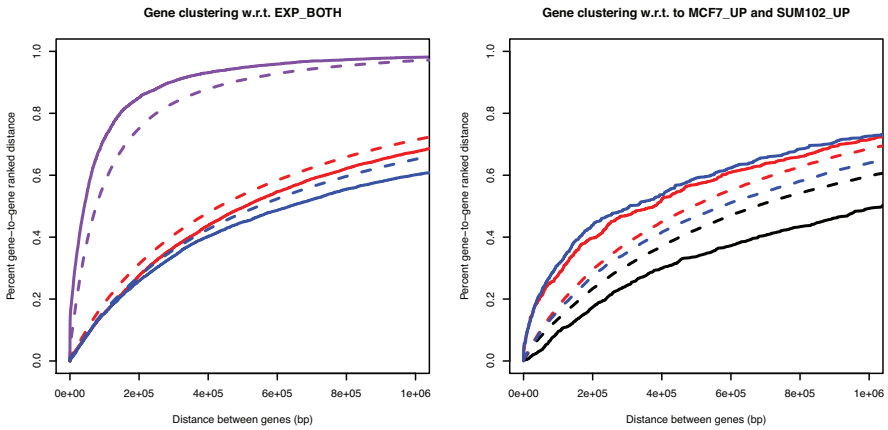

Figure 5.5: To determine whether clustering of subtype-selective FAIRE sites around comparably expressed genes was the result of their proximity to differentially expressed genes we calculated the distances for both within and between genes of each group. In the left panel the set of genes expressed in both cell types (excluding genes not expressed) were compared to each other (solid purple line) and to the genes up-regulated in MCF7 (solid blue line) and SUM102 (solid red line). For reference a similar iterative resampling of genes was carried out to determine the proximity that would be expected at random (dashed lines). Genes expressed in both cell types were found to be colocalized beyond what you would expect by chance (solid line is higher than dashed line). While those up-regulated in either cell type were found to be depleted (dashed line is higher than solid line). In the right panel, the set of gene up-regulated in either MCF7 (blue) or SUM102 (red) were found to be enriched within each group, but depleted between the two groups (black line). Together this suggests that genes with similar patterns of expression between the cell types are colocalized throughout the genome.

is capable of providing an added type of data that establishes the genomic context in which the set of active regulatory elements operate and serves as an another molecular characteristic that can be employed in the classification of samples.

5.4.4 FAIRE is capable of detecting subtype-specific active regulatory elements in clinical tumor samples

Given that FAIRE was able to detected relevant differences in regulatory elements between the luminal and basal-like subtypes using cell lines, we next performed FAIRE

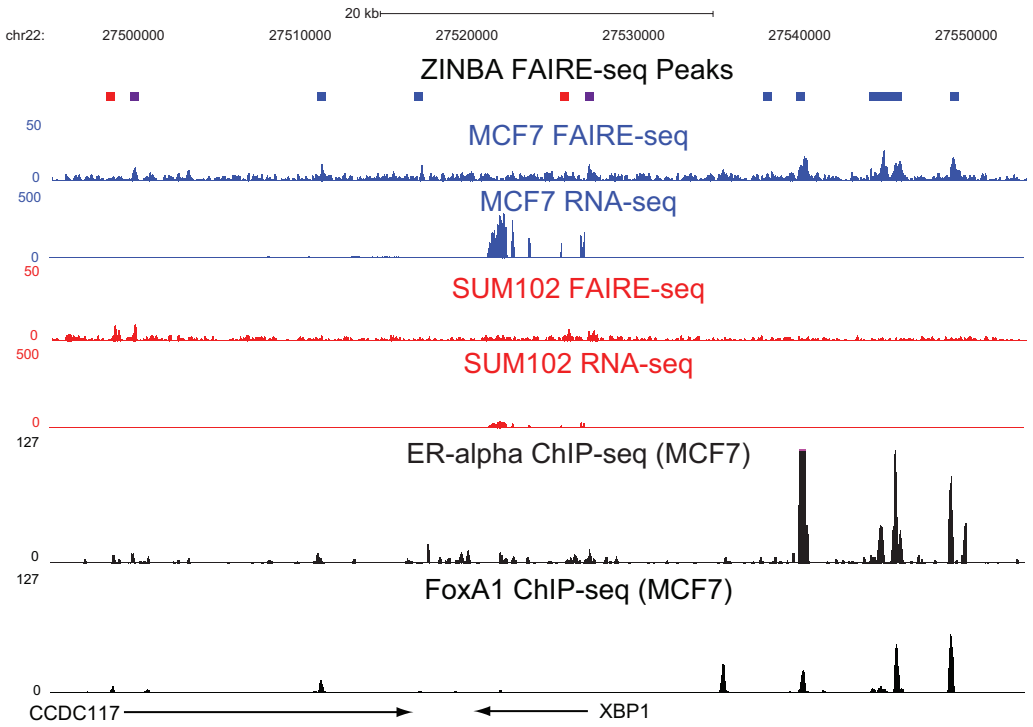

Figure 5.6: Shown is a 60 kb region on chromosome 22 containing the XBP1 gene, which is up-regulated in MCF7 cells. All tracks show the number of extended reads from the given assay overlapping each base pair. Peaks called by ZINBA are shown as the top track and colored based on being detected only in MCF7 (blue), only in SUM102 (red) or detected in both (purple). Genes are indicated as arrows, which point in the direction of transcription.

on a set of clinical breast tumors. FAIRE was performed on 10 luminal (A and B) and 8 basal-like tumors. The FAIRE-seq data from each of the tumor samples was analyzed using ZINBA and the union set of peaks was combined for subsequent analyses. First we identified the set of active regulatory elements that distinguished the luminal and basal-like subtypes. In general there were more sites identified in common between subtypes than what was found with the cell lines. However the set of subtype-selective active regulatory elements did reflect functionally relevant differences, such as enrichment of ER binding sites in the regulatory elements of luminal tumors (Figure 5.11A). Whereas for the basal-like active regulatory elements were enriched for the AP-1 motif. We also

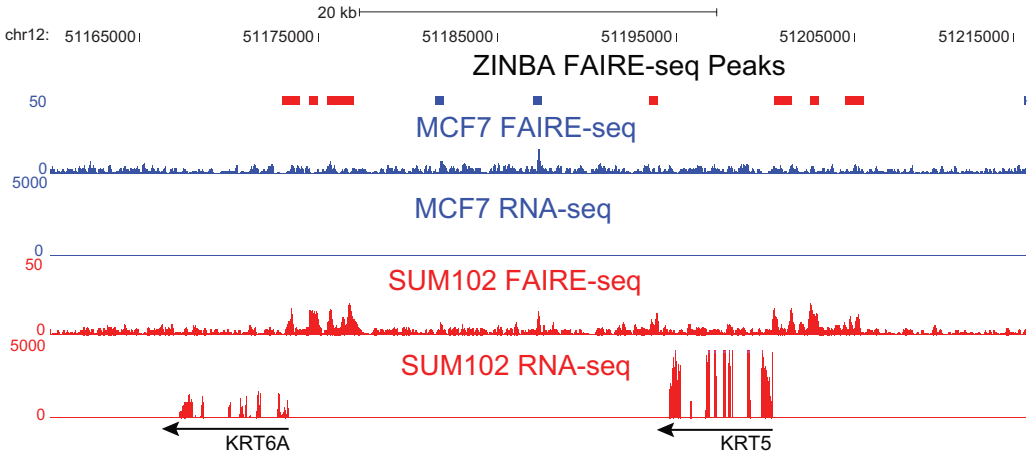

Figure 5.7: Shown is a 60 kb region on chromosome 12 containing the KRT5 and KRT6A genes, which are up-regulated in SUM02 cells. All tracks show the number of extended reads from the given assay overlapping each base pair. Peaks called by ZINBA are shown as the top track and colored based on being detected only in MCF7 (blue), only in SUM102 (red) or detected in both (purple). Genes are indicated as arrows, which point in the direction of transcription.

found the subtype-selective regulatory elements clustered around the genes up-regulated in the respective subtypes (Figure 5.11B). Together these results indicate that FAIRE will be a powerful tool for discovery of the molecular characteristics underlying cancer and that FAIRE holds promise as a clinical diagnostic tool.

5.4.5 Discovery of transcription factor binding sites within subtype-selective FAIRE sites

Finally, we determined the set of transcription factor binding motifs that distinguished breast cancer subtypes using HOMER [63]. We identified the set of motifs enriched between the luminal and basal-like subtypes. Here we compared the subtype-selective FAIRE sites from the cell lines and the tumors independently. Many of the motifs that were identified in a given subtype were consistent between the tumors and cell lines. The set of motifs that were identified in both are reported in Figure 5.12.

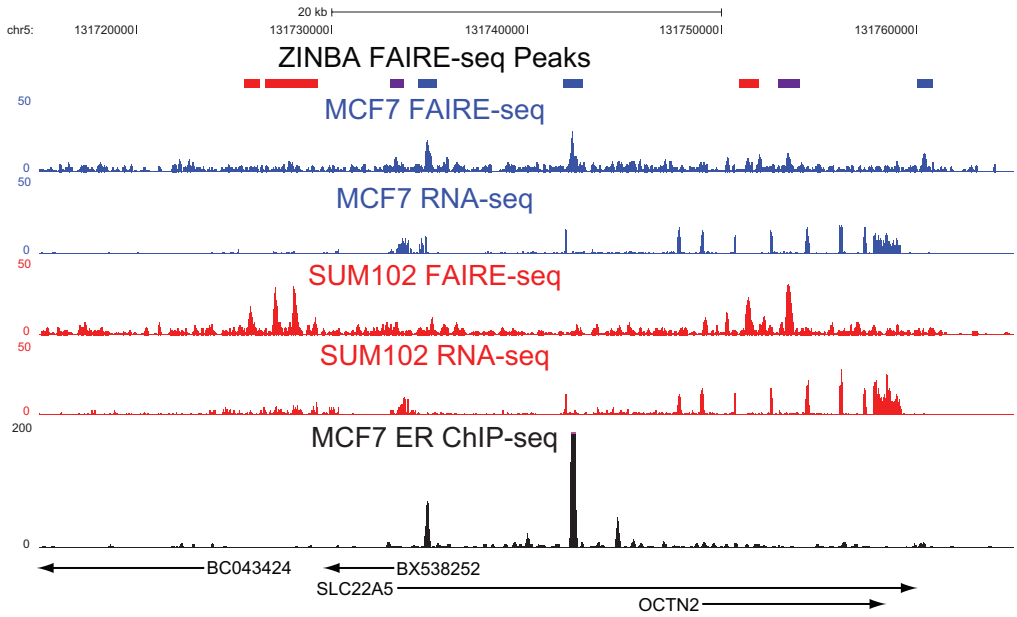

Figure 5.8: Shown is a 60 kb region on chromosome 5 containing the SLC22A5 gene, which is expressed in both cell types at a comparable level. All tracks show the number of extended reads from the given assay overlapping each base pair. Peaks called by ZINBA are shown as the top track and colored based on being detected only in MCF7 (blue), only in SUM102 (red) or detected in both (purple). Genes are indicated as arrows, which point in the direction of transcription.

The set of motifs identified in the luminal subtype samples included several factors known to play a role in the regulation of genes in the luminal cells and are important for estrogen-dependent regulation of gene expression, including the estrogen responsive element (ERE), FOXA1, GATA, E2F, NRF1, EGR1, Sp1 (KLF7), ZBTB3 and CTCF (Figure 5.12A). Several of these motifs serve as the binding sites for transcription factors included within the gene signature for the luminal subtype [121] and are known to be active in luminal breast cancer samples [28, 89, 104, 114].

Whereas for the basal-like FAIRE sites the set of motifs included AP-1, NFAT(RHD), CEBP(bZIP), AARE(HLH), RUNX1(Runt), STAT6, Pax3, Pbx1, HoxA9, and ETS 1/2 . Given that the basal-like is principally defined by its triple negative status there is little known about the status or activity of transcription factors. For the AP-1 complex

dimerization between cJun/Fra1 has been identified in the basal-like subtype [5] and Fra1 expression has been associated with high grade estrogen-receptor negative breast tumors by immunohistochemistry [100]. Although NFAT has not been directly implicated in basal-like breast cancer it does play a role in angiogenesis and is the target of the immunosuppressive drug, tacrolimus, being explored as a treatment of breast cancer [146]. ETS1 was recently identified to regulate alphaB-crystallin, a pro-survival factor overexpressed in basal-like tumors [17]. This information will hopefully serve to guide subsequent work to identify the set factors that binds to these sites and even possibility establish an affirmative marker of the basal-like subtype.

5.5 Discussion

Genome-wide maps of active regulatory elements give us a better understanding of how the availability of sequence-based regulatory elements are coordinated with the regulation of factors that utilize them across breast cancer subtypes. Here the set of active regulatory elements discovered between the luminal and basal-like subtypes reflect the established differences in the presence and activity of regulatory factors. However the established set of regulatory factors is far from exhaustive and the set of regulatory motifs encoded within subtype-selective sites offers the ability to identify additional candidate regulatory factors. These findings also underscore the importance of understanding not only which genes are differentially expressed, but what are the regulatory events governing its expression. This information will be particularly important for reconciling differences in clinical outcomes for seemingly identical tumors based on existing molecular characteristics. The ability to detect CNVs using FAIRE offers the ability to capture two sets of molecular data in a single assay, similar to how SNP genotyping arrays are used to estimate CNVs. Along the same line, given the FAIRE-seq samples are directly read, as sequencing efficiency continues to increase and cost come down it

will soon be possible to affordably obtain sufficient coverage for reliable genome-wide determination of genotypes too. Detection of CNVs also offers the ability to understand the genomic context with which genes and regulatory elements operate. These findings support the utility of FAIRE as a clinical diagnostic for the genome-wide detection of functional *in vivo* regulatory elements.

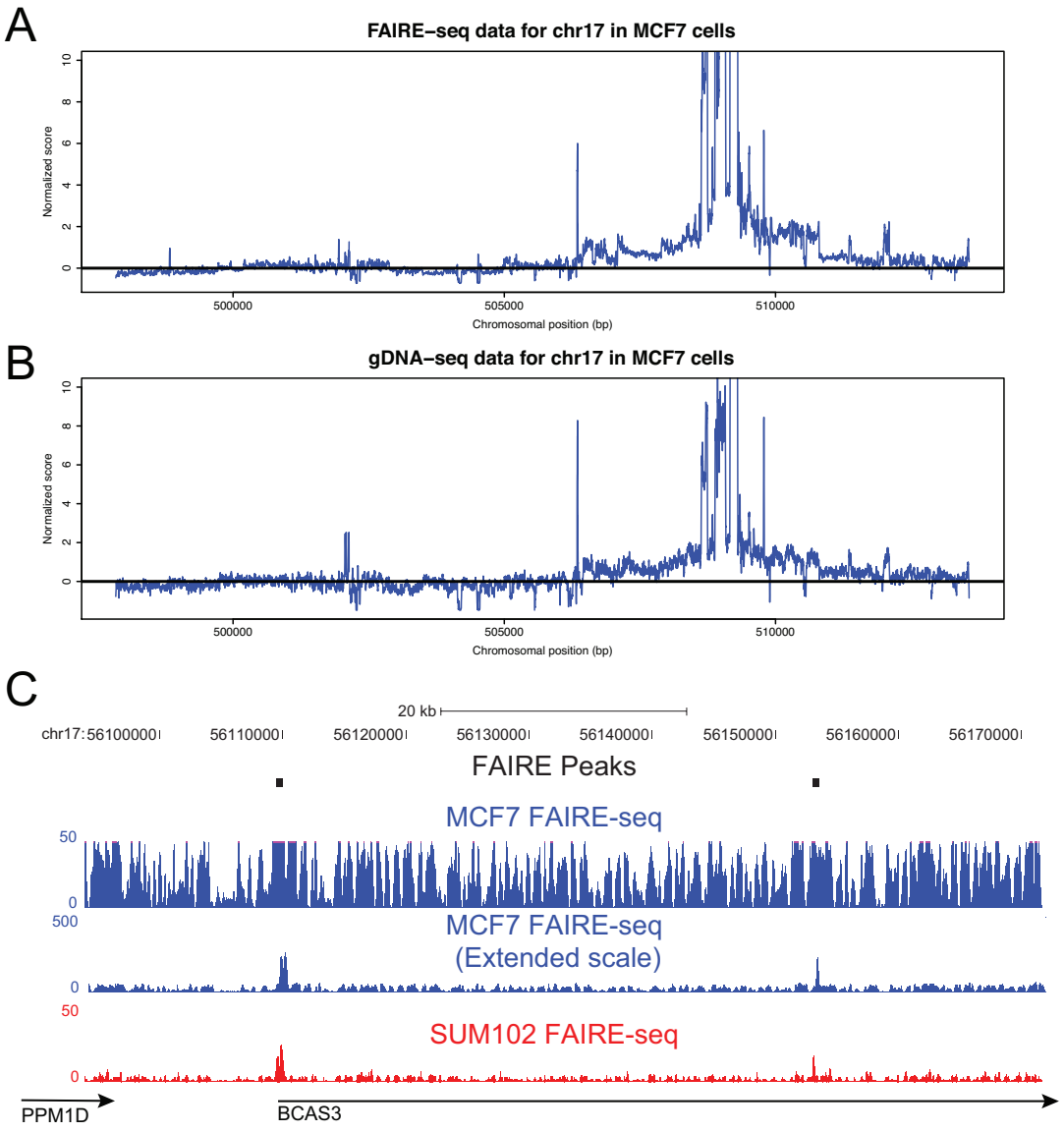

Figure 5.9: FAIRE-seq (A) and genomic DNA-seq (gDNA-seq, B) data is shown for all of chromosome 17. The data is represented as the median 500 bp window within 50 kb windows sliding across the chromosomes. The value of each 500 bp window is calculated as the number of reads per alignable base. Data from all 50 kb windows throughout the genome were normalized by median centered. (C) Here is an example from within the large amplified region on chromosome 17. The top track shows the set of peaks called by ZINBA. For the first MCF7 track the scale of the y-axis is 0-50, which is normally appropriate for viewing the FAIRE-seq data in these cells, but is saturated due to a genomic amplification. In the second MCF7 track the y-axis has been extended to 0-500. It is now possible to see the set of peaks, which are also present in the SUM102 data.

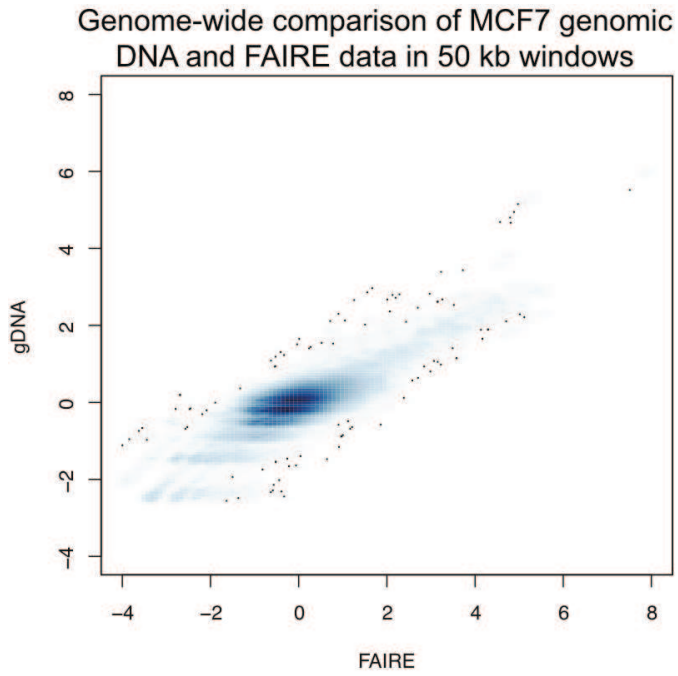

Figure 5.10: Here the genome-wide set of 50 kb windows from the FAIRE-seq (x-axis) and gDNA-seq (y-axis) are plotted. Blue indicates the density of points, with darker indicating a higher density and white being none.

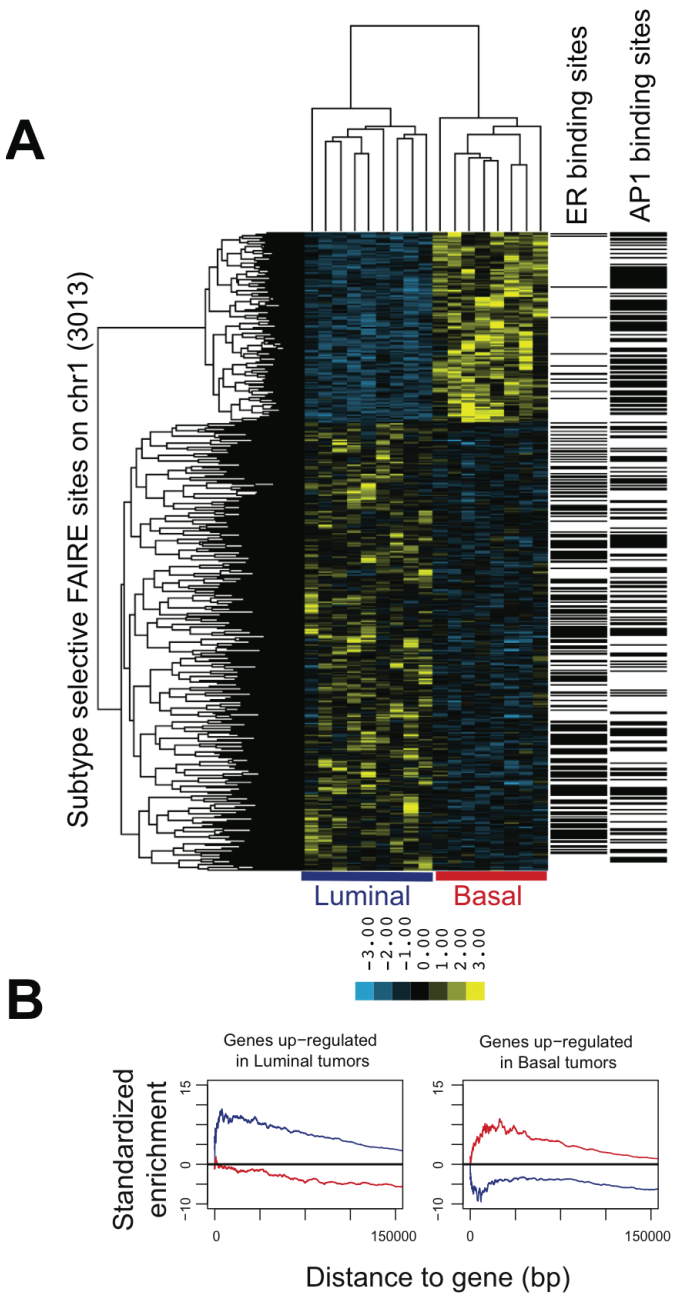

Figure 5.11: FAIRE-seq data from luminal and basal-like breast tumors was analyzed using ZINBA to identify a union set of enriched regions. Those regions that best distinguished the two subtypes were then selected based on having the greatest difference in mean values and the lowest within group variance. Shown in (A) are the 3013 subtype-selective FAIRE sites (rows) from chromosome 1 that were identified in the 18 tumor samples (columns). For comparison these sites were compared to estrogen receptor (ER) binding sites from MCF7 cells and the occurrence of the AP-1 motif. Black bars indicate an overlap between the subtype-selective FAIRE sites and the feature. In (B), the set of subtype-selective FAIRE sites for luminal (blue line) and basal-like (red line) were compared to the set of genes up-regulated in the luminal (left) and basal-like subtypes.

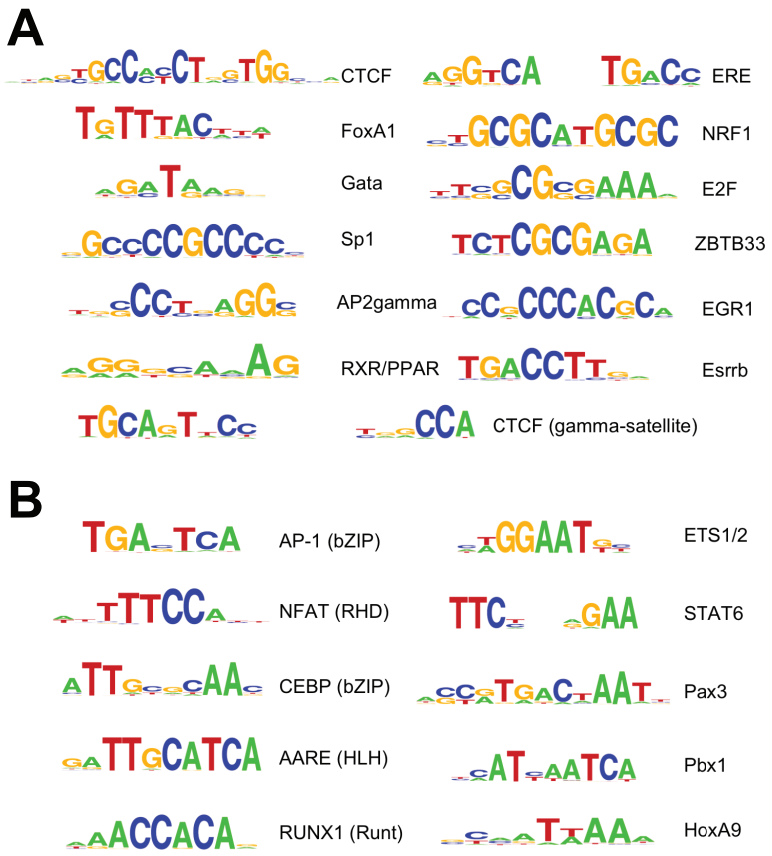

Figure 5.12: The set of FAIRE sites that distinguished luminal and basal-like subtypes were compared using HOMER to identify enrichment of known transcription factor binding motifs. The set of motifs reported for each subtype were those that were identified in both the cell lines and tumors. (A) Shows the set of motifs enriched in the luminal FAIRE sites. (B) Shows the set of motifs enriched in the basal-like FAIRE sites.

Chapter 6

Identification of regulatory elements in the transformation of mammary epithelial cells

This work was carried out in collaboration with Kevin Struhl at Harvard Medical School. All of the tissue culture for the timecourse samples and the generation of the gene expression data was performed by Heather Hirsch. Isolation and fixation of the cancer stem cells was performed by Marianne Lindahl Allen. All sequencing of the FAIRE samples was carried out at the UNC high-throughput sequencing facility.

6.1 Introduction

Carcinogenesis is the result of genetic and epigenetic alterations occurring in somatic cells leading to the errant activity of oncogenes, tumor suppressors and microRNAs [57, 33, 47]. These alterations cause reprogramming of the regulatory mechanisms controlling apoptosis and cell division resulting in the formation of a tumor. The resulting tumors are often composed of a heterogeneous population of cells, characterized

histopathologically by various degrees of differentiation, proliferation, vascularity, inflammation and invasiveness. Heterogeneity is attributed to both the emergence of clonal populations resulting from genetic instability and the presence of cancer stem cells.

Cancer stem cells are functionally defined based on their ability to seed tumors in mice, cellular plasticity and expression of markers for normal stem cells. Cancer stem cells were first implicated as the originating cell type for acute myeloid leukemia [16]. The discovered cells were capable of initiating tumor formation in nude mice, had similar cell surface markers to hematopoietic stem cells, had the capacity for self-renewal and possessed the capacity to differentiate. A population of tumor initiating cells was also isolated from solid breast tumors based on the presence of the $CD44^{+}CD24^{-/low}$ cell surface markers [2]. These too had the capacity to form tumors in mice and the resultant tumors had the same complexity of cellular phenotypes as the original breast tumor.

The origins and even existence of cancer stem cells within tumors is still a matter of debate. Do normal stem cells simply undergo oncogenic transformation or can partially or fully differentiated cells acquire stem-like properties? Given the heterogenous population of cells found in tumors what are the molecular characteristics that distinguish cancer stem cells? Are the diverse types of cells derived from a single clonal cell or acquired through the cultivation of a tumor microenvironment?

A set of recent studies [66, 70] has provided both a framework to investigate these types of questions and have provided some promising initial findings into the formation and treatment of breast cancer. Here the authors stably transfected an inducible form of the Src oncogene into MCF10A cells, which are an estrogen-receptor negative spontaneously immortalized mammary epithelial cell line. Upon induction the cells

undergo an almost immediate and irreversible transformation to a cancerous phenotype, which form mammospheres [54, 98] and are capable of seeding tumors in mice. Using both biochemical and genomics approaches the authors have identified the set of components responsible for inducing an epigenetic switch, which includes activation of the components of the inflammation pathway. In addition they identified a alteration of metabolic pathways by gene expression analysis that are held in common across different forms of cancer. Most importantly they found that after 36 hours of induction $\sim$10% of the transformed cells expressed the $CD44^{+}CD24^{-/low}$ cell surface markers of cancer stem cells. As has been reported previously these were in fact the cells responsible for seeding tumors in mice [2]. They also found them to be relatively resistant to the chemotherapeutic agents and served as the founder cells for cancer recurrence [66]. Ultimately they found that these cells could be selectively targeted and destroyed in a mouse model using a drug called metformin, which has long been used as a treatment of diabetes. The drug effectively decreases systemic glucose levels, which in this case results in the destruction of cancer stem cells when used in combination with a reduced dose of chemotherapy.

Here we have performed FAIRE throughout a timecourse of Src induction in MCF10A cells. We also performed FAIRE in the cancer stem cells isolated by flow-cytometry based at the final time point. The goal of this research was to identify and characterize the set of regulatory elements which varied throughout the transformation. We also sought to characterize the set of regulatory elements that distinguished the cancer stem cells. Initially it was thought that the MCF10A cells were a fairly homogeneous population of cells and the creation of the cancer stem cells were affected through an independent epigenetic pathway in a portion of the cell population. However, as will be described below, we found that the cancer stem cell population was in fact a separate cell population to begin with and there is evidence that the creation of the cancer stem

cell population was achieved through the differential activation of CD44 in these cells (Figure 6.1).

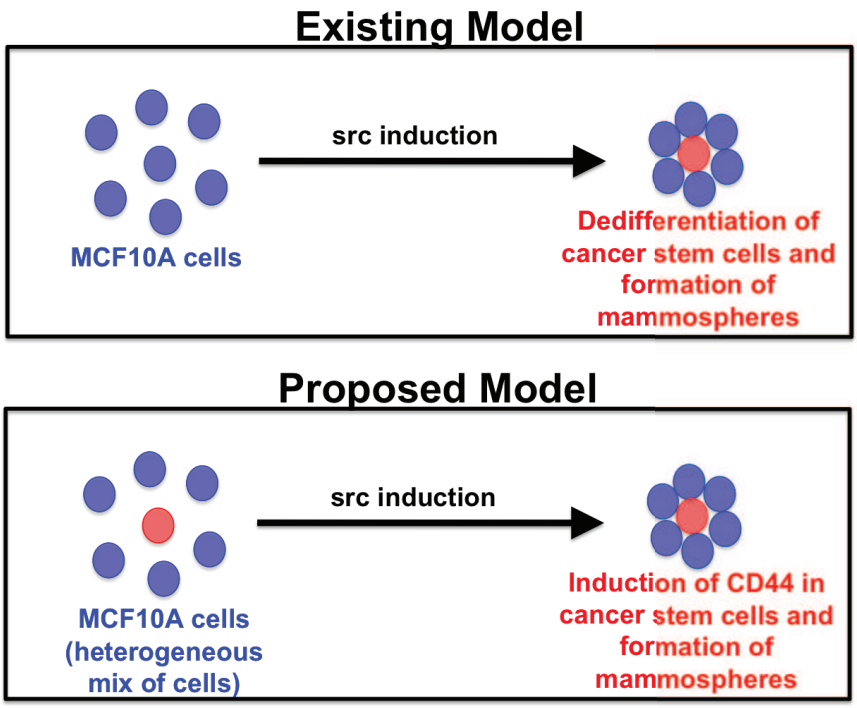

Figure 6.1: The original model proposed by the authors viewed the MCF10A cells as a homogeneous mix of cells, whereby a sub-population is reprogrammed to become cancer stem cells and form mammospheres. Whereas the data presented here supports at least two cell populations are present in the MCF10A cells, which undergo differential activation of CD44 and acquisition of the properties of cancer stem cells.

6.2 Materials and Methods

6.2.1 Cell lines

MCF10A cells are mammary epithelial cells derived from fibrocystic breast tissue that was obtained from a mastectomy of a 36-y-old woman with no family history of breast cancer and no evidence of disease [154]. Genetic analysis did not reveal any amplification of HER2/neu oncogene or mutations in H-Ras oncogenes, and these cells do not

express estrogen receptor (ER). The experiments here use a derivative of MCF10A containing an integrated fusion of the v-Src oncoprotein with the ligand-binding domain of ER.

MCF10A were cultured in a 5% CO_2 humidified incubator at 37°C in growth media (DMEM/F12 supplemented with 5% horse serum, 20 ng/ml epidermal growth factor (EGF), 10 μg/ml insulin, 0.5 μg/ml hydrocortisone, 100 ng/ml cholera toxin, and antibiotics) [36]. MCF10A were stably transfected with ER-Src using a "virus incubation cocktail", which is comprised of growth media and a viral titer capably of delivering a MOI of 3 to 5. Following 5 hours of infection cells are supplemented with additional growth media and the complete media is replaced following 18 hours of infection. After 36 to 48 hours postinfection cells are selected using a drug marker and passaged once to recover from drug treatment. The Src kinase was induced with 1 ?mol/L 4OH-tamoxifen dissolved (Sigma) in ethanol. Morphologic changes, phenotypic transformation, and foci formation occurred 24 to 36 h after tamoxifen addition, and were monitored by phase-contrast microscopy.

Cancer stem cells were isolated by flow cytometric cell sorting of transformed cell populations using single-cell suspensions. Cells were stained with CD44 antibody (FITC conjugated; 555478, BD Biosciences) and with CD24 antibody (phycoerythrin conjugated; 555428, BD Biosciences). Cancer stem cells (CD44high/CD24low) were isolated from no-stemtransformed cells (CD44low/CD24high).

6.2.2 Expression analysis

RNA was extracted from MCF10A cells using standard Trizol purification through RNeasy columns (RNeasy Clean- Up Kit, Qiagen). Biotinylated cRNA were prepared according to [52] and hybridized to Affymetrix Human U133 2.0 A expression arrays for 16 to 18 hours at 45°C, washed on fluidics station and scanned using Affymetrix Gene

Chip Scanner 3000. ComBat [75] was used to remove non-biological experimental variation or batch effects between batches of microarray experiments. All gene expression data was normalized and summarized using the RMA algorithm [73] with an updated Entrez gene probeset definition. The set of differentially expressed genes were identified by fitting a linear model to the expression data between the 0 hour and each of the time points. The set of residual values were recorded for comparison, which represents the extent to which the expression between time points deviates from equal. The residuals were standardized and the set of differentially expressed genes were identified as those with a score >2.5 for any time point, similar to [149].

6.2.3 Analysis of sequencing data

The raw reads for each FAIRE-seq sample (3 replicates at each time point) were aligned to the UCSC hg18 build of the human genome [45] using bowtie [90]. Each aligned position was allowed to occur up to four times throughout the genome, where one was selected at random for each those positions that were not unique. The set of enriched regions were then identified by ZINBA [129], using 300 bp windows with 75 bp offsets. The background and enriched components were modeled using G/C content and an interaction term between mappability and local background estimate. No peak refinement was included in this analysis. Overlap between datasets and with genomic features were carried using a suite of tools called BEDTools [126].

6.2.4 Identify differentially activated FAIRE sites

The union set of sites identified as enriched throughout the timecourse were combined and analyzed for those that there either gained or lost during the experiment. The occurrence of each active regulatory element was evaluated to determine the extent to which it differed with respect to the other time points, either present at one timepoint

and absent at the others or vice versa. The count of reads aligning to each region identified in the union set was recorded for each timepoint and normalized by the total number sequencing read for that sample. Here the normalized values for each timepoint were compared to the mean value of the remaining timepoints and divided by the standard deviation of the values for the remaining timepoints. All those regulatory elements with a difference score >2.5 were identified as being changed throughout the timecourse.

6.2.5 Motif discovery

The set of sequences that correspond to the active regulatory elements that changed throughout the timecourse were analyzed to see whether any motifs were enriched. Enriched motifs were identified using MEME [6] to compare the sequence at active regulatory elements to a set of background sequences, which were defined as the 5 kb of sequence flanking each active regulatory element. The set of derived motifs were compared to several databases of known motifs using TOMTOM [56]. Finally, the set of active regulatory elements that did not change were scanned for the presence of any of the discovered motifs using FIMO [53].

6.2.6 FAIRE sites enrichment around differentially expressed genes

For each gene identified as being differentially expressed throughout the timecourse the five nearest active regulatory elements were polled to identify their group membership, which was either identified only in the cancer stem cells, only in the timecourse or identified in both datasets. To determine what would be expected by chance the position of the active regulatory elements were iteratively shuffled throughout the genome and the analysis was repeated. Enrichment was calculated as the log ratio of the average

observed numbers of a given class versus background estimate.

6.2.7 Identification of large-scale differences in genomic content

Copy number variation was estimated using large sliding windows (50 kb) with 10 kb steps across the genome. The value of each 50 kb window was calculated as the median value for all 500 bp windows falling within the larger window. The value of each 500 bp window was the number of reads per mappable base within the window, excluding windows where <25% of the bases were mappable. The set of amplified and deleted regions were determined using cnv-seq [170], where the set of 50 kb windows from the timecourse were compared to those from the cancer stem cells directly.

6.3 Results

6.3.1 Relatively few changes in active regulatory elements detected throughout the transformation

The set of active regulatory elements identified by FAIRE throughout the transformation of the MCF10A cells revealed only a minority (∼5%) were either gained or lost (Figure 6.2A), which is consistent with the relatively few number of differentially expressed genes (∼1500 genes, Figure 6.3A). Although there were fluctuations in amplitude of FAIRE signals at many of the sites, particularly at 4 hours post-induction, which suggests changes in the activity of regulatory factors at these sites. For the set of FAIRE sites that did change throughout the transformation the majority of changes involved the gain of new regulatory elements, especially again at the 4 hour time point (Figure 6.2B).

We performed motif discovery using the minority of sites that were either gained or lost, which detected among others the AP-1, Oct4 and NFAT motifs (Figure 6.2C). The AP-1 is a heterodimeric complex composed of Jun, Fos and Maf protein families, which can form both homo- and heterodimers and depending on the cellular context can either function as a tumor suppressor or promote cellular proliferation [39]. AP-1 is activated in response to a variety of stimuli, including cytokines, growth factors and viral infection. Oct4 is a homeodomain transcription factor encoded by the POU5F1 gene and is involved in the self-renewal of embryonic stem cells [115]. While NFAT is a family of transcription factors conatining a REL homology domain (RHD) that are primarily expressed in the immune system [30], some of which are activated by calcium signaling through calmodulin activation of the phosphatase calcineurin. However these motifs were not restricted to the set of FAIRE sites that changed throughout the timecourse and were found with FAIRE sites throughout the genome (Figure 6.2D). Together these results are consistent with those found by the Stuhl lab indicating that the activation of the inflammation pathway and the presence of factors involved in stem cell function.

6.3.2 Cancer stem cells have a distinct set of open chromatin sites

FAIRE performed in the isolated cancer stem cells revealed a distinct set of regulatory elements than those found from FAIRE performed in the timecourse samples (Figure 6.4A). Although this degree of divergence is less than what has been seen for two cell types from separate lineages. Given that the FAIRE data from 36 hour time point contained these samples it would appear that we were unable to detect these regulatory elements when they are present in only $\sim$10% of the cells in a population. Next we assessed the extent to which the set of genes differentially expressed throughout the

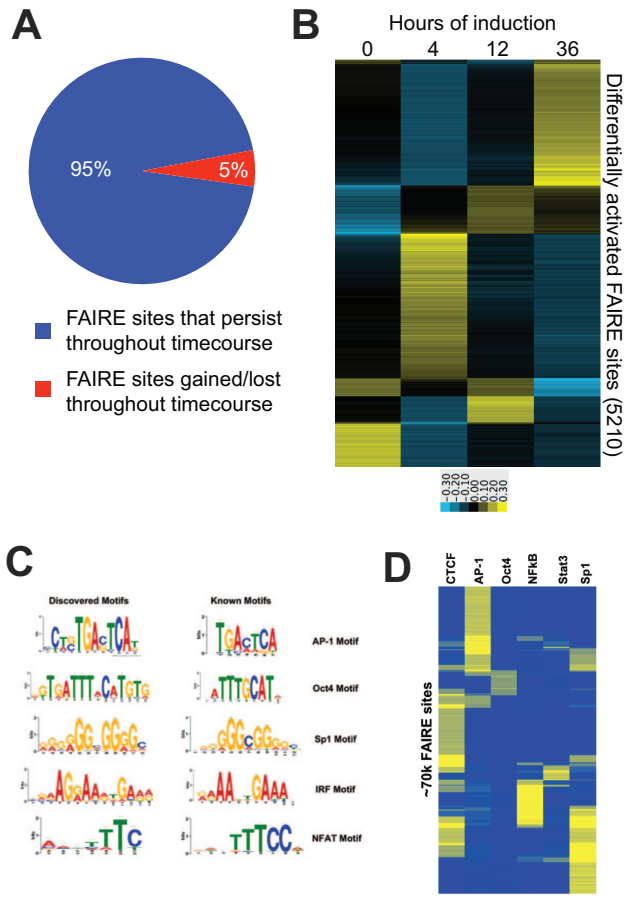

Figure 6.2: FAIRE was performed on MCF10A cells after 0, 4, 12 and 36 hours of induction. (A) ZINBA identified ∼100K FAIRE sites across all of the timepoints. Only ∼5% of these sites were found to be gained or lost throughout the timecourse. (B) Here the set of FAIRE sites gained/lost (rows) in the timecourse (time points, columns) were clustered using a self organizing map. The largest number of changes in FAIRE sites occurred at 4 and 36 hours. In general relatively few sites were lost. (C) Motif discovery for the set of FAIRE sites that changed was performed using MEME (left column) and were matched to known motifs using TOMTOM (right column). (D) The set of motifs along with CTCF, NFkB and Stat3 were mapped to the 100k FAIRE sites identified throughout the timecourse using FIMO. Approximately 70k of the FAIRE sites (rows) contained at least one motif (columns). The set of FAIRE sites were clustered, where yellow indicates the presence of the motif and blue the absence.

timecourse were enriched for FAIRE sites found only in the cancer stem cells, only in the timecourse or found in both samples. Here we found that all of the peaks from cancer stem cells were enriched around the differentially expressed genes, whereas the

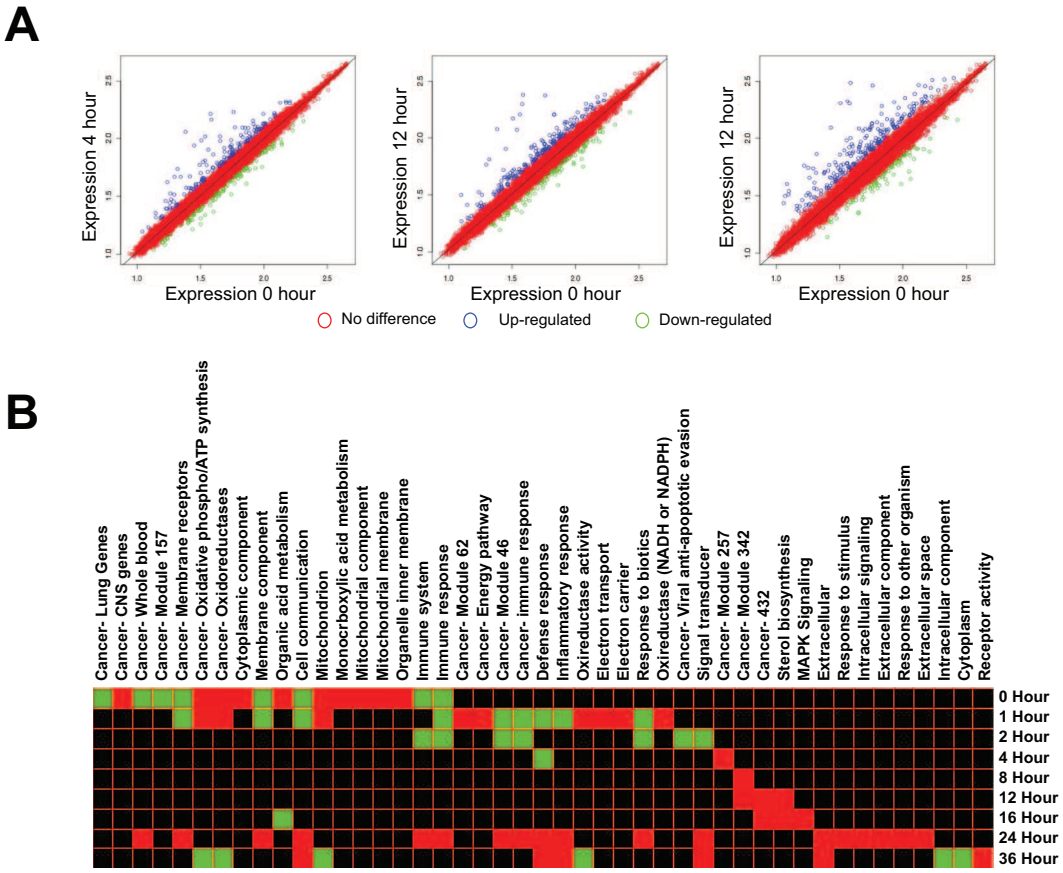

Figure 6.3: Gene expression was measured using Affymetrix arrays at 0, 1, 2, 4, 8, 12, 16, 24 and 36 hours of induction. The expression data was normalized using RMA and differentially expressed genes were identified as those that deviated from a linear model. (A) The 4, 12 and 36 hour expression data was compared to 0 hour values. Genes that were either up- or down-regulated relative to the 0 hour data are indicated as blue or green points, respectively. Red points indicate no difference in expression. (B) The set of differentially genes were analyzed to identify pathways or modules effected throughout the timecourse. The set of gene groups that were collectively up- (red) or down-regulated (green) throughout the timecourse are displayed in the columns with the time points indicated on the rows.

FAIRE sites only present in the timecourse were depleted around the differentially expressed genes (Figure 6.4B). An example of this can be seen at the IL6 gene, which was shown to be essential for the formation of mammospheres [70], there is a prominent peak from the cancer stem cells (Figure 6.4C).

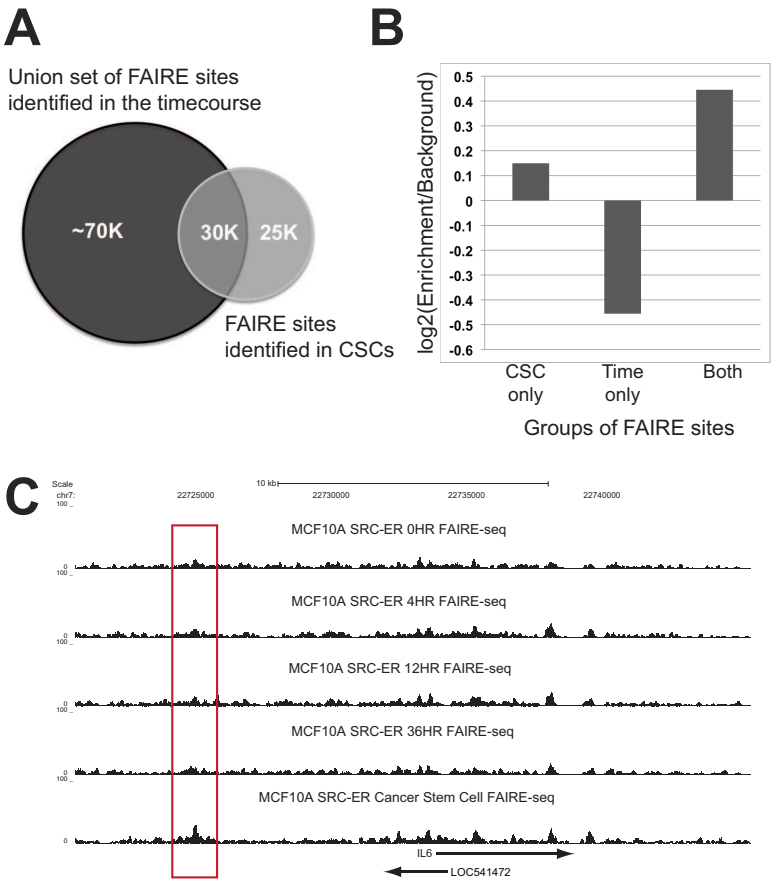

Figure 6.4: Following the transformation a sub-population (∼10%) of cells are created with characteristics of cancer stem cells. (A) Comparison of the set of FAIRE sites from the cancer stem cells with those from the timecourse revealed approximately half of the sites are unique to the cancer stem cells. (B) For each differentially gene the class of five nearest FAIRE sites were polled. the classes included FAIRE sites found only in cancer stem cell, only in the timecourse or those found in both. Enrichment was calculated by comparing the observed relationship to what would be expect by chance. (C) At the IL6 locus, which is differentially expressed and essential for he formation of mammospheres, there is a distinct FAIRE site in cancer stem cells (red box).

6.3.3 Cancer stem cells are derived from a independent cell population

Next we performed an analysis to identify any large-scale differences in genomic content between the FAIRE samples from the timecourse and cancer stem cells. We observed several regions with considerable variation in genomic content between the two cell

types that corresponded to functionally relevant differences between the two cell types. First was an amplification of the HIF1A gene on chr14 in the cancer stem cells, which appeared to be diploid in the timecourse samples (Figure 6.5A). HIF1A is in fact highly expressed in the cancer stem cells and given that these cells typically occupy the internal portion of the mammosphere, which is relatively hypoxic, this would be advantageous. Whereas a deletion was detected in the cancer stem cells on chr10 at the PTEN locus (Figure 6.5B), which is a known tumor suppressor and its loss would certainly confer a more carcinogenic state. Another region deleted in the cancer stem cells occurred on chr17 and encompassed STAT5B, STAT5A and STAT3 (Figure 6.5C). The STAT genes are transcriptional regulators that participate in a number of cellular functions, typically in response to cytokine and growth factor stimulation. They are involved in cell growth and apoptosis and are activated in breast cancer. On chr8 we detected a region in the cancer stem cells that is either a heterozygous deletion or heterochromatic region encompassing several tumor supressor genes (Figure 6.5D).

Although unlikely all of these changes could have resulted from alterations in the genomic content of the cancer stem cells following transformation. However, we also identified a deletion on chr9 in the timecourse samples that was not deleted in the cancer stem cells (Figure 6.6). This occurred at the CDKN2A gene locus, which is a cell cycle inhibitor and is frequently deleted cancers. In fact this locus is reported to be a homozygous deletion in MCF10A cells [78], which is the result of a set of translocations between chromosomes 3, 5 and 9 [106]. Given that it would not be possible for the cancer stem cells to gain this locus back during the transformation, the only way this observation is possible is if the cancer stem cells were present as a separate population of cells prior to transformation.

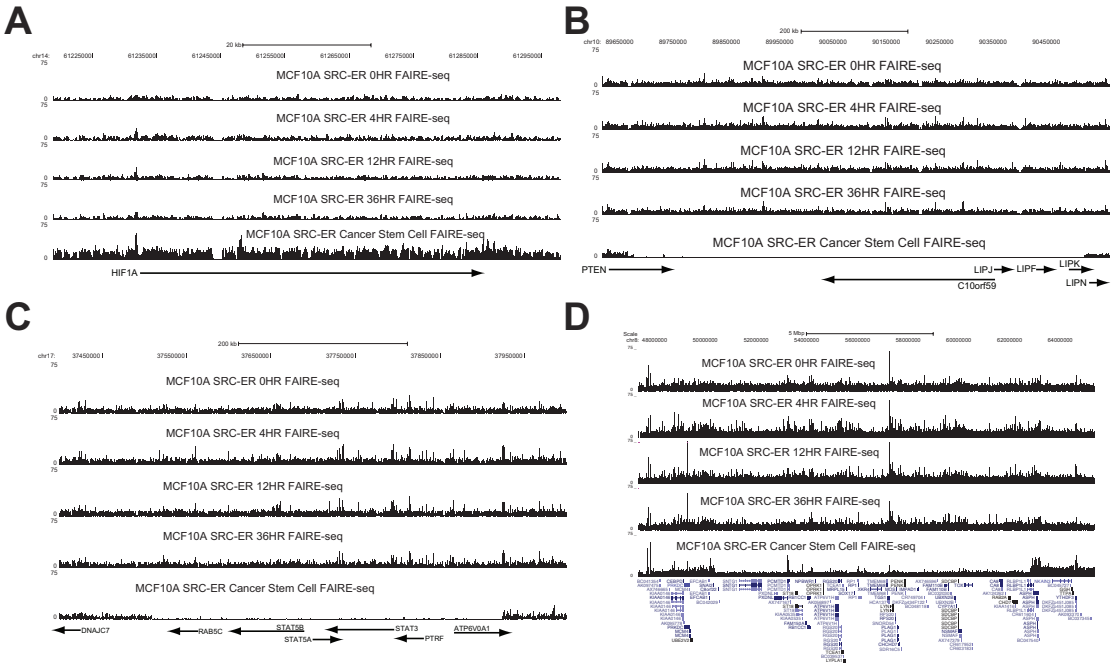

Figure 6.5: Amplified and deleted regions were identified using the count from FAIRE data in large (50 kb) sliding windows that were analyzed using cnv-seq [170]. Several amplifications and deletions were identified between the cancer stem and non-stem cells. (A) An amplified region in cancer stem cells on chr14 contained the HIF1A gene, which is overexpresed in the cancer stem cells. (B) A deletion in cancer stem cells on chr10 that included PTEN, a tumor suppressor. (C) Deletion of STAT3A, STAT5A and STAT5B genes on chr17 in cancer stem cells. (D) A region on chr8 with approximately half the FAIRE signal that surrounding regions contains several tumor suppressor genes.

6.3.4 Calcium-dependent signaling pathway drives higher expression of CD44 in cancer stem cells

CD44 is a hyaluronic acid receptor that is expressed in a variety of cell types and has been shown to play a role in cell migration, inflammation, immune response and even mammary gland development [62]. In mammary gland development it is first expressed during puberty and subsequently follow estrous cycles where it is expressed in the myoepithelium and to a lesser extent in luminal epithelial cells. During lactation CD44 is down-regulated, but is reactivated upon involution. Several signal transduction

events have been identified that lead to the induction of CD44 for a variety of cell types [48, 109], including via calmodulin/Ca^{2+} in PMA-stimulated T lymphoma cells [147] and by Egr1 in murine B cells [105]. Both the Egr1 and AP-1 motifs are present within the CD44 promoter and it has been shown that AP-1 acts through activation of CAMK2 whereas Egr1 is through JNK [109].

Prior to transformation the cancer stem cell population is not detectable using CD44 as a cell surface marker. Given that the cancer stem cells appear to be an independent population of cells present prior to transformation we wanted determine whether there was evidence for a distinct regulatory pathway enhancing expression levels of CD44 expression in the cancer stem cell population.

Several FAIRE sites were identified within 150 kb of the CD44, which exhibited distinct occurrences between cancer stem and non-stem cells (Figure 6.7). Examination of the Egr1 and JNK gene loci contained FAIRE sites predominantly from cancer non-stem cells (Figure 6.8). Whereas the calmodulin and CAMK2 gene loci, which are amplified in the cancer stem cells, contained several prominent FAIRE sites (Figure 6.9). Although some of the FAIRE sites were unique to cancer stem cells, the enhanced activation is likely the result of the amplification of these genes. Therefore, it appears that the higher level of expression of CD44 in cancer stem cells is likely due to amplification of components the calcium-dependent signaling pathway.

6.3.5 Alterations in mitochondria DNA content during the timecourse

In addition to isolating DNA from the nuclear genome, FAIRE is also capable of isolating DNA from the mitochondria. The mitochondrial genome is a short (16 kb) circular genome that exists at high copy numbers within each organelle. Regulation of the genes on the mitochondrial genome are controlled in part by nuclear encoded

transcription factors, but replication is carried out independent of mitosis. The mitochondrial genome is also not packaged into chromatin and is therefore highly enriched by FAIRE. Using the set of differentially expressed genes from the timecourse we looked to see whether they were enriched with respect to a specific cellular or biological function (Figure 6.3B) [142]. Several of the groups identified reflected what had been found in the initial studies [67, 70], including energy metabolism and inflammation response. We observed several groups related to mitochondrial function. Analysis of the FAIRE data revealed a sharp increase in the content of the mitochondrial genome at 4 hours post-induction, followed by a return at 12 hours to the levels seen for the 0 hour sample (Figure 6.10). The increase corresponds to the entire mitochondrial genome and not enrichment of a specific locus. We also observed that the mitochondrial DNA was relatively depleted in the cancer stem cell population (Figure 6.10).

These findings are potentially significant for a few reasons, first production of all the cellular components needed for uncontrolled cell division requires activation of anaerobic glycolysis (the Warburg effect) [20, 102]. Second the authors of the initial studies identified metformin as a drug that selectively targets the cancer stem cells for destruction [67, 71]. Metformin, which is used in the treatment of diabetes, depletes systemic glucose by decreasing gluconeogenesis in the liver, reducing glucose uptake in the intestine and sequestering glucose in muscles. Therefore, it likely targets the cancer stem cells by cutting off their food supply. Third, the formation of mammospheres, which cancer stem cells are the nucleus, likely creates a relatively oxygen-depleted environment, which is also reflected in the amplification and overexpression of HIF1A. The reduced mitochondrial genome content of the cancer stem cells is consistent with an increased glycolytic metabolism and greater susceptibility to the effects of metformin.

6.4 Discussion

Here we have provided the first genome-wide assessment of the genomic regulatory elements associated with the formation of breast cancer. We found that there were relatively few regulatory elements actually gained or lost. Using the set of FAIRE sites that changed were able to derived a set of transcription factor binding motifs. Interestingly these motifs did not simply occur in the set of sites that changed but were found at sites genome-wide. Some of these motifs serve as the binding sites for families, such as AP-1, which form both homo- and heterodimeric complexes to direct a wide variety of cellular functions. In the case of AP-1 depending of the composition of the heterodimer and the cellular environment can act to promote proliferation or serve as a tumor suppressor. Therefore, the events leading to the formation of cancer may not be result of widespread remodeling of chromatin, but may instead be through the aberrant activation of members of a regulatory families that use the existing regulatory information to induce transformation.

We have also identified the genome-wide set of active regulatory elements from cancer stem cells isolated from transformed MCF10A cells. Understanding the molecular events leading to the formation of these cells ultimately helps us understand how some (if not all) breast cancers are formed. Given the stark clinical outcomes attributed to the presence of cancer stem cells, including resistance to chemotherapeutic agents, metastasis and recurrence, it will also be important for effective treatment.However, several questions still remain regarding how cancer stem cells are defined and the exact properties they possess. Such as, aside from expression of the CD44 marker are there other properties of these cells that warrant the label of cancer stem cells. It has been hypothesized that the CD44 marker actually confers a general stickiness that is more amenable to adhering to the mammary fat pad in mice and would be a preferable substrate for the formation of mammospheres. Results from this work offer some molecular

characteristics that could be used in to explore the properties of tumors formed using cancer stem cells in mice. For example, following excision of the xenografted tumor in mice what is the genomic profile of the newly created cancer non-stem cells that make up the bulk of the tumor. Do these cells now have the deleted region on chr9? If so, is there evidence that they have differentiated into a new cell type other than cancer stem cell that seeded the tumor? Conversely, if CD44 were overexpressed in the 90% of cancer non-stem cells, would these now have the capacity to seed tumors in mice? Even if simply having the CD44 marker is shown to be the qualifying trait for tumor formation, this would be useful for determining candidate cell types of origin for breast cancer. Are there other compensatory functions these cells must also posses, such being able to operate in relatively hypoxic conditions?

Although the findings from this work were certainly unexpected and more questions remain unanswered than answered, it does highlight the utility of FAIRE as a tool for the investigation of breast cancer origins.

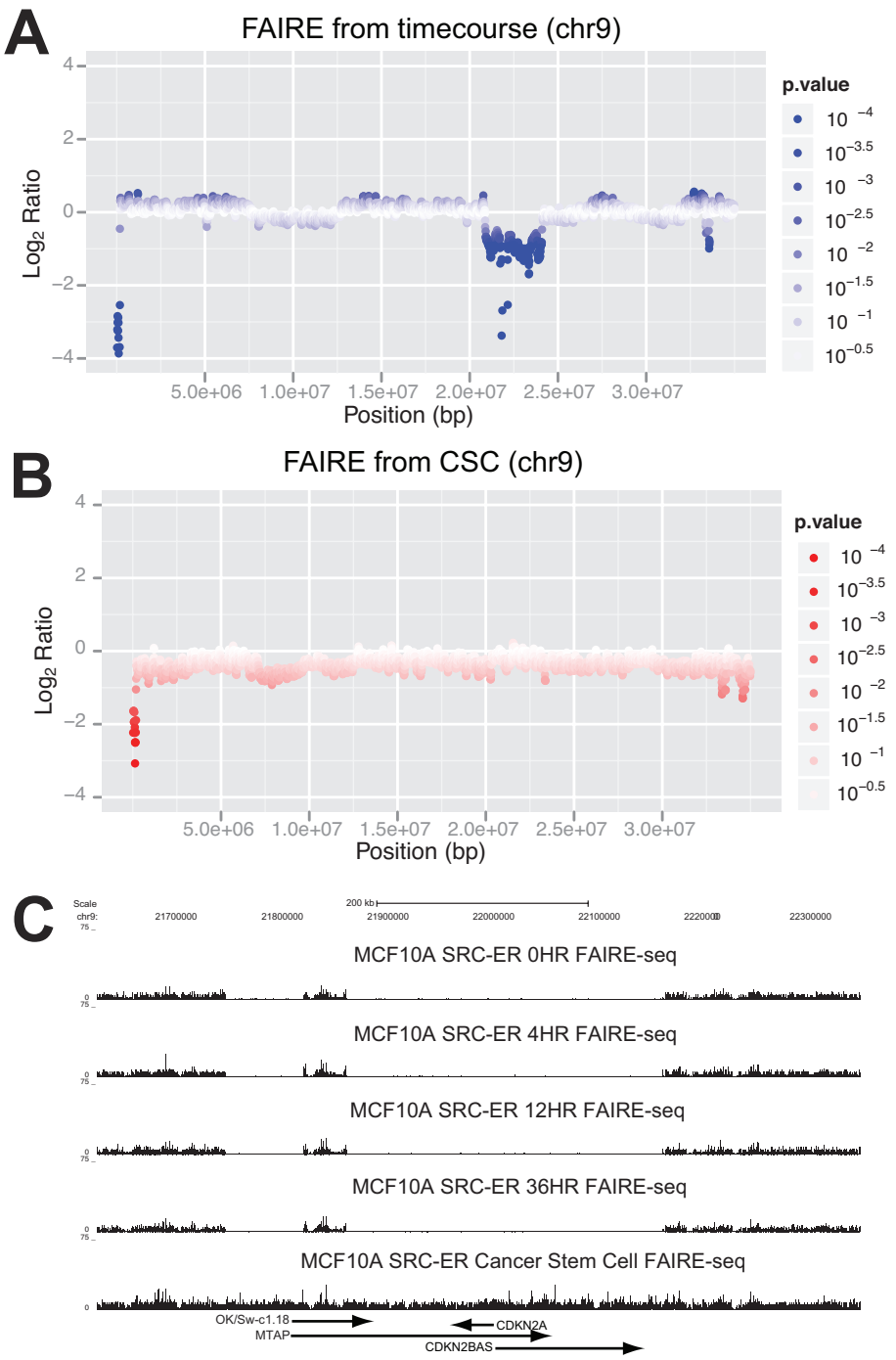

Figure 6.6: Amplified and deleted regions were identified using the count from FAIRE data in large (50 kb) sliding windows that were analyzed using cnv-seq [170]. A deletion on chr9 was identified for the timecourse samples (A), but was present in the cancer stem cells (B). (C) The region corresponds to the CDKN2A (p16) gene, which is a cell cycle inhibitor, that is reported to be a homozygous deletion in MCF10A cells [78].

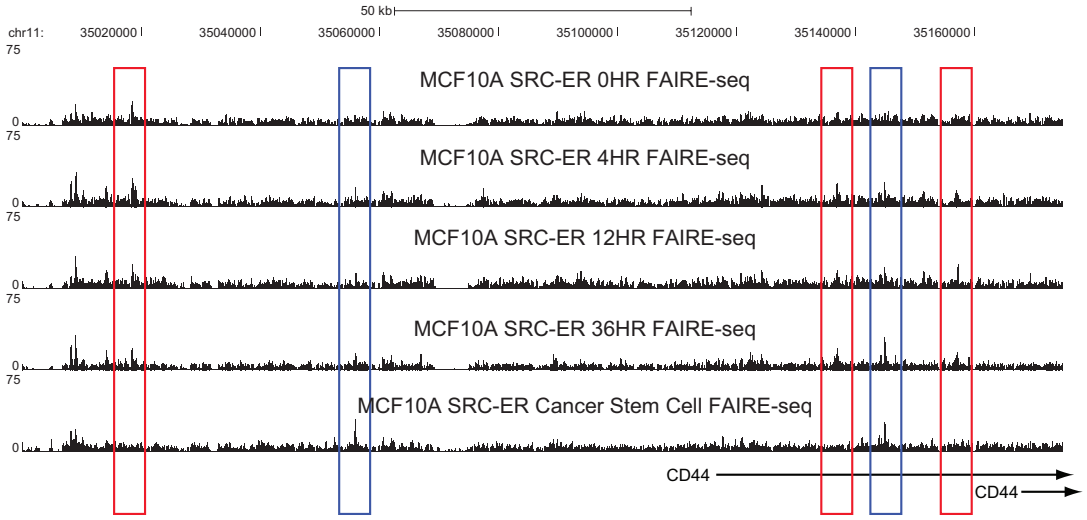

Figure 6.7: A 150 kb genomic locus containing the transcription start sites for the CD44 gene and it upstream region. The blue boxes indicate peaks in the cancer stem cells that were not present at the 0 hour time point. While the red boxes indicate peaks in the timecourse that are not present in the cancer stem cells, regardless of the presence at 0 hour.

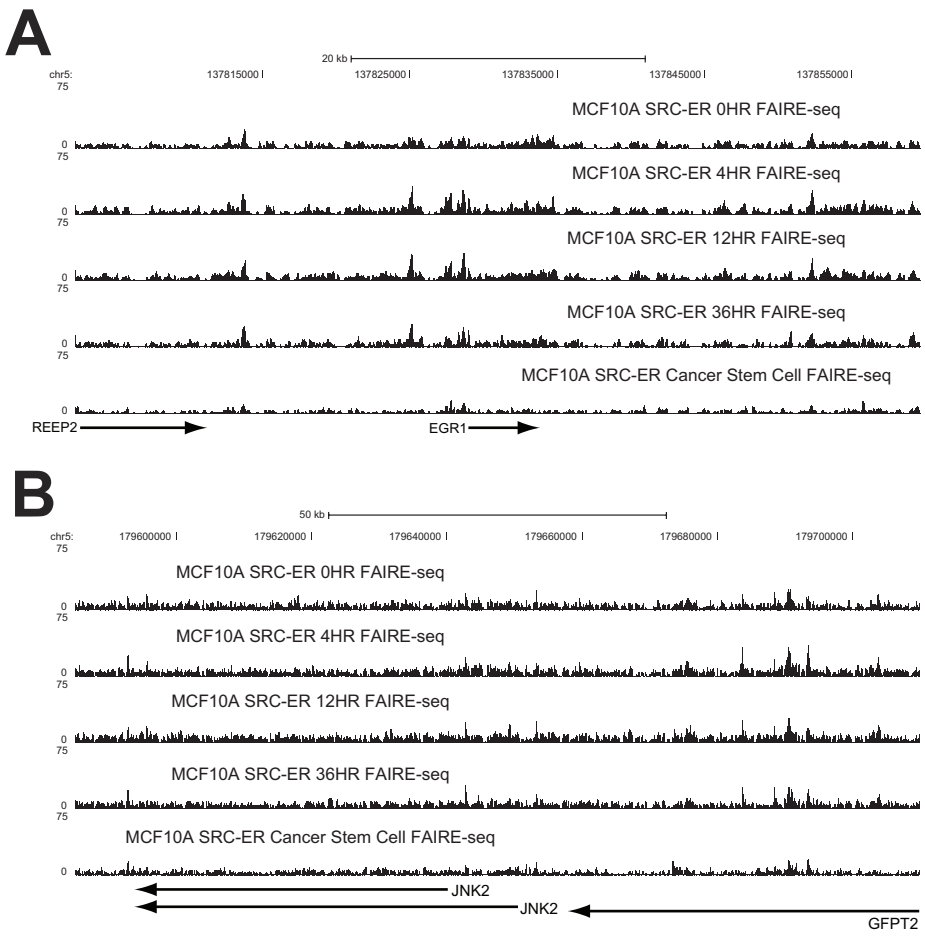

Figure 6.8: Genomic loci containing the (A) Egr1 and (B) JNK genes. The genes are represented as arrows pointing in the direction of transcription. The genomic loci containing these genes have FAIRE sites from the timecourse samples, but not the cancer stem cells.

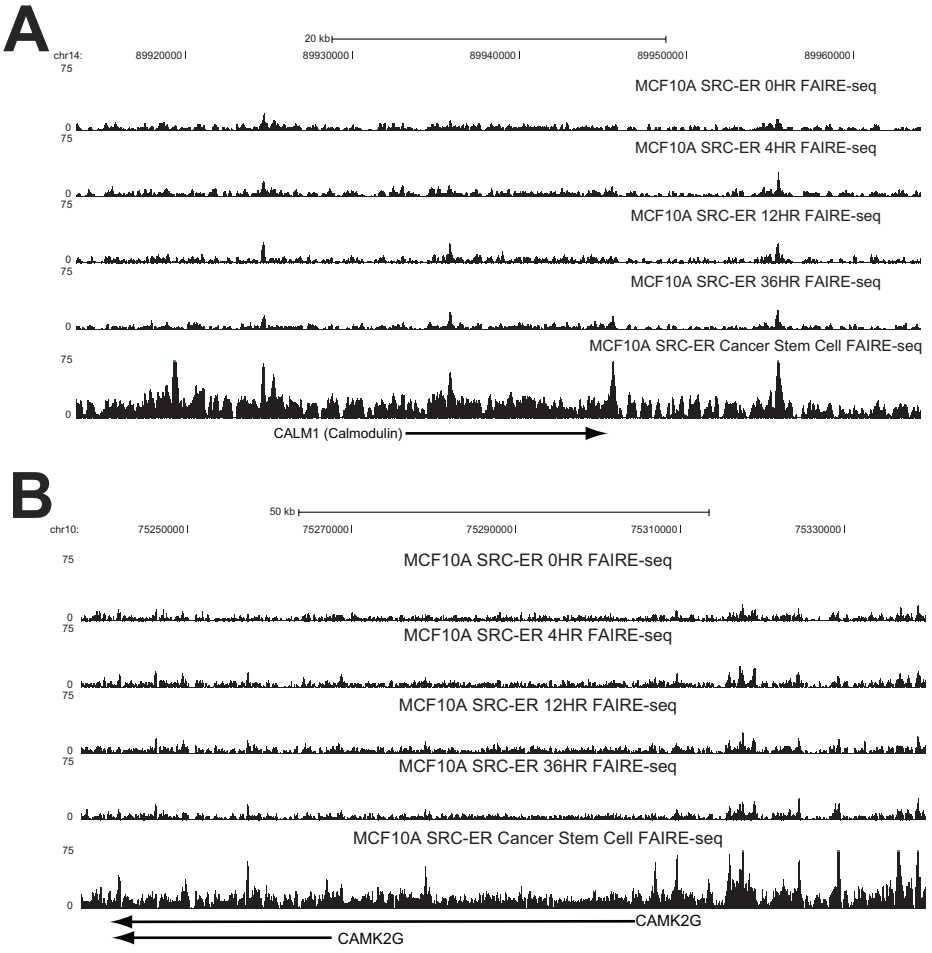

Figure 6.9: Genomic loci containing the the (A) calmodulin and (B) CAMK2G genes. Genes are represented as arrows pointing in the direction of transcription. The genomic DNA for cancer stem cells is amplified at these loci.

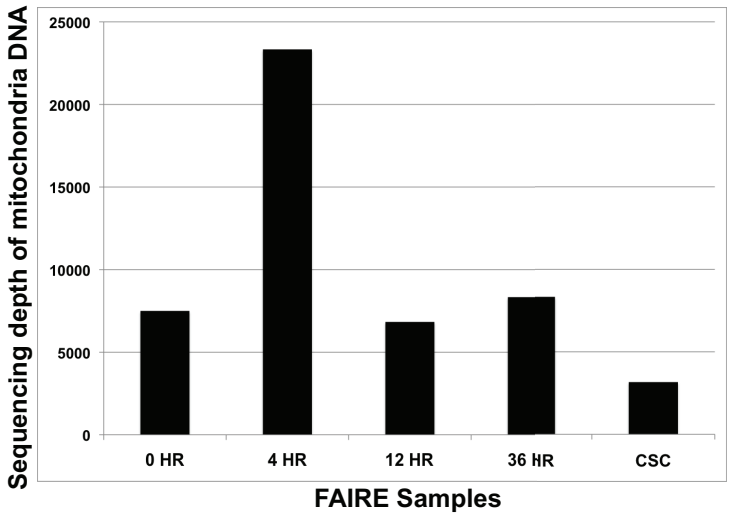

Figure 6.10: Mitochondrial genome is not packaged into chromatin and is therefore enriched by FAIRE. Depth of coverage of the mitochondrial genome is indicated for each of the time points, with a notable increase at 4 hours and depletion for the cancer stem cells.

Chapter 7

Discussions and perspectives

We have presented evidence that FAIRE is capable of isolating nucleosome-depleted DNA, a hallmark of active regulatory elements, from human chromatin. Genome-wide maps of active regulatory elements will allow a better understanding of how the availability of sequence-based regulatory elements are coordinated with the regulation of factors that utilize them in a given cellular environment. Understanding this relationship will be critical to constructing models of gene regulation in all eukaryotic cells. Here we have begun to functionally annotate the catalogue of regulatory elements in human breast cancer.

Several aspects of FAIRE make it a powerful genome-wide approach for detecting functional *in vivo* regulatory elements in breast cancer. FAIRE is amenable to clinical applications because it is relatively easy to perform and works with a limited amount of tissue samples ($\sim$20 mg). It requires little treatment of cells prior to the addition of formaldehyde and involves only a few reagents: formaldehyde, phenol, chloroform, and ethanol. The successful application of FAIRE on a limited numbers of cells expands its utility beyond what other DNA accessibility assays can accomplish, such as performing genome-wide assays of chromatin structure on cells grown in small-well plates for screening the effects of small molecules on chromatin. Identification of the genome-wide

set of active regulatory elements provides a means for understanding the mechanisms driving expression of genes in a given subtype, which will aid the refinement of subtype classification and evaluation of clinical outcomes. Finally, identification of the set of regulatory motifs that differentiate subtypes provides a functional basis for the selection of relevant transcription factor candidates.

Given the capacity of FAIRE for detection of regulatory elements in breast tumors, future studies will serve to expand our understanding of the clinical heterogeneity of breast cancer. In particular an expanded set of samples will offer the opportunity to identify the set of regulatory elements predictive of clinical outcomes. It will also be useful for determining a minimal set of diagnostic regulatory elements that could be used for clinical applications. One of the major challenges for transitioning techniques from the laboratory to the clinical is establishing standardized methodologies that minimize variability and offer a reliable diagnostic. Given the relatively simple nature of the FAIRE procedure it can readily be standardized using robotics to achieve a high degree of precision and reproducibly. Another major challenge for the study of breast cancer will be dissecting the cellular identity and molecular distinctions amongst the heterogeneous set of cells that comprise the tumor. Given that FAIRE is capable of working with limited cell numbers it would also be possible to generate the profiles of active regulatory elements for subsets of tumor cells.

However a more general challenge for the analysis and interpretation of FAIRE data from human cells will be to develop a more comprehensive understanding of the function of these regulatory elements. For instance the the vast majority of sites identified were far from any annotated gene. For the majority of these distal sites, it is not yet possible to ascribe a function, identify what factors might be bound or determine the set of target genes. One resource to address this challenge is the emerging set of consortium-based datasets, such as those derived from the ENCODE project, which will provide

a foundation for understanding the relationships among these factors, and be critical to constructing realistic models of gene regulation. Additionally, since FAIRE recovers the complete DNA fragments at regulatory elements it is possible to use this material directly in functional assays, such as with reporter vectors, which can interrogate the regulatory capacity of these fragments.

Altogether FAIRE is a promising new technique that offers tremendous potential for better understanding basic chromatin biology and the study of human disease.

Appendix A: FAIRE Protocol

FAIRE Cell Culture Protocol

1. **Crosslinking**

   (a) If cells are grown in suspension remove an aliquot to be used as an unfixed reference and place on ice. Otherwise, the reference sample can be obtained by removing an aliquot following sonication, reversing the crosslinks, and purifying the DNA.

   (b) Add 37% formaldehyde directly to media to a final concentration of 1%.

   (c) Incubate at 25°C for 5 min with shaking 80 rpm.

   (d) Add 2.5 M glycine to a final concentration of 125 mM, incubate 5 min at RT with shaking.

   (e) Spin at 700 x g for 5 min at 4°C.

   (f) Wash twice with ice cold 1xPBS, spin at 1000 rpm for 5 min at 4°C.

   (g) Cells can be snap frozen at this point and stored at -80°C.

2. **Cell lysis** (if frozen thaw cells on ice)

   (a) Resuspend cells in 1 ml of lysis buffer per 10^7 (or 0.4g) cells.

   (b) Add 1 ml 0.5 mm glass beads to rubber sealed 2 ml screw topped tube. Add 1 ml of cells in lysis buffer.

   (c) Lyse cells in the mini-beadbeater-8 for five 1 minute sessions, ice cells for two minutes between each session.

   (d) Recover the lysate by puncturing the bottom of 2 ml tube with 25G syringe and drain into 15 ml tube on ice. Filtered air can be used to force liquid through hole.

   (e) Add an additional 500 μl of lysis buffer to flush remaining sample.

   (f) Transfer 300 μl aliquots to 1.5 ml tubes and sonicate in Bioruptor for 15 minutes on HIGH using 30 second pulses and 30 seconds of rest, keep waterbath at a constant 4°C.

(g) Spin the extract at 15,000 x g for 5 minutes at 4°C to clear cellular debris. Transfer the supernatant to a new tube.

(h) Remove an aliquot equivalent to 500 ng genomic DNA and check fragment size on 1% agarose gel.

3. **Phenol/Chloroform**

(a) Add an equal volume of phenol/chloroform, vortex, and spin at 12,000 x g for 5 minutes, transfer aqueous phase to a new tube. NOTE: If aqueous phase is small add 500 μl of TE to interphase, vortex, spin down and recover aqueous phase.

(b) Add an equal volume phenol/chloroform to aqueous phase in fresh tube, vortex, spin down, and transfer aqueous phase to a fresh tube.

(c) Add an equal volume of chloroform-isoamyl alcohol (24:1), vortex, and spin 12,000 x g for 5 min.

(d) Add 1/10th volume of 3 M Sodium Acetate (pH 5.2) and 1 μl of 20 mg/ml glycogen, mix by inverting, and add 2X volume of 95% ethanol. Incubate at -20°C 1 hour to overnight.

(e) Pellet precipitated DNA at 15,000 x g for 30 min at 4°C and remove supernatant. Wash pellet with 500 μl 70% ethanol, spin at 15,000 x g for 5 min at room temp (25°C). Remove supernatant and dry pellet in speed-vac.

(f) Resuspend pellet in 50 μl 10 mM Tris-HCl (pH 7.4).

(g) Add 1 μl of 10 mg/ml RNase A and incubate at 37°C for 1 hour.

(h) Cleanup sample using spin column (must recover 75 to 200 bp DNA) or additional phenol/chloroform extraction and ethanol precipitation.

Lysis buffer
2% Triton X-100
1% SDS
100 mM NaCl
10 mM Tris-Cl ph 8.0
1 mM EDTA

Phenol/Chloroform- Sigma #P3803 phenol, chloroform, and isoamyl alcohol 25:24:1 saturated with 10mM Tris, pH 8.0, 1 mM EDTA

Checking fragment sizes after sonication
NOTE: Limit vortexing to avoid additional shearing

1. Add 1 μl of 10 mg/ml of RNase A, flick tube to mix, and incubate at 37°C for 1 hour.

2. Incubate at 65°C for 4 hours to overnight.

3. Add 1 μl of 10 mg/ml of Proteinase K, flick tube to mix, and incubate at 37°C for 1 hour.

4. Add 10 mM Tris-HCl (pH7.4) to a final volume of 250 μl. Add an equal volume phenol/chloroform, mix, and spin at 12,000 x g 5 minutes, transfer aqueous phase to a new tube.

5. Add an equal chloroform-isoamyl alcohol (24:1), mix, spin at 12,000 x g for 5 minutes, and transfer aqueous phase to a new tube.

6. Add 1/10th volume of 3M Sodium Acetate (pH 5.2) and 1 μl of 20 mg/ml glycogen, mix by inverting, and add 2X volume of 95% ethanol, incubate at -20°C for 1 hour

7. Pellet DNA at 15,000 x g for 10 minutes at 4°C, wash with 500 μl 70% ethanol, and spin at 15,000 x g for 5 min at room temp (25°C)

8. Dry pellet and resuspend in 10 μl 10 mM Tris-HCl (pH 7.4) and run on a 1% agarose gel.

NOTE: An ideal distribution is a smear from 1000 bp to 100 bp with an average size of 500 bp

Alternative Cell Lysis

NOTE: If bead-beater is not available, this procedure works for cells, not yeast

1. Add 10 ml of Buffer L1 per 10^8 cells and rock at 4°C for 10 min.
2. Spin cells at 1300 x g for 5 min at 4°C and remove supernatant.
3. Resuspend pellet in 10 ml of Buffer L2 per 10^8 cells and rock at RT for 10 min.
4. Spin cells at 1300 x g for 5 min at 4°C and remove supernatant.
5. Resuspend pellet in 3.5 ml of Buffer L3 per 10^8 cells. Proceed with sonication

Buffer L1- per 100 ml	
Volume	Reagent
5 ml	1M Hepes KOH, pH 7.5
2.8 ml	5M NaCl
0.2 ml	0.5 M EDTA (pH 8.0)
10 ml	100% Glycerol
5 ml	100% NP-40
0.25 ml	100% Triton X-100
76.7 ml	dH_20

Buffer L2- per 100 ml	
Volume	Reagent
4 ml	5 M NaCl
0.2 ml	0.5 M EDTA (pH 8.0)
0.1 ml	0.5 M EGTA (pH 8.0)
1 ml	1 M Tris (pH 8.0)
94.7 ml	dH_20

Buffer L3- per 100 ml	
Volume	Reagent
0.2 ml	0.5 M EDTA (pH 8.0)
0.1 ml	0.5 M EGTA (pH 8.0)
1 ml	1 M Tris (pH 8.0)
2 ml	5 M NaCl
1 ml	10% Na-Deoxycholate
500 mg	N-lauroyl sarcosine
1 ml	50X Protease Inhibitor
94.7 ml	dH_20

FAIRE Tissue Protocol

Crosslinking Fresh Soft Samples

1. Mince fresh tissue with scalpel and place in dounce with 1 ml of PBS per 10 mg of tissue along with 37% formaldehyde to a final concentration of 1%, swirl occasionally, and incubate at 25°C for 5 min.

2. Add 2.5 M glycine to a final concentration of 125 mM and incubate at 25°C for 5 min.

3. Disassociate cells with dounce, transfer to 1.5 ml tube, spin at 1,000 x g for 5 min at 4°C, discard supernatant, and wash 2x with ice cold PBS.

Crosslinking Frozen or Fibrous Samples

1. Add tissue sample to a 15 ml conical tissue grinder (VWR cat#47732-446), pre-cooled in liquid nitrogen bath, incubate 10 minutes in liquid nitrogen bath, and grind into powder.

2. Add 1 ml of PBS per 10 mg of tissue 37% formaldehyde to a final concentration of 1%, swirl occasionally, and incubate at 25°C for 7 min.

3. Add 2.5 M glycine to a final concentration of 125 mM and incubate at 25°C for 5 min.

4. Spin at 1,000 x g for 5 min at 4o C, discard supernatant, and wash 2x with ice cold PBS

Proceed with cell lysis, sonication, phenol/chloroform extraction, and ethanol precipitation outlined above.

NOTE: For especially difficult samples use large 2.8 mm ceramic or metal beads (Precellys CK28 or MK28) and perform additional rounds in bead-beater.

Appendix B: FAIRE-seq library preparation

BLUNTING THE FRAGMENTS (Epicentre END-IT DNA REPAIR KIT # ER0720)

1-34 μl	DNA (< 5 μg)
5 μl	10x End-it Repair Buffer
5 μl	2.5 mM dNTP mix
5 μl	10 mM ATP
1 μl	End Repair Enzyme mix
50 μl	Reaction volume (add H_2O to volume)

Incubate 45 minutes at room temperature
Cleanup using Qiagen PCR Purification Column:

Use 250 μl PBI buffer
Spin at 10,000 rpm
Elute with 35 μl EB

ADD A OVERHANG (NEB Klenow Exo-minus 50 U/μl #M0212M)

34 μl	DNA
5 μl	10x NEB 2
1 μl	10 mM dATP
1 μl	Klenow 50 U/μl
50 μl	Reaction volume (add H_2O to volume)

Incubate 30 minutes at 37°C
Cleanup using Qiagen Mini-Elute Purification Column:

Use 250 μl PBI buffer
Spin at 10,000 rpm
Elute with 11 μl EB

LIGATION OF ADAPTERS (Epicentre FAST LINK KIT #LK11025)
NOTE: Dilute adapters 1:10, for DNA sample <50 ng us 1 μl; for <100 ng use 2 μl; for $>$ 500 ng use 1 μl undilute

10 μl	DNA
3 μl	10x Fast-Link Buffer
1.5 μl	10 mM ATP
1 μl	Adapters
2 μl	Fast-Link DNA Ligase 2 U/μl
30 μl	Reaction volume (add H_2O to volume)

Incubate overnight at 16°C
Cleanup using Qiagen PCR Purification Column:

Use 200 μl PBI buffer
Spin at 10,000 rpm
Elute with 37 μl EB

PCR AMPLIFICATION (Stratagene PfuUltra™ II Fusion HS DNA Polymerase # 600670)

36 μl	DNA (use 50 ng)
2 μl	Illumina primers
10 μl	10x PfuUltra II reaction buffer
10 μl	2.5 mM dNTP
1 μl	Phusion polymerase
50 μl	Reaction volume (add H_2O to volume)

Cycling parameters:
98°C 30 sec
(98°C 20 sec, 65°C 30 sec, 72°C 30 sec) repeat 12 cycles
72°C 5 min
Hold at 4°C
Cleanup using Qiagen Mini-Elute Purification Column:

Use 500 μl PBI buffer
Spin at 10,000 rpm
Elute with 11 μl EB

SIZE SELECT LIBRARY

Loading buffer
50 mM Tris pH 8.0
40mM EDTA
40% (w/v) sucrose

Run sample on 2% agarose gel
3 μl Loading Buffer per 10 μl Sample
Run at 120 V for 1 hour
Excise brightest region +/- 100 bp
Purify using Qiagen Gel Extraction Column

Use 6x QG Buffer
Use 2x Isopropanol
Spin at 10,000 rpm
Elute with 51 μl EB

NOTE: Do not heat gel slice in QG buffer to the recommended 55°C

References

[1] Abbott, D W, V S Ivanova, X Wang, W M Bonner and J Ausió. 2001. "Characterization of the stability and folding of H2A.Z chromatin particles: implications for transcriptional activation." *J Biol Chem* 276(45):41945–9.

[2] Al-Hajj, Muhammad, Max S Wicha, Adalberto Benito-Hernandez, Sean J Morrison and Michael F Clarke. 2003. "Prospective identification of tumorigenic breast cancer cells." *Proc Natl Acad Sci U S A* 100(7):3983–8.

[3] Albertson, Donna G. 2006. "Gene amplification in cancer." *Trends Genet* 22(8):447–55.

[4] Ashurst, J L, C-K Chen, J G R Gilbert, K Jekosch, S Keenan, P Meidl, S M Searle, J Stalker, R Storey, S Trevanion, L Wilming and T Hubbard. 2005. "The Vertebrate Genome Annotation (Vega) database." *Nucleic Acids Res* 33(Database issue):D459–65.

[5] Baan, Bart, Evangelia Pardali, Peter ten Dijke and Hans van Dam. 2010. "In situ proximity ligation detection of c-Jun/AP-1 dimers reveals increased levels of c-Jun/Fra1 complexes in aggressive breast cancer cell lines in vitro and in vivo." *Mol Cell Proteomics* 9(9):1982–90.

[6] Bailey, Timothy L. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press pp. 28–36.

[7] Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev and Keji Zhao. 2007. "High-resolution profiling of histone methylations in the human genome." *Cell* 129(4):823–37.

[8] Benjamini, Y. and Y. Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1):289–300.

[9] Berdasco, María and Manel Esteller. 2010. "Aberrant epigenetic landscape in cancer: how cellular identity goes awry." *Dev Cell* 19(5):698–711.

[10] Bernstein, Bradley E, Chih Long Liu, Emily L Humphrey, Ethan O Perlstein and Stuart L Schreiber. 2004. "Global nucleosome occupancy in yeast." *Genome Biol* 5(9):R62.

[11] Bertucci, François and Daniel Birnbaum. 2008. "Reasons for breast cancer heterogeneity." *J Biol* 7(2):6.

[12] Bièche, Ivan and Rosette Lidereau. 2011. "Genome-based and transcriptome-based molecular classification of breast cancer." *Curr Opin Oncol* 23(1):93–9.

[13] Blows, Fiona M, Kristy E Driver, Marjanka K Schmidt, Annegien Broeks, Flora E van Leeuwen, Jelle Wesseling, Maggie C Cheang, Karen Gelmon, Torsten O Nielsen, Carl Blomqvist, Päivi Heikkilä, Tuomas Heikkinen, Heli Nevanlinna, Lars A Akslen, Louis R Bégin, William D Foulkes, Fergus J Couch, Xianshu Wang, Vicky Cafourek, Janet E Olson, Laura Baglietto, Graham G Giles, Gianluca Severi, Catriona A McLean, Melissa C Southey, Emad Rakha, Andrew R Green, Ian O Ellis, Mark E Sherman, Jolanta Lissowska, William F Anderson, Angela Cox, Simon S Cross, Malcolm W R Reed, Elena Provenzano, Sarah-Jane Dawson, Alison M Dunning, Manjeet Humphreys, Douglas F Easton, Montserrat García-Closas, Carlos Caldas, Paul D Pharoah and David Huntsman. 2010. "Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies." *PLoS Med* 7(5):e1000279.

[14] Boeger, Hinrich, Joachim Griesenbeck, J Seth Strattan and Roger D Kornberg. 2003. "Nucleosomes unfold completely at a transcriptionally active promoter." *Mol Cell* 11(6):1587–98.

[15] Bolstad, B M, R A Irizarry, M Astrand and T P Speed. 2003. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics* 19(2):185–93.

[16] Bonnet, D and J E Dick. 1997. "Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell." *Nat Med* 3(7):730–7.

[17] Bosman, Joshua D, Fruma Yehiely, Joseph R Evans and Vincent L Cryns. 2010. "Regulation of alphaB-crystallin gene expression by the transcription factor Ets1 in breast cancer." *Breast Cancer Res Treat* 119(1):63–70.

[18] Boyle, Alan P, Justin Guinney, Gregory E Crawford and Terrence S Furey. 2008. "F-Seq: a feature density estimator for high-throughput sequence tags." *Bioinformatics* 24(21):2537–8.

[19] Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey and Gregory E Crawford. 2008. "High-resolution mapping and characterization of open chromatin across the genome." *Cell* 132(2):311–22.

[20] Brand, K A and U Hermfisse. 1997. "Aerobic glycolysis by proliferating cells: a protective strategy against reactive oxygen species." *FASEB J* 11(5):388–95.

[21] Brutlag, D, C Schlehuber and J Bonner. 1969. "Properties of formaldehyde-treated nucleohistone." *Biochemistry* 8(8):3214–8.

[22] Buck, Michael J, Andrew B Nobel and Jason D Lieb. 2005. "ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data." *Genome Biol* 6(11):R97.

[23] Buck, Michael J and Jason D Lieb. 2004. "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." *Genomics* 83(3):349–60.

[24] Bulyk, Martha L. 2004. "Integrative functional genomics." *Genome Biol* 5(7):331.

[25] Carroll, Jason S, Clifford A Meyer, Jun Song, Wei Li, Timothy R Geistlinger, Jérôme Eeckhoute, Alexander S Brodsky, Erika Krasnickas Keeton, Kirsten C Fertuck, Giles F Hall, Qianben Wang, Stefan Bekiranov, Victor Sementchenko, Edward A Fox, Pamela A Silver, Thomas R Gingeras, X Shirley Liu and Myles Brown. 2006. "Genome-wide analysis of estrogen receptor binding sites." *Nat Genet* 38(11):1289–97.

[26] Chin, Koei, Sandy DeVries, Jane Fridlyand, Paul T Spellman, Ritu Roydasgupta, Wen-Lin Kuo, Anna Lapuk, Richard M Neve, Zuwei Qian, Tom Ryder, Fanqing Chen, Heidi Feiler, Taku Tokuyasu, Chris Kingsley, Shanaz Dairkee, Zhenhang Meng, Karen Chew, Daniel Pinkel, Ajay Jain, Britt Marie Ljung, Laura Esserman, Donna G Albertson, Frederic M Waldman and Joe W Gray. 2006. "Genomic and transcriptional aberrations linked to breast cancer pathophysiologies." *Cancer Cell* 10(6):529–41.

[27] Christensen, Brock C, Karl T Kelsey, Shichun Zheng, E Andres Houseman, Carmen J Marsit, Margaret R Wrensch, Joseph L Wiemels, Heather H Nelson, Margaret R Karagas, Lawrence H Kushi, Marilyn L Kwan and John K Wiencke. 2010. "Breast cancer DNA methylation profiles are associated with tumor size and alcohol and folate intake." *PLoS Genet* 6(7):e1001043.

[28] Cicatiello, Luigi, Margherita Mutarelli, Oli M V Grober, Ornella Paris, Lorenzo Ferraro, Maria Ravo, Roberta Tarallo, Shujun Luo, Gary P Schroth, Martin Seifert, Christian Zinser, Maria Luisa Chiusano, Alessandra Traini, Michele De Bortoli and Alessandro Weisz. 2010. "Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs." *Am J Pathol* 176(5):2113–30.

[29] Cooper, Sara J, Nathan D Trinklein, Elizabeth D Anton, Loan Nguyen and Richard M Myers. 2006. "Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome." *Genome Res* 16(1):1–10.

[30] Crabtree, Gerald R and Eric N Olson. 2002. "NFAT signaling: choreographing the social lives of cells." *Cell* 109 Suppl:S67–79.

[31] Crawford, Gregory E, Ingeborg E Holt, James C Mullikin, Denise Tai, Robert Blakesley, Gerard Bouffard, Alice Young, Catherine Masiello, Eric D Green, Tyra G Wolfsberg, Francis S Collins and National Institutes Of Health Intramural Sequencing Center. 2004. "Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites." *Proc Natl Acad Sci U S A* 101(4):992–7.

[32] Crawford, Gregory E, Sean Davis, Peter C Scacheri, Gabriel Renaud, Mohamad J Halawi, Michael R Erdos, Roland Green, Paul S Meltzer, Tyra G Wolfsberg and Francis S Collins. 2006. "DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays." *Nat Methods* 3(7):503–9.

[33] Croce, Carlo M. 2008. "Oncogenes and cancer." *N Engl J Med* 358(5):502–11.

[34] Dai, Zunyan, Audrey C Papp, Danxin Wang, Heather Hampel and Wolfgang Sadee. 2008. "Genotyping panel for assessing response to cancer chemotherapy." *BMC Med Genomics* 1:24.

[35] David, Lior, Wolfgang Huber, Marina Granovskaia, Joern Toedling, Curtis J Palm, Lee Bofkin, Ted Jones, Ronald W Davis and Lars M Steinmetz. 2006. "A high-resolution map of transcription in the yeast genome." *Proc Natl Acad Sci U S A* 103(14):5320–5.

[36] Debnath, Jayanta, Senthil K Muthuswamy and Joan S Brugge. 2003. "Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures." *Methods* 30(3):256–68.

[37] Dempster, A.P., N.M. Laird and D.B. Rubin. 1977. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

[38] Dohm, Juliane C, Claudio Lottaz, Tatiana Borodina and Heinz Himmelbauer. 2008. "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing." *Nucleic Acids Res* 36(16):e105.

[39] Eferl, Robert and Erwin F Wagner. 2003. "AP-1: a double-edged sword in tumorigenesis." *Nat Rev Cancer* 3(11):859–68.

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

[41] ENCODE Project Consortium. 2004. "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science* 306(5696):636–40.

[42] ENCODE Project Consortium. 2007. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature* 447(7146):799–816.

[43] Engle, L J, C L Simpson and J E Landers. 2006. "Using high-throughput SNP technologies to study cancer." *Oncogene* 25(11):1594–601.

[44] Esteller, Manel. 2007. "Cancer epigenomics: DNA methylomes and histone-modification maps." *Nat Rev Genet* 8(4):286–98.

[45] Fujita, Pauline A, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, Mark Diekhans, Timothy R Dreszer, Belinda M Giardine, Rachel A Harte, Jennifer Hillman-Jackson, Fan Hsu, Vanessa Kirkup, Robert M Kuhn, Katrina Learned, Chin H Li, Laurence R Meyer, Andy Pohl, Brian J Raney, Kate R Rosenbloom, Kayla E Smith, David Haussler and W James Kent. 2011. "The UCSC Genome Browser database: update 2011." *Nucleic Acids Res* 39(Database issue):D876–82.

[46] Garvie, C W and C Wolberger. 2001. "Recognition of specific DNA sequences." *Mol Cell* 8(5):937–46.

[47] Gazin, Claude, Narendra Wajapeyee, Stephane Gobeil, Ching-Man Virbasius and Michael R Green. 2007. "An elaborate pathway required for Ras-mediated epigenetic silencing." *Nature* 449(7165):1073–7.

[48] Gee, Katrina, Wilfred Lim, Wei Ma, Devki Nandan, Francisco Diaz-Mitoma, Maya Kozlowski and Ashok Kumar. 2002. "Differential regulation of CD44 expression by lipopolysaccharide (LPS) and TNF-alpha in human monocytic cells: distinct involvement of c-Jun N-terminal kinase in LPS-induced CD44 expression." *J Immunol* 169(10):5660–72.

[49] Gibbons, Francis D, Markus Proft, Kevin Struhl and Frederick P Roth. 2005. "Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization." *Genome Biol* 6(11):R96.

[50] Giresi, Paul G and Jason D Lieb. 2009. "Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements)." *Methods* 48(3):233–9.

[51] Giresi, Paul G, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer and Jason D Lieb. 2007. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin." *Genome Res* 17(6):877–85.

[52] Golub, T R, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield and E S Lander. 1999. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science* 286(5439):531–7.

[53] Grant, Charles E, Timothy L Bailey and William Stafford Noble. 2011. "FIMO: scanning for occurrences of a given motif." *Bioinformatics* 27(7):1017–8.

[54] Grimshaw, Matthew J, Lucienne Cooper, Konstantinos Papazisis, Julia A Coleman, Hermann R Bohnenkamp, Laura Chiapero-Stanke, Joyce Taylor-Papadimitriou and Joy M Burchell. 2008. "Mammosphere culture of metastatic breast cancer cells enriches for tumorigenic breast cancer cells." *Breast Cancer Res* 10(3):R52.

[55] Gunjan, Akash, Johanna Paik and Alain Verreault. 2005. "Regulation of histone synthesis and nucleosome assembly." *Biochimie* 87(7):625–35.

[56] Gupta, Shobhit, John A Stamatoyannopoulos, Timothy L Bailey and William Stafford Noble. 2007. "Quantifying similarity between motifs." *Genome Biol* 8(2):R24.

[57] Hahn, William C and Robert A Weinberg. 2002. "Rules for making human tumor cells." *N Engl J Med* 347(20):1593–603.

[58] Hampton, Oliver A, Petra Den Hollander, Christopher A Miller, David A Delgado, Jian Li, Cristian Coarfa, Ronald A Harris, Stephen Richards, Steven E Scherer, Donna M Muzny, Richard A Gibbs, Adrian V Lee and Aleksandar Milosavljevic. 2009. "A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome." *Genome Res* 19(2):167–77.

[59] Hanahan, D and R A Weinberg. 2000. "The hallmarks of cancer." *Cell* 100(1):57–70.

[60] Hanahan, Douglas and Robert A Weinberg. 2011. "Hallmarks of cancer: the next generation." *Cell* 144(5):646–74.

[61] Harrow, Jennifer, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James G R Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E Antonarakis and Roderic Guigo. 2006. "GENCODE: producing a reference annotation for ENCODE." *Genome Biol* 7 Suppl 1:S4.1–9.

[62] Hebbard, L, A Steffen, V Zawadzki, C Fieber, N Howells, J Moll, H Ponta, M Hofmann and J Sleeman. 2000. "CD44 expression and regulation during mammary gland development and function." *J Cell Sci* 113 (Pt 14):2619–30.

[63] Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh and Christopher K Glass. 2010. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." *Mol Cell* 38(4):576–89.

[64] Hillier, LaDeana W, Gabor T Marth, Aaron R Quinlan, David Dooling, Ginger Fewell, Derek Barnett, Paul Fox, Jarret I Glasscock, Matthew Hickenbotham, Weichun Huang, Vincent J Magrini, Ryan J Richt, Sacha N Sander, Donald A Stewart, Michael Stromberg, Eric F Tsung, Todd Wylie, Tim Schedl, Richard K Wilson and Elaine R Mardis. 2008. "Whole-genome sequencing and variant discovery in C. elegans." *Nat Methods* 5(2):183–8.

[65] Hinrichs, A S, D Karolchik, R Baertsch, G P Barber, G Bejerano, H Clawson, M Diekhans, T S Furey, R A Harte, F Hsu, J Hillman-Jackson, R M Kuhn, J S Pedersen, A Pohl, B J Raney, K R Rosenbloom, A Siepel, K E Smith, C W Sugnet, A Sultan-Qurraie, D J Thomas, H Trumbower, R J Weber, M Weirauch, A S Zweig, D Haussler and W J Kent. 2006. "The UCSC Genome Browser Database: update 2006." *Nucleic Acids Res* 34(Database issue):D590–8.

[66] Hirsch, Heather A, Dimitrios Iliopoulos, Amita Joshi, Yong Zhang, Savina A Jaeger, Martha Bulyk, Philip N Tsichlis, X Shirley Liu and Kevin Struhl. 2010. "A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases." *Cancer Cell* 17(4):348–61.

[67] Hirsch, Heather A, Dimitrios Iliopoulos, Philip N Tsichlis and Kevin Struhl. 2009. "Metformin selectively targets cancer stem cells, and acts together with chemotherapy to block tumor growth and prolong remission." *Cancer Res* 69(19):7507–11.

[68] Hogan, Gregory J, Cheol-Koo Lee and Jason D Lieb. 2006. "Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters." *PLoS Genet* 2(9):e158.

[69] Hurowitz, Evan H and Patrick O Brown. 2003. "Genome-wide analysis of mRNA lengths in Saccharomyces cerevisiae." *Genome Biol* 5(1):R2.

[70] Iliopoulos, Dimitrios, Heather A Hirsch and Kevin Struhl. 2009. "An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation." *Cell* 139(4):693–706.

[71] Iliopoulos, Dimitrios, Heather A Hirsch and Kevin Struhl. 2011. "Metformin decreases the dose of chemotherapy for prolonging tumor remission in mouse xenografts involving multiple cancer cell types." *Cancer Res* 71(9):3196–201.

[72] Iorio, Marilena V, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, Sylvie Ménard, Juan P Palazzo, Anne Rosenberg, Piero Musiani, Stefano Volinia, Italo Nenci, George A Calin, Patrizia Querzoli, Massimo Negrini and Carlo M Croce. 2005. "MicroRNA gene expression deregulation in human breast cancer." *Cancer Res* 65(16):7065–70.

[73] Irizarry, Rafael A, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf and Terence P Speed. 2003. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics* 4(2):249–64.

[74] Ji, Hongkai and Wing Hung Wong. 2005. "TileMap: create chromosomal map of tiling array hybridizations." *Bioinformatics* 21(18):3629–36.

[75] Johnson, W Evan, Cheng Li and Ariel Rabinovic. 2007. "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8(1):118–27.

[76] Johnson, W Evan, Wei Li, Clifford A Meyer, Raphael Gottardo, Jason S Carroll, Myles Brown and X Shirley Liu. 2006. "Model-based analysis of tiling-arrays for ChIP-chip." *Proc Natl Acad Sci U S A* 103(33):12457–62.

[77] Jones, Peter A and Stephen B Baylin. 2007. "The epigenomics of cancer." *Cell* 128(4):683–92.

[78] Kadota, Mitsutaka, Howard H Yang, Bianca Gomez, Misako Sato, Robert J Clifford, Daoud Meerzaman, Barbara K Dunn, Lalage M Wakefield and Maxwell P Lee. 2010. "Delineating genetic alterations for tumor progression in the MCF10A series of breast cancer cell lines." *PLoS One* 5(2):e9201.

[79] Kamakaka, Rohinton T and Sue Biggins. 2005. "Histone variants: deviants?" *Genes Dev* 19(3):295–310.

[80] Kao, Jessica, Keyan Salari, Melanie Bocanegra, Yoon-La Choi, Luc Girard, Jeet Gandhi, Kevin A Kwei, Tina Hernandez-Boussard, Pei Wang, Adi F Gazdar, John D Minna and Jonathan R Pollack. 2009. "Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery." *PLoS One* 4(7):e6146.

[81] Keene, M A and S C Elgin. 1981. "Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure." *Cell* 27(1 Pt 2):57–64.

[82] Keene, M A, V Corces, K Lowenhaupt and S C Elgin. 1981. "DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription." *Proc Natl Acad Sci U S A* 78(1):143–6.

[83] Kharchenko, Peter V, Michael Y Tolstorukov and Peter J Park. 2008. "Design and analysis of ChIP-seq experiments for DNA-binding proteins." *Nat Biotechnol* 26(12):1351–9.

[84] Khoshnaw, S M, A R Green, D G Powe and I O Ellis. 2009. "MicroRNA involvement in the pathogenesis and management of breast cancer." *J Clin Pathol* 62(5):422–8.

[85] Kim, Tae Hoon, Leah O Barrera, Chunxu Qu, Sara Van Calcar, Nathan D Trinklein, Sara J Cooper, Rosa M Luna, Christopher K Glass, Michael G Rosenfeld, Richard M Myers and Bing Ren. 2005. "Direct isolation and identification of promoters in the human genome." *Genome Res* 15(6):830–9.

[86] Kim, Tae Hoon, Leah O Barrera, Ming Zheng, Chunxu Qu, Michael A Singer, Todd A Richmond, Yingnian Wu, Roland D Green and Bing Ren. 2005. "A high-resolution map of active promoters in the human genome." *Nature* 436(7052):876–80.

[87] Knudson, A G. 2001. "Two genetic hits (more or less) to cancer." *Nat Rev Cancer* 1(2):157–62.

[88] Koch, Christoph M, Robert M Andrews, Paul Flicek, Shane C Dillon, Ulaş Karaöz, Gayle K Clelland, Sarah Wilcox, David M Beare, Joanna C Fowler, Phillippe Couttet, Keith D James, Gregory C Lefebvre, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Pawandeep Dhami, Cordelia F Langford, Zhiping Weng, Ewan Birney, Nigel P Carter, David Vetrie and Ian Dunham. 2007. "The landscape of histone modifications across 1% of the human genome in five human cell lines." *Genome Res* 17(6):691–707.

[89] Kouros-Mehr, Hosein, Seth K Bechis, Euan M Slorach, Laurie E Littlepage, Mikala Egeblad, Andrew J Ewald, Sung-Yun Pai, I-Cheng Ho and Zena Werb. 2008. "GATA-3 links tumor differentiation and dissemination in a luminal breast cancer model." *Cancer Cell* 13(2):141–52.

[90] Langmead, Ben, Cole Trapnell, Mihai Pop and Steven L Salzberg. 2009. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biol* 10(3):R25.

[91] Leary, Rebecca J, Jimmy C Lin, Jordan Cummins, Simina Boca, Laura D Wood, D Williams Parsons, Siân Jones, Tobias Sjöblom, Ben-Ho Park, Ramon Parsons, Joseph Willis, Dawn Dawson, James K V Willson, Tatiana Nikolskaya, Yuri Nikolsky, Levy Kopelovich, Nick Papadopoulos, Len A Pennacchio, Tian-Li Wang, Sanford D Markowitz, Giovanni Parmigiani, Kenneth W Kinzler, Bert Vogelstein and Victor E Velculescu. 2008. "Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers." *Proc Natl Acad Sci U S A* 105(42):16224–9.

[92] Lee, Cheol-Koo, Yoichiro Shibata, Bhargavi Rao, Brian D Strahl and Jason D Lieb. 2004. "Evidence for nucleosome depletion at active regulatory regions genome-wide." *Nat Genet* 36(8):900–5.

[93] Lee, Tong Ihn, Sarah E Johnstone and Richard A Young. 2006. "Chromatin immunoprecipitation and microarray-based analysis of protein location." *Nat Protoc* 1(2):729–48.

[94] Levy, A and M Noll. 1981. "Chromatin fine structure of active and repressed genes." *Nature* 289(5794):198–203.

[95] Lewandowska, Joanna and Agnieszka Bartoszek. 2011. "DNA methylation in cancer development, diagnosis and therapy–multiple opportunities for genotoxic agents to act as methylome disruptors or remediators." *Mutagenesis* 26(4):475–87.

[96] Li, Heng, Jue Ruan and Richard Durbin. 2008. "Mapping short DNA sequencing reads and calling variants using mapping quality scores." *Genome Res* 18(11):1851–8.

[97] Li, Ruiqiang, Yingrui Li, Karsten Kristiansen and Jun Wang. 2008. "SOAP: short oligonucleotide alignment program." *Bioinformatics* 24(5):713–4.

[98] Liao, Mai-Jing, Cheng Cheng Zhang, Beiyan Zhou, Drazen B Zimonjic, Sendurai A Mani, Megan Kaba, Ann Gifford, Ferenc Reinhardt, Nicholas C Popescu, Wenjun Guo, Elinor Ng Eaton, Harvey F Lodish and Robert A Weinberg. 2007. "Enrichment of a population of mammary gland cells that form mammospheres and have in vivo repopulating activity." *Cancer Res* 67(17):8131–8.

[99] Livak, K J and T D Schmittgen. 2001. "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method." *Methods* 25(4):402–8.

[100] Logullo, Angela Flavia, Mônica Maria Ágata Stiepcich, Cintia Aparecida Bueno de Toledo Osório, Sueli Nonogaki, Fátima Solange Pasini, Rafael Malagoli Rocha, Fernando Augusto Soares and Maria M Brentani. 2011. "Role of Fos-related antigen 1 in the progression and prognosis of ductal breast carcinoma." *Histopathology* 58(4):617–25.

[101] Lopez-Garcia, Maria A, Felipe C Geyer, Magali Lacroix-Triki, Caterina Marchió and Jorge S Reis-Filho. 2010. "Breast cancer precursors revisited: molecular features and progression pathways." *Histopathology* 57(2):171–92.

[102] López-Lázaro, Miguel. 2010. "A new view of carcinogenesis and an alternative approach to cancer therapy." *Mol Med* 16(3-4):144–53.

[103] Luger, K, A W Mäder, R K Richmond, D F Sargent and T J Richmond. 1997. "Crystal structure of the nucleosome core particle at 2.8 A resolution." *Nature* 389(6648):251–60.

[104] Lupien, Mathieu, Jérôme Eeckhoute, Clifford A Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S Carroll, X Shirley Liu and Myles Brown. 2008. "FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription." *Cell* 132(6):958–70.

[105] Maltzman, J S, J A Carman and J G Monroe. 1996. "Role of EGR1 in regulation of stimulus-dependent CD44 transcription in B lymphocytes." *Mol Cell Biol* 16(5):2283–94.

[106] Marella, Narasimharao V, Kishore S Malyavantham, Jianmin Wang, Sei-ichi Matsui, Ping Liang and Ronald Berezney. 2009. "Cytogenetic and cDNA microarray expression analysis of MCF10 human breast cancer progression cell lines." *Cancer Res* 69(14):5946–53.

[107] Martens, Joseph A and Fred Winston. 2003. "Recent advances in understanding chromatin remodeling by Swi/Snf complexes." *Curr Opin Genet Dev* 13(2):136–42.

[108] McGhee, J D, W I Wood, M Dolan, J D Engel and G Felsenfeld. 1981. "A 200 base pair region at the 5' end of the chicken adult beta-globin gene is accessible to nuclease digestion." *Cell* 27(1 Pt 2):45–55.

[109] Mishra, Jyoti P, Sasmita Mishra, Katrina Gee and Ashok Kumar. 2005. "Differential involvement of calmodulin-dependent protein kinase II-activated AP-1 and c-Jun N-terminal kinase-activated EGR-1 signaling pathways in tumor necrosis factor-alpha and lipopolysaccharide-induced CD44 expression in human monocytic cells." *J Biol Chem* 280(29):26825–37.

[110] Morse, R H. 2000. "RAP, RAP, open up! New wrinkles for RAP1 in yeast." *Trends Genet* 16(2):51–3.

[111] Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer and Barbara Wold. 2008. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nat Methods* 5(7):621–8.

[112] Mulligan, Anna Marie, Dushanthi Pinnaduwage, Shelley B Bull, Frances P O'Malley and Irene L Andrulis. 2008. "Prognostic effect of basal-like breast cancers is time dependent: evidence from tissue microarray studies on a lymph node-negative cohort." *Clin Cancer Res* 14(13):4168–74.

[113] Nagy, Peter L, Michael L Cleary, Patrick O Brown and Jason D Lieb. 2003. "Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin." *Proc Natl Acad Sci U S A* 100(11):6364–9.

[114] Ngwenya, Sharon and Stephen Safe. 2003. "Cell context-dependent differences in the induction of E2F-1 gene expression by 17 beta-estradiol in MCF-7 and ZR-75 cells." *Endocrinology* 144(5):1675–85.

[115] Niwa, H, J Miyazaki and A G Smith. 2000. "Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells." *Nat Genet* 24(4):372–6.

[116] Ohtsuki, M, S Flanagan, I M Freedberg and M Blumenberg. 1993. "A cluster of five nuclear proteins regulates keratin gene transcription." *Gene Expr* 3(2):201–13.

[117] Palacios, José, Emiliano Honrado, Ana Osorio, Alicia Cazorla, David Sarrió, Alicia Barroso, Sandra Rodríguez, Juan C Cigudosa, Orland Diez, Carmen Alonso, Enrique Lerma, Lydia Sánchez, Carmen Rivas and Javier Benítez. 2003. "Immunohistochemical characteristics defined by tissue microarray of hereditary breast cancer not attributable to BRCA1 or BRCA2 mutations: differences from breast carcinomas arising in BRCA1 and BRCA2 mutation carriers." *Clin Cancer Res* 9(10 Pt 1):3606–14.

[118] Park, Peter J. 2009. "ChIP-seq: advantages and challenges of a maturing technology." *Nat Rev Genet* 10(10):669–80.

[119] Parker, Joel S, Michael Mullins, Maggie C U Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F Quackenbush, Inge J Stijleman, Juan Palazzo, J S Marron, Andrew B Nobel, Elaine Mardis, Torsten O Nielsen, Matthew J Ellis, Charles M Perou and Philip S Bernard. 2009. "Supervised risk predictor of breast cancer based on intrinsic subtypes." *J Clin Oncol* 27(8):1160–7.

[120] Pepke, Shirley, Barbara Wold and Ali Mortazavi. 2009. "Computation for ChIP-seq and RNA-seq studies." *Nat Methods* 6(11 Suppl):S22–32.

[121] Perou, C M, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, P O Brown and D Botstein. 2000. "Molecular portraits of human breast tumours." *Nature* 406(6797):747–52.

[122] Polach, K J and J Widom. 1995. "Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation." *J Mol Biol* 254(2):130–49.

[123] Prat, Aleix, Joel S Parker, Olga Karginova, Cheng Fan, Chad Livasy, Jason I Herschkowitz, Xiaping He and Charles M Perou. 2010. "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer." *Breast Cancer Res* 12(5):R68.

[124] Pruitt, Kim D, Tatiana Tatusova, William Klimke and Donna R Maglott. 2009. "NCBI Reference Sequences: current status, policy and new initiatives." *Nucleic Acids Res* 37(Database issue):D32–6.

[125] Quail, Michael A, Iwanka Kozarewa, Frances Smith, Aylwyn Scally, Philip J Stephens, Richard Durbin, Harold Swerdlow and Daniel J Turner. 2008. "A large genome center's improvements to the Illumina sequencing system." *Nat Methods* 5(12):1005–10.

[126] Quinlan, Aaron R and Ira M Hall. 2010. "BEDTools: a flexible suite of utilities for comparing genomic features." *Bioinformatics* 26(6):841–2.

[127] Rakha, Emad A, Jorge S Reis-Filho and Ian O Ellis. 2008. "Basal-like breast cancer: a critical review." *J Clin Oncol* 26(15):2568–81.

[128] Rao, Bhargavi, Yoichiro Shibata, Brian D Strahl and Jason D Lieb. 2005. "Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide." *Mol Cell Biol* 25(21):9447–59.

[129] Rashid, Naim, Paul G Giresi, Joseph G Ibrahim, Wei Sun and Jason D Lieb. 2011. "ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions." *Genome Biol* 12(7):R67.

[130] Reich, Michael, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo and Jill P Mesirov. 2006. "GenePattern 2.0." *Nat Genet* 38(5):500–1.

[131] Reinke, Hans and Wolfram Hörz. 2003. "Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter." *Mol Cell* 11(6):1599–607.

[132] Reis-Filho, J S and A N J Tutt. 2008. "Triple negative tumours: a critical review." *Histopathology* 52(1):108–18.

[133] 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. 2000. "Genome-wide location and function of DNA binding proteins." *Science* 290(5500):2306–9.

[134] Rhead, Brooke, Donna Karolchik, Robert M Kuhn, Angie S Hinrichs, Ann S Zweig, Pauline A Fujita, Mark Diekhans, Kayla E Smith, Kate R Rosenbloom, Brian J Raney, Andy Pohl, Michael Pheasant, Laurence R Meyer, Katrina Learned, Fan Hsu, Jennifer Hillman-Jackson, Rachel A Harte, Belinda Giardine, Timothy R Dreszer, Hiram Clawson, Galt P Barber, David Haussler and W James Kent. 2010. "The UCSC Genome Browser database: update 2010." *Nucleic Acids Res* 38(Database issue):D613–9.

[135] Ross, Jeffrey S. 2009. "Multigene classifiers, prognostic factors, and predictors of breast cancer clinical outcome." *Adv Anat Pathol* 16(4):204–15.

[136] Rouzier, Roman, Charles M Perou, W Fraser Symmans, Nuhad Ibrahim, Massimo Cristofanilli, Keith Anderson, Kenneth R Hess, James Stec, Mark Ayers, Peter Wagner, Paolo Morandi, Chang Fan, Islam Rabiul, Jeffrey S Ross, Gabriel N Hortobagyi and Lajos Pusztai. 2005. "Breast cancer molecular subtypes respond differently to preoperative chemotherapy." *Clin Cancer Res* 11(16):5678–85.

[137] Rozowsky, Joel, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder and Mark B Gerstein. 2009. "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." *Nat Biotechnol* 27(1):66–75.

[138] Sabo, Peter J, Michael Hawrylycz, James C Wallace, Richard Humbert, Man Yu, Anthony Shafer, Janelle Kawamoto, Robert Hall, Joshua Mack, Michael O Dorschner, Michael McArthur and John A Stamatoyannopoulos. 2004. "Discovery of functional noncoding elements by digital analysis of chromatin structure." *Proc Natl Acad Sci U S A* 101(48):16837–42.

[139] Sabo, Peter J, Michael S Kuehn, Robert Thurman, Brett E Johnson, Ericka M Johnson, Hua Cao, Man Yu, Elizabeth Rosenzweig, Jeff Goldy, Andrew Haydock, Molly Weaver, Anthony Shafer, Kristin Lee, Fidencio Neri, Richard Humbert, Michael A Singer, Todd A Richmond, Michael O Dorschner, Michael McArthur, Michael Hawrylycz, Roland D Green, Patrick A Navas, William S Noble and John A Stamatoyannopoulos. 2006. "Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays." *Nat Methods* 3(7):511–8.

[140] Schwabish, Marc A and Kevin Struhl. 2004. "Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II." *Mol Cell Biol* 24(23):10111–7.

[141] Schwarz, G. 1978. "Estimating the dimension of a model." *The annals of statistics* 6(2):461–464.

[142] Segal, Eran, Nir Friedman, Daphne Koller and Aviv Regev. 2004. "A module map showing conditional activity of expression modules in cancer." *Nat Genet* 36(10):1090–8.

[143] Segal, Eran, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K Moore, Ji-Ping Z Wang and Jonathan Widom. 2006. "A genomic code for nucleosome positioning." *Nature* 442(7104):772–8.

[144] Sekinger, Edward A, Zarmik Moqtaderi and Kevin Struhl. 2005. "Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast." *Mol Cell* 18(6):735–48.

[145] Sengupta, Surojeet, Catherine G N Sharma and V Craig Jordan. 2010. "Estrogen regulation of X-box binding protein-1 and its role in estrogen induced growth of breast and endometrial cancer cells." *Horm Mol Biol Clin Investig* 2(2):235–243.

[146] Siamakpour-Reihani, Sharareh, Joseph Caster, Desh Bandhu Nepal, Andrew Courtwright, Eleanor Hilliard, Jerry Usary, David Ketelsen, David Darr, Xiang Jun Shen, Cam Patterson and Nancy Klauber-Demore. 2011. "The Role of Calcineurin/NFAT in SFRP2 Induced Angiogenesis-A Rationale for

Breast Cancer Treatment with the Calcineurin Inhibitor Tacrolimus." *PLoS One* 6(6):e20412.

[147] Sionov, R V and D Naor. 1998. "Calcium- and calmodulin-dependent PMA-activation of the CD44 adhesion molecule." *Cell Adhes Commun* 6(6):503–23.

[148] Smith, Andrew D, Zhenyu Xuan and Michael Q Zhang. 2008. "Using quality scores and longer reads improves accuracy of Solexa read mapping." *BMC Bioinformatics* 9:128.

[149] Smyth, Gordon K. 2004. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Stat Appl Genet Mol Biol* 3:Article3.

[150] Solomon, M J and A Varshavsky. 1985. "Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures." *Proc Natl Acad Sci U S A* 82(19):6470–4.

[151] Song, Lingyun, Zhancheng Zhang, Linda L Grasfeder, Alan P Boyle, Paul G Giresi, Bum-Kyu Lee, Nathan C Sheffield, Stefan Gräf, Mikael Huss, Damian Keefe, Zheng Liu, Darin London, Ryan M McDaniell, Yoichiro Shibata, Kimberly A Showers, Jeremy M Simon, Teresa Vales, Tianyuan Wang, Deborah Winter, Zhuzhu Zhang, Neil D Clarke, Ewan Birney, Vishy R Iyer, Gregory E Crawford, Jason D Lieb and Terrence S Furey. 2011. "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity." *Genome Res* .

[152] Sopel, M. 2010. "The myoepithelial cell: its role in normal mammary glands and breast cancer." *Folia Morphol (Warsz)* 69(1):1–14.

[153] Sørlie, T, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P Eystein Lønning and A L Børresen-Dale. 2001. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." *Proc Natl Acad Sci U S A* 98(19):10869–74.

[154] Soule, H D, T M Maloney, S R Wolman, W D Peterson, Jr, R Brenz, C M McGrath, J Russo, R J Pauley, R F Jones and S C Brooks. 1990. "Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10." *Cancer Res* 50(18):6075–86.

[155] Sudarsanam, P and F Winston. 2000. "The Swi/Snf family nucleosome-remodeling complexes and transcriptional control." *Trends Genet* 16(8):345–51.

[156] Tirkkonen, M, M Tanner, R Karhu, A Kallioniemi, J Isola and O P Kallioniemi. 1998. "Molecular cytogenetics of primary breast cancer by CGH." *Genes Chromosomes Cancer* 21(3):177–84.

[157] Trinklein, Nathan D, Shelley J Force Aldred, Alok J Saldanha and Richard M Myers. 2003. "Identification and functional analysis of human transcriptional promoters." *Genome Res* 13(2):308–12.

[158] Tsukiyama, T and C Wu. 1995. "Purification and properties of an ATP-dependent nucleosome remodeling factor." *Cell* 83(6):1011–20.

[159] *U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2007 Incidence and Mortality Web-based Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute*. 2011.
URL: *http://www.cdc.gov/cancer/dcpc/data/women.htm*

[160] van Beers, Erik H and Petra M Nederlof. 2006. "Array-CGH and breast cancer." *Breast Cancer Res* 8(3):210.

[161] van Leeuwen, Fred and Bas van Steensel. 2005. "Histone modifications: from genome-wide maps to functional insights." *Genome Biol* 6(6):113.

[162] Varga-Weisz, P. 2001. "ATP-dependent chromatin remodeling factors: nucleosome shufflers with many missions." *Oncogene* 20(24):3076–85.

[163] Visvader, Jane E. 2009. "Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis." *Genes Dev* 23(22):2563–77.

[164] Visvader, Jane E. 2011. "Cells of origin in cancer." *Nature* 469(7330):314–22.

[165] Wallrath, L L, Q Lu, H Granok and S C Elgin. 1994. "Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures." *Bioessays* 16(3):165–70.

[166] Weinberg, R A. 1991. "Tumor suppressor genes." *Science* 254(5035):1138–46.

[167] Wong, G K, D A Passey and J Yu. 2001. "Most of the human genome is transcribed." *Genome Res* 11(12):1975–7.

[168] Wu, C. 1980. "The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I." *Nature* 286(5776):854–60.

[169] Wu, C, Y C Wong and S C Elgin. 1979. "The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity." *Cell* 16(4):807–14.

[170] Xie, Chao and Martti T Tammi. 2009. "CNV-seq, a new method to detect copy number variation using high-throughput sequencing." *BMC Bioinformatics* 10:80.

[171] Yu, L and R H Morse. 1999. "Chromatin opening and transactivator potentiation by RAP1 in Saccharomyces cerevisiae." *Mol Cell Biol* 19(8):5279–88.

[172] Yuan, Guo-Cheng, Yuen-Jong Liu, Michael F Dion, Michael D Slack, Lani F Wu, Steven J Altschuler and Oliver J Rando. 2005. "Genome-scale identification of nucleosome positions in S. cerevisiae." *Science* 309(5734):626–30.

[173] Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li and X Shirley Liu. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol* 9(9):R137.

[174] Zhang, Zhengdong D, Joel Rozowsky, Hugo Y K Lam, Jiang Du, Michael Snyder and Mark Gerstein. 2007. "Tilescope: online analysis pipeline for high-density tiling microarray data." *Genome Biol* 8(5):R81.