Title
Glucocorticoid receptor DNA occupancy and transcriptional regulation across cell types

Author
Cooper, Samantha B.

Publication Date
2010

Peer reviewed|Thesis/dissertation

Glucocorticoid receptor DNA occupancy and transcriptional regulation across cell types

by

Samantha Cooper

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgments

Keith R. Yamamoto and R. Kip Guy guided me through my graduate career and supported me during the personal tumult of these last few years. Graduate school feels like a long slog during the slow periods, and their support and encouragement dragged me out of occasional scientific doldrums.

In the Guy lab, I had the pleasure of working with Peter Madrid, Anang Shelat, Jeremy Mallari, and David Smithson. They and the others in the lab created a lively, engaging environment for scientific exchange. Kip regularly made rounds through the lab, advising and encouraging his students. I don't think I have ever heard anyone who has worked with Kip give any less than a glowingly positive account of their interactions with him, and I second all those reviews. I was distraught when Kip announced his plans to leave UCSF for a position at St. Jude Children's Research Hospital, but Kip helped me make the transition to a new lab and he continued to support and advise me for the rest of my graduate career. It is a measure of the devotion of Kip's students, post-docs, and other staff that just about everyone else in the lab followed him to Memphis. I was the only student who stayed behind, and if I didn't have a husband and a house tying me to the Bay Area, I might well have followed also.

Keith's lab turned out to be an ideal new home after the dissolution of the Guy

lab. It was an honor and a joy to work with Keith. He let me chart my own scientific course but kept me focused on the important questions. He always managed to restate my ideas in ways that made them sound much cleverer than they were to begin with. Keith's warmth, his love of science, and his commitment to mentorship made him a superb advisor.

The best thing about the Yamamoto lab, besides Keith himself, was the extraordinary people Keith filled the lab with. I was honored to be among them, and they contributed enormously to my scientific development and to my enjoyment of graduate school. My baymates, Eric Bolton, Stefan Taubert, and Miles Pufall, patiently fielded many basic molecular biology questions and provided both scientific advice and companionship during long hours in the lab. I will miss Eric's dry sense of humor and his stack of qPCR plates, as well as his bottomless knowledge of laboratory methods and resources on the UCSF campus. Miles holds himself and others to a high standard of both scientific rigor and good lab citizenship, and he took a leadership role in the lab almost as soon as he joined. His gentle and not-so-gentle ribbing produced many smiles but did little to hide his genuine warmth and caring for the people around him. Stefan was generous with his knowledge, advice, and laughter, and our discussions about science made me a more sophisticated consumer of the scientific literature.

Outside my baymates, the other major influences on my work in the lab were Sebastiaan Meijsing, whose advice, incisive comments, and help with methods were enormously valuable; Meghan Holdorf, whose friendship and companionship in the trenches made my progress swifter and more fun; Tony Gerber, who provided a sounding board for ideas and computational approaches. Wally Wang, Carlos Pantoja,

Lisa Watson, Karin Buser, Jason Huff, Alex So, Brian Feldman, and Jordan Ward all provided ideas and discussions that kept me thinking and learning throughout my graduate years.

No discussion of the Yamamoto lab would be complete without a bow to Jenny Banaszek. She is the person who made Keith accessible to all of us. She also kept me smiling with many talks of child rearing and visions of what to expect to expect in the years ahead, and she sympathized with the rigors of working all day after comforting children for half the night. Before Jenny, Val Dougherty filled the role of keeping Keith in line, and, like Jenny, she managed to command the respect, fondness, and gratitude of everyone in the lab. I know that both Jenny and Val had multiple responsibilities beyond fitting graduate students onto Keith's schedule, but they always managed to cheerfully squeeze me in when I needed a little extra help or support.

The Bioinformatics program at UCSF would not be the same without Patsy Babbitt, my academic advisor. She provided wise counsel throughout my academic career and was a central force shaping the program. Julia Molla, the program coordinator, kept things humming smoothly and was endlessly patient with my lack of attention to administrative details. My colleagues in the Bioinformatics program, including Alex Adai, Holly Atkinson, Brian Tuch, Debbie Lin, David Williamson, Mike Kim, and Libusha Kelly, added enormously to my educational experience. Mike and Libusha, in particular, are superb speakers who helped me improve my speaking skills.

My work was strongly influenced by my association with Joe DeRisi's lab. I was always drawn to the work in his lab, and quickly become something of a DeRisi lab groupie. As I started doing extensive microarray work, I became a frequent visitor

to his lab and to the Center for Advanced Technology, a shared facility that Joe had a major role in forming. I eventually started sitting in on the Derisi lab meetings to stay current with the latest technology, which Joe's lab was always on the forefront of adopting and improving upon. My work was strongly influenced my interactions with Joe and with members of the his lab, particularly Katherine Sorber, Michelle Dimon, and Kael Fischer. Hao Li and members of his lab, including Brian Tuch and Christina Chaivorapol also provided invaluable insights at critical junctures in my research.

The Center for Advanced Technology (CAT) was instrumental in my graduate career. Adam Carroll, the director when I first start at UCSF, was a font of useful knowledge. He provided a sounding board for my experimental plans and kept me up to date with the latest tools in the CAT. Very early in my graduate school career, when I had never even used a spin column to purify DNA, he patiently taught me how to run microarrays. He and Paige Nittler, who later took over the running of the CAT, shepherded me through my first MEEBO microarray print run. Paige, like Adam, is both an invaluable scientific resource and a good friend.

No discussion of my scientific development would be complete without mention of Ellen Judd, my friend, study partner, and housemate during my formative college years in the Berkeley Physics department. I have no illusions that I would have made it through those first grueling three semesters of honors physics and two semesters of quantum mechanics classes without Ellen. Those first two years of physics left me with the confidence that, no matter how hard a subject appeared on first glance, with sufficient determination I could master it. Without Ellen, I may well have been defeated by those early classes and walked away with a very different sense of myself.

There is one early science lesson that stands out in my memory. In 8th grade, in Mr. Zakrzewski's science class, I was introduced to the connection between math and science. It was the first time that we were really asked to connect the physical world with the mathematical world, and it initially stumped me and many of my classmates. A short tutoring session from my father brought complete clarity to the subject, and ultimately I was one of the few students in the class who had a firm grasp of the material. I am fairly certain that this is the only time in my entire school career that I got any help with homework from my father, but when I look back it stands out as a turning in my intellectual development. For the record, it is my mother who deserves credit for every single other academic assist over the years. Beyond this, it goes almost without saying that I owe to my parents everything that I have achieved and all that is best in my life.

Without a doubt, the most important influence on my graduate career was my husband, Philip Soffer. He helped me make the decision to leave the professional world and return to graduate school. He supported me financially and emotionally throughout. He kept me focused on the goal (a degree and a return to industry) and on my strengths (the interface of technology, math, and biology). He coached me in presentation skills and in my elevator pitch. His confidence in me may often be undeserved but it is always appreciated. And his tolerance of my periodic lapses into work obsession, and to my failure during these times to attend to the most basic daily tasks of life, is invaluable.

No doubt the most significant drain on my productivity during graduate school has been my two pregnancies and three resultant children. While I cannot claim that they helped my scientific achievement in any way, Aaron, Caroline, and Josiah are

no doubt the best product of the last few years. Their shrieks, their laughter, and the pitter-patter of their feet fill our house now, and they remind me every day that science is about making the world a better place for our children and our children's children.

I cannot close my summation of the last half dozen years without remembering and acknowledging my sister-in-law and friend, Abigail Soffer. Her loss is a shadow that hangs over the last six years, and her absence will always be with me. She made me a gentler, more compassionate person, and she showed me how precious a gift happiness is. It is to Abby that I dedicate this work.

Abstract

In response to glucocorticoids, the glucocorticoid receptor induces a cell type specific transcriptional program through both direct and indirect interactions with glucocorticoid response elements (GREs). The association between GREs and regulated genes is often unclear, because a GRE can be tens of thousands of base pairs from the gene it regulates. To help understand the mechanisms driving GR-DNA binding and transcription, we looked at the genome-wide binding and transcriptional response to glucocorticoids in three cell lines. Consistent with previous reports, we found that binding is more closely associated with upregulation than with downregulation. Interestingly, we found that GR binding regions (GRBRs) that contain a canonical GR motif, called a GR binding sequence (GBS), are better predictors of upregulation than those lacking a GBS, while the reverse is true of downregulation. To help understand the determinants of cell type specific binding, we compared the sequences of the GRBRs in the three different cell types and found that binding motifs for both the glucocorticoid receptor and for other transcription factors are overrepresented to different degrees in each cell type. The canonical GR motif is present in 66% of the U2OS sites, verses 14% of the Nalm6 sites; similarly, the AP-1 binding motif is present in 22% of A549 sites and 6% of Nalm6 sites. This difference in the sequence composition of binding sites suggests that the determinants of binding may be encompassed

in the local sequence of a binding site. Supporting this, we found that the cell type specificity of binding could be recapitulated in reporters containing approximately 400 base pairs of DNA.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

The glucocorticoid receptor (GR) is a ligand activated transcription factor that is expressed in virtually every cell in the human body. Its ligand, cortisol, is a circulating hormone that varies in level throughout the day and in response to various external stimuli. Cortisol stimulation elicits different phenotypic responses in different cell types, including: gluconeogenesis in hepatocytes, suppression of bone formation by osteoblasts, differentiation or apoptosis in immature lymphocytes, and anti-inflammatory programs in immune system cells. It also induces apoptosis or proliferation in various cancer cell lines.

In the absence of ligand, GR resides in the cytoplasm bound to a chaperone complex. Once bound to ligand, GR translocates into the nucleus and binds directly to DNA or tethers to DNA through another DNA-bound transcription factor. GR can recruit cofactors that activate transcription through interactions with the basal transcription machinery or that remodel chromatin to enable binding by additional

transcription factors. GR can also repress transcription by recruiting corepressors or by displacing another DNA-bound or tethered transcription factor.

A genomic region occupied by GR in a given cellular context, called a GR binding region (GRBR), that confers a particular program of glucocorticoid regulation on an associated gene is referred to as a glucocorticoid response element (GRE). A GRE can be thousands or tens of thousands of base pairs from the transcription start site (TSS) of the gene it regulates.

In vitro, GR binds directly to a 15 base pair stretch of DNA, called a GR binding sequence (GBS), but in a genomic context a GR binding region is considerably more complex than this. GRBRs and other hormone receptor binding regions typically contain additional transcription factor binding sequences,[26, 53] consistent with reporter studies demonstrating that hormone response depends on both the GBS and the surrounding sequence.[37] Also, a significant fraction of GRBRs lack a GBS – 32% in one study [53] – and reporter studies demonstrate that such binding sites are often hormone responsive.[56, 31]

Before genome wide technologies, studies of GR function typically focused on a handful of genes, often chosen based on prior knowledge of what genes play a role in the phenotype associated with GR. Searches for DNA regulatory elements were typically confined to regions no more than a few thousand base pairs from TSS, or to regions that contained a canonical GR binding motif. More recently, microarrays have demonstrated that in most cell types hundreds of genes are direct targets of GR. Furthermore, unbiased searches for binding sites have demonstrated that many binding sites are far from the nearest transcription start site (TSS). Genome wide technologies also have better resolution at each individual binding sites; that is, the

location of GR binding is more readily apparent, and the extent and structure of the site is visible.

A wealth of data accumulated over many years has demonstrated that the different phenotypic programs enacted in different cell types is driven by differences in the transcriptional responses across cell types. However, the underlying drivers of the cell-specific transcriptional responses are not well characterized. In the current work, I aim to characterize how GR occupancy varies from cell type to cell type, to what degree these differences explain the differences in transcriptional response, and how the sequence composition of binding sites differs across cell types.

1.2 Methods

The methods relevant to this discussion are standard molecular biology techniques. They are described briefly here:

Gene Expression Microarrays - A ubiquitous technology for measuring the level of gene expression of all known genes. In this report, I used HEEBO arrays - glass slides spotted with 40,000 oligonucleotide probes corresponding to all the known human genes, with some genes covered by multiple probes. One of the limitations of this technology is that genes expressed at very low levels cannot be detected.

Chromatin Immunoprecipitation (ChIP) - A method for detecting the DNA binding sites of a protein of interest. Cells are treated with formaldehyde to crosslink the proteins to DNA, then lysed and sonicated to shear the DNA. An antibody specific to the protein of interest is used to immunoprecipitate the

protein and bound DNA fragments, then cross links are reversed. Typically qPCR is used to identify enriched fragments of DNA. ChIP can capture both direct and indirect GR-DNA interactions, though with different efficiencies. ChIP based methods offer only a snapshot of DNA occupancy at a moment in time in a population of cells.

ChIP-chip - A alternative to qPCR detection of enriched DNA fragments in a ChIP sample. The immunoprecipitated ChIP sample is amplified and then hybridized to a microarray containing probes for genomic DNA regions. Due to limitations in the number of probes that fit on a microarray, only part of the genome is typically probed with this technology.

ChIP-Seq - A more recent alternative to ChIP-chip. Rather than hybridizing the ChIP sample to a microarray, the sample is amplified and then sequenced with a deep sequencing technology, such as the Genome Analyzer from Illumina. This allows true genome-wide detection of binding sites at a reasonable cost. As with ChIP-chip, both the sonication and library preparation steps may introduce artifacts in the results. Careful controls help minimize these effects.

Luciferase reporters - A method for determining whether a stretch of DNA bound by a ligand activated transcription factor has regulatory capability in isolation. A stretch of DNA is cloned into a plasmid with a luciferase reporter. The plasmid is transfected into cells and the cells are treated with ligand or vehicle control. After 12-24 hours, the level of luciferase is assayed in treated and untreated cells. The ratio, normalized for independently measured tr ansfection efficiency, determines the regulatory strength of the cloned piece of DNA outside

of its native genomic context. In a reporter, the putative regulatory region is located close to the promoter, although in a genomic context the regulatory region may be thousands or tens of thousands of base pairs distant from the promotor. Also, reporters are not chromatinized in the same way as genomic DNA. Thus the regulatory activity of a stretch of DNA in a reporter is not definitive evidence of a regulatory role in the genome.

1.3 Mechanistic understanding of GR

A GR binding sequence (GBS) is a sequence motif that is bound with high affinity by GR in vitro. The motif is a 15 base pair imperfect palindrome to which GR binds as a dimer. Only a small fraction of the millions of GBSs in the genome are occupied by GR in a given cellular context, and many sites that do not contain a GBS are occupied by GR. Genomic sites occupied by GR are called glucocorticoid binding regions (GRBRs) and are specific to the cellular context. A primary glucocorticoid response element (GRE) is a GBR that confers a particular program of glucocorticoid regulation on a nearby gene. It is not clear how to definitively determine whether a given GBR acts as a GRE at a given gene. The most commonly used metric is a combination of proximity to the presumptive target gene and activity of the region of DNA in a reporter.

Some GREs contain multiple functional GBSs that operate in concert to produce the full transcriptional response. Wang et al. reported that the GRE 1500 base pairs upstream of the transcription start site of the strongly regulated GILZ gene contains four matches to the canonical GR motif. Individually mutating three of the four motif matches partially knocked down the transcription response of a luciferase

reporter, while mutating the fourth had no effect on the transcriptional response. Simultaneously mutating all three of the functional GBSs completely knocked out the transcriptional response of the reporter.[58] Similarly, So et al. demonstrated that knocking out one of the two GR motifs in a GRE only partially knocked down the transcriptional response, while mutating both completely knocked it down.[53]

As with the GILZ GRE, virtually all primary GREs are composite elements, composed of binding sites for GR as well as additional DNA-binding regulatory factors that act in conjunction with GR in the particular context of each particular GRE. Pearce et al. showed that at one such site the distance between GR and an AP-1 binding site governed the activity of that site. At one spacing, the GRE activated transcription, while at another spacing the GRE repressed activity or was non-functional.[37] Diamond et al. demonstrated that the same stretch of DNA could act as a negative or positive GRE, depending on the relative levels of expression of genes encoding subunits of the AP-1 family.[13]

In addition to direct recognition of a GBS motif, GR can occupy some GREs through tethering – that is, context-specific interaction with another protein that is in turn directly bound to DNA. Although tethering by GR was initially associated with repression,[10, 31] both tethering and direct binding have been shown to mediate both the activation and repression activities of GR.[34, 55]

While many reports distinguish between tethering of GR and direct interaction of GR with DNA, it is important to keep in mind that it can be difficult to distinguish these two modes of interaction. Even in the absence of a canonical GR motif, GR may bind to a degenerate sequence with assistance from a stabilizing interaction with another transcription factor. While some reports of tethering attempt to distinguish

between these two modes,(for example, [31]) others take the lack of a canonical GR motif as an indication that GR does not directly bind DNA.

Despite early studies of long-range regulation by transcriptional enhancers, and the realization that metazoan transcriptional regulatory elements generally operated in this mode,[60, 9] most studies of mammalian transcriptional regulatory elements, including GREs, focused on genomic DNA a kilobase or less upstream of transcription start sites. It is now clear from studies in drosophila that transcriptional regulation in metazoans commonly operates over long distances.[50, 11]. In the case of mammalian GR mediated regulation, Hakim and coworkers used chromosome conformation capture to demonstrate that a GRE near the Lcn2 promoter interacts with the Ciz1 gene nearly 30 kb away, apparently acting to regulate both genes.[19] This is particularly significant in light of the fact that ChIP-chip and ChIP-Seq studies repeatedly find many GRBRs to be located far from the transcription start sites of putative target genes.[53, 6, 26]

Additional and often neglected factors in the GR mediated response to glucocorticoids are the presence of multiple isoforms and multiple post translational modifications. The dominant isoform is GRalpha. Two alternative splice variants, GRbeta and GRgamma, may play a role in various diseases: GRbeta in glucocorticoid resistant asthma and GRgamma in childhood acute lymphoblastic leukemia.[14] Phosphorylation has been reported to affect nuclear localization and to exert gene-specific effects on regulation.[62] Sumoylation has been reported to modulate the transcriptional response in a context-dependent way.[20, 27]

1.4 Global DNA binding of GR

So et al. use ChIP-chip to map the GR-DNA binding sites in 100 kb regions surrounding the TSSs of a set of known transcriptional targets of GR and AR.[53] They found 73 GR binding sites, the majority of which were more than 10 kB from the nearest TSS. They demonstrated that genes regulated in A549 were significantly more likely to have an associated binding site than unregulated genes. This relationship held when comparing genes regulated in A549 to genes that are hormone responsive exclusively in another cell type. This suggests that cell-specific GR occupancy is the driver of cell-specific GR responsiveness.

Reddy et al. mapped GR occupancy and transcriptional response in A549 cells with ChIP-Seq and RNA-Seq (transcription profiling using deep sequencing). Similar to So et al, they find that the majority of sites are more than 10 kB from the nearest TSS. They also show that up regulated genes are much more likely than downregulated or unregulated genes to have an associating binding site. The strong correlation between the results in these two A549 data sets suggests that ChIP-Seq and ChIP-chip are comparable techniques.[40]

John et al. explain some of the cell-type specific differences in transcriptional response by comparing the chromatin landscape with GR- DNA occupancy. They find that the vast majority of GR binding sites lie in regions of DNA that are nuclease-accessible before hormone treatment, suggesting that the chromatin landscape determines much of the cell type specificity of the glucocorticoid response.[23] On the other hand, Visel et al. mapped cell type specific binding sites of the p300 protein and found that reporters recapitulated many of the cell type specific binding profiles. These results suggest that the cell type specific code is entirely encapsulated in short

stretches of DNA.[57]

Two closely related receptors, the estrogen receptor (ER) and androgen receptor (AR), have also been studied with ChIP-chip and ChIP-Seq. In the case of ER, Krum et al. compared binding sites on chromosomes 21 and 22 in MCF7 and USOS-ERalpha cells. They found only a 15% overlap between binding sites in the two cell types, similar to the level of overlap of regulated genes between the two cell types. The AP-1 motif was highly enriched in the binding sites in both cell types, whereas the FoxA1 motif was only enriched in MCF7 cells. They also showed that up regulated genes were enriched for nearby ERBRs, whereas downregulated genes were not. Their results support the idea that cell-specific recruitment is a major determinant of cell-specific gene regulation.[26]

Bolton et al. used ChIP-chip to identify androgen receptor binding sites in HPr-1AR cells, derived from normal human prostate epithelium. They found that that 69% of the AREs contained the motif canonical motif.[6] In contrast, Lin et al found with ChIP-Seq on PC3-AR, a prostate cancer cell line transfected with AR, that fewer that 3% of the AR binding sites contained the canonical AR motif.[29] The discrepancy between these two sets results may be due to the cell specific differences in AR binding, though it is also possible that the Lin et al study was tainted by a high number of false positives. Their published data contains a strikingly large number of 28 base pair sites, which may be an artifact of the ChIP-Seq technique. Since their raw data is not published, it is difficult to assess the quality of their results. Nonetheless, both studies identified additional motifs present in AREs, including AP-1, HNF-4alpha, OCT family proteins, PU.1.

1.5 Thesis outline

This thesis attempts to answer some basic questions about the cell-specific response to glucocorticoids.

Chapter 2 covers some of the technical issues in an increasingly popular genome-wide technique for detecting protein-DNA binding sites, ChIP-Seq. While based closely on the Illumina genomic DNA sample preparation method, this protocol uses reagents purchased from other sources in order to minimize expense. This method had some key advantages for the present study. First, it allowed the preparation of multiple libraries at a reasonable cost, so that the best libraries could be selected for sequencing. Second, it enabled a modification of the library preparation procedure for the addition of bar codes, which allows multiple samples to be sequenced in a single lane on a flow cell, again containing the total cost. This chapter also includes a detailed discussion of issues specific to working with the low concentration samples generated by ChIP.

Chapter 3 focuses on glucocorticoid induced GR-DNA binding (ChIP-Seq) and mRNA expression (microarrays) in three cell types. In order to understand the cell-type specificity of GR-mediated regulation, many studies have focused on global transcriptional responses alone. The premise of the work discussed here is that we need to move one step upstream, to the binding event that governs transcriptional response, and study common patterns in the relationship between cistrome and transcriptome across cell types, in order to understand the determinants of the transcriptional response to glucocorticoids. Understanding the cell-specific determinants of this first step is critical in understanding how GR ultimately enacts its cell-specific phenotypic effects. The connections between sequence and binding, and between binding and

transcription, are not simple: many GBSs are not bound by GR in vivo; binding does not always lead to transcription; binding sites may be far from the genes they regulate; many genes have multiple associated GR binding sites, only some subset of which may be significant in a given context. By comparing the cistrome and transcriptome across cell types, we can gain a better understanding of the relationship between sequence, binding, and transcription. The analysis of this data set integrates these three sources of information to help uncover some of the determinants of the GR mediated response and raises new questions that will require additional investigation.

Chapter 4 is a discussion of GR ligands that have different degrees of agonist/ antagonist behavior in different cell types. A set of ligands based on an arylpyrazole scaffold was developed by Shah et al. at UCSF [47] and demonstrated by Wang and coworkers in the Yamamoto lab to have desirable cell type-specific behavior.[59] That is, some of the ligands act as a stronger glucocorticoid in one cell type than in another. The proposed mechanism of the cell type specificity was differences in the strength of the transcriptional response at specific genes. For instance, ligand A induces a stronger response in gene X than ligand B, but a weaker response at gene Y than ligand B. The analogy to selective estrogen receptors is compelling: in the case of SERMs, the different conformational changes induced in the estrogen receptor by SERMs such as tamoxifen and relaxofine affect receptor-cofactor interactions, thereby producing an estrogen-like response at some genes but not at others, and thus an estrogenic response in some tissues but not in others. With these ideas in mind, the project discussed in this chapter was initiated with the intention of comparing the transcriptional responses induced by these arylpyrazole ligands. Preliminary results, combined with a reexamination of the results from the original paper, suggested that

the underlying premise of the project was weak. An alternate explanation for the cell specific effects of these GR ligands is discussed in this chapter, though further work will be required to determine which hypothesis is correct.

Chapter 2

ChIP-Seq sample prep and data analysis

2.1 Introduction

Chromatin immunoprecipitation followed by deep sequencing, or ChIP-Seq, is a relatively new method for detecting genome-wide binding sites of DNA binding proteins. This chapter will cover some of the issues in ChIP-Seq sample prep for the Illumina Genome Analyzer and some of the methods used in the data analysis. The focus will be on sample prep issues not covered elsewhere, including a sample quality control method designed specifically for ChIP samples.

2.2 ChIP-Seq Sample Prep

The goal of Illumina ChIP-Seq sample prep is to end up with a different piece of DNA on either side of each DNA fragment in the original sample. The basic procedure for

preparing double stranded DNA, starting with a sample containing DNA fragments of about 200 bp, is:

1. Blunt the ends of the DNA
2. Add an A overhang and a 5' phosphate group
3. Ligate on adaptors
4. PCR amplify the sample, priming off the adaptor sequences
5. Purify/size select to get the final library.

The particular difficultly in the preparation of libraries from ChIP samples is working with extremely low amounts of input DNA. For the ChIP libraries prepared in our lab using a polyclonal antibody for the glucocorticoid receptor, we typically have 50-500 pg of starting material. While it is possible to increase the amount of starting material by scaling up the ChIP reactions, we wanted to optimize the protocol for use on primary tissues where limited numbers of cells are available.

It is essential to start with a good ChIP sample. A good sample is sonicated or MNase digested to around 200 bp and shows strong enrichment for a known binding site over background. This enrichment requirement is discussed further below. (2.3)

It is important not to exceed a total of 18 PCR cycles in the entire library prep. If too much PCR is required to generate enough total material for a library, it means that there was insufficient starting material. Because not every molecule is amplified with equal efficiency, the final library will contain more copies of some molecules than of other molecules, leading to duplicate sequence reads on the Genome Analyzer. This is an artifact of sample prep, not a feature of the original sample, so these duplicates

are discarded before peak detection. In a good ChIP-Seq sample, 15-50% of the reads may be duplicates. In a sample that has been amplified too much, this number can be 95% or higher.

2.2.1 Adaptors

To save money and increase the flexibility of our library preparation, we chose to use our own library preparation protocol, based closely on published ChIP-Seq protocols, rather than the Illumina sample prep kit. Details of the protocol can be found at http://samanthacooper.dreamhosters.com/mywiki/index.php?title=SolexaChIP.

A schematic of the Solexa library preparation is shown below. INS represents a molecule of DNA from the starting sample – generally a fragment around 200 bp long. INS-A represents the insert with an A overhang added, as described in the sample prep protocol. The adaptors are shown as double stranded DNA with T overhangs that are ligated onto either side of INS. The primers for amplification are shown above and below the adaptors. The sequencing primer used during the Solexa sequencing run is also shown, though it is not used in the library preparation procedure.

```
                                                                    (PCR primer2)
                                                          3'TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC
                  5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT    INS-A P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG 3'
                  3' TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAG-P A-INS    TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC 5'
5'AATGATACGGCGACCACCGACACTCTTTCCCTACACGACGCTCTTCCGATCT
(PCR primer1)
(sequencing primer) 5'CACTCTTTCCCTACACGACGCTCTTCCGATCT
```

In order to run multiple biological replicates in a single lane of a Solexa flow cell, we added bar codes to the adaptors. The following schematic shows where a bar code (XXXX/YYYY) is placed in the adaptor. The insert and the right hand adaptor

are not shown here. Note that the bar code comes immediately after the sequencing primer, so it is always sequenced first. The bar codes we used are: XXXX(YYYY) = TCAT(ATGA), GACG(CGTC), AGTC(GACT), CTGA(TCAG).

```
                5'  ACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXT
                3'  TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAYYYY-P
5' AATGATACGGCGACCACCGACACTCTTTCCCTACACGACGCTCTTCCGATCT
(PCR primer1)
(sequencing primer) 5' CACTCTTTCCCTACACGACGCTCTTCCGATCT
```

We decided to join our adaptors together at the blunt ends. This means that the ligation reaction produces a circular piece of DNA. There are two reasons for doing this. One is to ensure that we get different Solexa adaptors on either end of our DNA (otherwise only 50% of the fragments will have different adaptors on both ends). The other is to increase ligation efficiency. The idea is that once one end of the adaptor ligates to a DNA fragment, circularization to ligate the other end is a very favorable reaction. Although this strategy has not been tested side by side with the Solexa sample prep kit, we have verified that it produced acceptable ChIP-Seq results.

To ensure that the PCR does not continue multiple times around the circularized DNA, we put a dU (deoxy-Uridine) between the two adaptors. After ligation, we digest the dU with UDG. This removes the uracil from the dU. DNA polymerase cannot cross the dU after digestion with UDG. This strategy is often used to prevent PCR cross contamination (some dU is included in all PCR reactions in the lab, and then UDG or UNG digestion is done at the start of every PCR reaction).

The adaptors used for sample prep are shown below. These are the sample sequences as shown above the sample prep schematic, but with the blunt ends connected with a dU deoxy-Uridine (represented as a U)

```
5’ P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG U ACACTCTTTCCCTACACGACGCTCTTCCGATCT
 3’ TCTAGCCTTCTCGAGCATACGGCAGAAGACGAAC U TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAG-P   5’
```

2.2.2 Ligation

The most critical reaction in the library prep is the ligation step. To optimize this reaction, we used the lowest possible reaction volume, we used a high concentration ligase, and we incubated the reaction overnight. It is important to use an appropriate ratio of adaptor to sample. Because there is inevitably some adaptor-adaptor ligation, too high a concentration of adaptor will result in a high fraction of adaptor-adaptor product. Then, when the sample is amplified with PCR, this adaptor-adaptor product will dominate the reaction, and very little of the desired product will be produced.

A 2:1 molar ratio or adaptors to insert (ChIP sample) seem to work well. To optimize this ratio for an antibody/target of interest:

1. Measure the sample concentration using Picogreen
2. Measure the adaptor concentration with Picogreen. Note that picogreen and the nanodrop yield different results for adaptor concentration, for unknown reasons.
3. Do ligation and PCR (18 cycles) at 1:1, 2:1, 5:1, and 10:1 ratios of adaptor to insert.
4. Run the product on a gel

The adaptor-adaptor product is 95-100 base pairs long. The library should be around 200-400 base pairs. The ideal adaptor to insert ratio is the one that yields the highest absolute quantity of library. The adaptor band can be almost completely removed using AMPure beads (Agencourt).

2.3 Sample quality control

This section discusses the theoretical detection limits of a ChIP-Seq run on a Solexa. Given a library prepared from a ChIP sample, it shows how to determine whether a given known binding site will be detectable in a sequencing run. Note that binding site enrichment generally increases after library preparation, so the analysis described here should be done on the final library rather than on the original ChIP sample.

Approach

Suppose that one lane on a flow cell yields 10 million reads, of which 5 million map uniquely to the human genome (standard numbers). Furher, suppose that at least 10 sequence reads from a given binding site are required to detect that site. For simplicity, assume that in the vicinity of a binding site, the sequence is sufficiently unique that all possible sequences map uniquely. Then, if using a cutoff of 10 sequence reads for calling a peak, we need at least 10 of the 10 million reads to be associated with that particular binding site. All things being equal, this means that at least 10 out of every 10 million molecules in the library sample must come from that binding site. So we need to calculate the fraction of DNA molecules in the library that come from a given site of interest. If this fraction is 1/1,000,000 and we get 10 million reads total, the site will be right at the limit of detection. As the fraction increases (say, to 1/100,000 = 100/10,000,000), the likelihood of detection increases.
The determine this fraction:

- Determine the molarity of DNA molecules in the sample (use a gel to estimate size and nanodrop to estimate concentration)

- Determine the molarity of DNA molecules that come from the binding site of interest (use qPCR with absolute standard)

Assumptions and approximations:

- We're taking an average fragment size to get the total number of DNA molecules.
- We're not taking into account that fact that some of the fragments pulled down in ChIP for a particular binding site will not be detected by the primer pairs we're using. This will lead to underestimating detectability. The degree that we underestimate depends on how well the primer pair is centered on the binding site and how short the amplicon is (centered and short is best).
- We're not taking into account the fact that adaptor-only sequences get preferentially sequenced. If many sequence reads are adaptor only, the denominator has to be adjusted accordingly

2.3.1 Calculations

Step 1) Calculate the molarity of DNA molecules in the library: First estimate the average fragment length by running the library on a gel or on an Agilent Bioanalyzer DNA chip. The average molecular weight of a base pair is 650, so 1 ng/ul (=0.001 g/l) of input DNA is equivalent to

$$\frac{10^{-3}}{(650 * fragment.length)}$$

Example: For an average fragment length of 350,

$$1ng/ul = \frac{10^{-3}}{(650 * 350)} = 4.4 * 10^{-9}M = 4.4nM$$

Thus the molar concentration of 10 ng/ul of DNA would be 44 nM.

Step 2) Determine the molarity of DNA molecules that map to a given binding site:

Use a primer pair that is capable of specifically detecting molecules that come from the given binding site (that is, a standard ChIP qPCR primer pair). Assume that the primer pair is virtually 100% efficient, a standard typically required of qPCR primers. Then include a control reaction that contains a known amount of purified amplicon with its corresponding primer pair. Use the starting concentration and the assumption of perfect efficiency to calculate the DNA concentration in the control reaction when it crosses the threshold (Ct). Then assume that all the reactions cross the Ct at the same total DNA concentration (should be true as long as starting DNA concentration is much less than the threshold concentration). Then use the amplicon length to calculate the molarity of the amplicon at this Ct. And then back calculate, based on perfect amplification efficiency, the starting molarity of the site of interest in the sample.

For controls, run a reaction with each amplicon and primer pair. You can optionally run a dilution series of both. But I've found that for amplicons of equal length, my dilution series show a perfect efficiency. I typically just one run amplicon at a starting concentration of 1e-7 ng/ul, and I load 0.1 ul in each reaction. I keep my amplicon stock at 0.1 ng/ul and dilute by 1000 x 2 just before running.

Calculation details

The calculation is shown in detail here. While this can be done more elegantly, it is written out in this form to make the logic easier to follow.
Definitions:

- ref.ct is the Ct of the reference amplicon.
- ref.conc is the starting concentration of the reference amplicon, in ng/ul
- ref.mol is the starting molarity of the reference amplicon, in moles/liters.
- library.mol is the molarity of DNA in your library (ng/ul), as calculated above.
- library.avelen is the average length of the DNA molecules in you library
- site.ct is the Ct of the site of interest.
- site.mol is the starting molarity of the site of interest in your library – this is what we're try to calculate
- site.len is the length of amplicon of the site of interest.

$$ref.conc * 2^{ref.ct} = thresh.conc \tag{2.1}$$

$$site.conc * 2^{site.ct} = thresh.conc \tag{2.2}$$

Calculations Combine equations (2.1) and (2.2):

$$ref.conc * 2^{ref.ct} = site.conc * 2^{site.ct}$$

Solve for site.conc:

$$site.conc = ref.conc * 2^{ref.ct - site.ct} \tag{2.3}$$

site.conc is in ng/ul. Convert it to molarity. The messiness will divide out in another couple steps.

$$site.mol = site.con * \frac{10^{-3}}{650 * site.len}$$

$$library.mol = library.con * \frac{10^{-3}}{650 * library.avelen}$$

$$\frac{site.mol}{library.mol} = \frac{site.conc}{library.conc} * \frac{library.avelen}{site.len} \tag{2.4}$$

Substitute 2.3 into site.conc in 2.4 to get:

$$\frac{site.mol}{library.mol} = \frac{ref.conc * 2^{ref.ct - site.ct}}{library.conc} * \frac{library.avelen}{site.len} \tag{2.5}$$

site.mol/library.mol should be greater than 10/10,000,000

An example, with some real numbers from a ChIP-Seq library:

- The reference amplicon (ref.conc) is at 1e-7 ng/ul
- The reference reaction had a Ct of 26 (ref.ct)
- The library with primers for the site of interest had a Ct of 16 (site.ct)
- The average length of the fragments in the library is 350 bp(library.avelen)
- The concentration of the library is 10 ng/ul (library.conc)
- The length of the amplicon for the site of interest is 200 bp

Plugging into (2.5) yields an estimate of 178 sequence reads from this binding site in a run that produces 10 million reads total.:

$$\frac{site.mol}{library.mol} = \frac{1 * 10^{-7} * 2^{(26-16)}}{10} * \frac{350}{200} = 1.8 * 10^{-5} = \frac{1}{56,000} = \frac{178}{10,000,000}$$

2.4 Data analysis

This section contains some details of the data analysis that are not covered elsewhere in this document.

2.4.1 Eland pipeline

Image files from the Genome Analyzer were run through the standard Illumina pipeline, including aligment to the human genome using Eland. For the bar coded samples, the first five bases were excluded from the Eland alignment, and then the resulting alignment file was split into four separate files corresponding to each barcode.

2.4.2 Gory details of peak finding algorithm

For peak detection, I wrote my own software in R. At the time this project was starting, the peak detection programs were significantly less mature, so this approach gave me the best control of the results. It also allowed additional flexibility in comparing data sets from different cell types. The false discovery rate using my algorithm, based on peak detection in input DNA, is 1-2%.

The peak finding software is written in R and can be run either interactively in

R or, more easily, from the unix command line. The peak finding algorithm does not use a control lane, but later steps in the processing routine use a control lane to filter for false positives. Here is the rationale for this decision to decouple peak finding and filtering against the control:

- The peaks can always be filtered subsequently using the data from the control lane so we don't lose anything by not using the control lane at this step
- It's not always clear how a control lane should be used. For instance, suppose that the control lane has significantly fewer or significantly more reads than the sample lane. How should this be adjusted for?
- It is useful to be able to try a few different thresholds for filtering out false positives. This is easier to do if we decouple peak detection and filtering based on the negative control.
- There are cases where we might want to use more than one sample as the negative control. For instance, we might have both an input DNA control, a no treatment control, and an IgG control. This is easy to do when we decouple the steps.
- I use some heuristics for eliminating false positives. This involves identifying regions of the genome that repeatedly have many hits that don't localize into peaks of the standard size.

Peak finding steps:

- Remove sequences that do not map to exactly one location in the human genome (with up to 2 errors allowed). In other words, throw out sequences that map to

multiple locations in the genome.

- Eliminate exact duplicate hits. The motivation here is that we assume that duplicates (that is, sequence reads that map to the same location and same strand) are artifacts of the PCR amplification, and that they just confuse peak detection.

- Remove hits that fall in certain "bad" regions of the genome, where "bad" regions were identified with a heuristic. These bad regions typically have many hits in different sample types, and the hits don't look like ChIP peaks. The regions can span thousands of base pairs. Typically about 2% of hits are removed in this step.

- Remove hits in satellite repeat regions. There's a strong overlap between "bad" regions and satellite repeat regions, so I elected to just remove all satellite repeats. Other types of repeats are not removed. Another 2% of hits are removed here.

- Identify regions that have at least T (threshold) number of hits within a region containing N base pairs. N=400 and T=10 are reasonable values.

- Go through the peaks one by one and trim each one down to its optimal size. This step is designed to avoid including one or two extra hits that are hundreds of base pairs away from the bulk of the hits.

Filtering peaks to remove false positives:

- Remove peaks of less than a 250 base pair width.

- For each peak, count the number of sequence reads in each of the three input DNAs that lie under the peak. Remove any peaks that have less than 4 times as many hits as any of the input DNAs.

Note that the input DNA is not scaled by the number of sequence reads. All the input DNAs have more unique sequences that map to the human genome than the corresponding input DNAs. Some peak finding algorithms would scale the input DNA down, but I do not think this is optimal. Also note that in this method, I use all three input DNA as independent controls for every one of the ChIP DNAs. This is designed to minimize false positives.

False discovery rate

To calculate the false discovery rate, I ran the peak detection algorithm of each of the three input DNAs in turn. For the filtering step, I used the other two input DNAs and the ChIP sample corresponding to the target input DNA. The false discovery rates are 1-2% for each of the samples (253, 79, and 208 peaks for U2OS, A549, and Nalm6 respectively), and there were no more than 5 peaks with more than 15 sequence reads in any of the input DNA samples.

We also ran a vehicle control ChIP sample. Only 10 peaks were detected in the control sample, and they were all removed in the filtering step.

Chapter 3

Global similarities and differences in GR occupancy and transcription across cell types

3.1 Abstract

In response to glucocorticoids, the glucocorticoid receptor induces a cell type specific transcriptional program through both direct and indirect interactions with glucocorticoid response elements (GREs). The association between GREs and regulated genes is often unclear, because a GRE can be tens of thousands of base pairs from the gene it regulates. To help understand the mechanisms driving GR-DNA binding and transcription, we looked at the genome-wide binding and transcriptional response to glucocorticoids in three cell lines. Consistent with previous reports, we found that binding is more closely associated with upregulation than with downregulation. Interestingly, we found that GR binding regions (GRBRs) that contain a canonical GR motif, called a GR binding sequence (GBS), are better predictors of upregulation than those lacking a GBS, while the reverse is true of downregulation. To help understand the determinants of cell type specific binding, we compared the sequences of the GRBRs in the three different cell types and found that binding motifs for both the glucocorticoid receptor and for other transcription factors are overrepresented to different degrees in each cell type. The canonical GR motif is present in 66% of the U2OS sites, verses 14% of the Nalm6 sites; similarly, the AP-1 binding motif is present in 22% of A549 sites and 6% of Nalm6 sites. This difference in the sequence composition of binding sites suggests that the determinants of binding may be encompassed in the local sequence of a binding site. Supporting this, we found that the cell type specificity of binding could be recapitulated in reporters containing approximately 400 base pairs of DNA.

3.2 Introduction

The glucocorticoid receptor (GR) is a ligand activated transcription factor that is expressed in virtually every cell type in the human body. Its ligand, cortisol, induces different phenotypic responses in different cell types, including: gluconeogenesis in hepatocytes, suppression of bone formation by osteoblasts, differentiation or apoptosis in immature lymphocytes, and anti-inflammatory programs in immune system cells.

In the absence of ligand, GR resides in the cytoplasm bound to a chaperone complex. Once bound to ligand, GR translocates into the nucleus and binds directly to DNA or tethers to DNA through another DNA bound transcription factor. GR can recruit cofactors that activate transcription through interactions with the basal transcription machinery or that remodel chromatin to enable binding by additional transcription factors. GR can also repress transcription by recruiting corepressors or by displacing another DNA-bound or tethered transcription factor.

A genomic region occupied by GR in a given cellular context, called a GR binding region (GRBR), that confers a particular program of glucocorticoid regulation on an associated gene is referred to as a glucocorticoid response element (GRE). A GRE can be thousands or tens of thousands of base pairs from the transcription start site (TSS) of the gene it regulates.

GR binds directly to a 15 base pair stretch of DNA, called a GR binding sequence (GBS). So et al found that 32% of the GRBRs lack a GBS,[53] consistent with numerous reporter studies demonstrating a hormone response from binding sites lacking a GBS. GRBRs and other hormone receptor binding regions typically contain additional transcription factor binding sequences, [26, 53] also consistent with many reporter studies demonstrating that hormone response depends on both the GBS and

the surrounding sequence.

Various studies have looked at genome wide binding and transcription of nuclear hormone receptors. [53, 40, 26, 6] These studies have various common themes: upregulated genes are more likely than unregulated genes to have nearby binding sites, but the same is not necessarily true of downregulated genes; the majority of sites have a canonical hormone receptor binding motif, or GBS; additional cofactor motifs are overrepresented in the set of binding sites; the majority of binding sites are more than 10 kB from a TSS; binding is more prevalent than regulation.

To our knowledge, only one study directly compared binding and regulation across multiple cell types. Krum et al. compared ER alpha binding regions (ERBRs) on chromosomes 21 and 22 in MCF7 and USOS-ERalpha cells. They found a 15% overlap between ERBRs in the two cell types, similar to the level of overlap of regulated genes between the two cell types. The AP-1 motif was highly enriched in the binding sites in both cell types, whereas the FoxA1 motif was only enriched in MCF7 cells. They also showed that upregulated genes were enriched for nearby ERBRs, whereas downregulated genes were not. Their results support the idea that cell-specific recruitment is a major determinant of cell-specific gene regulation. [26]

It remains unclear why only a subset of GBSs act as GRBRs in a given context, and why only a fraction of GRBRs act as GREs. Comparing across cell types allows us to consider these questions by looking for common themes in GRBRs and GREs in different contexts. In the present study, we assess genome-wide GR occupancy with ChIP-Seq and transcriptional response with microarrays in three cell types. We use these data to look at the relationship between sequence and GRBRs and between GRBRs and regulation. Ultimately this type of cross cell type analysis will be crucial

in understanding the tissue specific effects of corticosteroids.

3.3 Results

3.3.1 Overview of ChIP-Seq

To compare GR occupancy and glucocorticoid responsiveness in different cell types, we used three GR-expressing human cell lines: A549, derived from a lung carcinoma; U2OS-GR, derived from a bone osteosarcoma; and Nalm6, derived from acute lymphoblastic leukemia. We used microarrays to identify genes regulated by glucocorticoids and ChIP-Seq to identify GR occupancy.

ChIP-Seq replicates

We performed chromatin immunoprecipitation (ChIP) with a GR antibody, then library preparation based on the Illumina protocol (with some variations described in materials and methods), followed by deep sequencing on a Genome Analyzer II (ChIP-Seq, [24]). We did biological replicates for all three cell types, as described below. As controls, we used input DNA for each cell type. We threw out all duplicate sequence reads and all reads that did not map uniquely to the human genome with 2 or fewer mismatches, and we refer to the set of sequence reads that remain as the usable sequence reads. We used a peak finding algorithm similar to that described by ([24]), but with additional criteria for filtering out false positives (materials and methods).

(3.1a) and (3.1b) demonstrate the consistency of the results from biological replicates. Each point corresponds to a ChIP-Seq peak, where the x coordinate is the

Cell type	ChIP-Seq tags	ChIP Peaks	Input DNA Peaks
U2OS	5.1 million	28,722	253
A549	1.4 million	4,332	79
Nalm6	3.9 million	17,771	208

Table 3.1: ChIP-Seq results: Number of usable sequence reads and number of peaks found in ChIP DNAs and input DNA controls.

Cell type	up-regulated	down-regulated
U2OS	1211	770
A549	380	257
Nalm6	153	27

Table 3.2: Microarray results: Number of genes up and down regulated in each cell types, at an adjusted p-value of 0.05 and absolute value of the log2 fold change greater than 0.6.

number of tags in one replicate and the y coordinate is the number in the other replicate. For the U2OS biological replicates, we ran independent U2OS ChIPs and sequencing on different days. One of the samples produced a much stronger signal, as indicated by the differences in the scale of the axes. Notably, the correlation is excellent between these two data sets, though the smaller data set produced about 1400 peaks while the larger data set produced 28,000. The data sets had, respectively, 0.68 and 5.1 million usable sequence reads, and 11.5% and 21.4% of these tags fell within peaks. These numbers represent differences in ChIP efficiency, library prep variability, and sequencing depth. The larger data set was used exclusively in the subsequent analysis.

For the Nalm6 and A549 (not shown) biological replicates, the independent ChIP samples were barcoded (materials and methods) and run in a single lane on a flow cell. (3.1b) is representative of the correlation in both sets of replicates. The Nalm6

replicates were relatively consistent in quality: 5792, 4704, 7141, 5373 peaks were identified in the replicates. Taking the first data set (with 5792 peaks) as a reference, we note that the other replicates identified 70%, 86%, and 75% of the 5792 peaks. This gives a rough estimate of the overlap to expect in similarly sized data sets. The A549 replicates were more variable in number of peaks identified (854, 1944,85, 991), but the correlations were similar to that shown in (3.1b). For the subsequent analysis of Nalm6 and A549, the data from all four biological replicates was pooled.

We identified more than 28,000 U2OS, 16,000 Nalm6, and 4,000 A549 binding sites with a false discovery rate of 1 to 2%. (table 3.1). Because the number of peaks identified in ChIP-Seq experiments is a function of ChIP efficiency, library preparation, and sequencing depth, the number of peaks in each cell type does not necessary reflect biological differences. The U2OS, A549, and Nalm6 data sets contained, respectively, 5.1, 1.4, and 3.9 million usable sequence reads, with 21%, 8.5%, and 18.5% of three reads, respectively, falling within one of the regions identified as a peaks.

Comparison to similar data sets

Using a cutoff of 10 sequence reads to define a binding region, we identified 83% (61/73), of the peaks found by ([53]). Lowering the detection threshold from 10 to 5 hits per peak, all but 3 of the 73 peaks (10.2, 13.1, and 17.4 in ([53]) were found in our data set.

GR occupancy across cell types

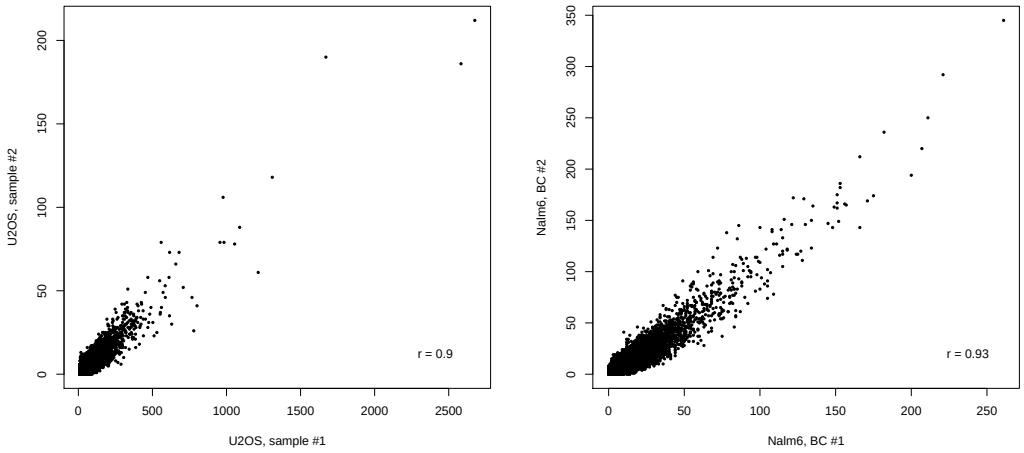
We found moderate overlap between binding sites in A549 and U2OS: 55% of the A549 binding sites were detected in U2OS and 9.5% of the U2OS binding sites were detected

in A549. The difference between the two numbers was due to the larger relative size of the U2OS data set: 28,000 versus 4,000 peaks, and 5.1 million versus 1.4 million usable sequence tags. To roughly estimate how big the overlap would be if the two data sets were of similar size, we applied to the U2OS data a detection threshold of 3.6 (=5.1 million/1.4 million) times our standard 10 hit threshold. Defining an overlap as those A549 binding regions that have at least 36 hits in U2OS, we find an overlap of 30%. This observation serves as a reminder of the slippery definition of an overlap when the applying cutoff thresholds to a data set.

There was strikingly little overlap between sites in Nalm6 and the two other cell types: 8.0% of Nalm6 sites were found in U2OS (6.0% adjusted for data set size as with the U2OS/A549 comparison) and 4.4% of U2OS sites were found in Nalm6. 5.7% of the A549 peaks were detected in Nalm6 (2.8% adjusted for data set size) and 1.8% of Nalm6 sites in A549. These overlaps were independent of distance from the nearest TSS (data not shown). Only 274 sites were identified in all the three cell types. Figures (3.1c) and (3.1d) show a visual comparison between the ChIP-Seq results in different data sets.

3.3.2 Stucture of peaks

To visualize the binding sites, we directionally extended each tag 200 bp, based on the average library fragment size, and calculated the overlap at each position as in ([42]). We opted to use the same extension length for all the data sets, rather than finding a best size based on each data set (described by [61]). While many sites had the classic triangular shape of ChIP binding sites, we noted numerous exceptions: sites with long tails and sites with two, three, and four peaks at varying distances



(a) U2OS vs U2OS

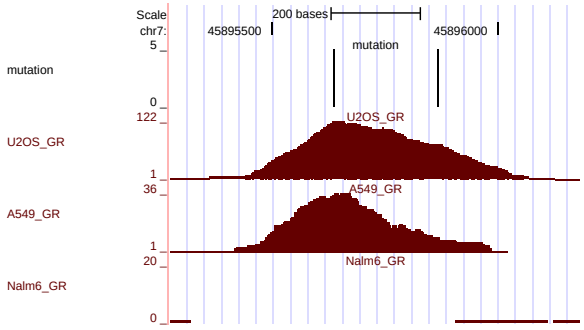
(b) Nalm6 vs Nalm6

(c) U2OS vs A549

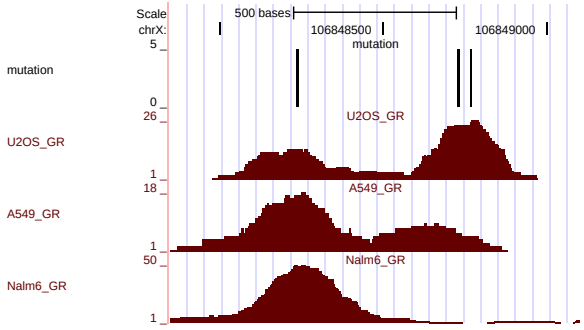
(d) U2OS vs Nalm6

Figure 3.1: ChIP-Seq versus ChIP-Seq plot: Each point represents the number of sequence reads in one peaks in one sample versus another sample. (a) Two biological replicates of U2OS ChIP, each run in one lane on a flow cell. (b) Two Nalm6 biological replicates, with immunoprecipitation and library prep done in parallel and run (with barcodes) in the same lane on a flow cell (c) U2OS vs. A549, calculated from peaks identified in A549 and (d) U2OS vs Nalm6, calculated from peaks identified in Nalm6.

apart. Two previous reports have shown that multiple GBSs contribute to the total hormone response at particular GREs in A549 cells.[58, 53] ChIP-Seq data for these GREs and the location of the GBSs previously shown to contribute to the hormone response are shown in figure 3.2. In this case, the long tail in (3.2a) and the double peak in (3.2b) reflect the regulatory contribution of multiple GBSs.



(a) IGFBP1



(b) GILZ

Figure 3.2: Binding sites with multiple functional GBSs: ChIP-Seq data from binding sites with previously published reporter data. Mutated sites that decrease reporter response are shown in the top track. For both these sites, complete knock-down of reporter function requires multiple mutations. (a) IGFBP1 binding site initially identified in ChIP-chip experiment on A549 cells ([53]) and (b) GILZ binding site initially identified in a ChIP scanning experiment. ([58])

We noted through visual exploration of our data in the UCSC genome browser that a significant fraction of the overlapping peaks in Nalm6 and the other two cell types had distinct profiles in Nalm6. To roughly quantify this, we counted the fraction of

overlapping peaks whose centers are at least 200 base pairs apart. We find that only 6% of the A549/U2Os overlapping peaks differ by at least this amount, while 25% of both the Nalm6/U2Os and Nalm6/A549 overlapping peaks exceed this threshold. Some examples are shown in the supplementary data (3.18).

3.3.3 Transcription

The transcriptional response (table 3.2) is most pronounced in U2OS cells, with 1969 regulated genes (adjusted p-value of 0.05 and absolute value of the fold change greater than 0.6), followed by A549 cells (637 regulated genes), and least pronounced in Nalm6 cells (180 regulated genes). Only 15% of the genes regulated in Nalm6 are repressed, whereas 39% and 40% of the U2OS and A549 genes, respectively, are repressed. Notably, the overall magnitude of the transcriptional response does not correlate with the total number of GR binding sites. While U2OS has both the strongest transcriptional response and the highest number of binding sites, Nalm6 has an intermediate level of binding sites but the weakest overall transcriptional response.

The overlaps of regulated genes are shown in figure 3.3. Some of the overlap is due to genes that are repressed in one cell type and activated in another. This accounts for 21 of the 68 overlapping genes Nalm6 and U2OS, 9 of the 36 overlapping Nalm6 and A549 genes, and 50 of the 294 overlapping A549 and U2OS genes. The 25 genes that are regulated in all three cell types are shown in supplementary figure 3.10.

Distribution of binding with respect to transcription start sites

For each cell type, we determined the distance of each binding site to the nearest annotated transcript, regardless of whether the transcript was regulated. Strikingly,

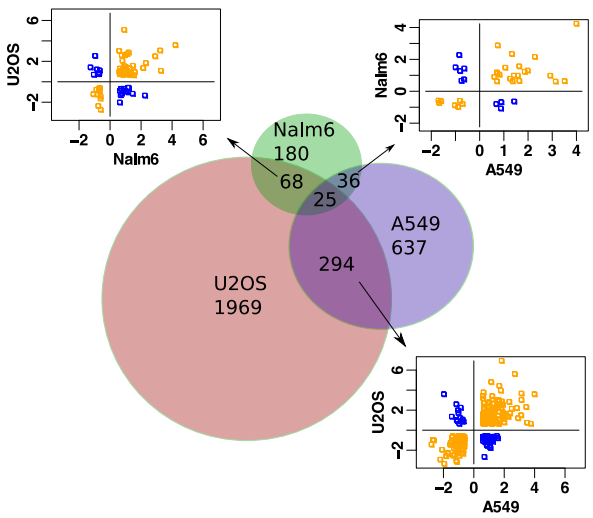


Figure 3.3: Overlap of transcriptional response to dexamethasone: Genes identified as regulated in each cell type. Some genes are up-regulated in one cell type and down-regulated in another.

we find that Nalm6 binding sites are much more likely to be located near a transcription start site (supplementary figure 3.11). 18% of Nalm6 binding sites are within 1,000 bp of a TSS, whereas the same is true for only 3.7% of A549 and 3.1% of U2OS binding sites. For a 10,000 bp window, these numbers are 42% for Nalm6, 25% for A549, and 20% for U2OS. This result could be due permissive binding facilitated by increased chromatin accessibility near transcription start sites, combined with greater detection sensitity of weak binding sites in Nalm6. Partially supporting this idea, we note that smaller peaks are more likely to be found near transcription start sites (supplementary figure 3.13). However, two features of this figure suggest that this is not the whole story: first, there is no trace of the relationship seen in Nalm6 between proximity to transcription start site and peak size in A549 or U2OS; and second, even the strongest Nalm6 peaks are more likely to be near a transcription start site than the peaks in A549 and U2OS.

Correlation of binding and transcription

To correlate binding and transcription, we first needed to choose an algorithm for counting how many binding sites were associated with a given gene. The best method for this was not obvious, for several reasons: enhancers may act at very long distances ([19]); HEEBO oligonucleotide microarrays do not distinguish among multiple transcripts for a single gene; and nuclear hormone receptor binding sites and responsive genes have been shown to occur in clusters along a chromosome, meaning that a single binding site may regulate multiple genes ([6]). We elected to use the following algorithm: for a given gene and a threshold window W, we counted the number of binding sites within W base pairs of the transcription start site of each of the transcripts for the gene. We selected the highest number of binding sites and assigned it to that gene for the window W. We repeated the calculation for windows from 1 kB to 250 kB, in 1 kB increments. Note that this algorithm allows a single peak to map to multiple genes.

We used receiver operator curves (ROCs) to show the predictive value of a binding site near a TSS for up or down regulation of that gene (3.4). In general, an ROC curve shows how the false positive rate and true positive rate change as some threshold is varied. In this case, the threshold that we varied was the window W around the TSS used for assigning a peak to a gene. The presence of a binding site within this window was used to predict that a gene would be regulated. Thus false positives were genes that had an associated binding site but were not regulated. The false positive rate was then fraction of unregulated genes that had an associated binding site (that is, the fraction of unregulated genes that were incorrectly predicted to be regulated). Similarly, true positives were regulated genes that had an associate binding site. The

true positive rate was the fraction of regulated genes that had an associated binding site (that is the fraction of genes that were correctly predicted to be regulated).

Another way to think about the ROC curve is to consider a fixed distance from the TSS, say 10 kb. In (3.4), the grey squares on the ROC curves and the associated inset bar graph show the false positive and true positive rates at a 10 kB threshold. The ROC curve is generated by changing this threshold from 0 kb (the point at (0,0) to 250 kb (the point at (1,1)). As the threshold changes, both the false positive and true positive rates increase.

In ROC curves, the diagonal line represents a completely uninformative predictor. An ideal predictor would extend all the way up to the upper left corner, with a true positive rate approaching 1 and a false positive rate near 0. The area under the ROC is one measure of the predictive power of the classification: as the area increases from 0.5 to 1, the predictive power increases.

Figure (3.4) shows that binding is, as expected, a good predictor of upregulation in all three cell types. It is a much poorer predictor, especially in A549 and U2OS, of downregulation. One important issue in generating these ROC curves is determining which genes should be included in the regulated and unregulated sets. For the curves shown here, we used the top 150 genes in each set. The supplmentary data shows similar results for genes selected with different stringency thresholds (figures 3.15 and 3.16). The unregulated genes are all the genes on the array that are not classified as either up or down regulated (at an adjusted p-value of 0.2). These curves show that binding is more predictive of upregulation than of downregulation in all three cell types and of the three cell types, binding is most predictive of downregulation in Nalm6.

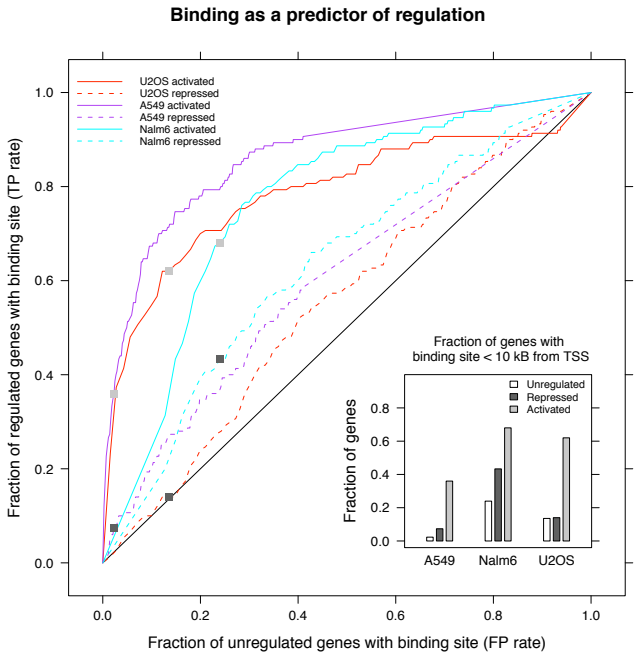


Figure 3.4: ROC curve using occurrence of a ChIP-Seq peak within different distances of the transcription start site. Solid red, purple, and cyan lines shows U2OS, A549, and Nalm6 upregulated genes, respectively. Dashed lines show down-regulated genes. The light and dark grey squares on the ROC curve, and the corresponding light and dark gray bars on the inset bar plot, represent a 10 kb window on either side of the transcription start site. The ROC curve is constructed by varying this window size and plotting the fraction of regulated verses unregulated genes with at least one binding site within that window of the TSS.

One noticeable feature of the Nalm6 ROC curve for upregulated genes is that it does not rise as sharply from the origin as the A549 and U2OS curves. As discussed earlier, Nalm6 binding sites are more likely to fall near a transcription start site than A549 or U2OS binding sites. The relatively low slope of the Nalm6 curve near the origin is due to the fact that the Nalm6 binding sites near transcription start sites are relatively less likely to be associated with a regulated gene. Figure (3.4 shows that

13% of unregulated genes in Nalm6 cells have a binding site within 1 kB of the TSS, while 31% of up-regulated and 20% of down-regulated genes have a binding within 1 kB. For comparison, the numbers in U2OS are 1.5% of unregulated genes, 23% of up-regulated, and 1.3% of down-regulated. We noted above that the binding sites near the TSSs in Nalm6 are on average smaller than the more distal sites. If we use a higher cutoff for calling a binding site in Nalm6, the Nalm6 ROC curve appears more like the others (supplementary figure 3.16)

A small fraction of genes with binding sites are regulated

We looked next at the fraction of genes with or without a nearby binding site that were regulated. Using a fixed window of 20 kB around the transcription start site, we found that only 19% of U2OS genes with an associated binding site were regulated, and 3.8% of the genes without a binding site were regulated. Equivalently, the positive predictive value of a binding site is 19% and the negative predictive value is 100%-3.8% = 96.2%. For A549, 19% of genes with a binding site were activated and 1.5% of genes without a binding site were activated. For Nalm6, these numbers were 2.1% and 0.37%.

While Nalm6 appears to be an outlier, we noted that these numbers were strongly dependent on the ChIP-Seq detection threshold used. We identified different detection thresholds in each cell type that produced more consistent results across cell types. In Nalm6, using a ChIP-Seq peak size threshold of 60 sequence reads (verses the default of 10) and a loose gene expression threshold (adjusted p-value of 0.1), we found that 12% of the genes with binding sites were regulated and 0.93% of genes without binding sites were regulated. Similarly, for an A549 and U2OS peak size

threshold of 10 and a stringent gene expression threshold (adjusted p-value of 0.01 and minimum of 1.8 fold change), the corresponding numbers were 13% and 0.82% in A549 and 13% and 2.2% in U2OS, similar to the Nalm6 results.

Binding and regulation in common across cell types

We observed that some genes that were regulated the same way in two cell types had different binding sites in those cell types. For example, thioredoxin-interacting protein (*txnip*) is activated in all three cell types. Looking in an 80 kb window around the transcription start site, *txnip* has a total of six associated binding sites across the three cell types, but different subsets of those sites appear in each cell type (supplemental figure 3.12). The sole overlapping binding site among all three cell types is the most distal one, 11 kb upstream of the TSS.

To gauge the extent to which binding sites near commonly regulated genes overlap, we looked at four sets of genes: genes that are not regulated in either cell type; genes that are up-regulated in two cell types; genes that are down-regulated in two cell types; and genes that are up-regulated in one cell type and down-regulated in another. For computational reasons, we used a simpler method of mapping peaks to genes: each peak is mapped to the gene with the closest transcription start site within an 80 kb window. The results were similar regardless of the window size used (not shown).

For each pair of cell types, we used two different methods to assess the overlap in binding in the four sets of genes: (1) we identified all the binding sites associated with the genes in the first cell type and counted how many of those were also bound by GR in the second cell type (table 3.3) and (2) we identified which genes had at least one associated binding site in the first cell type and then counted what fraction

	Nalm6-U2OS	Nalm6-A549	U2OS-A549
not reg	5.3% (170/3188)	1.1% (47/4300)	11% (247/2287)
up-up	27% (25/94)	21% (14/68)	34% (100/292)
down-down	5.9% (1/17)	- (0/12)	13% (8/62)
up-down	11% (5/45)	- (0/10)	35% (18/51)

Table 3.3: Overlap of binding sites near regulated and unregulated genes: For each pair of cell types, the percentage of binding sites that overlap in each set of genes is shown. The set of genes are: genes not regulated in either cell type (defined as absolute value of the log fold change less than 0.2); genes up regulated in both cell types; genes down regulated in both cell types; genes up-regulated in one cell type and down-regulated in the other. For entries where the denominator is less than 15, percentages are not shown.

of those genes at least one site that was also bound in the second cell type. 3.4)

To illustrate, we describe the calculations for two of the entries in tables 3.3 and 3.4. Of the 36 genes activated in both Nalm6 and U2OS, there are 94 Nalm6 binding sites near 26 of the genes. 25 of the 94 binding sites (27%, 2nd row, 1st column of 3.3) are also found in U2OS, compared to 170 of the 3188 binding sites (5.3%, 1st row, 1st column of 3.3) around genes that are not regulated in either cell types. Looking on a gene-by-gene basis, 14 of the 26 genes (54%, 2nd row, 1st column of 3.4) with at least one Nalm6 binding site have at least one binding site in common with U2OS, compared to 154/1918 (8.0%, 1st row, 1st column of 3.4) of the unregulated genes.

Overall, tables (3.3) and (3.4) show that there is more overlap between binding sites in two cell types near genes that are up-regulated in both the cell types than near unregulated genes. Although the numbers are small because relatively few down-regulated genes have nearby binding sites, it appears that the commonly down-regulated are less like than commonly up-regulated genes to share binding sites.

	Nalm6-U2OS		Nalm6-A549		U2OS-A549	
not reg	8.0%	(154/1918)	1.8%	(46/2544)	15%	(202/1373)
up-up	54%	(14/26)	50%	(8/16)	61%	(68/112)
down-down	-	(1/8)	-	(0/5)	18%	(7/38)
up-down	-	(3/12)	-	(0/7)	50%	(13/26)

Table 3.4: Overlap of at least one binding site near regulated and unregulated genes: For each pair of cell types, the percentage of genes in each set with at least one overlapping binding site is shown. Groups are defined as in table 3.3. Percentages are not shown where the denominator is less than 15.

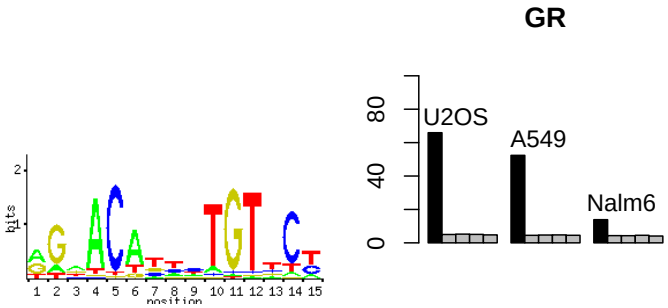
3.3.4 Motifs in binding sites

We did both an unbiased search for new motifs using Meme [3] and a search based on databases of known motifs using perl scripts and the TFBS package [28]. All of the motifs we identified in the Meme search were similar to a motif in the database, so we focus here on the searches starting with known databases.

For the database-based motif searches, we used 200 base pair windows around the center of the ChIP-Seq peak. The negative controls were 200 base pair windows offset by 0.5, 1, 2, and 10 kb from the ChIP-Seq peaks in each cell type. We defined a 5% false positive threshold based on finding 1 or more motif in at least 5% of the 2 kB offset U2OS sequences. The different sets of controls yielded a similar threshold cutoff value for the GR motif, but for a handful of motifs the Nalm6 controls were different due to the different sequence composition near transcription start sites. These motifs were discarded from further analysis.

Canonical GR motif

The canonical GR binding motif is consists of two 6 base pair palindromic half sites connected by a loosely specified three base pair linker. To determine whether cell



(a) canonical GR motif (b) % peaks with motif

Figure 3.5: Percentage of GR binding sites containing at least one GR motif in the central 200 base pairs. Black bars represent binding sites in indicated cell types. Gray bars represent motifs in control sites, which are 200 base pair windows offset by 0.5, 1, 2, and 10 kB from the ChIP-Seq peaks.

specific binding preference for different variants of the canonical GR motif accounted for any of the cell-type differences in the GR binding sites, we constructed a GR motif from each cell type independently. We searched for matches in the GR occupied sites to the canonical motif from ([53]) at the 5% false positive level and constructed a new motif for each cell type. As shown in (figure 3.14), we did not find obvious major differences in the motif.

We next asked what fraction of the GR binding sites in each of the cell types contains a canonical GR motif. We focused on a 200 bp window around the center of the peak of the binding site and used a 5% false positive threshold, defined as the motif score at which 5% of the 200 bp control sequences contain at least one motif. Figure 3.3.4 shows that the binging sites in U2OS cells are most likely to contain GR motifs (66%), followed by A549 (52%), and last by Nalm6 (14%) to contain GR motifs. We see similar trends with more and less stringent motif cutoffs (supplementary figure 3.17) and different sequence window sizes (not shown).

To determine whether the presence of a GR motif is correlated with the number

of tags in a binding site, we plotted the fraction of binding sites containing a motif as function of the number of tags in the binding site. We found (figure 3.6) that in U2OS cells, the fraction of sites with a motif plateaus at around 25 tags, while in A549 and Nalm6 sites, the fraction plateaus much later, if at all. One possible explanation for this data is that the fraction of peaks containing a canonical motif is a measure of the specificity of the peak-finding algorithm and smaller peaks are more likely to be false positives. Given that the increase in motifs continues up to at least 50 tags in A549 and at least 80 in Nalm6, this explanation is unlikely. Results were similar using a different peak finding algorithm (data not shown)([61]). A more probable explanation for these results is that weaker peaks are more likely to be tethering sites, where GR is bound to another factor that is itself bound to the DNA; these types of sites are thought to ChIP less efficiently. It is also consistent with the hypotheses that GR binds a highly degenerate form of the canonical motif or to a different motif altogether, albeit more weakly than the canonical motif.

To determine whether binding sites with the GR motif are more likely to be associated with regulated genes, we divided our peaks into two sets, one containing the canonical GR motif and the other lacking the motif. Then we plotted ROC curves for both sets of peaks and calculated the areas under the curves and above the diagonal (3.7). Interestingly, we found that peaks with canonical GR motifs are better predictors of activation than peaks without the motif in all three cell types, while peaks lacking the canonical motif are better predictors of repression. For this analysis, we used the top 150 genes in each set to have enough genes for statistical power but few enough to avoid secondary effects. Results are similar with more and less stringent sets of genes.

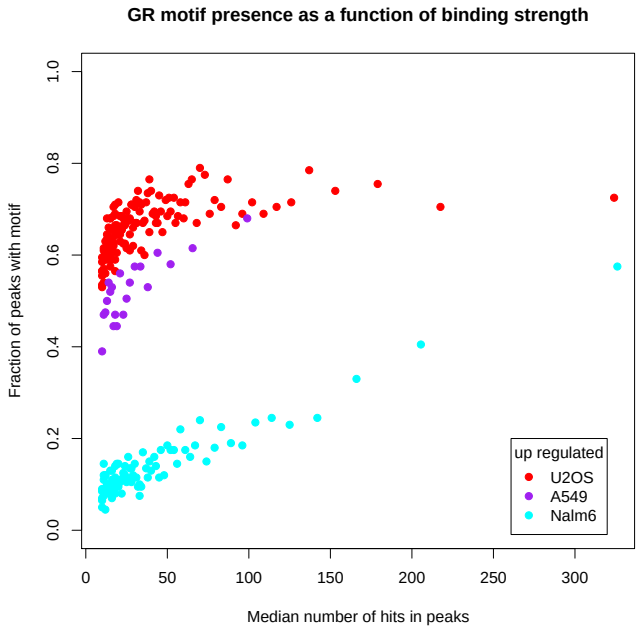


Figure 3.6: Fraction of binding sites with canonical motif: Peaks were ordered by number of sequence reads and binned into groups of 200. The median number of sequence reads in each group is shown on the x-axis and the fraction of peaks in the bin that contain the GR motif is shown on the y-axis. At all peak strengths, U2OS peaks are most likely to contain a motif, followed by A549 and then, at a much lower frequency, Nalm6.

Overrepresented motifs

To search for other overrepresented motifs in the GRBRs, we started with motifs from three different databases: Jaspar [45], Transfac 7.0 Public [32], and UniPROBE[1, 5]. We searched using the TFBS extension for BioPerl ([28]) for each motif at the 5% false positive level, defined as above, in the three ChIP-Seq data sets and in all the controls.

Figure 3.8 shows examples of motifs that are overrepresented to different degrees in the binding sites from the different cell types. Additional cell type specific motifs are shown in supplementary figures 3.20, 3.21, and 3.22. Since many transcription factors

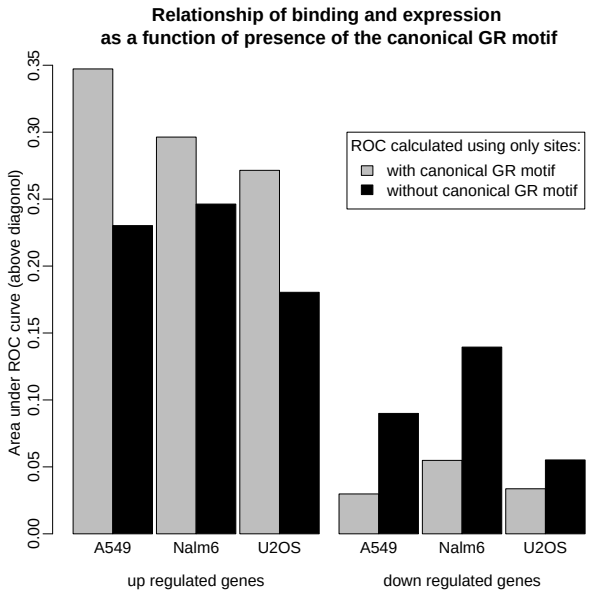
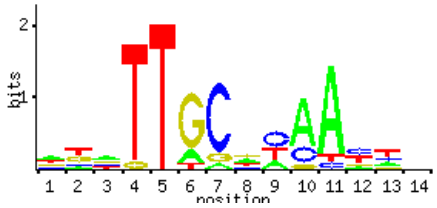
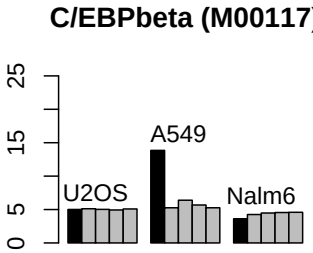


Figure 3.7: Area under under the ROC curve and above the diagonal. Black bars represent ROC curves generated using only the ChIP-Seq peaks containing canonical GR motifs, while grey bars represent curves generated using only peaks that do not contain the motif. For all the cell types, the area is higher (that is, the predictive capability is better) for up-regulated genes using only peaks that contain the GR motif. For down-regulated genes, the reverse is true in A549 and Nalm6 cells.

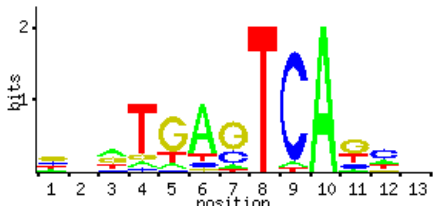
from within the same family have highly similar binding sites, it is difficult to identify which transcription factor binding site accounts for the observed overrepresentation, but it is clear that the composition of the binding sites is different in different cell types. We have looked at three additional ChIP-Seq data sets from four mouse cell lines (unpublished data, Wally Wang and Gordon Hager) and see a different subset of motifs overrepresented in each of those data sets, but with a high degree overlap with the motifs we identified in this study.



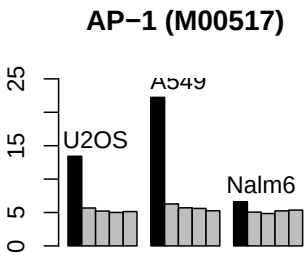
(a) Similar motifs: HLF



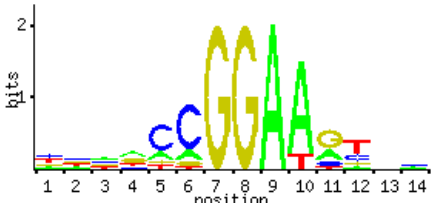
(b) % peaks with motif



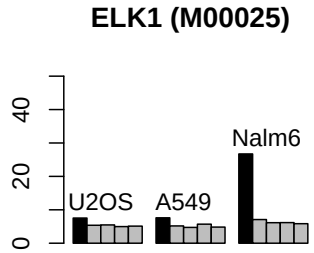
(c) Similar motifs: NFE2, Bach2, Fos, CREB



(d) % peaks with motif



(e) Similar motifs: c-ets-1(p54), GABPA, NRF-2, ELK4



(f) % peaks with motif

Figure 3.8: Selected overrepresented motifs. Black and grey bars are as described in figure 3.3.4. (a,b) C/EBPbeta is overrepresented only in A549. (e,f) AP-1 is overrepresented in U2OS and A549, but not in Nalm6. (g,h) ELK1 is overrepresented in Nalm6 only.

3.3.5 Reporters

To determine whether the cell-specificity of GR occupancy is determined solely by the local sequence composition of a binding site, we tested reporters corresponding to GRBRs specific to A549 and U2OS. We found that the reporter response tracked with the ChIP-Seq occupancy (3.9). The results in [23] suggest that chromatin structure is a major determinant of cell type specificity of GR occupancy. Since transiently transfected reporters are generally considered to be poorly chromatinized, our results suggest that either (1) cell type specific chromatin structure is not a major factor in differential binding or, alternatively, (2) that the relevant chromatin structure can be replicated in a reporter with a relatively short piece of DNA (about 400 bp).

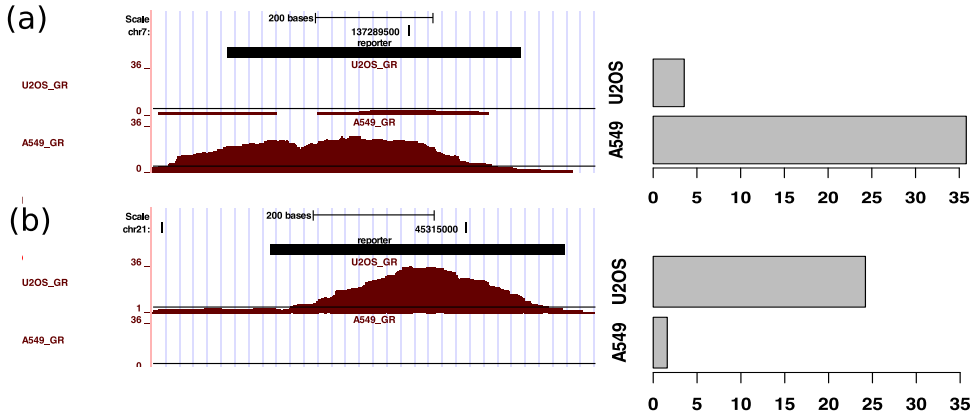


Figure 3.9: Cell type specific binding sites and corresponding reporters. Data is the average of two independent experiments. Both replicates are similar. (a) Reporter made from A549-specific binding site shows much stronger regulation in A549 than in U2OS. (b) Reporter made from U2OS-specific binding site shows much stronger regulation in U2OS than in the A549.

3.4 Discussion

The question of what governs the different GR-mediated responses in different cell types is fundamental to understanding the role of glucocorticoids in human physiology. We used a global approach to assess the GR-DNA occupancy and glucocorticoid-induced transcription across cell types. A comparison of the resultant data sets led to the following conclusions: (1) Binding is similarly predictive of upregulation and much less predictive of downregulation in all three cell types. Moreover, the presence of a GBS in a binding site improves the predictive power of the binding site for upregulation, but not for downregulation. (2) The majority of binding sites are not shared across cell types (3) The sequence composition of GBRSs is cell type specific, with different prevalences of the canonical motif and a different set of additional motifs in each cell type. However, the canonical GR motif is similar in all the cell types. (4) Genes that are activated in two cell types are more likely to share binding sites across the cell types than unregulated genes, but the same is not true of down-regulated genes. We discuss each of these conclusions in turn.

3.4.1 Relationship between binding and regulation

We looked at the relationship between binding and regulation in two different ways. First, we used ROC curves to compare the fraction of regulated genes that have binding sites (the TPR) to the fraction of unregulated genes that have binding sites (the FPR). Next, we compared the fraction of genes with binding sites that are regulated to the fraction of genes without binding sites that are regulated. These two methods give complementary information.

Using ROC curves, we varied the window around the transcription start site that

we used to assign a binding site to a gene. These plots helped us compare data sets with different levels of binding and regulation and enabled us to evaluate the effect of eliminating subsets of binding sites from consideration. In Nalm6 cells, we found that removing the smaller binding sites (the 58% of the peaks with 30 or fewer sequence reads) yielded a more informative ROC curve. In all the cell types, we found that binding sites that contain a canonical GR motif are more predictive of upregulation than binding sites that lack this motif. Taking a random subset of binding sites will decreases the area under the ROC curve, with a smaller subset causing a larger drop. Thus the fact that the relationship between motif presence and area under the ROC curves holds in all three cell types is particularly informative, because the fraction of sites with the motif ranges from 14% in Nalm6 to 66% in U2OS. Adding to the strength of this finding is the fact the trend is reversed for down-regulated genes, at least in Nalm6 and A549. While the area under the ROC curve does not appear to follow the same trend in U2OS down-regulated genes, it is important to note that the set of sites without the motif is much smaller in U2OS than the set with motifs; thus if this partition was completely uninformative with respect to downregulation, we would expect the area under the curve to be smaller for the smaller set. This kind of analysis can be used to test other proposed methods for partitioning the binding sites into potentially informative and uninformative sites.

Looking at binding and regulation from a different angle, we asked what fraction of the genes with or without nearby binding sites are regulated. Here we used a fixed window around the transcription start site to assign genes, though we found that varying this window did not alter the basic pattern. We observed that only 2% to 20% of genes with nearby binding sites are regulated, while between 0.4% and 3.8%

of genes without nearby binding sites are regulated, depending on the cell type and the thresholds used for ChIP-Seq and microarray expression data analysis. Notably, we find that using a looser expression cutoff and more stringent ChIP-Seq cutoff in Nalm6 (as compared to the other two cell types) produced comparable percentages of genes with and without binding sites that are regulated.

We propose two possible explanations for this difference between Nalm6 and the other cell types: (1) it is an experimental artifact from an extremely low backgr microarray experiment with less ability to detect small changes in expression; and (2) it is due to the lower efficiency in Nalm6 of some step between binding and regulation, perhaps due to the absence or low activity of a critical cofactor. Microarrays at a longer time point or additional replicates to increase statistical power might provide more insight into these options.

3.4.2 Overlap of binding sites across cell types

We found that between 30% and 55% of A549 binding sites and less than 8% of the Nalm6 binding sites are found U2OS. This is less than (Nalm6/U2OS) or roughly equal to (A549/U2OS) the amount of overlap in transcription between cell types. These numbers are consistent with the idea the binding is the critical step governing the cell-specific response to glucocorticoids ([53]). Nonetheless, we note that there are examples of common binding across cell types with different transcription response.

In addition to the broad differences in location of binding sites, we found that many binding sites that we classified as overlapping in Nalm6 and the other two cell types had a different shape in Nalm6. This is particularly intriguing in light of the observation that the shapes of two of our ChIP-Seq peaks reflect the contribution of

multiple GBSs shown to be functional in reporter assays. Some possible sources of these differences include: (1) differences in chromatin accessibility (2) the presence of different cofactors stabilizing binding; (3) stalling of the assembly of a binding complex in an intermediate configuration. It would be interesting to determine whether the differences in the shape of the peaks has any bearing on the regulatory potential of the binding site.

3.4.3 Sequence composition of binding sites

We found that in addition to differences in the actual location of binding sites in different cell types, there are differences in the sequence composition of binding sites in the three cell types. Surprisingly, the GR motif was much less common in Nalm6 than in the other cell types. It is present in 66% of U2OS, 52% of A549, and 14% Nalm6 sites. Although the number of Nalm6 sites increase with increasing number of hits in the peak, it never reaches the level of U2OS. While Nalm6 appears to be an extreme outlier, we note that we have seen other intermediate cell types – for instance, 38% of the binding sites in a pituitary cell line (unpublished data, Hager lab) contained a GBS.

Meijsing et al. show in vitro that different GBSs induce different conformations in the GR DNA binding domain. Further, they demonstrate that different GBSs induce different degrees of transcriptional response unrelated to the affinity of GR for the GBS. These results suggest that GBS sequences contain information beyond a simple affinity for GR. ([33]) Similarly, So. et al show that even the bases of the GBS that are only loosely constrained by the motif show strong evolutionary conservation across species in some GRBRs. ([53]) Taken together, these results suggest the possibility

that cell type specific preferences of GR for variants of the canonical motif might account for some of the cell type specificity of GR occupancy. With this hypothesis in mind, we identified thousands of GBS in GRBRs in each cell type and constructed cell type specific motifs (figure 3.14). We failed to identify major differences in the canonical motif across cell types, though it is possible that a careful statistical test would identify a subtle preference.

We found that different secondary motifs were overrepresented in the three different cell types. We have also looked at GR ChIP-Seq data sets in additional cell types produced by other labs (Hager, Wang, unpublished) and found different sets of overrepresented motifs. This may explain part of the cell type specificity of binding sites. An alternative possibility is that GR simply binds to open chromatin that contains a favorable binding sequence, and that these motifs are a reflection of what chromatin is open in each cell type – passengers, that is, rather than causal agents. However, the fact that the reporter constructs (generally considered to be poorly chromatinized) that we tested recapitulated the cell type specificity of the ChIP-Seq data suggests that the chromatin structure is not the whole story. It will be interesting to see whether we can identify the exact determinants of cell type specificity within these reporters.

The specific motifs we found overrepresented make sense in terms of the biology of the different cell types. We touch on a few of them here: Many ETS factors (most prevalent in Nalm6 sites) are involved in hematopoietic development ([48]). There are multiple reports of cross talk between ETS factors and GR (for example, [16, 17, 15, 35]). Interestingly some leukemic gene fusions involve ets domains ([18], such as TEL-AML1(RUNX1) and ETV6-AML1. These gene fusions are not present in

Nalm6 cells. Pax5 (Nalm6 only) is essential for B-cell proliferation. Downregulation of Pax5 is reported to play a role in growth arrest and apoptosis of Nalm6 cell in response to dexamethasone. [39] C/EBPbeta (A549 only) is known to play an important role in inflammation in lung epithelial cells. The GR C/EBPbeta tethering site upstream of DUSP1 described in [22] was detected in all three cell types, but interestingly, DUSP1 is only up-regulated in A549. AP-1 (U2OS and A549) is known to have significant cross talk with GR (for example, [56] Interestingly, one report suggests that AP-1 binding activity does not play in role in glucocorticoid sensitivity in ALL ([2]), consistent with our finding that AP-1 is not overrepresented in Nalm6 binding sites.

3.4.4 Binding near genes that are regulated in multiple cell types

We found that genes that are activated in two cell types are much more likely to have binding sites in common than unregulated genes. For example (table 3.4), there are 26 genes that have associated binding sites in Nalm6 and are up-regulated in both Nalm6 and U2OS. 54% of these genes share at least one of their Nalm6 binding sites with U2OS. The corresponding number for genes that are not regulated in either cell type is 8%. Looking at individual binding sites near these same 26 genes, we found that 27% of the Nalm6 binding sites are also present in U2OS, while for the unregulated genes this number 5.3%.

It is intriguing that two cell types have both overlapping and non-overlapping binding sites near commonly regulated genes. This suggests two different possibilities: (1) Only the common binding site plays an active in role in GR mediated activation

of the gene in question, (2) the non-overlapping binding sites in different cell types serve as modifiers of regulation, leading to more complex cell-type specific response than microarrays at a single time point can detect.

The same increased overlap in binding sites across cell types is not observed in down-regulated genes. Although the numbers are relatively small, there is some suggestion in the data (tables 3.3 and 3.4) that binding sites near genes that are up-regulated in one cell type and down regulated in another are more likely to have overlapping binding sites than genes that are down-regulated in both cell types. Diamond et al. demonstrated that the same GRE could activate or repress transcription based on the presence of other regulatory factors. ([13]) Our data suggest that this may be a relatively common phenomenon. The lack of overlap between binding sites near genes that are down-regulated in two cell types is consistent with our finding that relationship between binding and downregulation is relatively weak.

We note that it is difficult to assign GRE status to a GRBR with confidence. The standard assay is to clone the GRBR behind a promoter in a reporter vector and see if the reporter is hormone sensitive. The combination of a hormone sensitive reporter and the relative proximity of a binding site to a transcription start site is used to support the claim that a GRBR is a GRE that regulates the nearby gene. However, this method assumes that the behavior of an isolated fragment of DNA next to a promoter in a reporter construct mimics the behavior of that stretch of DNA in the genomic context, situated potentially tens of kB away from the corresponding promoter. Our findings suggest a way to determine with a higher degree of confidence that a particular GRBR is a GRE – that is, whether a site that binds GR actually plays an active role in the regulation of a particular gene. We propose that a GRBR

found in two cells near a gene activated in both those cell type is more likely than a random GRBR near a regulated gene to play a role in the regulation of that gene.

3.4.5 Future Directions

We have three discussed three levels of information: genomic sequence, DNA binding, and transcription. The three corresponding genomic elements are: GBSs, defined solely by sequence; GRBRs, defined by occupancy; and GREs, defined by the relationship between occupancy and transcription. The present study used a comparison of binding and regulation across cell types to advance our understanding of the relationship among these three elements. It also raised some interesting questions, including: What is the relative contribution of local sequences at occupied sites verses longer range determinants of chromatin structure on the cell-type specificity of GR? Is the same gene typically regulated with the same binding sites in different cell types? What is the predominant mode of glucocorticoid induced gene repression? Answering these questions will require mechanistic studies in reporters and bacterial artificial chromosomes along with genome-wide occupancy and expression profiling in additional cell types.

3.5 Materials and Methods

Cell culture

A549 cells were obtained from the UCSF cell culture facility. U2OS cells were derived by stably transfecting GRalpha into an osteosarcoma cell line that lost GR alpha some time during the tumorigenesis or cell line derivation. [43] Nalm6 cells were obtained

from St. Jude Children's Research Hospital. A549 and U2OS were maintained in DMEM with 5% FBS and Nalm6 were maintained in RPMI with 10% FBS. All the cells were maintained in a humid incubator with 5% carbon dioxide. The cells were treated at confluence for expression analysis or ChIP. Nalm6 were considered confluent at 2 million cells per milliliter.

Microarrays

The arrays were whole genome spotted oligo nucleotide arrays (HEEBO probe set, Invitrogen) printed at UCSF. Confluent cells were treated with 1 uM dexamethasone or EtOH for three hours. Microarrays were hybridized in a MAUI hybridization chamber (BioMicro Systems), scanned on a GenePix 4000B scanner (Molecular Devices), and gridded with SpotReader (Niles Scientific). The array data was analyzed with limma package in BioConductor ([41, 52, 51]. In brief: we applied the normalizeWithinArrays function with background correction method RMA. For each cell type, we required that a given probe be detected as present by SpotReader in at least 2 of the 3 arrays to be included in further analysis. We used lmFit and eBayes with the default parameters, including the Benjamini and Hochberg's method to adjust the p-value for multiple hypothesis testing. In this method of controlling the false discovery rate, a p-value cutoff of 0.05 will produce less that 5% false positives. We considered multiple different cutoffs for calling genes regulated, as described in the text.

ChIP-Seq

ChIP was performed as described in ([6]) with the following differences: cells were treated with 0.01% ethanol vehicle or 1 uM dexamethasone for 1.5 hours. Suspension cells (Nalm6) were centrifuged briefly prior to washing and lysis. Sonication was done with a Diagenode Bioruptor in 0.5 ml tubes in ice water for a total sonication time of 12 minutes. Protein G coupled magnetic beads from ActiveMotif were pre-incubated with N499, a polyclonal GR antibody ([31]) for one hour. Chromatin was incubated with rotation at 4 C for 2-4 hours. Three washes with each of the follow were used: 500 mM NaCl RIPA wash (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 500 mM NaCl, 5% glycerol, 0.1% sodium deoxycholate, 0.1% SDS, 1% Triton X-100, 0.5 mg/ul BSA) and an LiCl wash (20mM Tris, pH 8.0 1mM EDTA, 250mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate).

Solexa library preparation is described in detail in the supplementary material. The primary differences from the Illumina protocol are: different adaptors and bar coding method; longer ligation time; lower adaptor to DNA ratio; Ampure bead (Agencourt) purification in place of size selection. Single end read sequencing (36 base pairs) was done on the Illumina Genome Analyzer II. Results were processed with the Illumina pipeline and aligned to the genome with Eland (30-32 base pair fragments). We wrote a perl script to pull out the results for the individual bar codes. To visualize the Solexa data, we used the UCSC genome browser and MochiView ([21]), a fast and versatile tool for browsing large genomic data sets.

We discarded all duplicate sequence reads and all reads that did not map uniquely to the human genome with ELAND (Illumina). The sequences remaining after this filtering are defined as the usable sequence reads.

Data analysis

We wrote scripts in perl and R to do the peak detection and subsequent analysis. The peak detection algorithm looks for at least 10 sequence reads in a 400 base pair window, with at least 4 times as many reads in the ChIP sample than in any one of the three input DNA samples. All the input DNA samples had more usable sequence tags than ChIP samples. Peaks less than 250 base pairs in width were discarded. Peak profiles were constructed by extending each sequence read 200 base pairs along the appropriate strand and summing up the number of reads at each position in the peak. The center of a peak was defined as the location where the peak profile has its maximum value.

To calculate overlaps between ChIP-Seq binding regions, we identified all the peaks in one cell type and then counted the number of sequence tags within the peak boundaries in a second cell type. If this number exceeded 10 hits, our standard threshold for peak detection, we considered it an overlap. We also considered the effects of different thresholds, as described in the text.

We used the TFBS package ([28]) to search for matches to position weight matrices for known transcription factor binding sites. We modified the C code to optimize speed for our particular application. For an unbiased motif search, we used the central 100 bases of five sets of 400 randomly selected peaks from each cell type. To compare motifs from the databases and Meme motifs to each other, we used the motif analysis tools in MochiView.

To map the probes on the microarrays to known genes, we used the UCSC knownGene and kgXref tables and searched Ensembl for genes on the array not found in that table. Both these sources contain information for all the known transcripts for

each gene, which we used in mapping peaks to genes. For reasons of computational tractability, we used two slightly different mapping algorithms for different parts of the analysis. In both cases, distance from the transcription start site was calculated from the center of the peak to the closest transcript for a given gene. In the first method, used for generating the ROC curves and for calculating the fraction of genes with associated peaks that were regulated, peak mapping was done solely on the basis of proximity to the TSS, ignoring the location of other genes. All the peaks within a given window of the transcription start site were counted for each transcript, and the maximum count was used as the value for that gene. A single binding site could be assigned to multiple genes. In the second method, used to determine the distance of binding sites from the nearest gene and the overlap of binding sites near genes regulated in multiple cell types, we calculated the distance from every binding site to the nearest transcription start site, and used that as the distance to the associated gene.

3.6 Supplementary data

3.6.1 Figures

		Log FC, 3 hr, 1 uM dex		
symbol	**name**	**a549**	**u2os**	**nalm6**
ARID5A	AT rich interactive domain 5A (MRF1-like)	-0.52	-1.29	NA
ARID5A	AT rich interactive domain 5A (MRF1-like)	-0.74	-0.9	0.65
BCL6	NA	NA	NA	NA
BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	0.74	0.77	-0.97
BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	0.89	NA	-1.09
BCL6	B-cell CLL/lymphoma 6 (zinc finger protein 51)	0.79	0.68	-1.04
CALD1	caldesmon 1	1.28	0.51	0.54
CALD1	caldesmon 1	0.87	0.37	0.69
CALD1	caldesmon 1	2	1.31	1.3
DDIT4	DNA-damage-inducible transcript 4	1.07	-1.18	1.48
DENND3	NA	NA	0.84	1.05
DENND3	KIAA0870 protein	0.6	NA	NA
DUSP10	dual specificity phosphatase 10	-0.8	-0.47	1.26
DUSP10	NA	NA	NA	NA
DUSP10	dual specificity phosphatase 10	-0.76	-0.59	1.18
ENC1	ectodermal-neural cortex (with BTB-like domain)	-1.69	-2.74	-0.59
ENC1	ectodermal-neural cortex (with BTB-like domain)	-1.8	-2.6	NA
FAM43A	family with sequence similarity 43, member A	1.65	0.62	1.56
FKBP5	FK506 binding protein 5	2.96	1.44	1.04
FKBP5	FK506 binding protein 5	2.93	1.18	0.97
HSPA1B	heat shock 70kDa protein 1B	-0.7	-1.25	-0.79
IER2	immediate early response 2	-1.67	-2.38	-0.74
MAP3K6	mitogen-activated protein kinase kinase kinase 6	0.73	0.92	0.62
MAP3K6	mitogen-activated protein kinase kinase kinase 6	0.59	0.57	NA
MGC17330	HGFL gene	-0.83	-1.36	2.27
NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in	1.59	2.6	0.97
PER1	period homolog 1 (Drosophila)	3.52	0.73	0.63
PHLDA1	pleckstrin homology-like domain, family A, member 1	-0.81	1.2	-0.82
PHLDA1	ESTs^^^AA436592|Exon1_251^^-22^^	-1.01	1.97	-0.89
PRG1	NA	NA	NA	NA
PRG1	proteoglycan 1, secretory granule	0.83	1.09	1.04
PRG1	proteoglycan 1, secretory granule	0.84	1.08	1.13
RASSF4	Ras association (RalGDS/AF-6) domain family 4	0.91	2	2.26
RASSF4	Ras association (RalGDS/AF-6) domain family 4	1.34	1.83	2.33
SOCS1	suppressor of cytokine signaling 1	2.3	1.34	2.15
SOX4	SRY (sex determining region Y)-box 4	-0.8	-1.64	-0.6
SOX4	SRY (sex determining region Y)-box 4	-0.91	-1.12	-0.52
STARD13	NA	NA	NA	NA
STARD13	altSplice.a1^^scl0008162^scl0008162.1_316^exon_1800	1.26	-1.38	0.72
STARD13	START domain containing 13	1.27	-1.15	0.78
STARD13	NA	NA	NA	NA
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	1.44	1.1	-0.65
TSC22D3	TSC22 domain family 3	3.16	NA	3.5
TSC22D3	TSC22 domain family 3	4.01	3.58	4.23
TSC22D3	TSC22 domain family 3	3.28	3.03	3.3
TXNIP	thioredoxin interacting protein	1.87	2.86	1.22
ZFP36L2	zinc finger protein 36, C3H type-like 2	0.73	-1.04	1.18

Figure 3.10: Genes regulated in all three cell types: Down regulated genes are shown in green, up regulated in red. For genes with multiple probes on the array, results from all the probes are shown. NA indicates that the array data for that probe was removed by the gridding software, generally because the spot intensity was too low for accurate measurement. Note that BCL6, PHLDA1, and STARD13 have multiple probes on the array that show different regulation in the different cell types.

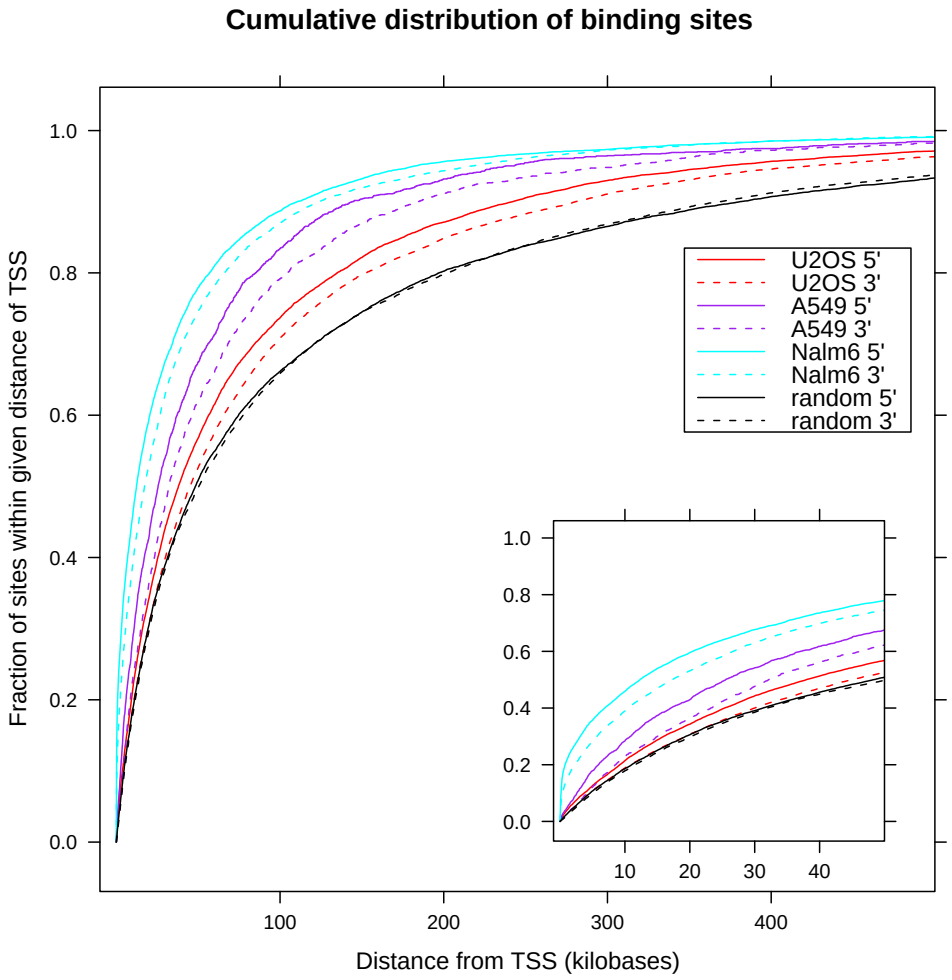


Figure 3.11: Distribution of binding sites with respect to annotated transcription start sites. The random distribution was generated by randomly selecting locations of GBSs, with the number chosen from each chromosome matching the total number of U2OS sites on that chromosome. All the non-random sets are slightly preferentially distributed 5' of the TSS. Nalm6 binding sites are distributed closest to TSSs, folowed by A549 and the then U2OS. This ordering persists even when a higher cutoff threshold is used for calling Nalm6 sites. It also persists if all sites within 2 kB of the transcription start site are removed from the three data sets.

Cell type	1 kBP	10 kBP
U2OS	3%	20%
A549	4%	25%
Nalm6	18%	42%
Random	2%	18%

Table 3.5: Percentage of binding sites in each cell type within 1 kb and 10 kb of transcription start sites. 3.11)

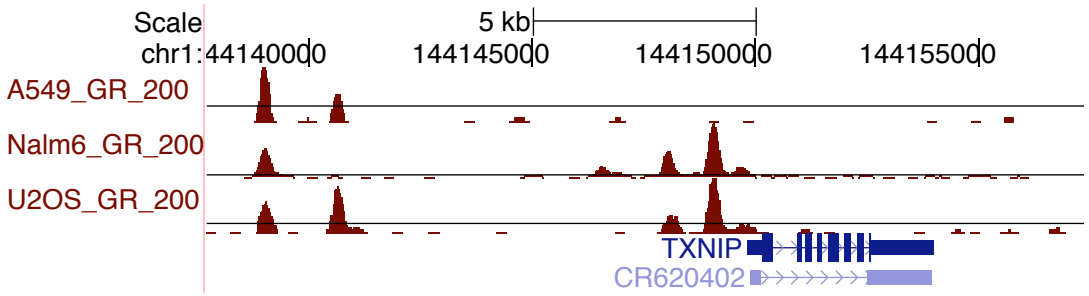


Figure 3.12: Binding sites near TXNIP: Thioredoxin-interacting protein (TXNIP) is up-regulated in Nalm6, U2OS, and A549. The associated binding sites in each cell type are shown here. Note that only the most distal binding site is shared by all three cell types.

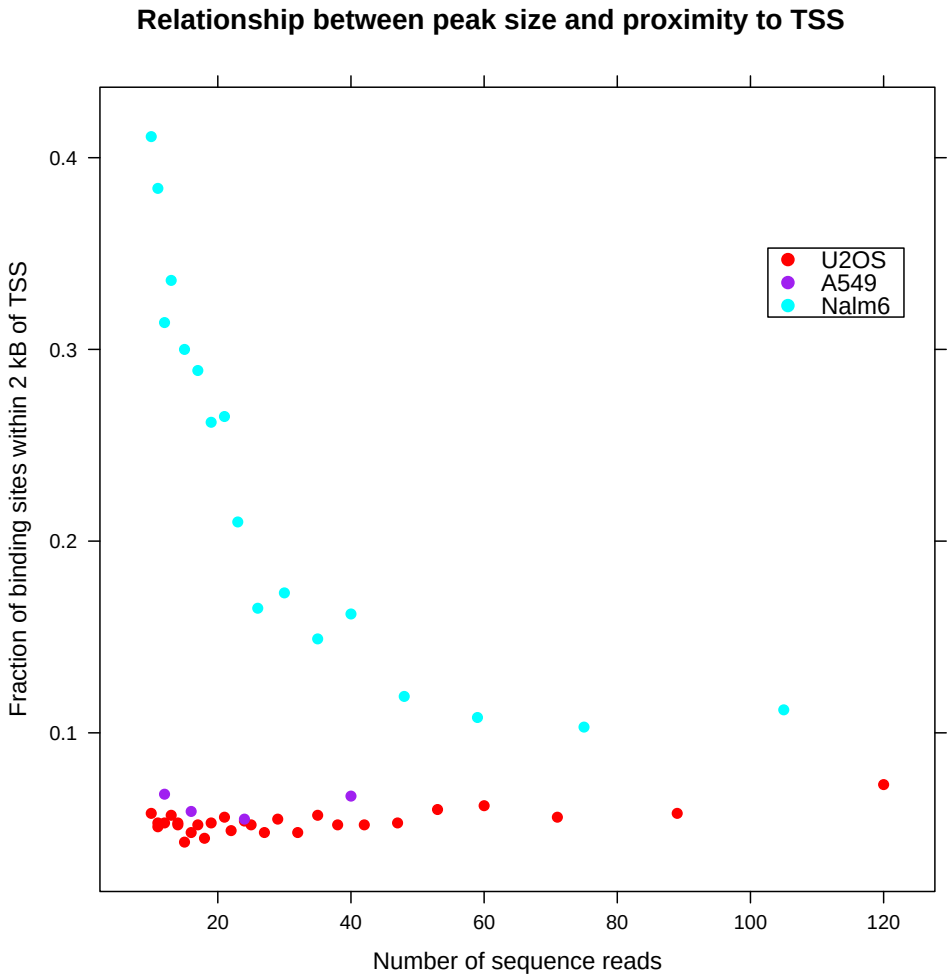
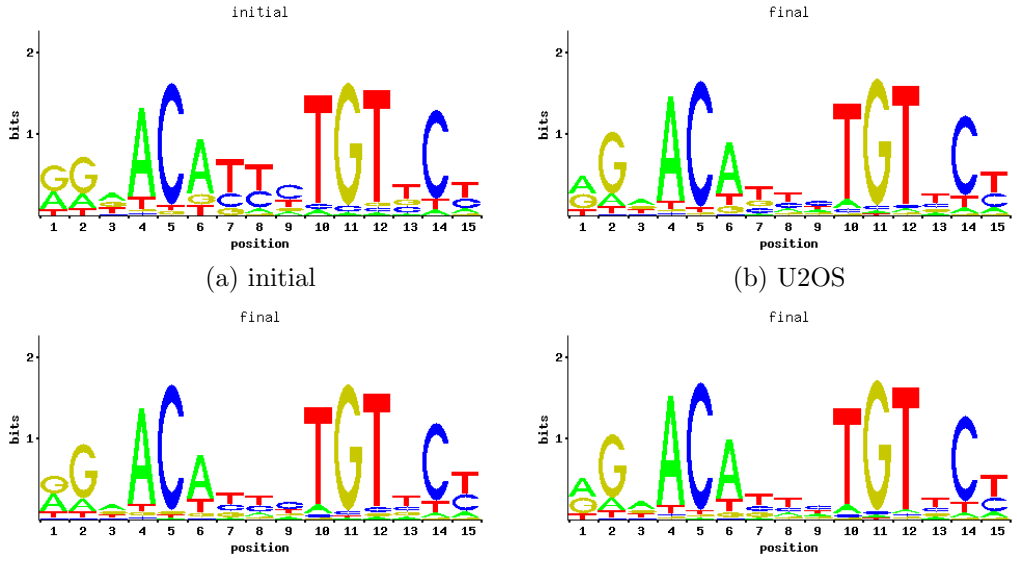


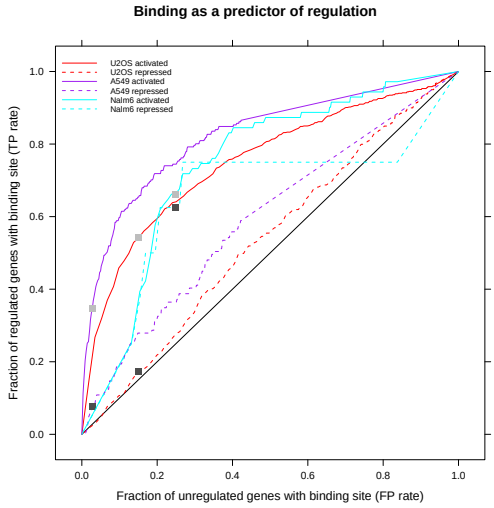
Figure 3.13: Fraction of ChIP-Seq peaks within 2 kB of transcription start site as a function of the strength of the peak. Peaks were ordered based on the number of sequence reads in the peak and binned into groups of 1000. The median number of sequence reads in each bin is plotted on the x-axis and the fraction of peaks within 2 kB of an annotated transcription start site is shown on the y-axis. Only Nalm6 shows a clear association between peak size and proximity to TSS. Even at very strong peaks (more than 50 sequence reads) the Nalm6 peaks are more likely to be close to a transcription start site than the A549 or U2OS peaks.



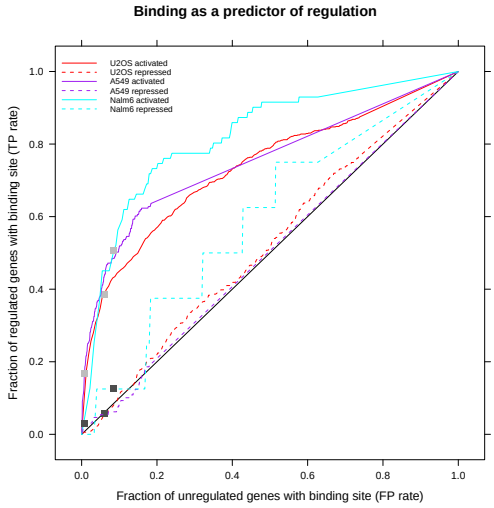
(a) initial (b) U2OS

(c) Nalm6 (d) A549

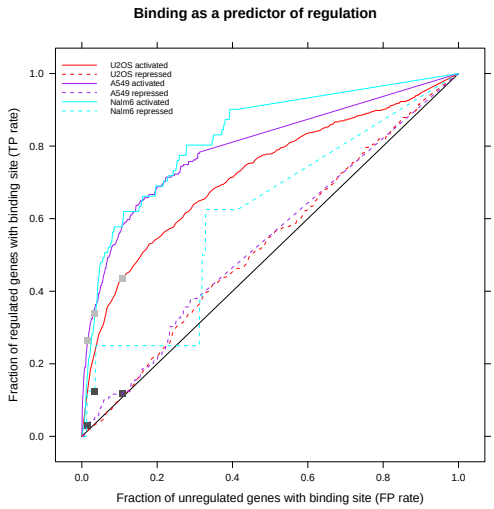
Figure 3.14: Canonical GR motif derived from each cell type. The initial PWM (a) was derived with a Meme search on a random subset of 400 U2OS peaks. This PWM was used to search the central 200 base pairs at a 5% false positive cutoff of all the (b) U2OS, (c) Nalm6, and (d) A549 ChIP-Seq peaks and a new motif was derived from each cell type.



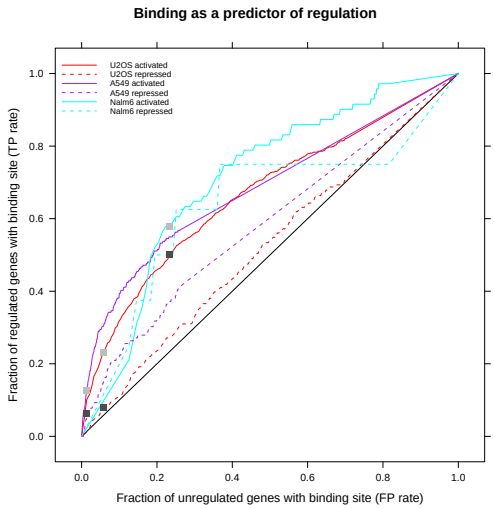
(a) Using all the binding sites



(b) Using only binding sites with at least 30 sequence reads

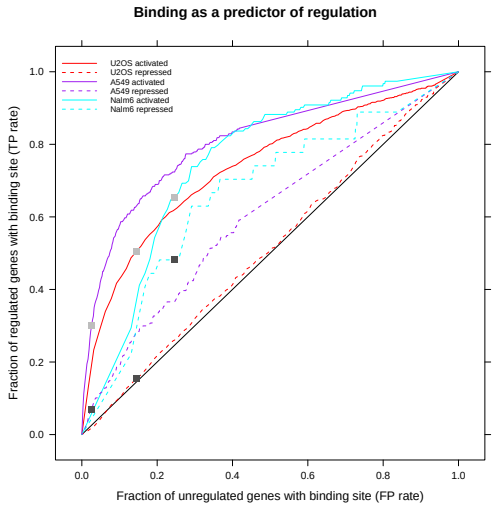


(c) Using only binding sites that contain a GBS

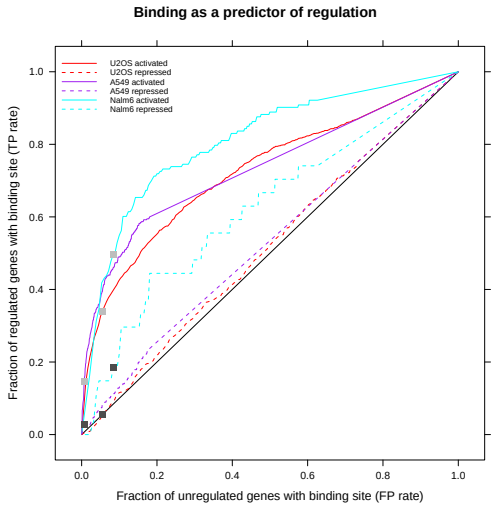


(d) Using only binding sites that lack a GBS

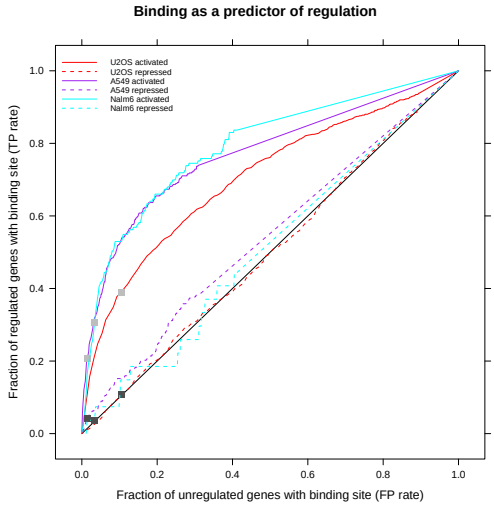
Figure 3.15: ROCs with genes in the most stringent set (adjusted p-value less than 0.01, fold change greater than 1.8 or less than 0.56). Unregulated genes are all the genes on the array that are not up- or down-regulated at this threshold. The ROCs were generated with different subsets of binding sites, as indicated in the caption of each figure.



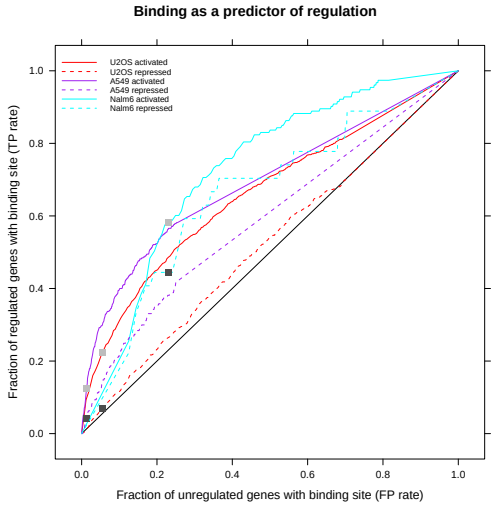
(a) Using all the binding sites



(b) Using only binding sites with at least 30 sequence reads



(c) Using only binding sites that contain a GBS



(d) Using only binding sites that lack a GBS

Figure 3.16: ROCs with genes in the looser set (adjusted p-value less than 0.05, fold change greater than 1.5 or less than 0.67). Unregulated genes are all the genes on the array that are not up- or down-regulated at this threshold. The ROCs were generated with different subsets of binding sites, as indicated in the caption of each figure.

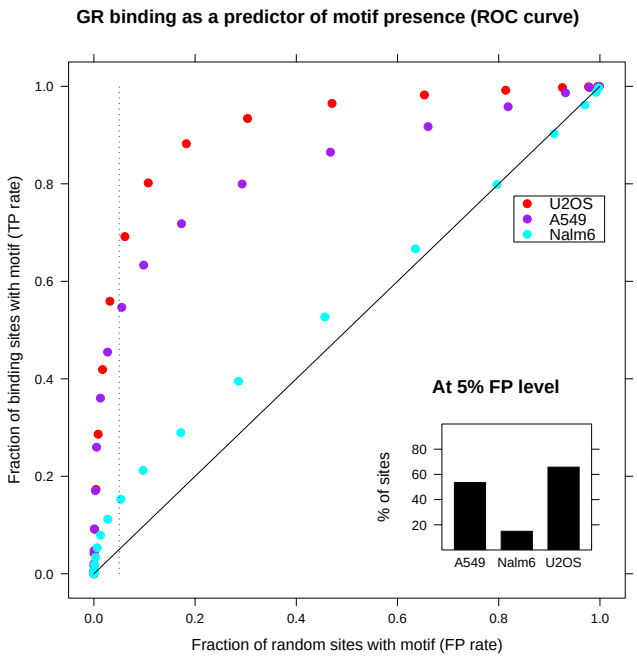
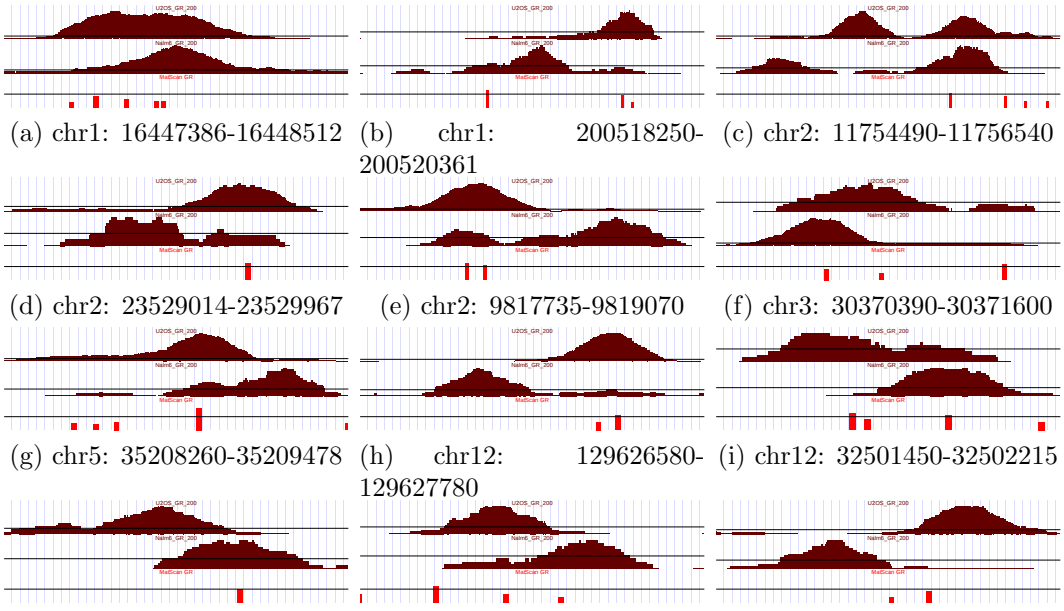


Figure 3.17: GBS occurrence in binding sites: ROC curve with the false positive rate defined as the fraction of control sites that contain a motif at a given threshold and true position rate defined as the fraction of binding sites that contain a motif at that threshold. The threshold is varied from 0% to 100% of the maximum possible score to generate the curve.



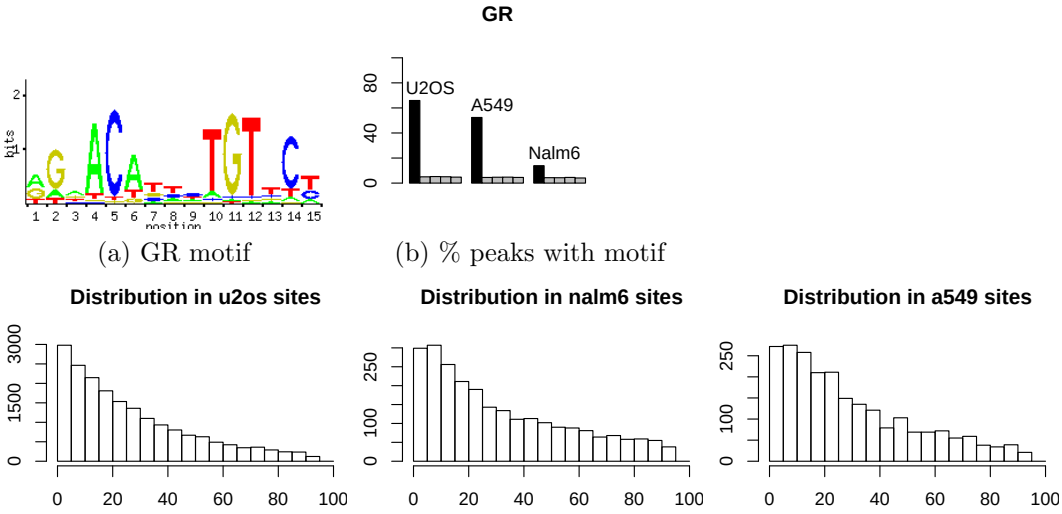
(a) chr1: 16447386-16448512 (b) chr1: 200518250-200520361 (c) chr2: 11754490-11756540

(d) chr2: 23529014-23529967 (e) chr2: 9817735-9819070 (f) chr3: 30370390-30371600

(g) chr5: 35208260-35209478 (h) chr12: 129626580-129627780 (i) chr12: 32501450-32502215

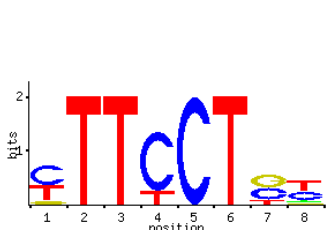
(j) chr14: 104607416-104608404 (k) chr17: 18169005-18170050 (l) chr19: 16407890-16408960

Figure 3.18: Examples of overlapping peaks in Nalm6 and U2OS with centers more than 200 base pairs apart. In each figure, the first track shows the U2OS peak, the second track shows the Nalm6 peak, and the third track shows the GBSs. The black horizontal line is at a fixed height to 5 in the peak tracks to make visual comparison between the figures easier. In the GBS track, the height of each red bar indicates how well the site matches the canonical motif.

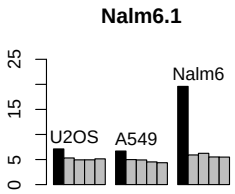


(a) GR motif (b) % peaks with motif

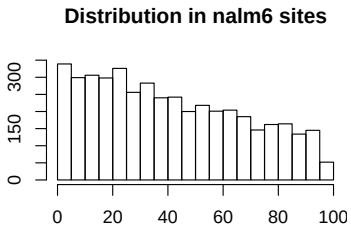
Figure 3.19: Occurence of GR motif in multiple cell types. (a) GR motif logo (b) Percentage of GR binding sites containing at least one GR motif in the central 200 base pairs. Black bars represent binding sites in indicated cell types. Gray bars represent control sites – 200 base pair windows offset by 500, 1000, 2000 and 10,000 base pairs from the real binding sites.(d-f) Distrition of motifs within binding sites for (d)U2OS, (e)A549, and (f)Nalm6. For each binding site containing a GR motif, the distance was calculated as the absolute value of the distance from the best motif to the center of the binding site. GR shows a strong central distribution in all the cell types.



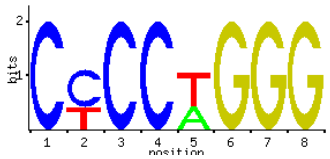
(a) Similar motifs: ETS factors Elf3 and SFPI1



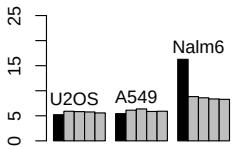
(b) % peaks with motif



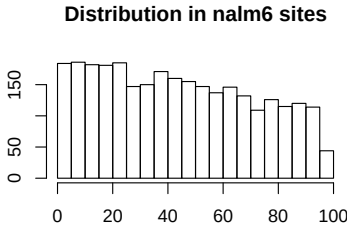
(c) Distance of motif from peak center



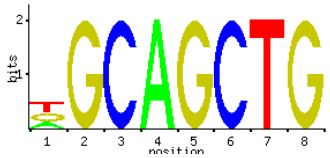
(d) Similar motifs: Zic family of C2H zinc fingers



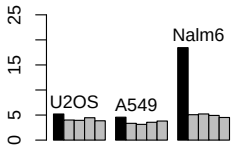
(e) % peaks with motif



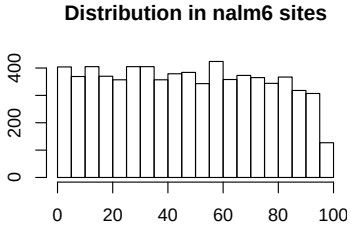
(f) Distance of motif from peak center



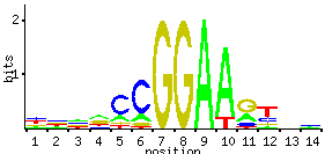
(g) Similar motifs: bHLH factors HEN1, NHLH1, ASCL2, E47



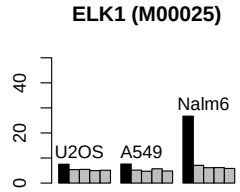
(h) % peaks with motif



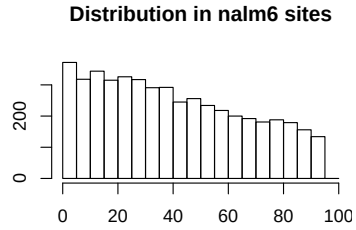
(i) Distance of motif from peak center



(j) Similar motifs: c-ets-1(p54), GABPA, NRF-2, ELK4

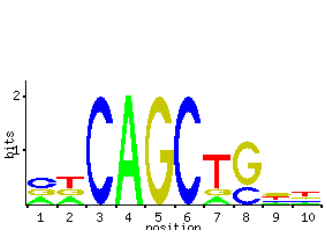


(k) % peaks with motif

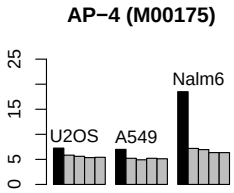


(l) Distance of motif from peak center

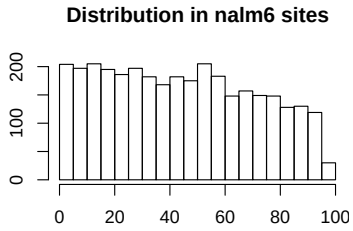
Figure 3.20: Motifs overrepresented in Nalm6: Plots are as described in figure 3.19. The first three motifs were initially identified with a Meme search. Some of the motifs (for example, c) show evidence of a central distribution, while others (for example, i) do not show evidence of a central distributed.



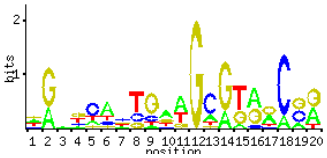
(a) Similar motifs: Myf, MyoD, E47, HEN1



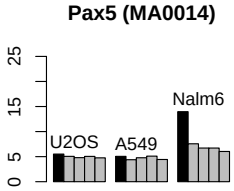
(b) % peaks with motif



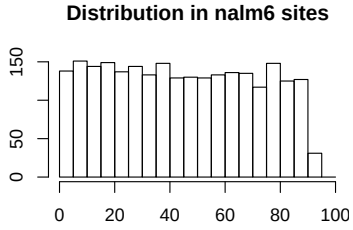
(c) Distance of motif from peak center



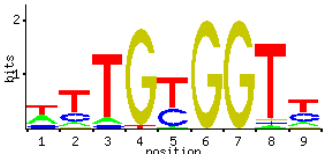
(d) Similar motifs: other Pax factors



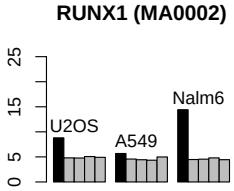
(e) % peaks with motif



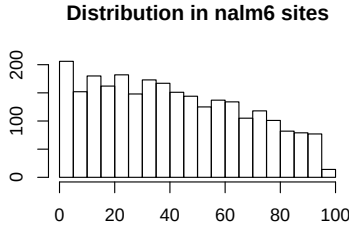
(f) Distance of motif from peak center



(g) Similar motifs: none. Synonym - AML1

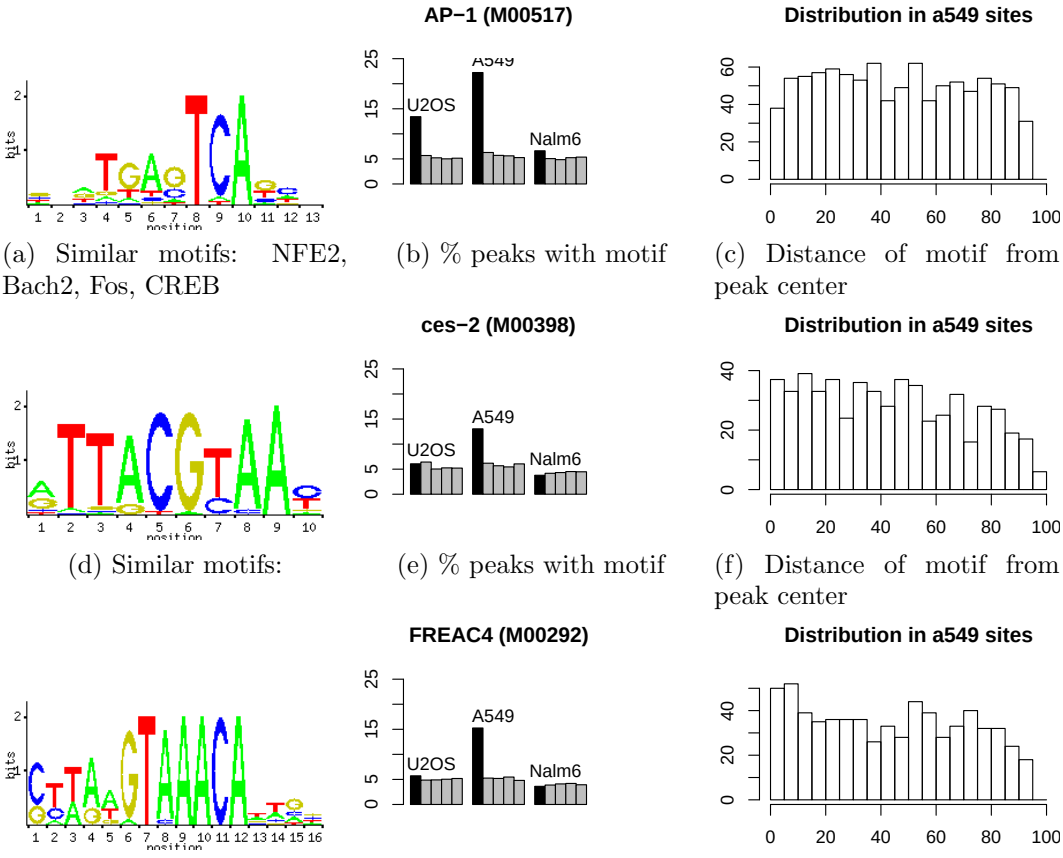


(h) % peaks with motif



(i) Distance of motif from peak center

Figure 3.21: Motifs overrepresented in Nalm6, continued from figure 3.20



(a) Similar motifs: NFE2, Bach2, Fos, CREB

(b) % peaks with motif

(c) Distance of motif from peak center

(d) Similar motifs:

(e) % peaks with motif

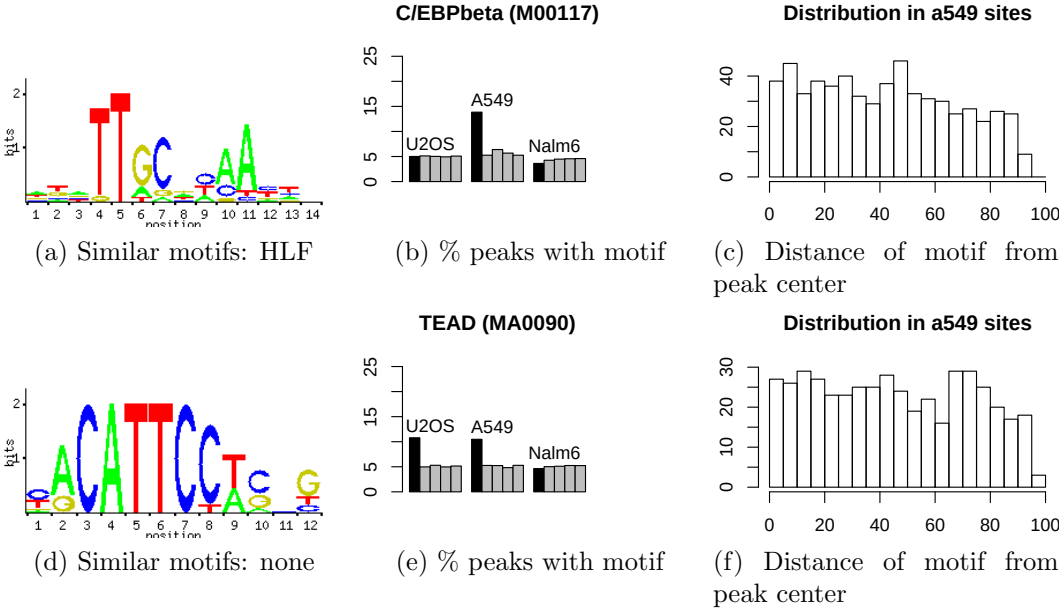
(f) Distance of motif from peak center

(g) Similar motifs: FOX family

(h) % peaks with motif

(i) Distance of motif from peak center

Figure 3.22: Motifs overrepresented in A549. See caption of figure 3.20 for a description.



(a) Similar motifs: HLF

(b) % peaks with motif

(c) Distance of motif from peak center

(d) Similar motifs: none

(e) % peaks with motif

(f) Distance of motif from peak center

Figure 3.23: Motifs overrepresented in A549, continued from figure 3.22

Chapter 4

Selective glucocorticoid receptor modulators

4.1 Introduction

Glucocorticoids are used clinically as anti-inflammatories and immunosupressants. Treatment with glucocorticoids comes with steep costs, however: due to the broad physiological spectrum of glucocorticoids, chronic therapeutic application triggers numerous adverse side effects on metabolism, cardiovascular function, and the neuroendocrine system. An ideal drug candidate would selectively activate only a desired GR activity, e.g., anti-inflammation. Researchers and pharmaceutical companies have long sought such molecules, called selective glucocorticoid receptor modulators (SEGRMs), that decouple different GR mediated response. This chapter first discusses the reasoning behind this effort, then presents data from one such set of molecules and argues that the original claims about how these molecules produce cell-specific effects are not supported by the evidence. Finally, it proposes some guidelines for how to

evaluate whether a molecule is a potential selective glucocorticoid receptor modulator.

4.2 Background

The GR has multiple different surfaces through which it interacts with other molecules. Most probably remain to be discovered, but a few have been well-defined genetically, biochemically or structurally:

AF1 – within the N-terminal region, coregulatory factors are thought bind to this surface, although these have not been systematically identified; multiple covalent modifications are also associated with the GR N-terminus.

DNA binding domain (DBD) - This domain binds to the canonical 15 base pair GBS motif.

Dimerization surface - This surface resides within the DBD. It is required for dimerization of GR on the DNA

Ligand binding domain (LBD)

AF2 - This surface resides within the LBD; coregulatory factors bind to this surface.

Based on numerous studies of isolated GR response elements in reporter plasmids, there are multiple modes by which GR is believed to affect transcription: direct binding to canonical motif; tethering through another factor; competion with another factor. In any of these binding modes, different combinations of GR surfaces serve as determinants of the transcriptional response at different response elements in different cell types. Rogatsky et al. demonstrated this elegantly by mutating different

surfaces of the receptor and measuring the change in the transcriptional response mediated by a set of GR response elements.[44] They demonstrated that, for instance, mutating AF1 knocks out one subset of transcriptional responses, while mutating the dimerization domain knocks out a different subset of responses. Thus it seems that different properties of the glucocorticoid receptor are responsible for transcriptional responses at different genes.

Ligands that decouple different phenotypic responses were first described for the estrogen receptor (ER): Tamoxifen acts as an ER agonist in uterus and bone, but as an antagonist in breast. Raloxifen acts as an ER agonist in bone, and an antagonist in breast and uterus.[36] Such ligands are collectively referred to as selective estrogen receptor modulators (SERMS). In a given cell type, different SERMs activate different sets of genes.[4, 25] While some of the tissue specific effects might be due to the intersection of the SERMS with different isoforms of the ER – that is, ERalpha vs ERbeta, the SERMS produce different transcriptional responses even in the presence of only one ER isoform.[25]

There is a clear structural basis for the different transcriptional responses mediated by the SERMS. Binding by tamoxifen and raloxifen cause different changes in the positioning of helix 12, part of the AF-2 surface, upon binding to ER.[8, 49] These differences lead to different cofactor recruit at this surface, and ultimately different transcriptional responses.[30, 7] While SERMs clearly have an important clinical use, it is not clear whether the in vivo selective agonist activity of these molecules is primarily due to different effects of the ligands on each of two isoforms or on selective effects of SERMs on a single isoform.[38]

Various studies have identified potential selective glucocorticoid receptor molecules that bind to GR in the ligand binding pocket.[47, 54, 12, 46] I chose to focus on a closely related series of nonsteroidal arylpyrazole compounds, which provoke GR activities that do not differ dramatically in simple reporters[47]. However, these compound display selective effects on certain endogenous genes, and on glucocorticoid-like phenotypic outcomes in particular cultured cell lines.[59]

4.3 Assessing gene specific regulation

The original paper on which this work was based assessed ligand responses in three cultured cell lines: A459, a proinflammatory human lung carcinoma line whose proliferation is inhibited by glucocorticoids; 3T3L1, a mouse preadipocyte cell line that differentials into adipocytes in response to a cocktail containing glucocorticoids; and MC3T3-E1, a mouse preosteoblast line, where glucocorticoids inhibit differentiation to osteoblasts. The present study was initiated using S49 cells, a mouse lymphoma cell line that apoptoses in response to corticosteroid. The experiments described here using 3T3L1 cells were done by Carlos Pantoja in connection with a different project.

In search of gene-specific effects of different GR ligands, we ran microarrays of S49 cells treated with dexamethasone, a synthetic steroidal glucocorticoid; corticosterone, the endogenous GR ligand in mice; RU486, a GR antagonist or partial agonist; and 3T3L1 cells treated with ligand 17 of arylpyrazole series described above (nonsteroidal ligand 17, or NSR-17) Figure 4.1 shows the results, plotted on a log-fold change over vehicle control. Each point represents a single gene. In (4.1a) and (4.1c), the top 200 regulated genes are shown. In (4.1b), all the genes that pass a stringent array quality control cutoff are shown. Note that in all the plots, there is a strong

linear correlation and there are not points that are clear outliers. Stringent statistical tests also indicate that there are no points that fall outside this linear relationship. A formal statistical test using the Limma package in Bioconductor does not identify any differentially regulated genes at a p-value of less that 0.05 (adjusted for multiple hypothesis testing). These data suggest the following: in S49 cells, RU486 is a pure antagonist; in S49 cells, corticosterone has a weaker overall effect than dexamethasone, but not gene specific effect; and in 3T3L1 cells, NSR-17 is weaker than dexamethasone but does not have gene specific effects. Additional data (not shown) from NSR-5 in 3T3L1 cell, with lower quality arrays and thus a higher threshold for detection of gene specific effects, also fail to demonstrate any clear gene-specific effect.

In light of these preliminary results, we revisited the earlier data on gene specific regulation in A549 cells, shown (4.2 as published in ([59]) and plotted it on log fold change vs log fold change axes (figure 4.3). Each point represents a single gene assayed by qPCR, and the asterisk marks the origin. The linear least square line fitted to the data is plotted. We define the slope of this line to be a measure of the global transcriptional efficacy of each ligand in A549 cells. To determine whether this global transcriptional efficacy of this fitted line is related to the phenotypic outcome, we plotted the phenotypic results in A549 cells verses the global transcriptional efficacy. The results clearly show that the phenotypic response in A549 correlates well with this measure. Interestingly, the correlation between the phenotypic outcome in two other cell types (differentiation of 3T3L1, reported in [59] and cell death in S49) and the global transcription efficacy (figure 4.4) is less clear. There are a points that appear to fall outside the linear relationship – NSRs 3, 5 and 8 in S49 cells and NSRs 3,5,6,7, and 10 in 3T3L1 cells.

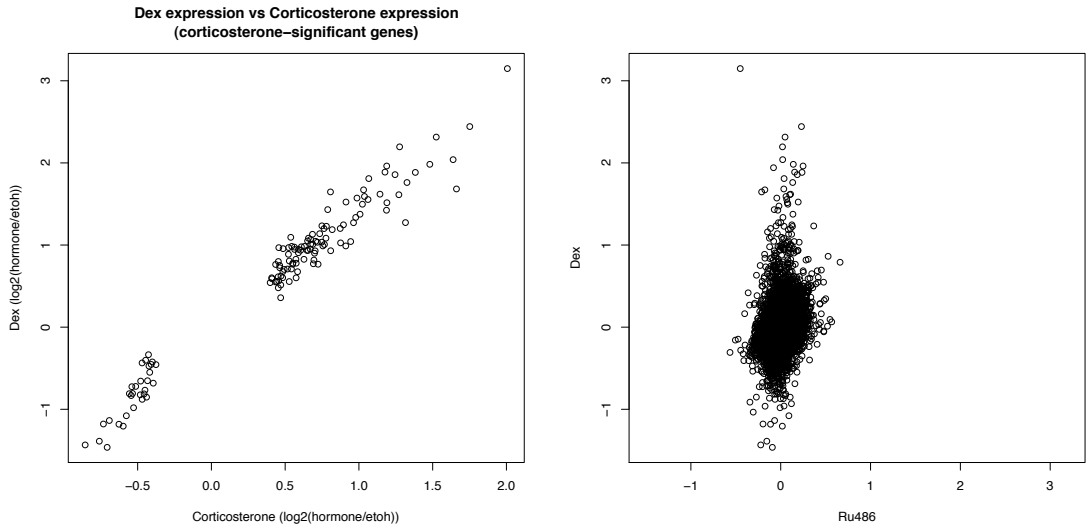
To see if the transcriptional response to each ligand corresponds to less than saturating dosages of dexamethasone, PCR primers were designed for a randomly selected set of genes identified on the microarrays as hormone responsive, plus a gene identified on the arrays as non-responsive. Then cells were treated in triplicate for three hours at 10 dosages of dexamethasone ranging from 1e-10 to 1e-6, as well 1e-6 M corticosterone, NSR-2, NSR-8, and NSR-12. The results are shown in figure 4.5. The response to each ligand very closely matches the response to a particular dosage of dexamethasone: corticosterone and NSR-2 to dex at 10 nM; NSR-8 and NSR-12 to dex at 3.2 nM. Note that one of the outliers in figure 4.4c, ligand 8, has a transcriptional response that is almost identical in strength to NSR-12 in S49, as well as a phenotypic response identical to NSR-12 in this cell type.

4.4 Conclusions

Based on the new data discussed here and the reanalysis of previously published data, the claim that the arylpyrazole ligands have gene specific transcriptional response is not supported by the data. On the contrary, it appears the different ligands produce transcriptional responses that are very similar to dexamethasone, but globally weaker to varying degrees. Two different methods are suggested here for assessing this response: (1) Plotting the log fold transcriptional change of a new ligand verses dex and measuring the slope of the line fitted to these points, and (2) Assaying the transcriptional response of the cells at a range of dexamethasone dosages and identifying which dosage best matches the response of the cells to the new ligand. Based on the fact that different genes have different dose response curves to dexamethasone, the second method may be more appropriate.

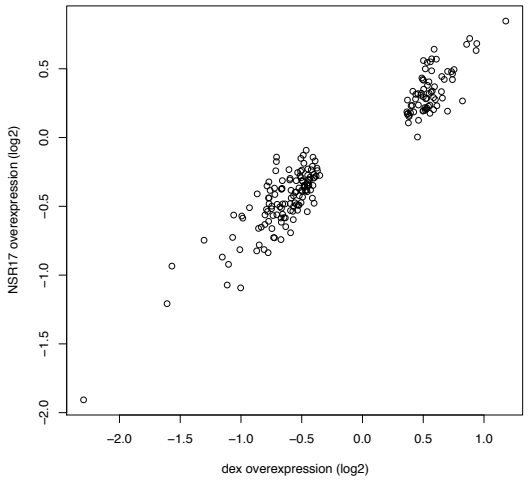
Using a measure of global transcriptional efficacy of each ligand in A549 cells, we show that this measure correlates very well with the phenotypic response in these cells. The measure correlates less well with the phenotypic outcome in two other cell lines, leading to the question of whether the global transcriptional efficacy measured in the other cell lines would better correlate with the phenotypic outcome in these cell lines. Results with a subset of ligands in S49 cells suggest that this may indeed be the case.

This work has two important implications for future work in assessing novel ligands. First, it is important to compare transcriptional responses to novel ligands to dose-response data for dexamethasone. Second, the most effective way to screen potential selective glucocorticoid receptor modulators for cell specific effects may be to assess transcriptional response of a handful of genes in multiple cell types. A selective ligand should show a different magnitude of transcriptional response (as a fraction of the magnitude of the dexamethasone response) in different cell types. This type of screen would be simpler and would be applicable to a wider array of cell types than screening for phenotypic outcomes.



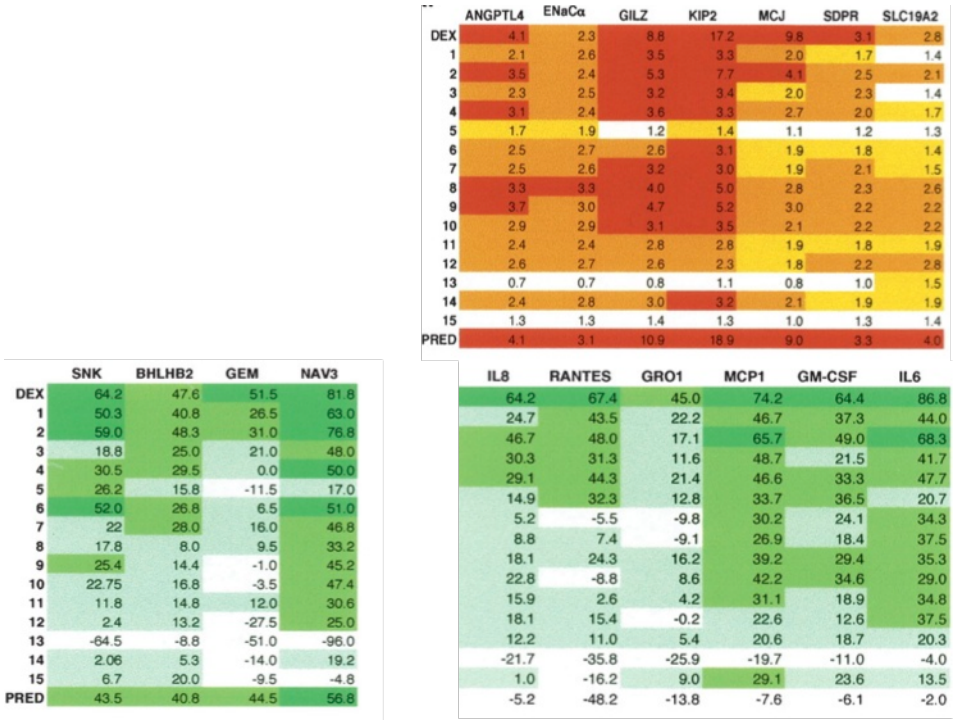
(a) Dex vs corticosterone in S49

(b) Dex vs Ru486 in S49



(c) NSR17 vs Dex in 3T3L1

Figure 4.1: Transcriptional response of S49 and 3T3L1 cells to different GR ligands. Cells were treated for three hours with a saturating dosage of ligand (1e-6M) or a vehicle control. Each point represents the log base 2 fold change of a single gene in response to the treatments shown on the x and y axes. (a) Only genes that are significantly regulated in response to corticosterone are shown; results are similar when genes significantly regulated by dexamethasone are shown (b) All genes that are clearly detectable on at least 2 of the three arrays in each condition are shown. (c) Only genes that are significantly regulated in response to dexamethasone are shown; results are similar when genes that are significantly regulated in response to NSR17 are shown.

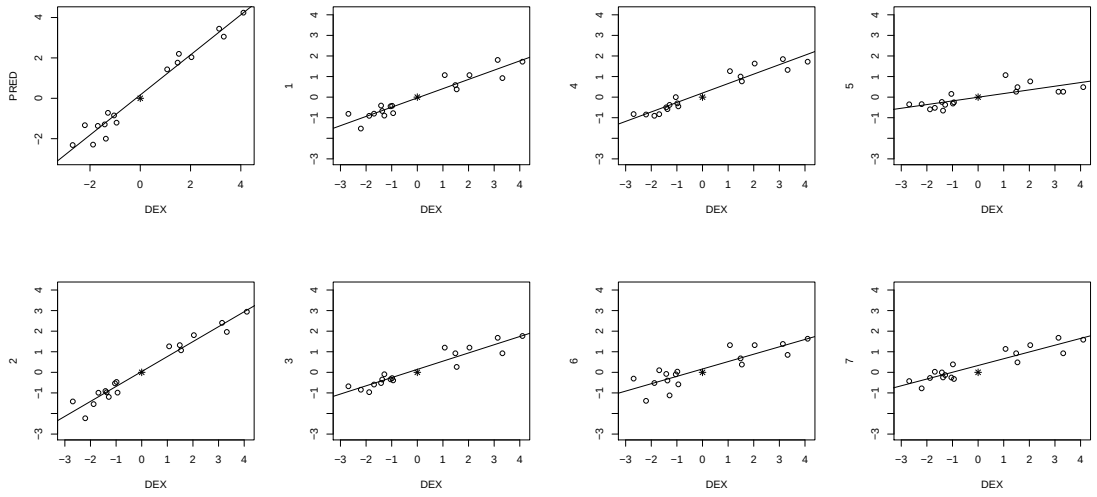


	ANGPTL4	ENaCα	GILZ	KIP2	MCJ	SDPR	SLC19A2
DEX	4.1	2.3	8.8	17.2	9.8	3.1	2.8
1	2.1	2.6	3.5	3.3	2.0	1.7	1.4
2	3.5	2.4	5.3	7.7	4.1	2.5	2.1
3	2.3	2.5	3.2	3.4	2.0	2.3	1.4
4	3.1	2.4	3.6	3.3	2.7	2.0	1.7
5	1.7	1.9	1.2	1.4	1.1	1.2	1.3
6	2.5	2.7	2.6	3.1	1.9	1.8	1.4
7	2.5	2.6	3.2	3.0	1.9	2.1	1.5
8	3.3	3.3	4.0	5.0	2.8	2.3	2.6
9	3.7	3.0	4.7	5.2	3.0	2.2	2.2
10	2.9	2.9	3.1	3.5	2.1	2.2	2.2
11	2.4	2.4	2.8	2.8	1.9	1.8	1.9
12	2.6	2.7	2.6	2.3	1.8	2.2	2.8
13	0.7	0.7	0.8	1.1	0.8	1.0	1.5
14	2.4	2.8	3.0	3.2	2.1	1.9	1.9
15	1.3	1.3	1.4	1.3	1.0	1.3	1.4
PRED	4.1	3.1	10.9	18.9	9.0	3.3	4.0

	SNK	BHLHB2	GEM	NAV3
DEX	64.2	47.6	51.5	81.8
1	50.3	40.8	26.5	63.0
2	59.0	48.3	31.0	76.8
3	18.8	25.0	21.0	48.0
4	30.5	29.5	0.0	50.0
5	26.2	15.8	-11.5	17.0
6	52.0	26.8	6.5	51.0
7	22	28.0	16.0	46.8
8	17.8	8.0	9.5	33.2
9	25.4	14.4	-1.0	45.2
10	22.75	16.8	-3.5	47.4
11	11.8	14.8	12.0	30.6
12	2.4	13.2	-27.5	25.0
13	-64.5	-8.8	-51.0	-96.0
14	2.06	5.3	-14.0	19.2
15	6.7	20.0	-9.5	-4.8
PRED	43.5	40.8	44.5	56.8

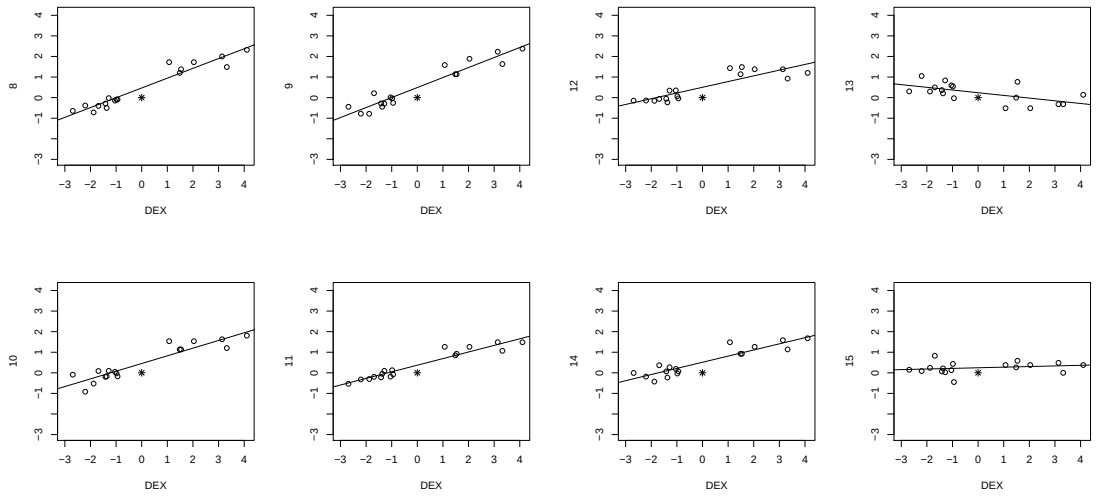
IL8	RANTES	GRO1	MCP1	GM-CSF	IL6
64.2	67.4	45.0	74.2	64.4	86.8
24.7	43.5	22.2	46.7	37.3	44.0
46.7	48.0	17.1	65.7	49.0	68.3
30.3	31.3	11.6	48.7	21.5	41.7
29.1	44.3	21.4	46.6	33.3	47.7
14.9	32.3	12.8	33.7	36.5	20.7
5.2	-5.5	-9.8	30.2	24.1	34.3
8.8	7.4	-9.1	26.9	18.4	37.5
18.1	24.3	16.2	39.2	29.4	35.3
22.8	-8.8	8.6	42.2	34.6	29.0
15.9	2.6	4.2	31.1	18.9	34.8
18.1	15.4	-0.2	22.6	12.6	37.5
12.2	11.0	5.4	20.6	18.7	20.3
-21.7	-35.8	-25.9	-19.7	-11.0	-4.0
1.0	-16.2	9.0	29.1	23.6	13.5
-5.2	-48.2	-13.8	-7.6	-6.1	-2.0

Figure 4.2: Figure 2 from ([59]). Downregulated genes are expressed as a percentage of dexamethasone response. Up regulated genes are expressed as log fold change over vehicle control. Red scale colors denote different levels of up regulation and green scale colors different levels of down regulation.



(a) prednisolone and NSRs 1-3

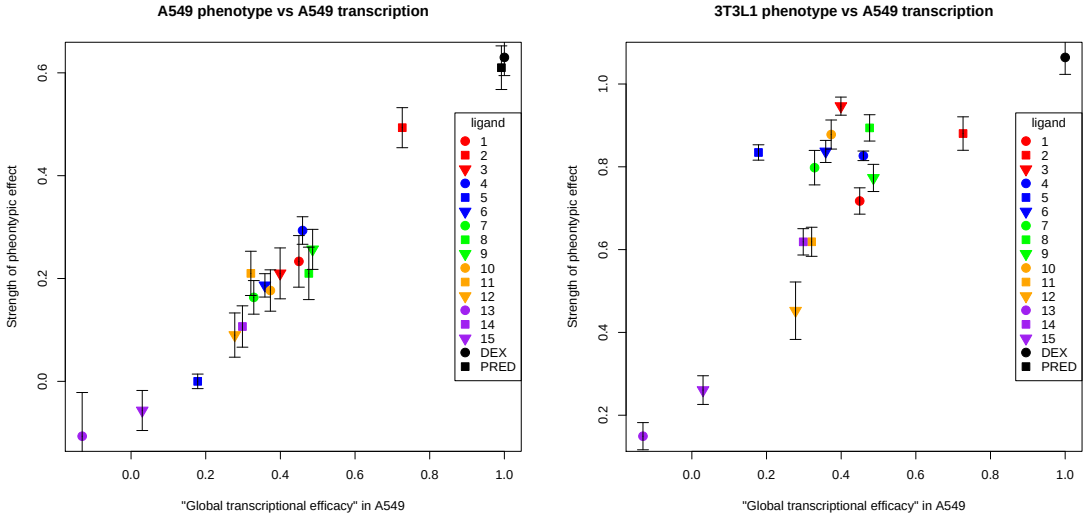
(b) ligands 4-7



(c) ligands 8-11

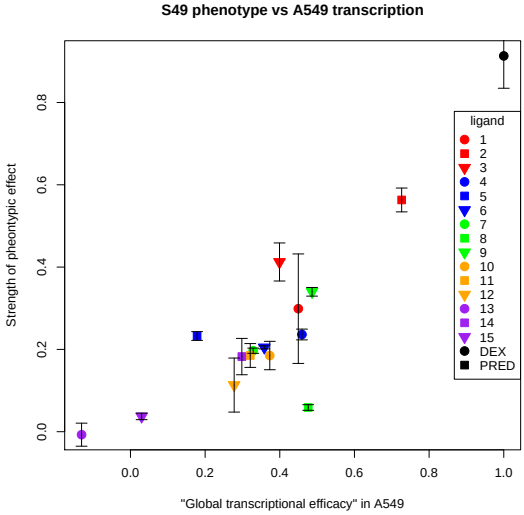
(d) ligands 12-15

Figure 4.3: Data from figure 4.2, replotted on a log-log scale, where each point represents a single gene).



(a) Proliferation in A549

(b) Differentiation of 3T3L1



(c) Cell death in S49

Figure 4.4: Relationship of strength of phenotypic response in three different cell types to the global transcriptional response in A549. The phenotypic response is normalized to a range from 0 (response to vehicle) to 1 (response to dexamethasone).

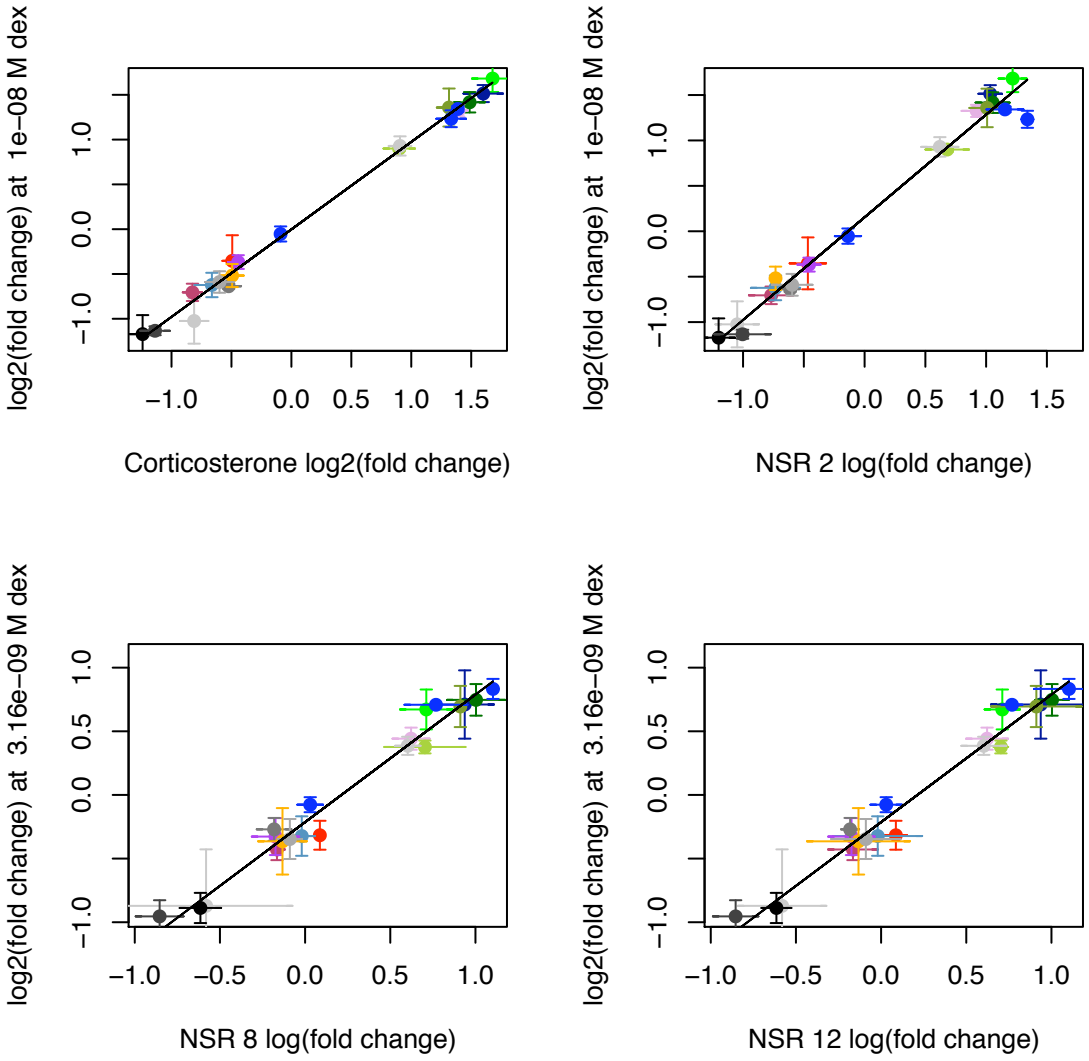


Figure 4.5: The transcriptional response of each of the ligand is plotted against the dosage of dexamethasone that it most closely matches. Axes are log fold change in expression of each gene over a vehicle control.

Chapter 5

Conclusions and future directions

5.1 Conclusions

GR occupancy is highly predictive of upregulation, and less so of downregulation. However, only a small fraction of the genes with nearby GR binding sites are regulated by GR in a given cellular context

If we were to use binding within a 20 kb window around the transcription start site to predict what genes are upregulated, we would correctly predict 61% of U2OS genes as regulated based on binding, while we would incorrectly predict 16% of unregulated genes as upregulated. Looked at from another angle, however, only 19% of genes with an associated binding site in U2OS are upregulated. This apparent discrepancy is due to the fact that only a small percentage of genes in a given cell type are regulated. It is interesting to speculate about the 81% of genes with an associated binding site that are not regulated. Some of these genes might be regulated in different conditions in U2OS cells, when another cofactor is present. Or these genes might show a delayed

response to glucocorticoids, only present at later time points. This large set of binding sites constitutes a set of GRBRs which are likely not GREs – at least not at the closest associated gene.

Turning our attention to down regulation, it is clear that the association between binding and down regulation is much weaker than between binding and upregulation. Recent reports have suggested little or no statistical association between GR and ER binding and down regulation. The results presented here suggest that, although weaker than the association with upregulation, GR occupancy is associated with down regulation, particularly in Nalm6 cells. This result is consistent with observations in reporters. Both activation and repression have been demonstrated in reporters to be the direct result of GR presence, although down regulation in a reporter is less common than upregulation.

Combining sequence and binding data improves the predictive power over binding alone.

The results presented here, as well as those in many other studies, show that a significant fraction of the GR binding sites do not contain a GBS, the canonical GR binding motif. GR is known to regulate transcription through tethering to another DNA bound transcription factor, so the lack of a GBS at some sites is expected. To explore the function of the binding sites lacking a GBS, the binding sites in each cell type were partitioned into two groups based on the presence of absence of this motif. Then the association of each group of binding sites with transcriptional regulation was assessed. Strikingly, binding sites with a GBS are more closely associated with upregulated genes and binding sites lacking a GBS are more closely associate with

downregulated genes. Thus binding data combined with sequence data has more predictive power than binding data alone.

No other clear associations were seen between the presence of particular sequence motifs in a binding site and transcriptional response. However, a larger data set with more cell types and more time points for transcriptional response may help uncover additional associations. For instance, it is possible that some secondary motifs are associated with an earlier transcriptional response and others with a later response. Additionally, the distance of the motif from the transcription start site might be related to the kinetics of the transcriptional response.

The information required for cell type specificity is contained within a relatively small piece of DNA.

Given the observation that the majority of binding sites are cell type specific, the next obvious question is whether this cell type specificity can be recapitulated in reporters. Based on past experience with reporters, it was not clear that this recapitulation would work, particularly for binding sites that contain a GBS. There are several reasons that this recapitulation might be expected to fail: (1) in a reporter, the site is no longer in its native chromatin context, removing the influence of chromatin accessibility; (2) in a reporter, the binding site is located right next to the promoter, removing any influence of secondary structure of DNA in the genomic context, and (3) even an isolated GBS placed next to an SV40 promoter in a reporter is in generally glucocorticoid responsive. In short, transcriptional response in reporters is more promiscuous than it is in the genomic context.

Two reporters were selected for testing, one for a site that was U2OS specific and

another for a site that was A549 specific. Both sites were approximately 400 base pairs long and both contained a GBS. The reporters show a striking cell type specificity, indicating that the determinants of cell type specific binding and regulation lie within the 400 base pair region used to construct the reporter. It will be interesting to narrow down the cell type specificity to a smaller region within these 400 base pairs. In addition, it will be interesting to find out whether, as a general rule, cell type specific GRBRs will generate cell type specific reporters. Taken together, this information will help unravel the cis-regulatory determinants of cell type specificity.

Another piece of evidence suggesting that the determinants of cell type specificity reside in the sequence region immediately surrounding the GR occupied sites comes from sequence analysis of these regions. There is a striking difference in the prevalence of secondary transcription factor motifs within these regions. For example, AP-1 sites are strongly overrepresented in A549 and U2OS binding sites, but not at all overrepresented in Nalm6 sites, while the reverse is true of RUNX1 sites. The relative levels or activities of these and other transcription factors in each cell type could thus govern the sites at which GR is able bind.

The canonical GR motif is indistinguishable in different cell types.

One possible explanation for differences in GR binding in different cell types could be a preference of GR for slightly different canonical GR motifs in different cell types. Given the evidence that DNA is an allosteric modulator of GR,[44] we might expect that different GBSs would favor different GR conformations, that would then favor different binding partners, leading to a preference for particular variants of the canonical GBS in different cell types. In fact, the GR motifs derived independently

from each of the three cell types studied are not noticeably different. If there is cell type specificity of the GBS, it is more complicated than a simple PWM can detect.

A gene regulated in two cell types is more likely than an unregulated gene to have binding sites in common, even when the gene is activated in one cell type and repressed in the other.

Comparing U2OS and A549, a gene that is upregulated in one cell type and down regulated in another cell type is almost as likely to share a common binding site (50%) as a gene that is upregulated in both cell types (61%). A gene that is unregulated in both cell types is much less likely(15%) to have a binding site in common between cell types. In this analysis, only genes that have at least one binding site in at least one cell type are considered.

The fact that 39% of the genes that are upregulated in both cell types do not have a binding site in common suggests two different possibilities, both of which have interesting implications. The first possibility is that only the overlapping binding sites function as GREs. Thus this overlap data might be used to distinguish which GRBRs are GREs and which are not. The second possibility is that the non-overlapping sites are also GREs, suggesting that genes that are upregulated in different cell types are frequently regulated through different mechanisms.

Turning to the case of genes that are upregulated in one cell type and downregulated in another, the degree of overlap between the binding sites in this set of genes suggests the possibility that the same site could commonly inducing upregulation in one cell type and downregulation in another. This phenomena with GR was first demonstrated by Diamond et al.,[13] but it is unclear whether this is a common and

physiologically significant phenomena. The results presented here suggest that it may be quite common, though significantly more mechanistic work remains to confirm this hypothesis.

Cell type specific differences in response to different GR ligands are not explained by gene-specific effects.

Wang et al. demonstrate that a set of GR ligands based on the arylpyrazole scaffold have different degrees of GR-like effects on different cell types. They propose that these cell type specific effects are due to gene-specific effects of different ligands.[47] The underlying idea is that different ligands in the GR binding pocket produce slightly different conformational changes in GR and thus expose different binding surfaces to other coregulatory proteins. Since different surfaces of GR have different contributions to regulation at different genes [43]), the different GR ligands would in this model be expected to have different transcriptional effects at each gene. The results in chapter 4 undermines some of the evidence for this gene-specific effect in this line of GR ligands. Instead, the data suggest an alternate hypothesis, that the cell-type specific differences are explained by differences in global transcriptional efficacy of each ligand in a given cell type.

5.2 Future Directions

The present work opens up more new questions than it answers. A few interesting avenues for additional investigation are discussed here. These directions are all geared towards understanding the mechanisms behind cell type specificity. These cross-cell type comparisons are fundamentally a tool to reach a better mechanistic

understanding of all GR mediated regulation.

Determine what drives the different behavior in Nalm6.

Global patterns of binding are dramatically different in Nalm6 than in the other cell types. First of all, only a minority of Nalm6 binding sites contain a GBS. Second, Nalm6 binding sites tend to be close to transcription start sites than those in A549 and U2OS. And third, even sites that bind in both Nalm6 and one of the other cell types frequently have a distinctly different structure in Nalm6. It would be interesting to determine whether this pattern holds true in other ALL cell lines, in ALL patient samples, and in non-cancerous lymphocytes. Some possible mechanistic explanations for this Nalm6 behavior include: expression of different isoforms of GRs; different post-translational modifications of GR; and different levels of coregulators. One interesting avenue would be to focus on binding sites that have different structures in Nalm6, and to look at how these sites act in a reporter context.

Investigate cell type specific GRBRs in reporters.

It would be interesting to choose 10-20 cell specific GRBRs and clone them into reporters to determine whether cell type specificity of binding can in general be recapitulated in a reporter. Then these reporters could be edited piecewise to identify the precise sequence determinants of cell specificity. Once the cis determinants are known, it will be easier to identify the trans factors involved.

Knock down putative cofactors in different cell types and compare resultant transcriptional and binding profiles.

While the results presented here suggest that particular cofactors play a role at particular binding sites, the best confirmation of this will be to knock down these putative cofactors and look for changes in GR occupancy at the predicted sites. This will be particularly instructive with cofactors not generally associated with GR, such as ELK1. Based on the results presented here, knocking down ELK1 is expected to have a major effect in Nalm6 but little effect in the other cell types. However, it is important to note that all the Ets factors share a similar binding site with ELK1, so it might be necessary to knock down more than on Ets factor to find the significant GR partner.

Look at common sites near genes that are up in one cell type and down in another.

There are a number of genes that are upregulated in one cell type but downregulated in another cell type, yet share a common binding site between the two cell types. Diamond et al. identify a GRE that activates or represses transcription depending on the relative levels of components of AP-1. It would be interesting to determine whether this phenomena is a common one. To investigate this, reporters could be constructed from these sites and tested in different cell types.

Use bacterial artificial chromosomes (BACs) to determine the contribution of multiple binding sites to the regulation of a gene.

Bacterial artificial chromosomes, or BACs, allow the insertion and editing of large genomic regions into a reporter system. In particular, it would be interesting to start with a genomic region that contain multiple binding sites associated with a regulated gene and ask which binding site(s) contribute to regulation. The example of TXNIP, which is upregulated in all three cell types, is an interesting one: in a 10 kB region upstream of this gene, there are multiple binding sites in all three cell types studied, but only one site is common to all three cell types.

5.3 Final remarks

As genome wide data sets become easier to obtain, the challenges in thinking about and analyzing these data is growing. The work presented here represents an attempt to ask a biologist's questions using statistical tools. These methods tie into more focused mechanistic work in a couple ways. First, they suggest how prevalent a phenomenon previously seen at a handful of genes – for instance, long distance regulation – is likely to be. Second, they generate strong associations that require a more mechanistic approach to validate, such as the apparent role of a particular cofactor in one cell type over another. And third, they identify new phenomena – for instance, the relative paucity of GBSs in Nalm6 GRBRs – that have not been seen before. It is hoped that the approaches and methods used this work demonstrate a way of grasping these large data sets and teasing out the biologically interesting information.

Bibliography

[1] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324:1720–1723, Jun 2009.

[2] S. Bailey, A. G. Hall, A. D. Pearson, and C. P. Redfern. The role of AP-1 in glucocorticoid resistance in leukaemia. *Leukemia*, 15:391–397, Mar 2001.

[3] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, 34:W369–373, Jul 2006.

[4] L. J. Ball, N. Levy, X. Zhao, C. Griffin, M. Tagliaferri, I. Cohen, W. A. Ricke, T. P. Speed, G. L. Firestone, and D. C. Leitman. Cell type- and estrogen receptor-subtype specific regulation of selective estrogen receptor modulator regulatory elements. *Mol. Cell. Endocrinol.*, 299:204–211, Feb 2009.

[5] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Pea-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid,

W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk, and T. R. Hughes. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133:1266–1276, Jun 2008.

[6] E. C. Bolton, A. Y. So, C. Chaivorapol, C. M. Haqq, H. Li, and K. R. Yamamoto. Cell- and gene-specific regulation of primary target genes by the androgen receptor. *Genes Dev.*, 21:2005–2017, Aug 2007.

[7] S. C. Brooks and D. F. Skafar. From ligand structure to biological activity: modified estratrienes and their estrogenic and antiestrogenic effects in MCF-7 cells. *Steroids*, 69:401–418, Jun 2004.

[8] A. M. Brzozowski, A. C. Pike, Z. Dauter, R. E. Hubbard, T. Bonn, O. Engstrm, L. Ohman, G. L. Greene, J. A. Gustafsson, and M. Carlquist. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, 389:753–758, Oct 1997.

[9] V. L. Chandler, B. A. Maler, and K. R. Yamamoto. DNA sequences bound specifically by glucocorticoid receptor in vitro render a heterologous promoter hormone responsive in vivo. *Cell*, 33:489–499, Jun 1983.

[10] U. R. Chandran, B. S. Warren, C. T. Baumann, G. L. Hager, and D. B. DeFranco. The glucocorticoid receptor is tethered to DNA-bound Oct-1 at the mouse gonadotropin-releasing hormone distal negative glucocorticoid response element. *J. Biol. Chem.*, 274:2372–2378, Jan 1999.

[11] F. Clard, Y. Moshkin, F. Karch, and R. K. Maeda. Probing long-distance regulatory interactions in the Drosophila melanogaster bithorax complex using Dam identification. *Nat. Genet.*, 38:931–935, Aug 2006.

[12] M. J. Coghlan, P. R. Kym, S. W. Elmore, A. X. Wang, J. R. Luly, D. Wilcox, M. Stashko, C. W. Lin, J. Miner, C. Tyree, M. Nakane, P. Jacobson, and B. C. Lane. Synthesis and characterization of non-steroidal ligands for the glucocorticoid receptor: selective quinoline derivatives with prednisolone-equivalent functional activity. *J. Med. Chem.*, 44:2879–2885, Aug 2001.

[13] M. I. Diamond, J. N. Miner, S. K. Yoshinaga, and K. R. Yamamoto. Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science*, 249:1266–1272, Sep 1990.

[14] D. Duma, C. M. Jewell, and J. A. Cidlowski. Multiple glucocorticoid receptor isoforms and mechanisms of post-translational modification. *J. Steroid Biochem. Mol. Biol.*, 102:11–21, Dec 2006.

[15] W. Eberhardt, M. Schulze, C. Engels, E. Klasmeier, and J. Pfeilschifter. Glucocorticoid-mediated suppression of cytokine-induced matrix metalloproteinase-9 expression in rat mesangial cells: involvement of nuclear factor-kappaB and Ets transcription factors. *Mol. Endocrinol.*, 16:1752–1766, Aug 2002.

[16] C. D. Geng, J. R. Schwartz, and W. V. Vedeckis. A conserved molecular mechanism is responsible for the auto-up-regulation of glucocorticoid receptor gene promoters. *Mol. Endocrinol.*, 22:2624–2642, Dec 2008.

[17] C. D. Geng and W. V. Vedeckis. c-Myb and members of the c-Ets family of transcription factors act as molecular switches to mediate opposite steroid regulation of the human glucocorticoid receptor 1A promoter. *J. Biol. Chem.*, 280:43264–43271, Dec 2005.

[18] D. G. Gilliland. The diverse role of the ETS family of transcription factors in cancer. *Clin. Cancer Res.*, 7:451–453, Mar 2001.

[19] O. Hakim, S. John, J. Q. Ling, S. C. Biddie, A. R. Hoffman, and G. L. Hager. Glucocorticoid receptor activation of the Ciz1-Lcn2 locus by long range interactions. *J. Biol. Chem.*, 284:6048–6052, Mar 2009.

[20] S. R. Holmstrom, S. Chupreta, A. Y. So, and J. A. Iiguez-Lluh. SUMO-mediated inhibition of glucocorticoid receptor synergistic activity depends on stable assembly at the promoter but not on DAXX. *Mol. Endocrinol.*, 22:2061–2075, Sep 2008.

[21] O. R. Homann and A. D Johnson. MochiView: versatile software for genome browsing and DNA motif analysis. Submitted for publication. Software available at http://johnsonlab.ucsf.edu/.

[22] K. Johansson-Haque, E. Palanichamy, and S. Okret. Stimulation of MAPK-phosphatase 1 gene expression by glucocorticoids occurs through a tethering mechanism involving C/EBP. *J. Mol. Endocrinol.*, 41:239–249, Oct 2008.

[23] S. John, P. J. Sabo, T. A. Johnson, M. H. Sung, S. C. Biddie, S. L. Lightman, T. C. Voss, S. R. Davis, P. S. Meltzer, J. A. Stamatoyannopoulos, and G. L.

Hager. Interaction of the glucocorticoid receptor with the chromatin landscape. *Mol. Cell*, 29:611–624, Mar 2008.

[24] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316:1497–1502, Jun 2007.

[25] M. Kian Tee, I. Rogatsky, C. Tzagarakis-Foster, A. Cvoro, J. An, R. J. Christy, K. R. Yamamoto, and D. C. Leitman. Estradiol and selective estrogen receptor modulators differentially regulate target genes with estrogen receptors alpha and beta. *Mol. Biol. Cell*, 15:1262–1272, Mar 2004.

[26] S. A. Krum, G. A. Miranda-Carboni, M. Lupien, J. Eeckhoute, J. S. Carroll, and M. Brown. Unique ERalpha cistromes control cell type-specific gene regulation. *Mol. Endocrinol.*, 22:2393–2406, Nov 2008.

[27] Y. Le Drean, N. Mincheneau, P. Le Goff, and D. Michel. Potentiation of glucocorticoid receptor transcriptional activity by sumoylation. *Endocrinology*, 143:3482–3489, Sep 2002.

[28] B. Lenhard and W. W. Wasserman. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, 18:1135–1136, Aug 2002.

[29] B. Lin, J. Wang, X. Hong, X. Yan, D. Hwang, J. H. Cho, D. Yi, A. G. Utleg, X. Fang, D. E. Schones, K. Zhao, G. S. Omenn, and L. Hood. Integrated expression profiling and ChIP-seq analyses of the growth inhibition response program of the androgen receptor. *PLoS ONE*, 4:e6589, 2009.

[30] D. G. Lloyd, H. M. Smith, T. O'Sullivan, D. M. Zisterer, and M. J. Meegan. Synthesis, structure-activity relationships and antagonistic effects in human MCF-7

breast cancer cells of flexible estrogen receptor modulators. *Med Chem*, 1:335–353, Jul 2005.

[31] H. F. Luecke and K. R. Yamamoto. The glucocorticoid receptor blocks P-TEFb recruitment by NFkappaB to effect promoter-specific transcriptional repression. *Genes Dev.*, 19:1116–1127, May 2005.

[32] V. Matys, E. Fricke, R. Geffers, E. Gssling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Mnch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31:374–378, Jan 2003.

[33] S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen, and K. R. Yamamoto. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science*, 324:407–410, Apr 2009.

[34] T. Meyer, D. B. Starr, and J. Carlstedt-Duke. The rat glucocorticoid receptor mutant K461A differentiates between two different mechanisms of transrepression. *J. Biol. Chem.*, 272:21090–21095, Aug 1997.

[35] J. Mullick, H. K. Anandatheerthavarada, G. Amuthan, S. V. Bhagwat, G. Biswas, V. Camasamudram, N. K. Bhat, S. E. Reddy, V. Rao, and N. G. Avadhani. Physical interaction and functional synergy between glucocorticoid receptor and Ets2 proteins for transcription activation of the rat cytochrome P-450c27 promoter. *J. Biol. Chem.*, 276:18007–18017, May 2001.

[36] C. K. Osborne, H. Zhao, and S. A. Fuqua. Selective estrogen receptor modulators: structure, function, and clinical use. *J. Clin. Oncol.*, 18:3172–3186, Sep 2000.

[37] D. Pearce, W. Matsui, J. N. Miner, and K. R. Yamamoto. Glucocorticoid receptor transcriptional activity determined by spacing of receptor and nonreceptor DNA sites. *J. Biol. Chem.*, 273:30081–30085, Nov 1998.

[38] A. C. Pike, A. M. Brzozowski, R. E. Hubbard, T. Bonn, A. G. Thorsell, O. Engstrm, J. Ljunggren, J. A. Gustafsson, and M. Carlquist. Structure of the ligand-binding domain of oestrogen receptor beta in the presence of a partial agonist and a full antagonist. *EMBO J.*, 18:4608–4618, Sep 1999.

[39] M. Rahman, Y. Hirabayashi, T. Ishii, M. Watanabe, L. Maolin, and T. Sasaki. Prednisolone sodium succinate down-regulates BSAP/Pax5 and causes a growth arrest in the Nalm6 pre-B cell line. *Tohoku J. Exp. Med.*, 193:237–244, Mar 2001.

[40] T. E. Reddy, F. Pauli, R. O. Sprouse, N. F. Neff, K. M. Newberry, M. J. Garabedian, and R. M. Myers. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.*, 19:2163–2171, Dec 2009.

[41] M. E. Ritchie, J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G. K. Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23:2700–2707, Oct 2007.

[42] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He,

M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4:651–657, Aug 2007.

[43] I. Rogatsky, J. M. Trowbridge, and M. J. Garabedian. Glucocorticoid receptor-mediated cell cycle arrest is achieved through distinct cell-specific transcriptional regulatory mechanisms. *Mol. Cell. Biol.*, 17:3181–3193, Jun 1997.

[44] I. Rogatsky, J. C. Wang, M. K. Derynck, D. F. Nonaka, D. B. Khodabakhsh, C. M. Haqq, B. D. Darimont, M. J. Garabedian, and K. R. Yamamoto. Target-specific utilization of transcriptional regulatory surfaces by the glucocorticoid receptor. *Proc. Natl. Acad. Sci. U.S.A.*, 100:13845–13850, Nov 2003.

[45] A. Sandelin, W. Alkema, P. Engstrm, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32:D91–94, Jan 2004.

[46] H. Schcke, A. Schottelius, W. D. Dcke, P. Strehlke, S. Jaroch, N. Schmees, H. Rehwinkel, H. Hennekes, and K. Asadullah. Dissociation of transactivation from transrepression by a selective glucocorticoid receptor agonist leads to separation of therapeutic effects from side effects. *Proc. Natl. Acad. Sci. U.S.A.*, 101:227–232, Jan 2004.

[47] N. Shah and T. S. Scanlan. Design and evaluation of novel nonsteroidal dissociating glucocorticoid receptor ligands. *Bioorg. Med. Chem. Lett.*, 14:5199–5203, Oct 2004.

[48] A. D. Sharrocks, A. L. Brown, Y. Ling, and P. R. Yates. The ETS-domain transcription factor family. *Int. J. Biochem. Cell Biol.*, 29:1371–1387, Dec 1997.

[49] A. K. Shiau, D. Barstad, P. M. Loria, L. Cheng, P. J. Kushner, D. A. Agard, and G. L. Greene. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell*, 95:927–937, Dec 1998.

[50] L. Sipos and H. Gyurkovics. Long-distance interactions between enhancers and promoters. *FEBS J.*, 272:3253–3259, Jul 2005.

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

[52] G. K. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31:265–273, Dec 2003.

[53] A. Y. So, C. Chaivorapol, E. C. Bolton, H. Li, and K. R. Yamamoto. Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor. *PLoS Genet.*, 3:e94, Jun 2007.

[54] C. Stahn, M. Lwenberg, D. W. Hommes, and F. Buttgereit. Molecular mechanisms of glucocorticoid action and selective glucocorticoid receptor agonists. *Mol. Cell. Endocrinol.*, 275:71–78, Sep 2007.

[55] D. B. Starr, W. Matsui, J. R. Thomas, and K. R. Yamamoto. Intracellular receptors use a common mechanism to interpret signaling information at response elements. *Genes Dev.*, 10:1271–1283, May 1996.

[56] S. Teurich and P. Angel. The glucocorticoid receptor synergizes with Jun homodimers to activate AP-1-regulated promoters lacking GR binding sites. *Chem. Senses*, 20:251–255, Apr 1995.

[57] A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457:854–858, Feb 2009.

[58] J. C. Wang, M. K. Derynck, D. F. Nonaka, D. B. Khodabakhsh, C. Haqq, and K. R. Yamamoto. Chromatin immunoprecipitation (ChIP) scanning identifies primary glucocorticoid receptor target genes. *Proc. Natl. Acad. Sci. U.S.A.*, 101:15603–15608, Nov 2004.

[59] J. C. Wang, N. Shah, C. Pantoja, S. H. Meijsing, J. D. Ho, T. S. Scanlan, and K. R. Yamamoto. Novel arylpyrazole compounds selectively modulate glucocorticoid receptor regulatory activity. *Genes Dev.*, 20:689–699, Mar 2006.

[60] K. R. Yamamoto. Steroid receptor regulated transcription of specific genes and gene networks. *Annu. Rev. Genet.*, 19:209–252, 1985.

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

[62] J. Zhou and J. A. Cidlowski. The human glucocorticoid receptor: one gene, multiple proteins and diverse responses. *Steroids*, 70:407–417, 2005.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

3/30/2010

Date