DNA Shape and Sequence of Binding Sites Modulate Regulation of Gene Expression by the Glucocorticoid Receptor

Inaugural-Dissertation to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)
submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Stefanie Schöne
from Dresden

2016

The dissertation was prepared under supervision of Dr. Sebastiaan H. Meijsing and Prof. Morgane Thomas-Chollier at the Max Planck institute for molecular genetics in Berlin from October 2011 to May 2016.

Berlin, 31. Mai 2016

1. Gutachter: Dr. Sebastiaan H. Meijsing
2. Gutachter: Prof. Dr. Markus Wahl

Date of Disputation: 14.09.2016

Contents

Acknowledgements

I would like to express my very great appreciation to Sebastiaan Meijsing and Morgane Thomas-Chollier, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques. I am very grateful that I was so lucky to have both of you as my supervisors.

A special thanks I would like to offer to Marcel Jurk, who patiently guided me in the process of our joint "flanking site" project and introduced me to the fields of NMR spectroscopy. Samuel Collombet, I want to thank for the support and tips for the DEseq2 analysis. I want to thank Prof. Martin Vingron for allowing me to be part of his genial department, for the support and for the useful comments on my projects.

I am particularly grateful for the technical assistance given by Edda Einfeldt. You are indeed a Master of ChIPs and the "good Soul" of our lab supporting us with sweets, fruits and vegetable. Moreover, I want to thank all current and former members of the Meijsing group: Verena Thormann, Marina Borschiwer, Jonas Telorac, Stephan Starick, Katja Borzym, Sergey Prykhozhij and of the Vingron Department: Mike Love, Jose Muino, Juliane Perner, Ruping Sun for providing a nice working atmosphere and for useful comments and help in conducting biological and computational experiments.

I want to thank Dr. Joachim Forner for giving me feedback on my thesis manuscript.

I want to thank Prof. Markus Wahl for reviewing this work.

For the financial support to pursue my PhD thesis during my motherhood I want to thank the Christiane-Nüsslein-Volhard foundation and the L'Oréal-UNESCO For Women in Science program.

Bei meinem Partner, Manuel Utecht, möchte ich mich ganz besonders bedanken, da er mich vollkommen unterstützt hat und mir den Freiraum geschaffen hat meine Doktorarbeit zu schreiben. Danke, dass du mich liebevoll beim Schreibprozess mit technischen Support, Ideen und Hinweisen unterstützt hast.

Hiermit möchte ich mich auch ganz herzlich bei meiner Mama, meinem Papa und meinem Bruder für die Unterstützung und das Vertrauen in mich bedanken. Ihr habt mich bei all meinen Bestrebungen in meiner wissenschaftlichen Ausbildung

begleitet und wart immer für mich da.

Record of Publications

- Telorac, J., Prykhozhij, S. V., **Schöne, S.**, Meierhofer, D., Sauer, S., Thomas-Chollier, M., and Meijsing, S. H., *Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements.*, Nucleic acids research, **2016**, July.

- **Schöne, S.**, Jurk, M., Bagherpoor Helabad, M., Dror, I., Lebars, I., Kieffer, B., Imhof, P., Rohs, R., Vingron, M., Thomas-Chollier, M., and Meijsing, S. H., *Sequences flanking the core binding site modulate glucocorticoid receptor structure and activity*, Nature Communications, **2016**, September.

- **Schöne, S.**, Collombet, S., Thomas-Chollier, M., and Meijsing, S. H., *High-throughput and quantitative assessment of GR binding sequence activity by synthetic Starr-seq*, manuscript in preparation.

1. Abstract

1.1. Abstract

In my thesis, I studied the details of gene regulation by the glucocorticoid receptor (GR), a nuclear hormone receptor acting as a transcription factor. My aim was to better understand how sequence variants bound by the same transcription factor may play a role in the fine-tuning of gene expression. GR is an important therapeutic target in medicine and has been intensively studied. Despite our broad understanding of GR action, we still do not fully understand the details of gene regulation. Researchers found that different GR binding sequence (GBS) variants induce different levels of GR activity, which is seemingly disconnected from *in vitro* binding affinity of GR [1]. It was hypothesized that DNA allostery might explain the differences in GR activity. Additionally, classical enhancer studies, which have been mostly used so far, only allow low-thoughput studies of the effect of sequence variants on GR activity.

I developed and implemented a modified version of the STARR-seq (self-transcribing active regulatory region sequencing) method for quantitative massively parallel assessment of short synthesized GBS variants for their ability to fine-tune GR action. I found that GBS variants and sequences flanking a GBS indeed modulate GR activity. For example, I identified a new GBS variant, which is C-rich and leads to strong GR dependent up-regulation of a target gene. I also studied GR dependent activation of endogenous genes and found that the direct flanking bases have an important role. At these positions, A/T are associated with strong up-regulation, but G/C are associated with weak regulation. To understand the effect induced by GBS flanking sites on GR action, I conducted structural experiments of the DNA and the DNA:GR protein complex. I found that the difference in activation of target genes was uncorrelated with binding affinity *in vitro* and *in vivo*. Further, DNA structure prediction showed that the flanking bases induce structural changes adjacent to and within the GBS. For GR, the different structural experiments revealed that the conformation of the dimer partners, the dynamics and the relative position of the dimer partners are affected by flanking bases. Altogether, these findings suggest that DNA allostery might induce different structural conformations in GR that affect GR downstream of binding and modulate its action. In the last part of my thesis, I found that the composition of the GR response element influences the level of expression noise. For example, multiple GR binding sites yield high expression and noise, whereas composite binding sites harboring a GR binding site and a binding site for other transcription factors results in high expression with

relatively little noise.

In summary, I could show that GBS variants affect fine-tuning of gene expression by GR and that DNA allostery might play an important role in fine-tuning the expression of individual GR target genes.

1.2. Zusammenfassung

In meiner Doktorarbeit habe ich die Feinheiten der Genregulation durch den Glukokortikoidrezeptor (GR) untersucht. GR ist ein nuklearer Hormonrezeptor, welcher als Transkriptionsfaktor fungiert. Insbesondere wollte ich verstehen, welche Rolle die verschiedenen Sequenzvarianten, welche vom selben Transkriptionsfaktor gebunden werden, bei der Feinregulierung der Expression von Zielgenen spielt. In der Medizin ist GR ein wichtiges therapeutisches Zielobjekt und wurde bereits intensiv erforscht. Doch trotz unseres großen Wissens um die Wirkung von GR, verstehen wir viele Feinheiten der Genregulierung durch GR nicht im vollen Umfang. Wissenschaftler haben herausgefunden, dass die Wirkung von GR offensichtlich nicht mit der Bindungsstärke von GR an die DNS im Zusammenhang steht [1]. Daher wurde spekuliert, ob DNS Allosterie ein Mechanismus ist, der die Stärke der Genexpression in diesen Zusammenhang erklärt. Darüber hinaus wurden viele Studien zu GR mit klassischen Enhancer-Methoden mit niedrigen Testsequenz-Durchsatz durchgeführt.

Ich dagegen nutze eine neue Technik, genannt STARR-seq (self-transcribing active regulatory region sequencing), um im Hochdurchsatz kurze synthetisierte GR-Bindungssequenzen (GBS) auf ihre quantitative Wirkung auf die GR-Genregulation zu testen. Tatsächlich fand ich heraus, dass GBS-Varianten aber auch die flankierenden Sequenzen einer GBS zur Feinregulierung durch GR beitragen. Zum Beispiel, konnte ich eine neue GBS-Variante bestimmen, welche C-reich ist und zu starker Hochregulierung von GR-Zielgenen führt. Darüber hinaus habe ich auch die Wirkung von GR im endogenen Zusammenhang untersucht. Ich konnte hierber herausstellen, dass die direkten flankierenden Sequenzen einer GBS, wenn sie A/T enthält zur starken Hochregulierung führte, während schwache Regulierung mit G/C in Verbindung stand. Der Unterschied in der Genregulierung konnte nicht durch die Bindungsstärke erklärt werden. Um den Effekt zu verstehen, welcher durch die flankierenden Sequenzen verursacht wurde, habe ich mehrere strukturelle Experimente für die DNS und das gebundene GR-Protein durchgeführt. DNS-Strukturvorhersagen zeigten, dass die DNS-Struktur der GBS durch die flankierenden Sequenzen verändert wird. Die strukturellen Experimente für GR zeigten, dass die Konformation der Dimer-Partner, die Dynamik und die relative Positionierung der Dimer-Partner durch die flankierenden Sequenzen beeinflusst wird. Aufgrund dieser Befunde, nehme ich an, dass DNS Allosterie diese unterschiedlichen GR-Strukturen verursacht hat und die Wirkung von GR beeinflusst. Im letzten Teil meiner Doktorarbeit habe ich untersucht, wie die Zell-zu-Zell Variabilität der Genexpression durch den Aufbau der GR-gebundenen Regulationsregion beeinflusst wird. Ich fand heraus, dass zum

Beispiel mehrere GBSs zu einer hoher Genexpression mit viel Variabilität führt, während eine Kompositsequenz bestehend aus einer GBS mit einer Bindungsstelle für einen anderen Transkriptionsfaktors zu einer hoher Genexpression mit relativ wenig Variabilität führt.

Alles in allem, konnte ich zeigen, dass GBS-Varianten die Feinregulierung der Genexpression durch GR beeinflussen und dass DNS Allosterie eine wichtige Rolle in der Feinregulierung der Genexpression zufällt.

2. Introduction

Transcriptional regulation is a key process by which a cell controls gene expression. The complexity of higher organisms like mammals is probably not based on a larger number of genes but on a more complex gene regulation network compared to simple animals like nematodes. This process is orchestrated by transcription factors (TFs) that define when, where and how much of a gene is expressed. More than 6% of our protein-coding genes encode for TFs, which shows how important transcriptional regulation is [2]. Further, transcriptional regulation is not a static process; it is highly flexible and can respond to a large variety of internal and external signals like cell cycle, nutrition or differentiation. Thus, in eukaryotes the regulation of genes is a complex concert of many different factors.

Over the last decades researchers are more and more able to understand the details of transcriptional regulation by transcription factors. The development of chromatin-immunoprecipitation (ChIP) [3], enhancer assays, chromosome conformation capture (3C) [4] and modern high throughput sequencing technologies allow us to study functional binding and action of hundreds of different TFs in genomes. However, despite our broad knowledge, the specific details of transcriptional regulation still remain partly elusive. For example, we do not fully understand why not all binding events of TFs really lead to gene regulation. Also we are not able to fully explain the differing strength of gene expression between binding events of a TF at different genes.

The aim of my thesis was to study the details of gene regulation by the glucocorticoid receptor (GR), a nuclear receptor acting as a transcription factor. In more detail, I wanted to understand how sequence variants bound by the same TF may play a role in the fine-tuning of gene expression of individual target genes. To achieve this, I used novel technologies like STARR-seq to quantify GR action in a high throughput enhancer reporter assay and structural studies to determine if and how binding site variants influence the structure of GR. I found that DNA is an allosteric factor in this process and can modulate the action of GR.

2.1. DNA: Structure and Organization

DNA is built by a combination of four different nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T). Yet, DNA is not only a stretch of letters to be read and interpreted by other molecules. DNA forms a highly structured, but flexible double helix consisting of a minor and major groove (Figure 2.1) [5]. The DNA is present as chromatin in the cell nucleus. To form chromatin, the DNA is

wrapped around histones which leads to a further compaction of the DNA. Histone tails may be chemically modified and there exist a large variety of chemical modifications, associated with DNA-based processes including closing and opening of chromatin, repression and activation of gene expression and defining among other promoters and regulatory regions [6]. All in all, DNA does not only store information, it largely provides the chromatin context in which genes encoded in the DNA are embedded and plays a key role in facilitating cell-type specific gene regulation.

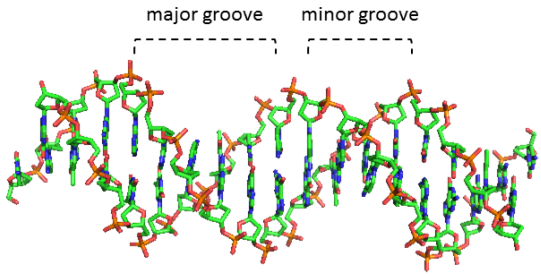


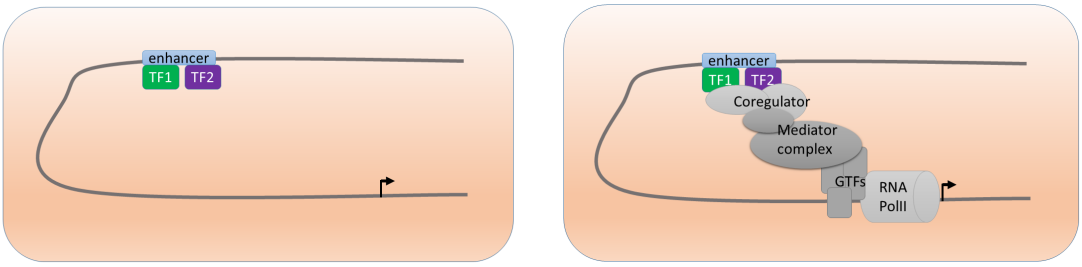
Figure 2.1.: Structure of a DNA double helix showing major and minor groove.

The human genome comprises around 3.2 billion bases, but only 2% of the DNA encode protein coding genes. The large majority (of 98%) are considered to be non-coding DNA. Parts of the non-coding DNA is transcribed into functional non-coding RNA molecules including ribosomal RNA, tRNA and microRNA [7]. Of particular interest for my thesis: non-coding DNA plays a key role in regulating the expression of genes via regulatory regions (promoter, enhancer, silencer). Other functions of non-coding DNA include pseudo-genes, scaffold attachment regions, centromeres and telomeres. In this way, 80% of human genomic DNA may have a biological function [8] and may not be "junk" as it was termed in the past [9].

2.2. Transcriptional Regulation

Regulatory regions in our genome can activate or repress gene expression upon binding of transcription factors. Regulatory regions that activate gene expression are called enhancers and regulatory regions that suppress gene expression are called silencers. They can be situated upstream or downstream of a gene or even within a gene. They can be located close to the transcription start site (TSS) (proximal promoter element) or as far away as 1 million base pairs (Mbp) (distal regulatory region) [10]. Interestingly, enhancers seem to function independent of orientation or location [11]. Yet, distance between TSS and enhancer seems to be important for the strength of gene activation [12, 13]. A single enhancer can regulate multiple genes [14], and a gene can be regulated by multiple enhancers. During development and varying cellular programs, different enhancers can be active at different stages and time points to regulate a single gene [15]. This diversity and complexity of regulatory elements in gene regulation is also the reason why research in this field

is rather challenging. Based on predictions, the mammalian genome may have as many as 1 million enhancers [8, 16]. Enhancers contain clusters of transcription factor binding sites (TFBS) for different TFs. TFs can be expressed in a cell-specific manner. Once an enhancer is bound by TFs the transcriptional regulation of associated genes is started. The TFs start recruiting coregulators and mediator complex (Figure 2.2), so called co-activators. Some co-activators also possess histone acetyltransferase activity, which acetylate nearby histones and increase access to DNA. As a last step, the transcription preinitiation complex is formed and RNA polymerase 2 (RNA PolII) starts transcribing DNA into mRNA [17]. In addition to transcription initiation, TFs can control transcriptional elongation and re-initiation.



(a) (b)

Figure 2.2.: Transcriptional regulation is orchestrated by TFs. Transcriptional activation begins with (a) binding of transcription factors (TFs) to an enhancer. (b) Next, coregulators and mediator complex are recruited by TFs before general TFs (GTFs) and RNA PolII assemble the transcription preinitiation complex.

However, the presence of a TFBS is not necessarily an indication that the site is really occupied by a TF. First of all, the TF has to be expressed or has to be in an active form to exercise its function. Second, only a small proportion of potential TFBS in a genome are bound, because many sites are not accessible for TFs. Other factors influence TF binding like openness of chromatin and competition with other chromosomal proteins and TFs [18, 19]. All this may result in developmental stage specific or cell-stage specific expression of genes. On the contrary, binding of a TF does not always lead to gene expression [20]. What distinguishes a productive from an apparent non-productive binding event remains largely elusive, although some chromatin features are associated with productive binding events (histone modifications H3K4me1 and H3K27ac) (reviewed in [21]).

2.2.1. Quantitative Understanding of Gene Regulation

Genes are not simply expressed in an on or off state, their expression is fine-tuned by integration of signals coming from the regulatory complexes, chromatin landscape and RNA processing machinery. A single transcription factor, like the glucocorticoid receptor, can control thousands of target genes. Interestingly, although these genes

are controlled seemingly by a similar mechanism, the expression level of these genes may vary. Until now, it is hard to predict if a gene is expressed or not. Even harder is the prediction at which level a gene is expressed [20, 22, 23]. Most of the scientific focus has been on attempts to understand which enhancers are responsible for the regulation of which genes in a qualitative way. Fewer studies have focused on understanding quantitatively how different enhancers might influence the expression of their target genes [12, 24, 25]. Yet, the expression level of a gene is important for organism survival. Higher organisms, for example, compensate for additional gene copies of the X chromosome to reach similar levels of gene product in males and females [26, 27]. In another case, researchers could show that heterozygous mice for the tumor suppressor p53 are highly susceptible to spontaneous tumors compared to wildtype and have a significantly lower survival rate [28]. For future research and medical development, it is important to also understand the details of fine-tuning of expression levels.

The level of gene expression can be influenced at several levels. One important part is played by the promoter structure (TATA-Box, GFT binding sites, proximal regulatory elements) in defining expression level and promoter structures are quite diverse [29, 30]. Another important feature is the structure and location of the regulatory sequence [12, 31]. An enhancer with multiple TF binding sites typically shows stronger gene expression than an enhancer with a single site. Also TFs in close proximity to a cofactor binding site show higher expression than an isolated TF. As mentioned before, distance between enhancer and TSS is very important for gene expression levels, as well. Yet, the effect of variants of a single transcription factor binding sequences on gene expression level is still unclear. The sequence of a functional TFBS can be quite variable within a genome. But why is it so variable? Possibly, because it may be linked to TF function and fine-tuning of gene expression. With my thesis, I wanted to study what impact sequence variants of TFBS may have on fine-tuning of expression levels and what could be the cause of this. The sequence of TFBS could be one feature that may help to improve prediction of gene expression levels from genome sequence.

In the end, gene expression level is only one part among others that defines final protein levels in the cell. Transcription and translation rates play important roles like mRNA and protein degradation rates to define protein levels. All these mechanisms are also adjustable depending on the cellular program and may vary between genes. Here, in my thesis I will focus only on the transcriptional regulation by TFs as a key mechanism for defining gene expression. I try to reduce the influence of the other mechanisms by using similar reporter constructs and the same cell line at nearly identical conditions.

2.2.2. The Glucocorticoid Receptor as a Model Transcription Factor

The glucocorticoid receptor (GR) is a transcription factor that regulates gene expression of a myriad of genes. It is a good model to study gene regulation, because it contains a ”switch” for activating it. This activation is a consequence of the

binding of glucocorticoids, a class of steroid hormones, to GR. Upon binding of the hormone molecule, GR can translocate from the cytoplasm, where it was bound and stabilized by chaperones, into the nucleus and start its genomic action. The genomic action of GR is quite diverse. First, GR can bind directly to DNA, alone or in combination with other TFs, or by binding indirectly to DNA through tethering to other TFs. Second, GR can both activate and repress gene expression depending on DNA sequence and involved cofactors. GR is expressed in almost any cell-type in our body and is involved in many cellular processes, for example in the anti-inflammatory response, metabolism, stress response and fetal lung maturation. The diversity of GR action may also be a consequence of different GR isoforms (GRα, GRβ, GRγ), which arise from alternative splicing [32]. Here in my thesis, I worked with the predominant GRα form only. In this way, GR can regulate a myriad of genes, for example 2450 genes alone in U2OS, an immortalized bone cell line [33]. GR is a widely used therapeutic target, which is targeted by treatment with glucocorticoids, to treat among others allergies, asthma, sepsis, cancer and heart diseases. Since GR is involved in many processes in our body, glucocorticoid treatment may cause many side effects. Research in improving treatment and reducing side-effects is still ongoing.

The classical and well-studied way of GR action is direct binding of GR to the DNA at GR binding sequences (GBS). A GBS is a typically 15 bp long sequence consisting of two inverted hexameric repeats separated by a 3 bp spacer (AGAACAxxxTGTTCT). GR recognizes a GBS as a dimer in a head-to-head arrangement (Figure 2.3a). The two GR dimer partners interact through the dimerisation interface to form a functional dimer. The helix 1 of each dimer partner is inserted into the major groove of the DNA to contact specific DNA bases and facilitate GBS recognition. Here at this point, I would like to introduce the numbering of GBS position that I will refer to throughout the thesis. Figure 2.3a shows the GR motif in combination with position numbering (-9 to 9) and the GR DBD-DNA crystal structure. The middle spacer position was set to 0 and each position up and downstream of the GBS spacer was numbered in this way.

GR belongs to the family of steroid receptors of the nuclear receptor superfamily that evolved from a common ancestor [34]. Each steroid receptor harbors a DNA binding domain, a ligand binding domain and a transactivation domain. Interestingly, the 3-Ketosteroid receptors, a subgroup of steroid receptors, show a strong homology of the DNA binding domain and share very similar DNA binding sequences (Figure 2.3b) [35]. Although each receptor has its own ligand by which it is activated and is involved in different cellular processes, it is paradox that they can potentially bind similar sequences. For example, the androgen receptor (AR) and GR have antagonistic roles in muscle cells. AR promotes muscle growth and GR promotes muscle wasting, but their preferred binding sequence is nearly identical (Figure 2.3b). How AR and GR know where to bind in the genome, when they are expressed and concurrently activated, is still up to debate. The same paradox is known for other TFs like the Hox family, which have similar binding sequence preferences, but nevertheless show unique binding characteristics and regulate Hox-

specific genes. One mechanism explaining this are interactions with specific cofactors that introduce conformational changes of the Hox protein and increase DNA binding specificity [36,37]. In summary, these studies show that a binding sequence alone cannot sufficiently explain if a binding site is bound and by which TF.

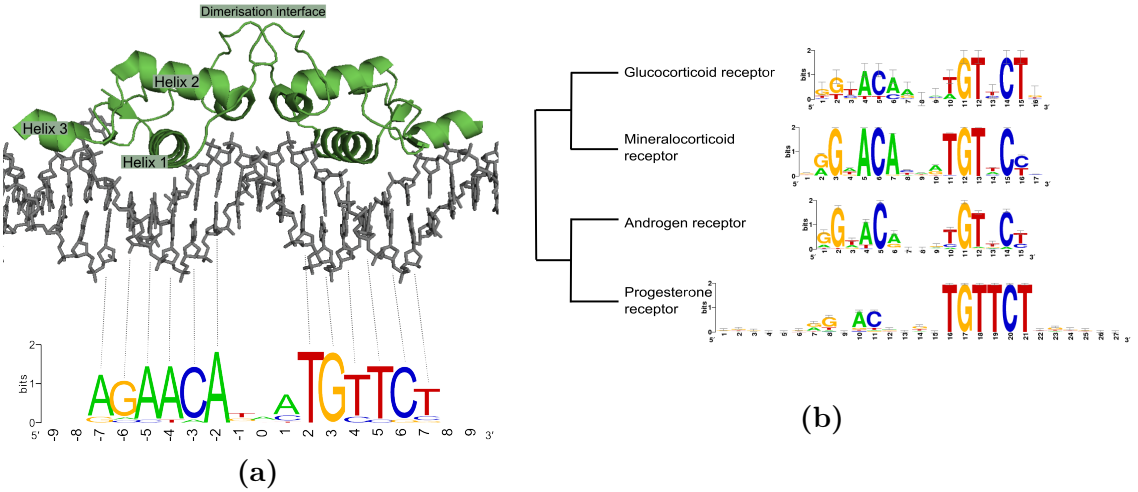


Figure 2.3.: GR and steroid receptor motifs. (a) Crystal structure of GR DBD dimer bound at DNA (PDB-ID:3G9M). (b) Comparison of motifs of the 3-Ketosteroid receptors in mammals (tree adapted from [35]). Logos were generated from Transfac and Jaspar motif matrices (M00205, MA0727.1, M00481, M00957).

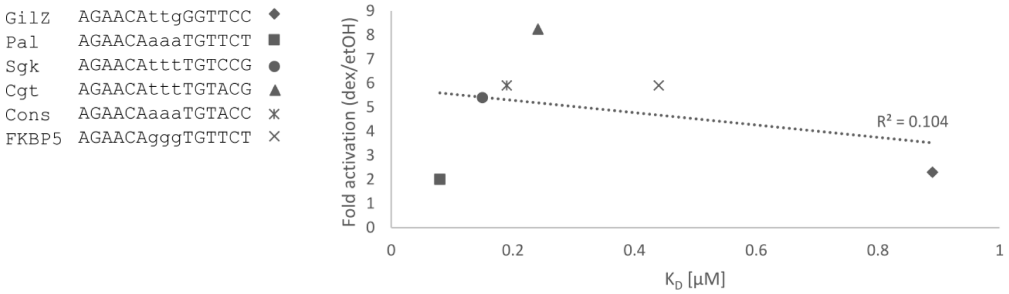


Figure 2.4.: Activity of GBS is uncorrelated to binding affinity K_D. Plot adapted from data from Meijsing et al. [1].

Understanding GR action on the bases of GBS is rather complex, because different GBS variants can activate different levels of gene expression. Yet, there is seemingly no correlation between binding strength of GR to a GBS *in vitro* and its activation rate *in vivo* (Figure 2.4). Though, it is an accepted dogma in the research community that TF activity is linked to binding strength of a TF to a sequence. For GR, this is seemingly only true to a certain extent. Of course, if a sequence is not a bound there cannot be activation of a target gene. Yet, the

strongest bound GBS, named Pal, is also the sequence with the lowest activation rate (Figure 2.4). So for GR, binding is seemingly disconnected from GR activity. To test how GBS sequence influences GR activity in a large scale, I adapted the STARR-seq method, a large-scale quantitative enhancer discovery method, to test thousands of GBS variants in parallel.

2.2.3. Genome-wide Quantitative Enhancer Discovery

Identification and characterization of enhancers has been challenging despite their importance in gene regulation. Classical methods study enhancers one-by-one through cloning a potential enhancer in a reporter construct in front of a minimal promoter and transfecting these constructs in cells to test enhancer function. I used one of these classical assays, the luciferase reporter assay, to study individual regulatory sequences. Over the last years, there have been many advances in technologies like DNA sequencing, which led to the development of high-throughput assays for the identification of enhancers in almost any cell type (reviewed in [16]). One of them is the STARR-seq method [38], which will be explained in more detail in the next paragraph, because I used this method in my thesis. However, other recently developed methods with similar application exist but exhibit differing assets and drawbacks. An alternative method to STARR-seq is MPRA (massively parallel reporter assays) [39], which requires DNA synthesis of test sequences and barcoding before transient transfection of multiple reporters in parallel. It provides quantitative information as a readout. There exist different subtypes of MPRA already, but the basic principle is similar in all (reviewed in [40]). Akhtar et al. developed the TRIP (Thousands of Reporters Integrated in Parallel) assay, which is basically a multiplexed reporter assay by random integration into the genome coupled with high-throughput sequencing [13]. It allows one to study enhancers in chromosomal context and provides quantitative information of enhancer activity. The FIREWATCh (Functional Identification of Regulatory Elements Within Accessible Chromatin) is used to identify enhancers from open chromatin regions by reporter integration and does not provide quantitative information, but is used to identify enhancers on a genome-scale [41].

For this thesis, I adapted the STARR-seq method. STARR-seq stands for self-transcribing active regulatory region sequencing [38]. In the STARR-seq method a regulatory region is integrated in a reporter construct between a reporter gene and a polyA site (Figure 2.5), assuming that enhancer function is not location-dependent [11]. This is different to all the other methods mentioned above, where enhancers are part of the promoter and not part of the transcribed sequence in the 3'UTR. In STARR-seq, the regulatory region drives expression of the reporter gene and of itself. The produced mRNA contains the sequence of the regulatory region and the amount of mRNA provides a quantitative information of the regulatory strength, both sequence identity and activity can be accessed by sequencing. Any arbitrary source of DNA can be used for STARR-seq to test its regulatory functions. In the original publication, the authors used genomic fragments of Drosophila genome and

human chromosomes as input material [38]. In this way, the ability to serve as an enhancer could be tested genome-wide. A modification of the STARR-seq method is the CapStarr-seq method, which preselects regions of interest in mammalian cells by capturing genomic fragments on a custom designed microarray before integration into the STARR-seq screening vector [42]. An advantage of STARR-seq is that no barcoding is needed, since the regulatory region serves as its own unique barcode. Other methods (MPRA, TRIP, ...) depend on barcoding and thorough sequencing of the input library beforehand to associate barcodes with enhancers [13, 43], which could constitute an additional source of errors.

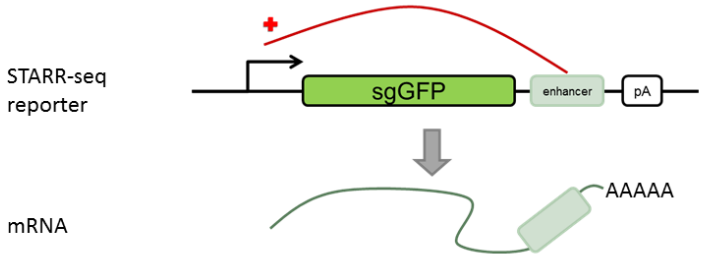


Figure 2.5.: STARR-seq concept and reporter set-up. An enhancer is inserted between reporter gene (GFP) and poly-adenylation site (pA). The enhancer transcribe itself. The + sign indicates transcriptional activation.

For my adaptation of the STARR-seq method, I applied synthesized DNA fragments as input for the STARR-seq library instead of genomic fragments. These fragments were rather short (<200 bp) and contained only a single GBS with partially randomized nucleotides to study the variability of GBS variants and their effect on GR activity.

2.2.4. How DNA is Read by Transcription Factors

When a transcription factor binds, it usually not only reads the bases of the DNA, but also the three-dimensional structure of the DNA and this might have consequences for the DNA and the TF. We have to keep in mind that DNA and TF are capable of a certain structural flexibility depending on the circumstances [44–46]. Base readout originates from the physical contact formed between amino acid side chains and the bases through hydrogen bonds, water-mediated hydrogen bonds and hydrophobic contacts. For the shape readout, a TF reads the global shape (bending, A-DNA, ...) and the local shape (major groove, minor groove, kink, ...) of the DNA. Of course, shape is also encoded in the DNA sequence, but the relationship between nucleotide composition and shape is more complex. Depending on the nucleotide composition, the DNA structure can differ, for example AT-rich tracks show a narrower minor groove or bending of DNA compared to GC-rich tracks [44, 45]. The majority of TFs read out both DNA features, base and shape. Position weight

matrices (PMW) are an easily comprehensible and intuitive way of representing the DNA binding sequences of a TF, but they represent mainly the DNA base readout of a TF [47]. PMWs are often graphically represented as logos (Figure 2.3). Interdependencies between nucleotide composition and shape is so far too complex to be represented as a PMW.

It was shown that Hox proteins, I mentioned earlier, upon interaction with their cofactor read out the structure of the DNA [36]. If the structure is a match for the complex of Hox protein and cofactor, they can interact with then DNA and start recruiting coregulators. These sites can have low binding affinity, but still confer high binding specificity [48]. So DNA structure could explain, next to other features like affinity, why some sequences are occupied by a TF while other are not although they contain a binding sequence.

When a protein binds a certain factor it can lead to conformational changes of the protein affecting not only the local structure of the interaction domain but more importantly it affects remote parts of the protein. This process is called allostery. For example, binding of the hormone molecule leads to conformational changes in GR, that bring GR in its active form. But there are also indications that allostery can occur through DNA on proteins and that this may be an important feature for the assembly of the transcriptional regulatory machinery [1, 49, 50]. In this model, DNA can induce structural changes of the bound protein. Therefore, one working hypothesis of my thesis was that DNA structure can differ between TFBS variants and that this DNA can induce different conformational changes in associated proteins. Following this idea, DNA allostery could explain the difference in gene regulation level induced by GBS variants and its disconnect with binding affinity. In the next paragraph, I will explain various methods that can be used to study DNA:protein interactions structurally.

2.3. Structural Study of DNA-TF Interaction

Research in genomic and structural biology help us to understand the complex relationship between TF DNA binding and function. In this thesis, I used different structural analysis methods to study GR-DNA interactions for GBS variants. These structural studies comprise nuclear magnetic resonance (NMR) spectroscopy, molecular dynamic (MD) simulations and DNA shape prediction. Co-crystal structure of GR DBD (generated through X-ray crystallography) bound to different GBSs have been published previously and showed structural differences induced by GBS variants in the lever arm of GR and differences in contact formation between DNA and GR [1]. Similarly, NMR analysis of GR DBD for wildtype and A477T mutant have been published previously, highlighting the importance of the dimerisation interface in recognition of spacer variations of the GBS [50]. In this way, the available structural data provided a good foundation for understanding the GR-DBD DNA interaction, and were used for preparation of new structural experiments by my colleagues Dr. Marcel Jurk, Mahdi Bagherpoor Helabad and myself.

NMR spectroscopy is based on the principle that many nuclei (atomic core) have a spin and all are electrically charged. If an external magnetic field is applied to a nucleus, a specific resonant frequency can be measured. The chemical shift is the resonant frequency of a nucleus relative to a standard in a magnetic field and used to generate a NMR spectrum. Commonly measured nuclei are ^{1}H, ^{13}C, ^{15}N and ^{31}P. As a result, information about the nucleus' chemical environment can be derived from its resonant frequency. This can provide information about the structure, dynamics and chemical environment of the molecule. In NMR, the molecule of interest is usually studied in solution. Different to this, for X-ray crystallography the molecule of interest is studied in a crystalline form. Most molecules, including biological molecules, are in principle able to form crystals. When a beam of X-ray is applied to a crystal, the atoms of the crystal diffract the beams in specific directions. From the angle and the intensities of this diffraction pattern a three-dimensional picture of the electron densities in the crystal can be made and used to determine atom positions and chemical bonding.

Both, NMR spectroscopy and X-ray crystallography need a highly concentrated and pure probe. However, for GR we are only able to purify the DNA binding domain (DBD) and not the whole protein, because it readily forms aggregates. That is why for both NMR and X-ray crystallography studies, we are limited in our findings to the GR DBD. Another drawback of both methods so far is that we cannot easily distinguish between the two dimer partners of a GR dimer. In the previous X-ray crystallography, GR-DNA formed a pseudo-continuous helix, but in an unorientated fashion, producing an average signal for both dimer partners [1]. Also in NMR, an average signal is produced for the amino acid residues of both dimer partners. One might be able to observe two or more signals for one residue in NMR, but the origin of the signals remains unknown without further experiments. In my thesis, this will play an important role, since I observed different behavior of the dimer partners induced by DNA sequence variants.

Most structural studies, like NMR and X-ray crystallography, are an approximation of the real life situation in a cell, since the data is acquired in *in vitro* experiments. Until now we are not able to track a single molecule and study its structure at atomic resolution in a living cell. But our understanding of chemical and physical interaction between molecules and the advancement in computer power also allows us now to simulate molecule behaviour *in silico* at various conditions [51]. That is why I also turned to MD simulations (CHARMM27 force field) [52]; they allow us to observe a molecule and its movement for milliseconds (10^{-3} s) at atomic resolutions. Together with my colleagues Dr. Marcel Jurk and Mahdi Bagherpoor Helabad, we analyzed MD simulations of GR bound to GBS variants, which were in parallel studied by NMR analysis. With MD simulations of GR and DNA, we were able to observe different behavior of the two dimer partners.

DNA structure can be studied by the previously mentioned methods as well, but only on a low-throughput scale. Remo Rohs and colleagues developed tools to predict DNA shape from a given sequence [44, 53]. The tool is based on DNA crystal structure, derived from hundreds of published protein bound DNA co-crystal

structures. Thanks to this tool and published R package [54,55], I was able to study and compare the different DNA shape features (minor groove width, propeller twist, helix twist and roll) of hundreds of GBS variants in parallel, showing the structural flexibility of GR binding sequences.

2.4. Noise in Gene Expression

GBS variants may influence gene expression levels. Next to gene expression level, expression noise is an important feature of expression and GBS variants could as well influence expression noise. That is why I also studied the influence of GBS variants on expression noise. Expression noise describes the cell-to-cell variability of expression levels within a homogeneous cell population. The noise can vary greatly between individual genes within a cell population. In a cell, the residence time of the transcriptional machinery at a promoter is rather short lived. A promoter cycles therefore between inactive and active promoter, producing a burst of transcription [56]. Expression noise originates from these bursts in transcription and bursts can vary in length and frequency. Noise should not be considered as a burden of gene expression. It is a way of introducing variability in a cell population, without changing the DNA code. Especially during development, it is a regulatory "cheap" and simple way to produce randomness for setting cell fate or making the organism fitter in reaction to sudden changes of the environment (so called bet-hedging). But not all genes show a high degree in expression variability. Some genes, for example important master regulators, are expressed at well-defined levels, whereas other genes are allowed to have a greater variation in a cell population [57]. The recent development of single-cell-analysis allows to study expression noise.

There exists a strong connection between noise and regulatory networks. Many genes exhibit an autoregulatory feedback loop, which may increase or decrease expression noise depending if it is an activating or inhibiting feedback loop (reviewed in [58]). Noisy expression of transcriptional regulators may also lead to noisy expression of target genes in a cell population and in this way the noise is propagated. However, through connected gene circuits, this noise may be buffered and reduced for the individual target genes.

Expression noise has been mostly studied for eukaryotes in yeast, with dual fluorescence reporter [12]. The influence of TFs and other promoter features on noise has revealed insights into the process of transcriptional regulation and why some promoters are noisier than others [56]. Due to advances in single-cell-technologies, single endogenous genes can now be studied, also in mammalian cells [59], which gives us insight into endogenous gene expression noise. However, here in my thesis, I used dual fluorescence reporters to study expression noise of GR response elements. The GR response element was not part of a promoter but part of an enhancer, which has not been done so far for GR and other transcription factor for studying expression noise.

2.5. Aim of this Thesis

This thesis was aimed at giving new insights into the understanding of fine-tuning of gene expression by transcription factor binding sequence variants. I used GR as a model transcription factor for this thesis. First of all, I wanted to understand how GBS sequence influences GR-dependent transcriptional output levels. Towards this goal, I established a high throughput method to study multiple variants in parallel. In the second part, I wanted to understand how GBS variants, and with a focus on the flanking sites, modulate GR action, when binding affinity is seemingly disconnected from this process. Therefore, I used a computational approach to find endogenous GBS variants that are associated with strong and weak GR responsive genes. Further, I applied different structural methods and used GR mutants to study DNA structure and GR conformation to find an explanation for the differences in gene expression levels. Additionally, my aim was to establish a method to study coregulator occupancy at GBS variants. In the last part of my thesis, my objective was to test if the composition of cis-regulatory elements also modulates cell-to-cell variability of gene expression.

3. Materials and Methods

3.1. Materials and Organisms

3.1.1. General Materials

If not stated differently, general chemicals and lab-ware were purchased from the following companies: Sigma-Aldrich, Merck, Roth, Fluka, Eppendorf, Greiner Bio one, BD Biosciences, TPP, and Sarstedt.

3.1.2. DNA Plasmids

pGL3 promoter vector	SV40 minimal promoter::Firefly luciferase	Promega
pcDNA3.1-rGR	Expression vector for rat GR	[60]
p6R	empty mammalian expression plasmid	[61]
pRL-CMV vector	Renilla Luciferase Control Reporter Vector	Promega
SAA-GFP	AAVS1 targeting vector	[62]
ZFN-plasmid	ZFN expression construct	[62]
human STARR-seq sequencing vector	a kind gift of the Alexander Stark lab	[38]

3.1.3. Antibodies

N499 - rabbit anti GR IgG	polyclonal, raised against the N-terminal amino acid sequence of the human GR (residues 1-499)
Sc-333 - anti Sam68	Santa Cruz - rabbit polyclonal IgG, epitope at C-terminus of protein
Sc-584 - anti p300	Santa Cruz
Sc-6098 - anti SRC1	Santa Cruz
rabbit anti actin IgG	I-19-R, Santa Cruz
HRP goat anti rabbit IgG	Invitrogen

3.1.4. Organisms

Human Cell Culture

Two human cell lines, U2OS and A549, have been used for this thesis. The U2OS cell line is an epithelial bone cell line derived from an osteosarcoma of a 15-years old female Caucasian. GR-18 was derived from U2OS and stably expresses rat

GRα [63]. A549, an epithelial cell line, was derived from a lung carcinoma of a 58-years old male Caucasian.

Bacterial Strains

For cloning and plasmid amplification I used the following Escherichia coli strains: DH5α, dam-/dcm- (NEB), MegaX DH10B T1R (ThermoFisher) and DB3.1 (Invitrogen). The strain DB3.1 is not sensitive for ccdB, a protein which interacts with the DNA gyrase and eventually kills sensitive bacteria. Hence, this strain was used to amplify plasmids carrying a ccdB gene like the human STARR-seq screening vector.

3.2. Cloning and PCR

3.2.1. Cloning of GR Binding Sites into pGL3 Promoter Vector

GR binding sequences were ordered as DNA oligomers (forward and reverse) with 5' KpnI overhangs and 3' XhoI overhangs. The annealed oligos were ligated into KpnI-HF (Fermentas) and XhoI (New England Biolabs) digested pGL3 promoter vector (Promega). Table B.1 on page 101 contains a list of oligomers used for cloning.

3.2.2. Cloning of GBS Reporter Plasmid for Stable Integration

Cloning of reporter plasmids for stable integration was done as previously described [62]. pGL3 promoter vector containing the GBS (section 3.2.1) were amplified via Phusion-HF (NEB) PCR using primer SS063 and SS064 (Table B.3). PCR product and SAA-GFP plasmid was digested with SalI (NEB) and KpnI (NEB) and purified with PCR clean-up extractII (Macherey-Nagel) before ligation.

3.2.3. Site-Directed Mutagenesis

The Pfu Ultra Polymerase (Agilent Technologies) was used for site-directed mutagenesis PCR and to amplify specific plasmids from the STARR-seq plasmid library. For a reaction mixture of a total volume of 25 μl, I applied 20 ng template plasmid, 2 μl dNTPs (2 mM each), 2.5 μl 10x Pfu Ultra Buffer, 1 μl of each primer (10 μM). For amplification of specific plasmids from a STARR-seq plasmid library, 50 ng plasmid library was used and 26 PCR cycles were run. Pfu Ultra polymerase (0.5 μl) was added after the initial denaturing phase.
PFU-PCR-Program:

Initial Denaturation:	95°C	60 sec	
add Pfu Ultra			
Denaturation:	95°C	30 sec	
Annealing:	55°C	60 sec	16 Cycles
Synthesis:	68°C	1 kb/min	
Final Synthesis:	68°C	10 min	
Cooling:	4°C	infinite	

After the PCR, 1 μl of DpnI was directly added to the reaction and incubated for 2 h at 37°C to remove template plasmid. 2.5 μl of the reaction were transformed into *E.coli* . Table B.2 on page 101 contains a list of oligomers used for site-directed mutagenesis.

3.2.4. Colony PCR

Colony PCR was used to analyze single *E.coli* colonies for containing the correct plasmid after the cloning procedure. Primers were designed in a way that the first primer targets the plasmid backbone and the second primer targets the insert. For a reaction mixture of a total volume of 20 μl, I applied 0.25 μl dNTPs (25 mM each), 0.25 μl Taq DNA Polymerase (EURx), 2 μl 10x Buffer C (EURx), 1 μl of each primer (10 μM) and a small amount of the *E.coli* colony.
Colony-PCR-Program:

Initial Denaturation:	95°C	3 min	
Denaturation:	94°C	30 sec	
Annealing:	52°C	30 sec	35 Cycles
Synthesis:	72°C	1 kb/min	
Final Synthesis:	72°C	3 min	
Cooling:	4°C	infinite	

Presence of correct PCR product was analyzed on a DNA agarose gel.

3.2.5. Sanger Sequencing of DNA

For Sanger sequencing of specific Plasmid-DNA and other DNA samples, DNA was send with appropriate primer to Eurofins Genomics (Ebersberg, Germany).

3.3. Bacterial Transformation, Bacterial Cell Culture and DNA Plasmid Preparation

"Mix & Go" competent cells (made competent using the Zymo research "Mix & Go" kit) were thawed on ice, mixed with up to 5 μl of plasmid DNA and directly plated out onto a pre-warmed LB-Agar-plate with antibiotics. Electro-competent cells were thawed on ice, mixed with up to 5 μl plasmid DNA and transferred into a pre-chilled 0.1 cm cuvette (Bio-rad). Gene Pulser (Bio-rad) was set to following conditions: 1.5 kV, 25 μF and 200 Ω. Freshly electro-transformed bacteria were

grown for 1 h in SOC medium (5 g/l yeast extract, 20 g/l bacto-tryptone, 10 mM sodium chloride, 10 mM magnesium chloride, 0.5 mM potassium chloride, 10 mM magnesium sulfate) before plating onto LB-Agar plates. *E. coli* liquid cultures were grown in the desired volume of LB medium (10 g/l sodium chloride, 10 g/l bacto-tryptone, 5 g/l yeast extract) overnight at 37°C at 190 rpm. For DNA plasmid Mini and Maxi preparation (Qiagen), I used 4 ml and 100 ml of the culture, respectively.

3.4. GR Protein Expression and Purification

For the EMSA and NMR experiments, rat GRα DNA binding domain (DBD, residue 440-525 for NMR, residue 380-540 for EMSA, ϵ=4470 $L \cdot mol^{-1} \cdot cm^{-1}$) was used, as well as the DBD of the GR dimerisation mutant A477T [60]. Protein expression and purification was done with supervision of Dr. Marcel Jurk. The codon-optimized protein expression plasmids pET28Fusion-rGRcoAlpha and rGRcoAlpha-A477T were made by Dr. Marcel Jurk.

3.4.1. GR DBD Protein Expression

For NMR

E. coli 'T7 express cells' transformed with the expression plasmid pET28Fusion-rGRcoAlpha or rGRcoAlpha-A477T were inoculated each in 50 ml 2xYT medium (16 g/l tryptone, 5 g/l NaCl, 10 g/l yeast extract) plus 50 μg/ml kanamycin and grown over night in a shaker at 37°C and 190 rpm. M9 minimal medium was freshly prepared (0.3 mM $CaCl_2$, 1 mM $MgCl_2$, 1x M9-salt, 1x TS2 solution, 20 μM Fe-(III)-citrate, 20 μM $ZnSO_4$, 6.7 μM EDTA, 3 mg/l Biotin, 3 mg/l Thiamin, 50 μg/ml Kanamycin, 80 g/l Glucose) and 1.5 g of $(^{15}N\text{-}NH_4)_2SO_4$ dissolved in 10 ml H_2O was added to 1 l of M9 medium. Two times 500 ml of M9 medium was inoculated with appropriate volume of pre-culture, so starting OD_{600} was at 0.2. Culture flask were incubated at 37°C and 210 rpm. At an OD_{600} between 0.8 to 1.0, temperature was set to 28°C and 30 minutes later protein expression was induced by addition of 0.5 mM IPTG (Amresco). After 4 to 5 hours, cells were harvested.

For EMSA

Protein expression of GR for EMSA was done in a similar way as described for NMR, but DH5alpha cells were used for protein expression in standard LB medium instead of M9 minimal medium.

3.4.2. GR DBD Protein Purification

The *E. coli* cell pellet was dissolved in buffer A (50 mM Tris pH 8.0, 500 mM NaCl, 15 mM Imidazol, 100 μM $ZNSO_4$, 5 mM β-ME) and cells were lysed with a french press. Lysed cells were centrifuged at 20000 rpm and 4°C for 30 min and crude

extract was cleared by passing it through a 0.45 μm filter. The cleared extract was run over a HisTrap FF column (GE Healthcare) to enrich His-tag containing GR DBD. In the next step, the His-tag was cleaved off in pooled fractions by incubation with TEV protease for 1 to 2 days. Successful cleavage of the His-tag was controlled on a protein gel. Protein extract was concentrated to 5 ml in Amicon Ultracel 3K (Millipore) and rebuffered in IEX buffer (20 mM Tris ph 7.6, 40 mM NaCl, 5 mM β-ME) with Zeba spin desalting columns (Thermofisher). Next, protein extract was run over a resource S column (GE Healthcare) to further purify GR DBD. Afterwards, pooled fractions were concentrated again with Amicon Ultracel 3K and with Zeba spin desalting columns rebuffered into NMR buffer (20 mM NaH_2PO_4 pH 6.7, 100 mM NaCl, 1 μM DTT) or storage buffer (25% glycerol, 20 mM Hepes pH 7.7, 100 mM NaCl, 1 mM DTT).

3.5. Human Cell Culture

3.5.1. Growth, Freezing and Thawing of Human Cell Lines

Both U2OS and A549 cells were grown in DMEM medium (Gibco) supplemented with 5 vol% FBS (Gibco) at 37°C and 5% CO_2 fumigation. For storage, cells were washed with PBS, trypsinized (Trypsin, Sigma-Aldrich), mixed again with DMEM-5 vol% FBS and centrifuged for 5 minutes at 1000 rpm. Supernatant was removed and the cell pellet was mixed with cryo-medium (FBS with 10-15 vol% DMSO). Aliquots were transferred into cryo-tubes and placed into a "Mr. Frosty" freezing container (ThermoFisher Scientific) and placed at -80°C overnight. For short-term storage, cells were kept at -80°C, or moved to liquid nitrogen tanks for long-term storage. To thaw cells, the tube was placed into a 37°C water bath, mixed with 10 ml DMEM-5 vol% FBS and immediately centrifuged at 100 g for 5 min. Supernatant was removed, the cell pellet was mixed with fresh DMEM-5 vol% FBS and transferred into a new culture bottle.

3.5.2. Transfection with Amaxa Nucleofector kit

Transfection with Amaxa Nucleofector kit resulted in high transfection rates of up to 90% of the living cell population. U2OS/GR18 cells were transfected using the Amaxa Nucleofector kit V (Lonza), using in total 2 μg plasmid DNA and 1 million cells. For the A549 cell line, I used the Amaxa Nucleofector kit T (Lonza), using in total 2 μg plasmid DNA and 1 million cells. Transfection was conducted according to the manufacturer's guideline with a Nucleofector device 2b (Lonza). Transfected cells were split equally into two wells of a 6-well microtiter plate. After four to six hours, cells were treated with 1 μM dexamethasone or as a control with 0.1% ethanol overnight.

3.5.3. Transfection with Lipofectamine for Luciferase Assay

U2OS cell line

Transfection with lipofectamine resulted in a lower transfection rate (20-40% of the living cell population). For each tested construct, four times 25000 cells were seeded into a well of a 48-well microtiter plate and incubated overnight. Cells were once washed with 500 μl PBS (Gibco) and next 100 μl serum-free DMEM was added. For 4 wells, I mixed in a tube 40 ng pGL3-promoter construct, 40 ng pcDNA3.1-rGR, 250 ng p6R and 0.4 ng pRL (in total 7 μl) with 3.2 μl PLUS reagent (Invitrogen). To each tube, 50 μl of serum-free DMEM (Gibco) was added. In parallel in a second tube, 1.6 μl lipofectamine (Invitrogen) were mixed with 50 μl serum-free DMEM (Gibco). After 15 minutes incubation at room temperature, 50 μl of "liofectamine-mix" were mixed with the "DNA-PLUS-mix". After another 15 minutes of incubation, 25 μl of this transfection mix was added to each of the 4 wells. After three hours, the medium was completely removed and 200 μl fresh DMEM-5 vol% FBS was added to each well. After another three hours, two of the four wells were treated with 1 μM dexamethasone and as a control two wells with 0.1% ethanol overnight.

A549 cell line

I seeded 60000 cells per well into a 48-well microtiter plate and incubated overnight. For 4 wells, I mixed in a tube 160 ng pGL3-promoter construct, 160 ng pcDNA3.1-rGR, 800 ng p6R and 16 ng pRL (in total 32 μl). To each tube, 100 μl of serum-free DMEM (Gibco) was added. In parallel in a second tube, 6 μl lipofectamine 2000 (Invitrogen) were mixed with 100 μl serum-free DMEM (Gibco). After 5 minutes incubation at room temperature, 100 μl of "liofectamine-mix" were mixed with the "DNA-mix". After another 20 minutes incubation, 50 μl of this transfection mix was added to each of the 4 wells. After six hours two of the four wells were treated with 1 μM dexamethasone and as a control two wells with 0.1% ethanol overnight.

3.5.4. Stable Integration at the *AAVS1* Locus

Stable integration of reporter was done as previously described [62]. The zinc-finger nuclease drives the integration of the GBS-reporter construct and a promoter-less GFP gene from the SAA-GFP vector into the "safe harbor" locus (AAVS1) [64], fusing the GFP gene to the *PPP1R12C* promoter.

10 μg of SAA-GFP plasmid (section 3.2.2) were co-transfected with 0.5 μg of zinc-finger-nuclease coding plasmid. Transfection was done in GR18 cell line as described in section 3.5.2. Seven days post transfection, the correct integration of the reporter construct was tested by PCR using primer R5 and LucNested (Table B.3). Around 3 weeks post transfection, cells were FACS sorted for GFP positive cells and from these cells clonal lines were derived. Each clonal line was again tested for correct integration of the reporter construct.

3.6. Luciferase reporter assay

Post transfection with lipofectamine, each cell line was treated for the luciferase reporter assay in the same way. I used the dual-luciferase reporter assay kit from Promega.

First, the media was completely aspirated and 65 μl of 1x passive lysis buffer (Promega) was added per well of a 48-well microtiter plate. During the lysis, the plate was incubated on a rocking shaker at room temperature for 15 minutes. Next, 2.5 μl of the lysed cell suspension was transferred into a white 384-well microtiter plate (Greiner). For the assay, I used a luminometer (LUMIstar Omega by BMG Labtech), gain was set to 3600. To measure the activity of firefly and renilla luciferase, I added 12 μl luciferase assay reagent II (LAR II), recorded the firefly luminescence signal, followed by addition of 12 μl Stop & Glo to record renilla luminescence signal. Reporter activity was specified as relative light units (RLU) = mean(firefly signal)/mean(renilla signal).

3.7. Electrophoretic Mobility Shift Assay (EMSA)

3.7.1. EMSA Assay

Electrophoretic mobility shift assay (EMSA) was used to test binding of GR to different DNA sequences and to measure binding affinity, K_D, of GR to these sequences. The forward DNA oligomer was labeled 5' with Cy5 fluorescent dye. Table B.3 on page 102 contains a list of DNA oligomers used for EMSAs. Forward and reverse oligomer were mixed in a ratio 10 to 11, respectively, before annealing in boiling water and diluted to 10 μM.

An EMSA-gel was casted containing 0.5x TBE-buffer (10x TBE: 1 M Tris-Borat, 20 mM EDTA) and 5% polyacrylamide. From the annealed DNA oligomer, 0.5 μl were mixed with 1 ml of 2x binding buffer (40 mM Tris pH 7.5, 4mM $MgCl_2$, 0.2 mM EDTA, 20% glycerol, 0.6 mg/ml BSA, 8 mM DTT, 200 mM NaCl). rGR*alpha* DBD (residue 380-540) was diluted with 1x binding buffer into 8 concentrations: 10 μM, 5 μM, 2 μM, 0.8 μM, 0.32 μM, 0.128 μM, 0.05 μM and 0 μM. For each concentration, 3 μl DNA, 3 μl poly(dI-dC) (0.2 μg/μl) and 6 μl GR were mixed and incubated at room temperature in the dark for 30 minutes. The EMSA gel was pre-run with 0.5x TBE-buffer at 250 V at 4°C for 30 minutes. After incubation, 10 μl of each sample was loaded onto the gel and the gel run for another 15 minutes.

Fluorescence bands in the gel were detected using a FLA 5100 scanner (Fujifilm) at an excitation wavelength of 640 nm, selecting the Cy5-filter and the photomultiplier tube was set to 800 V.

3.7.2. Determination of Binding Affinity K_D

Intensity of the band marking unbound DNA was measured for each protein concentration using the GelAnalyzer software. The data (F=unbound DNA band in-

tensity, P=protein concentration) was fitted to a model using non-linear regression with the formula F = 1 / (1 + (K_D /P)) using the software R. In this way, K_D could be extrapolated from the fitted model.

3.8. ChIP: Chromatin Immuno Precipitation

For each condition, approximately 5 million GR18 cells with stably integrated GBS-reporter (section 3.5.4) were grown in a 10 cm dish and treated for 1.5 hours with 1 μM dexamethasone or 0.1% ethanol as vehicle control. Cells were directly cross-linked with 1% formaldehyde at RT for 3 minutes. To quench cross-linking, glycine was added at a final concentration of 125 mM and cells incubated at 4°C for 10 minutes. Next, the cells were washed with PBS for 5 minutes and scraped into a 14-ml falcon tube. After centrifugation, the cell pellet was shock-frozen in liquid nitrogen and stored at -80°C.

To each cell pellet, 2 ml of ice-cold IP-lysis buffer (50 mM HEPES-KOH pH 7.4, 1 mM EDTA, 150 mM NaCl, 10% glycerol, 0.5% Triton X-100, 1:200 protease inhibitor cocktail) was added and nutated at 4°C for 30 minutes. Crude nuclei were pelleted by centrifugation at 4°C for 5 min at 1900 rpm. Nuclei were resuspended in 1000 μl ice-cold RIPA buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 150 mM NaCl, 5% glycerol, 0.1% sodium deoxycholate, 0.1% SDS, 1% Triton X-100, 1:200 protease inhibitor cocktail) and 300 μl transferred to each of three 1.5 ml tubes. Tubes were sonicated for 24 cycles, 30 s on and 30 s off, in ice water, at high intensity (Diagenode Bioruptor). The suspension was cleared by centrifugation at 4°C, max speed, for 15 minutes and the supernatants were pooled in a new 1.5 ml tube. For the immunoprecipitation, 2 μl N499 antibody was added to each sample and nutated at 4°C overnight. 33 μl from a 50% Protein A/G agarose beads slurry (Santa Cruz) were washed twice with 1 ml ice-cold RIPA buffer and nutated at 4°C overnight. RIPA-buffer was removed to reach again 50% beads slurry. Bead slurry was added to IP and nutated at 4°C for 4 h. Beads were washed 5 times with 1 ml RIPA-buffer containing 500 mM NaCl. Each IP was incubated with 75 μl proteinase K solution (TE pH 8.0, 0.7% SDS, 200 μg/ml proteinase K) for 3 hrs at 55°, followed by 65°C overnight. Each ChIP sample was purified using the "PCR clean-up" kit (Promega) and eluted in 100 μl elution buffer. Before qPCR, 5 μl ChIP sample was diluted with 15 μl water.

3.9. Determination of RNA, DNA and Fluorescence Protein Content in Human Cell Culture

3.9.1. mRNA Isolation and cDNA Preparation

From cells transfected with STARR-seq plasmid ($\sim$ 0.5 to 1 Mio cells, section 3.1.4), total RNA was purified using a "RNeasy Mini" kit with on-column DNase digest

(Qiagen) and eluted in 30 μl RNase-free water. For cDNA synthesis, 500 ng total RNA was mixed with 1 μl SS186 (2 μM), 1 μl SS189 (2 μM), 2 μl dNTP mix (2 mM each) and filled up to 14 μl with RNase free water. The reaction was incubated first at 65° for 5 minutes and immediately placed on ice for 1 minute. To each reaction, I added 4 μl 5x first-strand buffer, 1 μl 100 mM DTT and 1 μl SuperScript III (ThermoFisher Scientific). For the cDNA synthesis, the mixture was incubated at 50°C for 60 minutes followed by an incubation step at 70°C for 15 minutes. After cDNA synthesis, the 20 μl cDNA reaction was diluted with 230 μl water and used as template for qPCR.

3.9.2. Quantitative Real-Time PCR

Subsequent to ChIP experiments and cDNA synthesis, quantitative real-time PCR (qPCR) was used to measure relative DNA content of a specific DNA. Table B.3 on page 102 contains primers used for qPCR. qPCR was set up with 2 μl DNA, 5 μl SYBR-mix (100 mM Tris pH 8.3, 6 mM $MgCl_2$, 1 mg/ml BSA, 4 mM dNTPs, 0.66x SYBR-Green (Invitrogen), 1x ROX reference dye (Invitrogen), 0.2 U/μl perpetual Taq (EURx)) and 3 μl 0.66 μM primer mix (forward and reverse). Each sample was measured in duplicates. The delta cycle threshold (ΔCt=Ct(target)-Ct(*RPL19*)) was determined for each tested target DNA and compared to vehicle control to calculate fold activation with $2^{-\Delta\Delta Ct}$ and $\Delta\Delta Ct = \Delta Ct_{dex} - \Delta Ct_{etOH}$.

qPCR-Program on ABI 7900 HT (Applied Biosystems):

Initial Denaturation:	95°C	10 min		100% ramp rate
Denaturation:	95°C	15 sec		100% ramp rate
Annealing/elongation:	60°C	60 sec	40 Cycles	100% ramp rate
Dissociation stage				
Denaturation:	95°C	15 sec		100% ramp rate
Final annealing/elongation	60°C	15 sec		100% ramp rate
Denaturation:	95°C	15 sec		2% ramp rate

3.9.3. Fluorescence Detection in Human Cells

Fluorescence was detected in GR18 and A549 cells transfected with GBS-STARR-seq plasmid (section 3.14.1) together with pSV40::mCherry vector for transfection normalization. For transfection (described in section 3.5.3), the amount of material used per well was doubled and a 24-well microtiter plate seeded with 40000 cells was used.

Before loading cells onto the flow cytometer, cells were trypsinized and subjected to a cell strainer (BD). Fluorescence detection in human cells was conducted on a flow cytometer Accuri C6 (BD Biosciences). To measure mCherry (red fluorescence), the yellow laser at 552 nm and filter 610/20 were used and, to measure GFP, the deepblue laser at 473 nm and filter 510/20 were used. I analyzed only living cells, therefore I set a first gate to separate cells from cell debris and dirt particles.

A second gate was set to define mCherry positive transfected cells in comparison to cells transfected with empty p6R and showing only background fluorescence.

3.10. Measurement of Expression Noise

Transfection and fluorescence detection was done as previously described in section 3.9.3. To measure the expression noise of the activated STARR-seq construct (list of gblocks in table B.5), I set a third gate for GFP positive cells in comparison to cells transfected with empty p6R and showing only background fluorescence. I analyzed only cells showing both mCherry and GFP fluorescence. The mCherry signal was used to correct for extrinsic noise and therefore I divided the GFP signal by the mCherry signal. Relative expression of GFP ($\text{fluorescence}_{\text{GFP}}/\text{fluorescence}_{\text{mCherry}}$), from individual cells, was used to calculate mean expression μ and standard deviation σ of cell populations. After manual inspection of fluorescence signal distribution, the relative GFP fluorescence signal of the tested cell populations appeared to be log-normally distributed. In consequence, I chose the geometric coefficient of variation (CV) to measure expression noise. Expression noise is expressed as percentage.

$$CV = \sqrt{e^{s_{ln}^2} - 1} \cdot 100$$

s_{ln}: sample standard deviation of the data after a natural log transformation

It is important to note that the expression noise was recorded for hormone treated cells. Vehicle treated cells were always recorded as control, but exhibited no real GFP positive signal and therefore measurement of transcriptional noise was not possible. I repeated noise measurement for at least two biological replicates for each experiment to test if noise patterns were reproducible.

3.11. Semi-Quantitative Pull-down of DNA Binding Proteins

3.11.1. Nuclear Extract Preparation

Nuclear extract preparation was described previously [65]. GR18 cells were grown until 80% confluency and treated with 1 μM dexamethasone for 1.5 hours. Medium was removed from the cells which were washed once with ice-cold PBS. All following steps were conducted on ice. Next, 2 ml PBSI (PBS with 0.5 mM PMSF, 25 mM β-glycerophosphate, 10 mM NaF) were added to each 15 cm dish. Cells were harvested with a scraper, transferred into a falcon tube and pelleted at 550 g for 5 minutes. During that time, protease inhibitors (0.5 mM PMSF, 1 mM Na_3VO_4, 0.5 mM DTT, 1x protease inhibitor cocktail, 25 mM β-glycerophosphate, 10 mM NaF) were added to buffer A (10 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 300 mM Sucrose, 0.5 % Igepal), B (20 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 420 mM NaCl, 0.2 mM

EDTA, 2.5% glycerol) and D (20 mM HEPES pH 7.9, 100 mM KCl, 0.2 mM EDTA, 8% glycerol). Next, supernatant was removed, the cell pellet was transferred into a 1.5 ml tube and centrifuged again at 1500 g for 30 seconds. Supernatant was removed and the cell pellet was resuspended in 2x package cell volume of buffer A. After 10 minutes incubation on ice, the suspension was vortexed briefly and centrifuged at 2600 g for 30 seconds. Supernatant was removed again and the pellet, containing cell nuclei, was resuspended in 2/3 package cell volume of buffer B. The nuclei were sonicated for 5 s and centrifuged at 10,400 g for 5 min. Finally, the supernatant was diluted isovolumetrically with buffer D. The nuclear extract was split into aliquots and protein concentration was measured by Bradford assay "Roti-Quant" (Sigma-Aldrich). The aliquots were shock-frozen and stored at -80 °C.

3.11.2. Loading of Magnetic Pull-down Beads

Preparation of magnetic beads for DNA pull-down was done as described previously [66].

Preparation of biotinylated dsDNA

DNA oligomers were ordered as HPLC-purified and contained a PstI restriction site and a GR-binding sequence (see table B.4 on page 102). Oligomers were reconstituted at a concentration of 50 μM in annealing buffer (20 mM Tris-HCl pH 8.0, 10 mM $MgCl_2$, 100 mM KCl). 0.03 ml sense strand containing the biotin-TEG was mixed with 0.04 ml antisense strand. Oligomers were annealed using a PCR cycler (5 min 90°C, gradually to 65°C in 10 min, 5 min 65°C, switched off and allowed to cool to RT).

Loading of magnetic beads with dsDNA

1 mg Dynal MyOne C1 strepdavidin magnetic beads (Invitrogen) were washed twice with 0.4 ml TE buffer containing 0.01% Igepal (Sigma-Aldrich). Next, beads were washed twice with 0.75 ml DW buffer (20 mM Tris-HCl pH 8.0, 2 M NaCl, 0.5 mM EDTA, 0.03% Igepal). From the biotinylated dsDNA, 250 pmol (11 μl) was diluted in 0.4 ml DW buffer and mixed with washed beads. After 3 hours incubation at RT on a rotating mixer, dsDNA loaded beads were washed once with 0.4 ml TE buffer containing 0.02% Igepal and three times washed with 0.4 ml DW buffer. Beads loaded with dsDNA (200 pmol/mg) were stored in 0.1 ml DW buffer up to several weeks in a refrigerator.

3.11.3. DNA Pull-down of Proteins

DNA pull-downs were conducted as described previously [66]. For one pull-down experiment, 0.5 to 1 mg nuclear protein extract was used for 1 mg beads as starting

material. If no mass-spectrometry was conducted, one tenth of initial beads was sufficient for a western blot analysis.

1 mg DNA loaded beads were incubated with 1.3 ml blocking buffer (20 mM HEPES pH 7.9, 0.05 mg/ml BSA, 0.05 mg/ml glycogen, 0.3 M KCl, 0,02% igepal, 5 mg /ml polyvinylpyrrolidone) at RT on a rotating mixer for 1 h. At last, beads were washed once with 1.3 ml RE buffer (1x NEB buffer 3, 0.02% igepal, 2.5 % glycerol, 1 mM DTT, 0.2 mM PMSF, 1x protease inhibitors) and washed twice with 2.67 ml buffer G (20 mM Tris-HCl pH 7.3, 10% glycerol, 0.1 M KCl, 0.2 mM EDTA, 10 mM potassium glutamate, 0.04% NP40, supplemented freshly with 2 mM DTT, 0.4 mM PMSF, 1x protease inhibitor). Meanwhile, nuclear extract was cleared at 15000 g at 4° for 20 min. Next, supernatant was adjusted to 10 mM potassium glutamate and quickly diluted with one volume of poly dIdC (0.2 mg/ml, in buffer G). Again, nuclear extract was cleared at 15000 g for 10 min. Next, nuclear extract was incubated at a final concentration of 1.5 mg/ml with 1 mg blank beads (washed with TE buffer plus 0.02% NP40, buffer DW and equilibrated in buffer G) at 4°C for 1 hour on a rotating mixer. For semi-quantitative DNA pull-down, the nuclear extract was split equally and mixed separately with 1 mg probe magnetic beads and 1 mg control magnetic beads at a final concentration of 0.7 mg/ml. After 3 h incubation at 4°C on a rotating mixer, beads were washed twice with 1.8 ml (per mg beads) with buffer G and twice with 1.8 ml (per mg beads) buffer G without igepal.

The mass spectrometry sample preparation was done by the in-house mass spectrometry group of David Meierhofer. An on-bead digestion with trypsin was conducted followed by peptide precipitation. Mass spectrometry (MS) was done on a Q Exactive HF Orbitrap (Thermo Scientific). MS and data processing was done by David Meierhofer using Mascot server (Matrix Science) and the MaxQuant suite [67] and label-free quantification.

3.12. Polyacrylamide Gel Electrophoresis and Semi-dry Western Blotting

A 10% SDS-polyacrylamide gel (2.5 ml 1.5 M Tris-HCl pH 8.8, 3.3 ml 30% acrylamide, 0.1 ml 10% SDS, 4.1 ml water, 7.5 μl TEMED, 75 μl 10 % APS) plus stacking gel (2.5 ml 0.5 M Tris-HCl pH 6.8, 0.85 ml 30% acrylamide, 50 μl 10% SDS, 2.85 ml water, 5 μl TEMED, 40 μl 10 % APS) was casted. Protein sample was mixed with 6x sample buffer and heated to 95° for 5 minutes and previously cross-linked probes were heated for 20 min. After sample loading, the gel was run at 85 V in running buffer (25 mM Tris, 192 mM glycine, 0.1% SDS) until the separation gel was reached, then voltage was set to 110 V for 2 hours. At this stage, the protein gel was either transferred into Coomassie Brilliant Blue solution and protein bands stained or used for western blotting.

For western blotting, nitrocellulose membrane (0.45 μm; BIO-RAD) and blotting

paper were soaked in blot buffer (50 mM Tris, 40 mM glycine, 20% methanol). The blot sandwich was assembled on a trans-blot SD semi-dry transfer cell (Bio-Rad) using polyacrylamide gel, nitrocellulose membrane and blotting paper and run at 55 mA per gel for 1 hour. After transfer, the membrane was blocked in 5% BSA-TBST (20 mM Tris, 500 mM NaCl, 0.05% Tween 20) on a rocking shaker at room temperature for 1 hour. Primary antibody was diluted in 5% BSA-TBST (Anti-Actin 1:1000, N499 1:3000) before incubating membrane on a rocking shaker at 4°C over night. Membranes were washed three times with TBST for 5 minutes each and once with 5% BSA-TBST, before incubating with secondary antibody HRP goat anti rabbit (1:4000) on a rocking shaker at room temperature for 1 hour. Membranes were washed five times with TBST 5 minutes each. For HRP signal detection, SuperSignal West Dura Extended Duration Substrate (ThermoFisher Scientific) was used and HRP signal visualized with a LAS1000 camera (Fujifilm).

3.13. GR Protein NMR

3.13.1. DNA Oligomer Preparation for NMR

1 μmole DNA (table B.3) was dissolved in 500 μl ddH_2O overnight. DNA was run over a strong anion exchange column (MonoQ HR 16/10, GE) equilibrated with buffer MonoQ-A (10 mM NaOH). Next, DNA was eluted by running a gradient from 0% to 100% of buffer MonoQ-B (10 mM NaOH, 1 M NaCl) and fractions of main peak at 260 nm were pooled. DNA was frozen in liquid nitrogen and lyophilized for at least 24 h to remove surplus water. Lyophilized DNA was dissolved in 1 ml MonoQ-A and NAP-10 columns (GE) were used to exchange buffer to water. In the next step, the water was removed completely using a SpeedVac. DNA was dissolved in 300 μl ddH_2O and concentration was determined. Forward ssDNA and reverse complement ssDNA were mixed 1 to 1 in 1.5 ml reaction tubes and hybridized overnight in boiling water.

3.13.2. ^{1}H-^{15}N-HSQC

2D (^{1}H, ^{15}N) heteronuclear single quantum coherence spectroscopy (HSQC) was recorded and processed by Dr. Marcel Jurk.

For NMR experiments, purified GR-DBD (final concentration 40 μM, section 3.4.1) was mixed with hybridized DNA oligomer (final concentration 53 μM). ^{1}H, ^{15}N-HSQC spectra were recorded as SOFAST versions [68] at 35°C on a Bruker AV 600 MHz spectrometer (Bruker, Karlsruhe, Germany) equipped with a cryo-probehead. Data processing was done using TOPSPIN (version 3.1, Bruker).

3.14. STARR-seq with Synthetic GBS Library

The original STARR-seq method was developed by the group of Alexander Stark [38] and later CapStarr-seq was published [42]. I adapted parts of both methods for a STARR-seq version to test synthetically produce DNA fragments containing GR binding sites.

3.14.1. Generation of Input Libraries

DNA Fragment Design for Integration into human STARR-seq vector

DNA fragments were designed to contain adapter sequences for illumina sequencing primer with P5 and P7 and in-fusion target sites for integration into the human STARR-seq screening vector (Figure 3.1). Any sequence from an arbitrary source can be used as insert and I inserted either small sequences like a single GBS (29 bp) or large sequences derived from GR-ChIP-seq peaks (up to 750 bp).

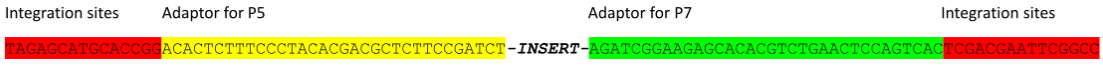


Figure 3.1.: DNA fragment design for integration in STARR-seq screening vector

GBS Fragment Preparation

To generate a GBS variant library, sequences containing N were ordered from IDT as "DNA Ultramer oligonucleotide" (table B.6 on page 103). The oligonucleotides were made double stranded in a one cycle Phusion-PCR (98°C 35 sec, 72°C 5 min) with primer SS194 (table B.3).

For testing of individual GBSs and GR ChIP-seq peaks, DNA fragments were ordered as "gblock gene fragments" from IDT (see table B.5). Because gblocks are already double stranded DNA fragments, they can be directly used for in-Fusion HD cloning reactions (Takara Clontech) with linearized human STARR-seq screening vector.

Preparation of Input Libraries

The human STARR-seq screening vector was digested with SalI-HF and AgeI-HF (NEB) at 37°C for 3 hours and linearized vector was purified after gel electrophoresis. For one in-Fusion HD cloning (Takara Clontech) reaction, I mixed 100 ng linearized STARR-seq vector, 25 ng GBS fragment and 2 μl 5x In-Fusion HD Enzyme premix (reaction volume 10 μl). Five reactions were set-up and incubated at 50° for 15 min before placing on ice. After pooling of reactions, plasmid DNA was cleaned-up using AMPure XP beads (Beckman Coulter) and eluted in 25 μl ddH_2O. For plasmid amplification, I used the highly electro-competent strain MegaX DH10B

(Invitrogen). 20 μl MegaX DH10B were transformed with 2.5 μl DNA, according to the manufacturer's protocol. Ten transformations were pooled and transferred in 1 l LB_{Amp} and incubated overnight in a shaker. The plasmid libraries were extracted using the "Plasmid Plus Maxi" kit (Qiagen).

3.14.2. Transfection into Human Cells

Transfection was done similarly to section 3.5.2 with the exception that 5 million cells (GR18, A549) were transfected with 5 μg plasmid input library for each condition and transferred into a 15 cm tissue culture dish. I tested two hormone incubation times, 4 hours and 15 hours. For the 4h time point, cells were transfected and on the next day treated with 1 μM dexamethasone and as a control with 0.1% ethanol vehicle for 4 hours. For the 15h time point, cells were transfected and after 4 to 6 hours treated with 1 μM dexamethasone and as a control with ethanol for 15 hours. To test if transfection affected input library composition, from the same batch 5 million cells were transfected with 5 μg plasmid input library in parallel to the experiment and cells were grown as long as cells for the experiment. Finally, the plasmid library was isolated from these cells using "Plasmid Mini" kit (Qiagen).

3.14.3. RNA Isolation and cDNA Preparation

Total RNA was extracted from treated cells using the "RNeasy midi" kit (Qiagen). The poly(A) RNA fraction was isolated from total RNA using 300 μl Dynabeads Oligo-dT25 (Invitrogen) per sample, according to the manufacturer's protocol. Poly(A) RNA was treated directly with turboDNase (Ambion) for 30 minutes at 37°C. Next, Poly(A) RNA was cleaned up with "RNeasy MinElute" kit (Qiagen) and eluted twice with 14 μl elution buffer. RNA was reverse transcribed using SuperScriptIII (Invitrogen) and a specific primer for STARR-seq mRNA (SS186, table B.3). For each condition, I prepared 10 reactions, each with 125 ng Poly(A) RNA as template and prepared it according to the manufacturer's protocol. After cDNA synthesis, reactions were pooled again and subjected to a RNaseA digest with 10 μg RNase A per 100 μl cDNA and incubated for 1 h at 37°. Finally, cDNA was cleaned up by the DNA clean up kit (Promega) and eluted in 50 μl elution buffer. cDNA content was measured using Qubit fluorometric quantitation (Thermo-Fisher).

3.14.4. Sequencing Library Preparation

To reduce STARR-seq plasmid background contamination and specially amplify cDNA, a 2-step nested PCR using "KAPA Hifi Hot Start Ready Mix" (KAPA biosystems) was conducted to prepare a sequencing library. 1 to 5 ng cDNA was used per reaction and amplified using two specific primers, SS192 and SS193 (table B.3). The entire cDNA was used in the first PCR step:

Initial Denaturation:	98°C	45 sec	
Denaturation:	98°C	15 sec	
Annealing:	65°C	30 sec	15 Cycles
Synthesis:	72°C	70 sec	
Cooling:	4°C	infinite	

PCR product was pooled and purified by Agencourt AMPureXP beads (Beckman Coulter) with a beads/PCR ratio of 0.8 and eluted in 50 μl TE-buffer. DNA content was determined using Qubit fluorometric quantitation. The purified PCR product served as template for the second PCR, using 5 ng DNA per reaction, the KAPA Hifi Hot Start Ready Mix and NEBNext Multiplex Oligos (NEB), for in total 8 reactions:

Initial Denaturation:	98°C	45 sec	
Denaturation:	98°C	15 sec	
Annealing:	65°C	30 sec	10 Cycles
Synthesis:	72°C	30 sec	
Cooling:	4°C	infinite	

PCR product was pooled and again purified by Agencourt AMPureXP beads (Beckman Coulter) with a beads/PCR ratio of 0.8 and eluted in 50 μl ddH_2O. Sample volume was reduced in a SpeedVac to 20 μl and DNA content was determined using Qubit fluorometric quantitation. Each library was multiplexed and sequenced on an Illumina HiSeq2500 platform, following the manufacturer's protocol.

Plasmid input controls were treated using the same condition as described above except using primer SS195 and SS196 for amplification in the first PCR step.

3.15. Computational Analysis

Data processing and analysis was done with the programming languages *Perl* and *R* (R version 3.2.0, Bioconductor version 3.2). Regulatory signals in DNA sequences were detected, processed and analyzed with "regulatory sequence analysis tools" (RSAT), a series of modular computer programs [69]. Specifically, I used *matrix-scan*, *matrix quality* and *peak-motifs* from RSAT. *Matrix-scan* was used to scan DNA sequences with a profile-matrix for matches [70]. *Matrix quality* was used to score the quality and compare score distribution of a matrix in datasets to background [71]. *Peak-motifs* is a pipeline for de-novo motif discovery in ChIP-seq data-sets [72]. The *bedtools* intersectBed [73] was used to intersect the genomic coordinates of two files.

3.15.1. Computational Preparation and Analysis of STARR-seq Data

The quality of the raw sequencing reads (*.fastq) was controlled by FastQC (Babraham Institute). Only reads with high quality according to FastQC and without Ns were extracted, 10 to 20 million reads depending on the sequencing run. The reads were aligned with a self-written *Perl* script and only sequences exactly matching the

input library in length and nucleotide composition were kept. In this way, around 17% of the reads had to be excluded. As a last step, the occurrence of each sequence variant was counted.

To compare sequence counts across experimental conditions and identify differentially expressed sequences, count data was analyzed with DESeq2 [74]. Due to the high number of outliers in DESeq2, outlier replacement was turned off by *DESeq* with *minReplicatesForReplace=Inf* and *results* with *cooksCutoff=FALSE*. To fit the dispersion curve to the mean distribution, local smoothed dispersion was preselected (*DESeq* with *fitType="local"*).

3.15.2. Identification of GR Binding Sequences Associated with Gene Regulation

From DNA microarray data in dexamethasone-treated U2OS cells (made by Samantha Cooper, stored at ArrayExpress: E-GEOD-38971), the differentially regulated genes were called using adjusted p-value <0.05 and were assigned to 2 different groups. The first group consisted of the 20% most up-regulated genes upon hormone treatment ($\log_2$fold change(dexamethasone/ethanol vehicle)= 1.91 to 7.86; 290 of 1447 genes). GRα ChIP-seq data was used from a previous publication [33] (E-MTAB-2731). Next, I extracted the ChIP-seq peaks (FDR=2%) from U2OS cells in a 40 kb window centered on the transcription start site of each gene (543 peaks in total from 290 genes of the most up-regulated genes). For a control group, I extracted a similar number of peaks (532) for the genes showing only weak regulation ($\log_2$(absolute foldchange)$\leq|0.72|$) upon hormone treatment independently of up or down regulation. For each group of peaks, I performed de-novo motif searches using *peak-motifs* (default settings and using the dyad-algorithm) from *RSAT* [72]. Detected motifs were analyzed and motifs matching GR consensus motif (M00205, TRANSFAC 2010.1) were manually extracted.

3.15.3. Analysis of GR ChIP-Seq Data

Enrichment Plot of flanked GBS

To score the enrichment of AT-flanked GBS and GC-flanked GBS in the peaks associated with strong up-regulation and weak regulation, *matrix-quality* from RSAT was used to compute normalized weight differences (NWD) [71]. The input matrices for matrix-quality, were generated from GR motif found by *peak-motifs* (see section 3.15.2) by enforcing only A/T or G/C at the flanking site position.

Comparison of Peak Height

To compare peak height of A/T and G/C flanked GBS, the GRα ChIP-seq peaks in U2OS were scanned for occurrence of GBS-match with RSAT-matrix-scan (p-value cut off: 10^{-4}, TRANSFAC matrix M00205) [70]. Peaks were grouped according to

flanking sites and peaks containing A/T flanked versus G/C flanked GBSs were plotted. The Boxplot of peak height was produced using R.

3.15.4. DNAshapeR: Predicting DNA Shape

For the second flanking site construct, I used DNAshapeR to predict DNA shape. DNA shape features (minor groove width, propeller twist, helix twist and roll) of a given sequence (*.fasta) were predicted by *DNAshapeR* [55] in R, which is based on the DNAshape prediction method [54]. DNAshape uses a sliding pentamer window on a DNA sequence to derive the structural features from all-atom Monte Carlo simulations [53].

DNA shape prediction for the first flanking site (see section 4.2.3) was done by Iris Dror based on the DNA shape prediction method [54]. For the group analysis of A/T and G/C flanked GBS, I extracted 83 GBS flanked by A/T from peaks associated with strong GR responsive genes and 75 GBS flanked by G/C from peaks associated with weak GR responsive genes. Next, I aligned the GBSs and set the middle spacer position to 0.

4. Results

4.1. Quantitative Massively Parallel Assessment of GBS Activity with STARR-seq

GR binding sequences show a large variability in nucleotide composition in the human genome. For instance, after scanning of GR bound ChIP-seq peaks in U2OS cells for occurrence of the GR motif (matrix-scan, transfac-matrix M00205, p-value=10^{-4}), I identified 14676 GBSs. Astonishingly, I found in these GBSs 12295 different GBS variants and more than 74% of the GBSs were unique. A similar number of unique GBSs was recorded by Watson et al. [50]. However, until now researchers were limited to testing a small number of GBS for GR activity, for instance by luciferase reporter assays, and they could not fully investigate the sequence variability of GBSs. The newly developed STARR-seq method may allow for the first time to test in parallel a large number of GBS variants and their effect on GR activity. I adapted the STARR-seq method to quantitatively assay the activity of a large number of GBS variants by generating a synthetic GBS library. In this way, I could assume that the only variable affecting GBS activity was the sequence composition itself.

4.1.1. Establishment of STARR-seq for GR

Figure 4.1.: Graphical representation of STARR-seq screening construct for testing of GR binding sequences. The region containing a GBS was inserted between a GFP reporter gene and a polyA site.

As a first step, I tested different variations of GBS inserts for the STARR-seq screening vector (Figure 4.1), since in the beginning I did not know whether a single GBS was sufficient to activate reporter gene expression or only a GBS surrounded by sequence as in the endogenous genomic context. Therefore, I tested three different constructs variants in GR18 cells and recorded GFP expression of the STARR-seq reporter gene in comparison to the constitutively active SV40 enhancer and to a random sequence (Figure 4.2). I co-transfected mCherry to normalize for transfection efficiency. The first construct contained a GR ChIP-seq peak sequence (211

bp) from near the *FKBP5* gene, containing the Fkbp5-2 GBS in an endogenous sequence context. *FKBP5* is a gene directly regulated by GR. In the second and third version, the Fkbp5-2 GBS was inserted as a triplet (3x) and as a single sequence (1x), respectively. The SV40 enhancer and the random sequence were supposed to be insensitive to hormone (1 μM dexamethasone) treatment. The SV40 enhancer showed without hormone treatment a strong GFP expression (>90% GFP positive cells), whereas the random sequence showed no GFP expression (1.2 to 2.4% GFP positive cells) compared to untransfected cells (2.2% GFP positive cells). The three Fkbp5-2 construct variants showed strong GFP expression only upon hormone treatment (63% to 70% GFP positive cells, Figure 4.2). The relative mean expression (GFP/mCherry) was similar for all three construct variants: FKBP5-peak was 0.09 RFU, 3xFkbp5-2 and 1xFkbp5-2 were 0.19 RFU. Altogether, I could show that a single GBS activates STARR-seq reporter gene expression in U2OS.

The reason for using STARR-seq is to quantitatively assess the activity of GBS variants. This relies on the ability of GBS variants to induce different levels of activation of the STARR-seq reporter. To test if this is the case, I tested two GBS variants, Fkbp5-2 and Cgt, and analyzed activation of expression by the two GBS variants both at the mRNA level and at the GFP level. Indeed, differences in activity were detectable, Fkbp5-2 showed both higher mRNA (6-times more) and GFP level (1.2-times more) than Cgt (Figure 4.3). Notably, the STARR-seq vector with SV40 enhancer and random sequence showed hormone-dependent regulation at mRNA level as well, indicating a slight hormone-dependent activity of the STARR-seq screening vector backbone. In summary, the STARR-seq method appeared to be usable for quantitative assessment of the activity of single GBS variants.

4.1.2. STARR-seq Experiment with Synthetic Library: Quality Control

The core GR binding sequence alone includes 15 nucleotides. Testing all possible variants of a 15 nt long sequence would result in testing around 1 billion (4^{15}) different variants. The STARR-seq method was designed to test millions of candidate sequences in parallel, but comprises three bottlenecks: firstly the plasmid transformation in *E. coli*, secondly the transfection into human cells and thirdly the sequencing depth. To completely cover 1 billion sequence variants would mean a tremendous amount of material, and might not be easily doable for one experimentalist. In order to reduce the number of variants, I chose to vary only parts of a GBS and its flanking regions. I was interested in the influence of the composition of the GR-halfsite sequence, the flanking sequence and the spacer sequence on GR's activity. That is why I chose to design three different STARR-seq screening libraries with short randomized DNA sequences (Figure 4.4). In the first library, called GBS-halfsite (GBS-HS) library, one halfsite of a GBS was fixed to a specific sequence (position -7 to -2) and the second halfsite was variable (position 2 to 7). The spacer contained one T (position -1) next to the fixed GBS-halfsite and spacer position 0 and 1 were variable. In total 8 consecutive nucleotides were chosen to be randomized, resulting in the following consensus AGAACAtnnNNNNNN and leading to

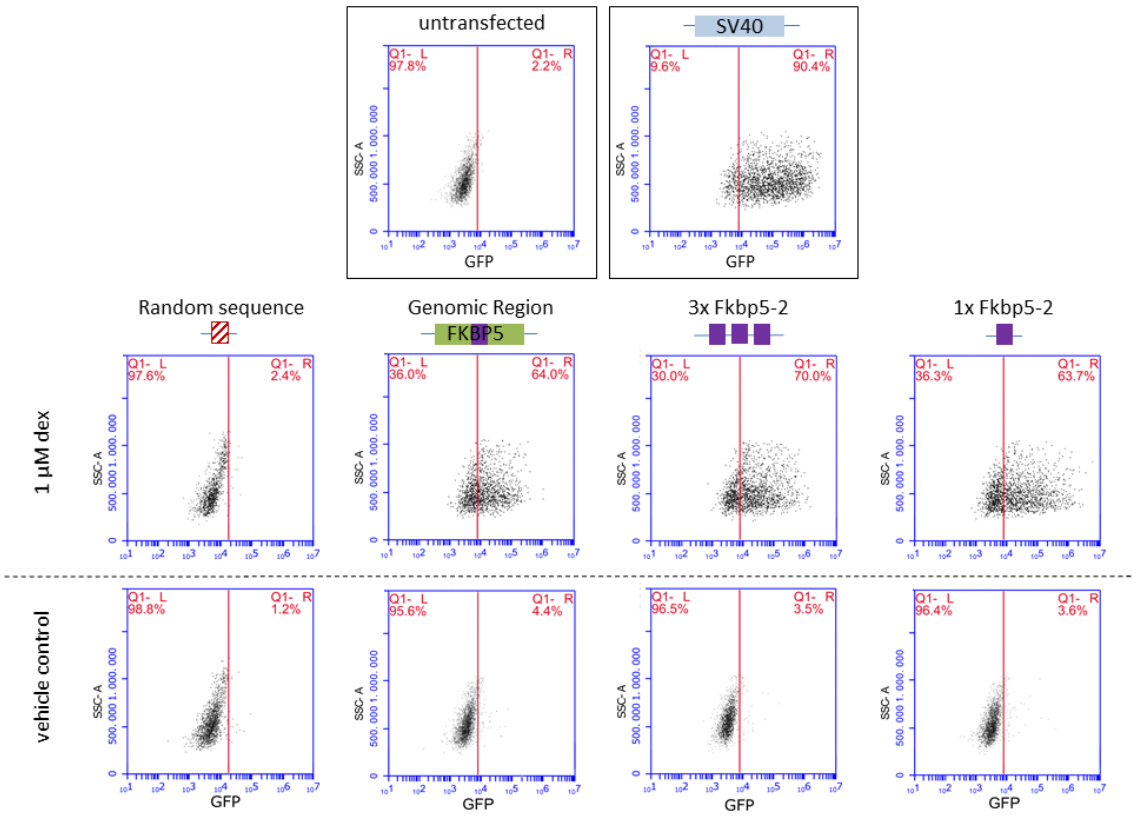


Figure 4.2.: Testing GFP expression of STARR-seq vector with different inserts. Inserts are SV40 enhancer, random sequence, *FKBP5* peak sequence including Fkbp5-2 GBS, 3xFkbp5-2 GBS and 1xFkbp5-2 GBS. GR18 cells were co-transfected with STARR-seq constructs and pSV40::mCherry for transfection normalization. "Untransfected" cells were transfected only with pSV40::mCherry and not with STARR-seq screening vector. GFP signal for mCherry-positive cells are shown. Percentage of Q1-R (top right corner) represents % of GFP positive cells for each construct. Between 1200 to 2000 GR18 cells were recorded in each experiment.

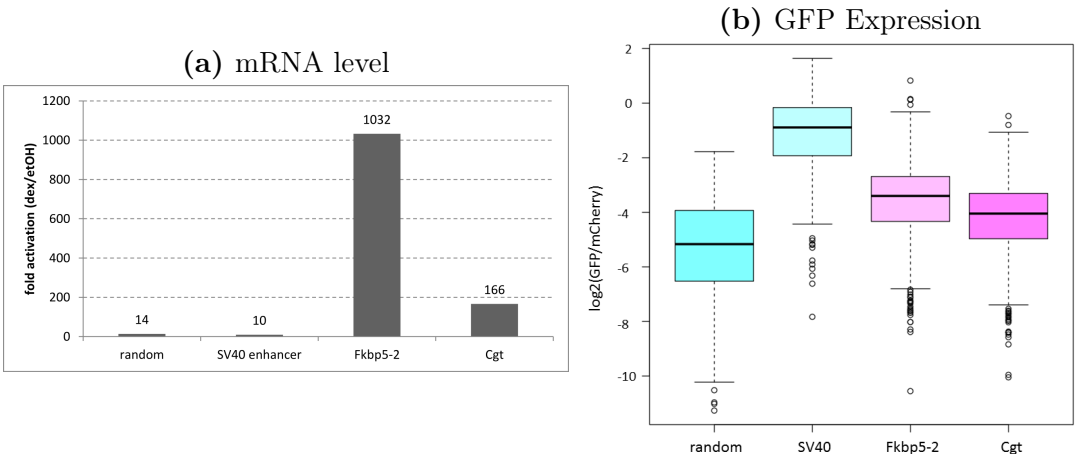


Figure 4.3.: Analysis of reporter gene expression in STARR-seq screening vector by comparing Fkbp5-2 GBS and Cgt GBS. qPCR measurement of mRNA of single experiment. Panel (a) shows that Fkbp5-2 GBS leads to 6-times more reporter gene expression upon hormone treatment than the Cgt GBS. Random sequence and SV40 enhancer show that the backbone of STARR-seq vector exhibits slight hormone dependent activation. Panel (b) shows that Fkbp5-2 GBS leads to 1.2-times higher GFP signal than Cgt GBS in hormone treated cells.

65536 (4^8) possible variants (Figure 4.4). The other two designed libraries, named Cgt-flank and Sgk-flank, consisted of two complete GBSs, Cgt (AGAACAtttTGTACG) and Sgk (AGAACAtttTGTCCG), respectively, and were flanked by 5 consecutive Ns downstream of the GBS, resulting each in 1024 (4^5) possible variants. Cgt-flank and Sgk-flank libraries were joined after the In-fusion reaction (see section 3.14.1) and were treated as one input library for transfections. Synthesis and integration of fragments into STARR-seq screening vector was specific and correct, when looking at the Sanger sequencing result of the STARR-seq libraries (Figure 4.4).

The fragments with Ns for the STARR-seq screening library generation were synthesized by the IDT company and for each N adenine, guanine, cytosine and thymine were incorporated at random. The random incorporation of nucleotides during synthesis worked well (Figure 4.4) and I recorded only a small bias in nucleotide incorporation (Figure 4.5a). Thymine incorporation happened a bit more frequently (30% of total) than the other 3 bases, whereas cytosine was incorporated less frequently (22% of total) than the other 3 bases. Throughout all experiments, I was able to recover sequencing information for more than 99% of all possible sequence variants for the GBS-HS library and even 100% for the Cgt/Sgk flank library. This shows a good experimental scale with a decent coverage of sequence variants.

After conducting the STARR-seq experiment, the nucleotide composition shifted only marginally in the reads from GBS-HS input library to cDNA reads of etoh control treated cells. In contrast, upon hormone treatment the nucleotide composition shifted. For example, the G at position 3 and the C at position 6 (Figure 4.5c) were

enriched compared to the etoh control, going from 28% to 32% in the dex sample and from 18% to 23%, respectively. This gave a first indication that the STARR-seq experiment for GR was successful, since I was expecting that upon hormone induction the consensus GR-halfsite would be enriched. The G3 and C6 correspond to the G and C in tGTtCt of the consensus GR-halfsite.

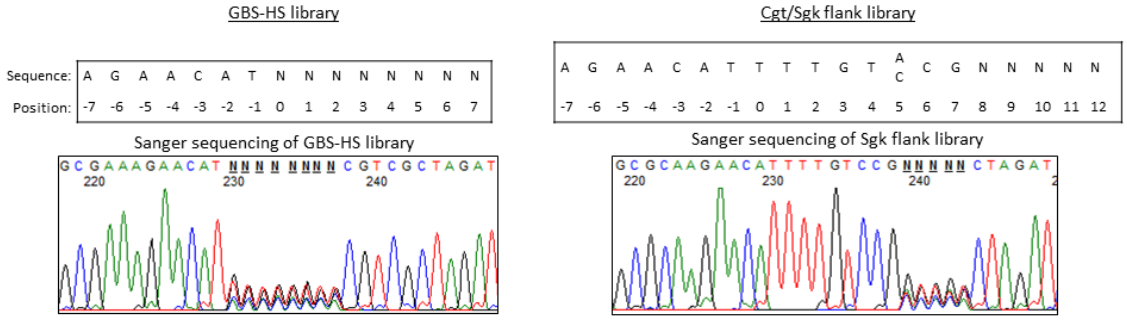


Figure 4.4.: Representation of STARR-seq screening libraries for GR. Library sequence and position are shown with Ns representing randomized part. Sanger sequencing of libraries shows specific and correct integration of synthesized fragments into the STARR-seq screening vector.

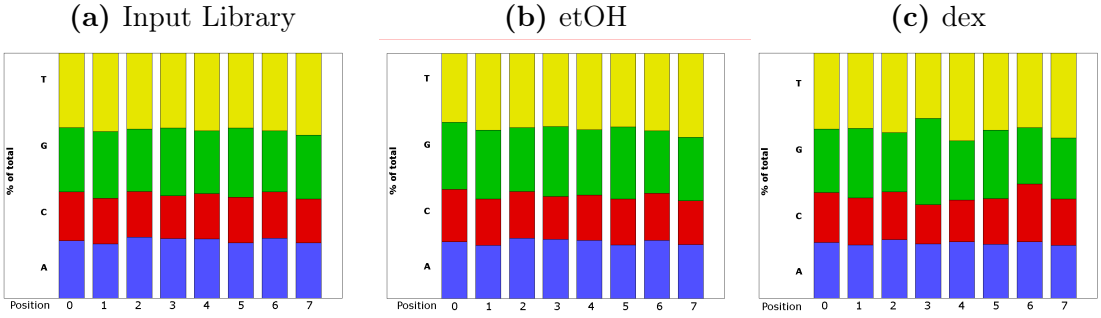


Figure 4.5.: Frequency plots for nucleotide composition of reads from (a) GBS-halfsite library before transfection, (b) from cDNA etoh control and (c) from cDNA dex sample.

After sequencing, I counted the occurrence of each sequence variant in each experiment and in the input library pre- and post-transfection. The transfection of the GBS-HS library into U2OS cells had little effect on the STARR-seq library composition (Figure 4.6a), since the STARR-seq library before and after transfection was nearly identical. By comparing count number of GBS-halfsite variants between hormone and etoh vehicle treated cells, I found that some sequences are mainly expressed in dex sample, some mainly in etoh control, but for most tested

sequences the count number was comparable between dex sample and etoh control (Figure 4.6b). The sequences specifically found to be expressed in the dex sample, likely correspond to sequences that are bound and activated by the GR protein. As expected, for most sequences expression did not change between dex and etoh, highlighting that most sequences are not functional GR target sequences.

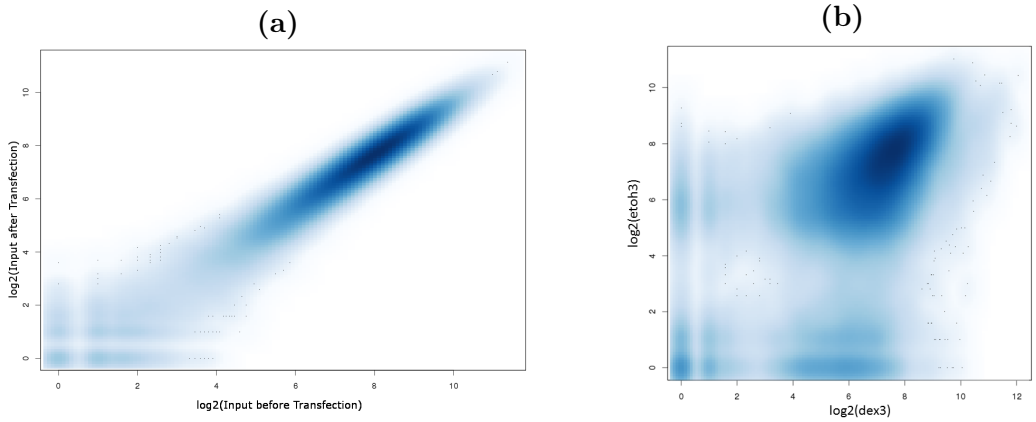


Figure 4.6.: Correlation of counts from STARR-seq GBS-HS library experiments Panel (a) shows a strong correlation between counts of input library pre- and post-transfection. (b) Correlation between counts of sequence variants were plotted for dex sample and etoh control from replicate Nr.3.

4.1.3. Summary of Conducted STARR-seq Experiments

This section lists a summary of all conducted STARR-seq experiments, conditions (treatment and incubation time of treatment) and replicate containing number of reads and number of variants for each tested library. I conducted experiments at two different hormone incubation times (4h and 15h). The number of aligned reads represents the sequencing depth for each experiment and contained only reads exactly matching the input library in length and nucleotide composition. The number of variants shows how many different sequence variants were identified.

cell line	Library	Treatment	time	# aligned reads	# variants
GR18	GBS-HS	dex1	15 h	16075045	61630
		etoh1		15467223	61025
		dex2		16090237	64440
		etoh2		13238403	63737
		input2		12360316	64531
		dex3		9583380	61407
		etoh3		9259587	57919
		input3		8429558	63341
GR18	GBS-HS	dex1	4 h	12059206	62700
		etoh1		10553965	61469
		dex2		13978745	61505
		etoh2		14775761	62222
A549	GBS-HS	dex1	15 h	14428380	63342
		etoh1		16462499	62463
		input1		10656798	64037
		dex2		14897589	63761
		etoh2		13049295	63444
GR18	Cgt/Sgk flank	dex1	4 h	8910038	2048
		etoh1		8795453	2048
		input1		8909858	2048
		dex2		17055208	2048
		etoh2		13715556	2048

4.1.4. Analysis of Sequence Activity of GBS-HS Library with DESeq2

To find sequences responding to hormone treatment, I conducted a differential analysis with DESeq2 of sequence count for three biological replicates of the GBS-HS library in GR18 after 15 hours hormone treatment. There was a large variability in the ethanol controls between replicates, which could be due to both technical and biological origin (Figure 4.7). The hormone treated samples cluster mostly together and argue for a high degree of reproducibility.

After merging the count data of all 3 replicates, I was able to extract count information for a total of 65510 variants from 65536 possible variants. The DESeq2 analysis revealed 1045 sequences to be differentially expressed with an adjusted p-value$<$0.01. From these sequences, 918 were up-regulated and 127 were down-regulated compared to ethanol control. In figure 4.8, I plotted the normalized counts for two example sequence, AGAACAT**TCCGTCCA** and AGAACAT**CACCTCTG**, representing up- and down-regulation, respectively. AGAACAT**TCCGTCCA** was expressed rarely in the ethanol control, but was strongly expressed after hormone treatment with a good agreement between replicates. The expression of AGAACAT**CACCTCTG** was reduced after hormone treatment and showed a strong down-regulation. The mean count of each differentially expressed sequence variant was plotted to its $\log_2$foldchange (Figure 4.9a) The mean count is the mean of normalized counts of all dex samples and etoh controls, normalized for sequencing depth. It can be seen that a larger cloud of sequences exhibits mean(count)$<$100 with high $\log_2$foldchange around -4 and 4. When looking at these sequences, they seem to show no apparent second GBS halfsite. Therefore, I named these sequences

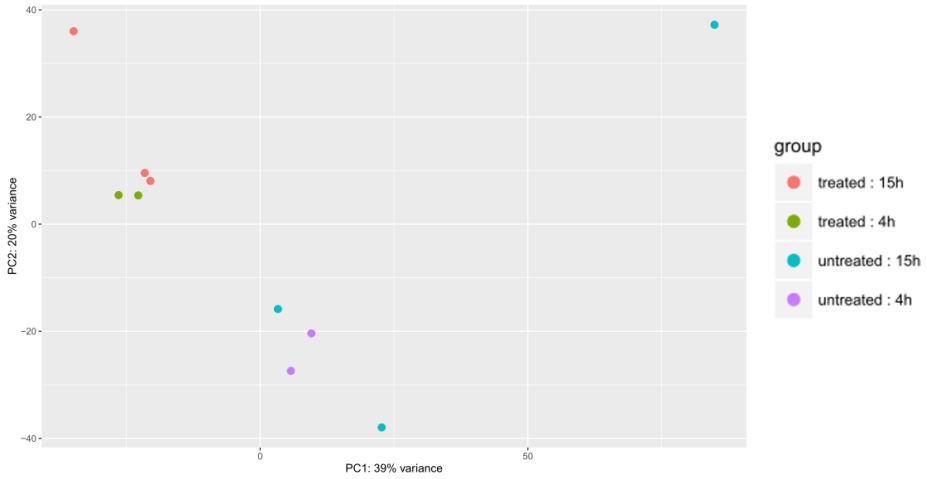


Figure 4.7.: Principal component analysis of biological replicates in comparison to treatment and time of hormone incubation. "Treated" corresponds to dex treated sample and "untreated" corresponds to ethanol control. Time of treatment was either 4 or 15 hours before RNA isolation.

"nonGBS" sequences. Expression of one example nonGBS, AGAACAT**GACTCTAA**, can be seen in figure 4.8. These nonGBS sequences were detected only rarely or if at all in one condition and a little more frequent in the other condition. Yet, these nonGBS sequences showed high reproducibility between replicates with an adjusted p-value$<$0.01. I plotted the $\log_2$foldchange and mean expression for all tested 65510 sequence and marked in red the differential expressed sequence and from these in orange the sequences with low mean count (Figure 4.9b). These sequences made-up 38% of the differential expressed sequences. To test if these nonGBS sequences were artifacts or indeed biological meaningful, I selected 9 "nonGBS" sequences, showing a $\log_2$foldchange between 5 and 6 in DESeq2 analysis and retested them by repeating the STARR-seq experiment with individual sequences instead of library context (Figure 4.9c). In contrast to the predicted regulation from the DESeq2 analysis, all tested nonGBS sequences were inactive when tested in isolation. This means that sequences with low mean counts that were identified by DESeq2 to be strongly differentially expressed were not reproducible and therefore are not biological meaningful. As a consequence, for future analysis, I removed all sequences with mean(count)$<$100 from the pool of differentially expressed sequences. I chose 100 mean(count) as a stringent cut-off to remove all sequences that correspond to the large cloud described in figure 4.9a and might be false positive. After removing the nonGBS sequences, 650 differentially expressed sequences remained. From these sequences, 592 were up-regulated and 78 were down-regulated.

As expected, most of the 65510 tested sequences were not differentially expressed upon hormone treatment. Figure 4.10a displays $\log_2$foldchange of all tested sequences ordered by function of foldchange. I grouped the sequences in 4 activator categories: negative, weak, medium and strong activator upon hormone treatment.

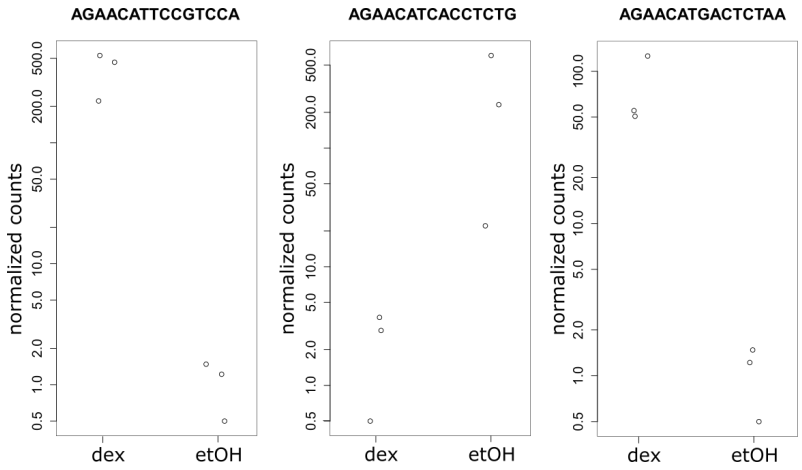


Figure 4.8.: Example of counts after normalization of replicates by DESeq2 for three sequences. AGAACATTCCGTCCA was the sequence with the best adjusted p-value ($2.3{\cdot}10^{-17}$). AGAACATCACCTCTG represents a good example of down regulation. AGAACATGACTCTAA represents an example with low mean count and high adjusted p-value ($1.8{\cdot}10^{-5}$), but was not differential regulated after validation.

From these four groups I selected 4 sequences each to be retested individually, taking into account that for the negative, strong and medium group, I selected only sequences with an adjusted p-value$<$0.01 (Figure 4.10b). The relative expression of selected sequences was in good agreement with their group activity in the STARR-seq library. A single sequence in the medium group showed high expression values ($\log_2$(relative expression)=4.5), indicating that some sequences might behave differently in library context compared to individual context. Logos and frequency plots were created for the aligned sequence for each group (Figure 4.10c). The logo of the strong and the medium activator group contained the canonical GR motif. The weak activator group showed no motif at all and the negative activator group showed no clear motif. GR might still be able to bind as a monomer to the sequences of the weak activator group, explaining maybe the slight up regulation. A color chart of all 592 differential up-regulated sequences, ranked by foldchange, showed a strong enrichment of G at position 3, of T at position 4 and of C at position 6 (Figure 4.11a). The bases at these positions corresponded to the preferences found for the classical GR motif (nntGTtCt), which recapitulated the identified logos of the strong and medium activator (Figure 4.10c). Notably, the top 19 sequences show the enrichment of a C-rich motif TnCGTnCc, marked by a T in the spacer at position 0 and by a C at position 2. I selected 4 C-rich GBS sequences (AGAACATTCCGTCCA, AGAACATTCCGTGAC, AGAACATCTCGTTCT, AGAACATTGAGATCC) and after individual retesting of gene expression, they showed strong up-regulation as predicted by DESeq2 (Figure 4.11b).

In summary, the STARR-seq experiment with a synthetic GBS library worked very well. I was able to rediscover the known GR motif in the differential expressed

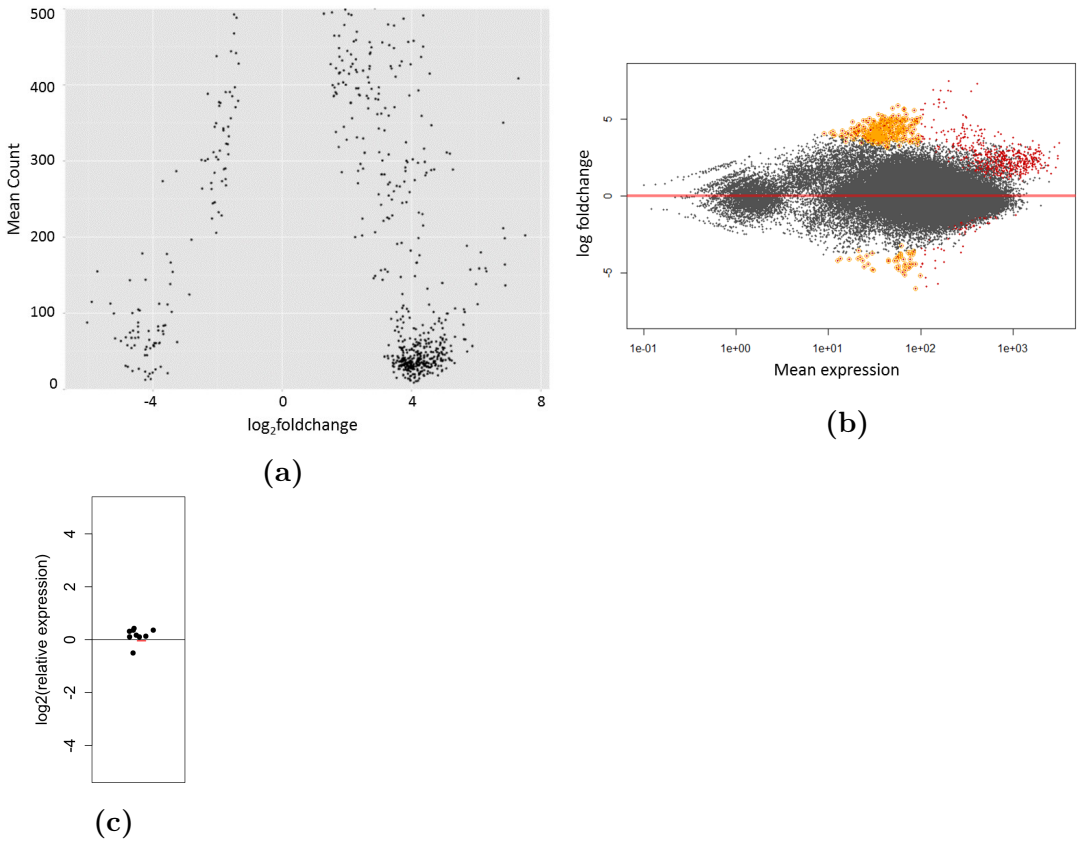


Figure 4.9.: Analysis of sequences with low mean count Panel (a) compares the mean count to the foldchange for all differentially expressed sequences. A large group with counts below 100 and strong fold change is revealed. Panel (b) shows MA-plot for all expressed sequences, red marks significantly differential expressed sequences (adjusted p-value<0.01), from these orange circle marks sequences with mean count below 100. (c) Relative expression of 9 "nonGBS" with mean(count)<100, after repeating STARR-seq for individual sequences and qPCR analysis. Relative expression was expression compared to background activity of STARR-seq screening vector in qPCR.

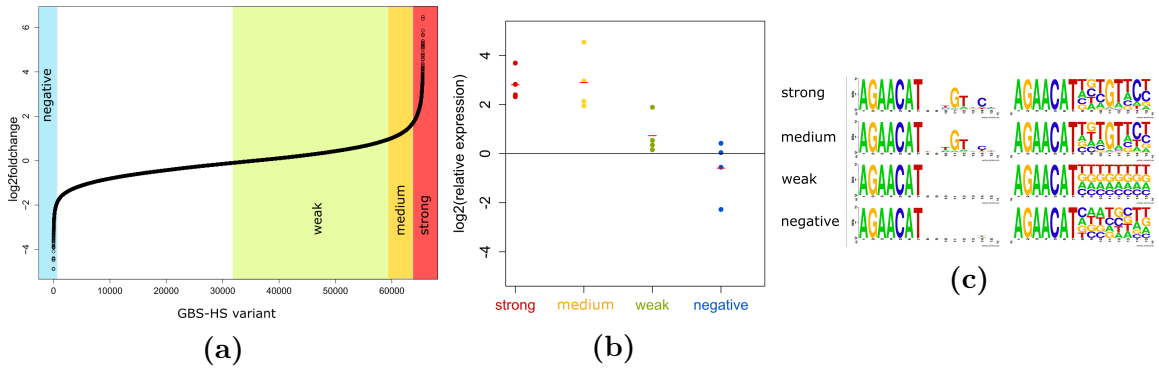


Figure 4.10.: Validation of STARR-seq experiment with synthetic GBS library. (a) Sequences were ranked in function to $\log_2$foldchange and grouped into 4 groups. (b) Relative expression of retested individual sequences from the 4 groups (c) (left) Logo and (right) frequency plot of sequences in each groups. Relative expression was expression compared to background activity of STARR-seq screening vector in qPCR.

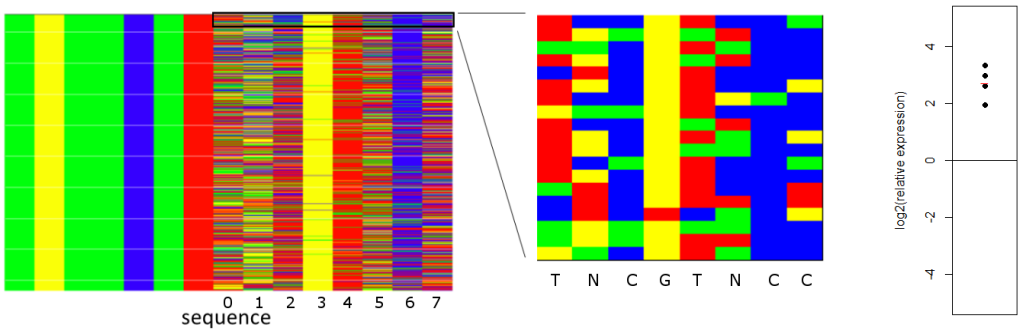
sequences (Figure 4.11a) and I discovered a second previously unknown GR motif, the C-rich motif (TnCGTnCc), which appeared to correlate with high activation. The nonGBS sequences revealed a small pitfall of the method and therefore for further analysis I removed the sequences with low mean(count). Yet, for most sequences, I believe, the STARR-seq library activity quantified by DESeq2 recapitulates well their real GR activity (Figure 4.10b).

4.1.5. Analysis of STARR-seq Sequence Activity for Different Conditions

In the next section I wanted to explore STARR-seq sequence activity for different conditions; I conducted STARR-seq experiments with 4h hormone induction, with the Cgt and Sgk flank library and in a different cell line.

GR action follows a complex kinetic regulation pattern, some genes are regulated very rapidly on minutes time-scale, some genes are rapidly induced but return soon back to background level and for some genes induction rates are slow but persist the complete induction period [75]. Other genes might be regulated indirectly by GR, because GR induces expression of a second transcription factor that activates or represses gene expression. Therefore, it might be advantageous to shorten hormone treatment, which might lead to fewer secondary effects by hormone induction. I compared the STARR-seq activity of sequences at the two tested time points, 4 hours and 15 hours, to be able to detect differences in potential kinetic patterns of induced sequences.

Further, not only is the GBS involved in defining GR action, but the regions flanking a GBS might also be involved in GR action. For other TFs, flanking bases were shown to be important for functional TF binding and action [43, 76]. To test



(a) (b)

Figure 4.11.: C-rich GBSs show strongest activation.(a) Color chart of the 592 differential up-regulated sequences were ranked in function to log_2foldchange (rising from bottom to top). The top 19 sequences contained a cluster of C-rich sequences with consensus sequence of TnCGTnCc. (b) 4 selected C-rich GBS show strong relative expression after individual retesting. Relative expression was expression compared to background activity of STARR-seq screening vector in qPCR.

the effect of flanking sequences on GR activity, I tested the STARR-seq Cgt and Sgk flanking libraries. Lastly, I tested the GBS-HS library with STARR-seq in a different cell line, A549, to detect possible cell type-specific GBS activity.

Comparison of Hormone Induction Time

The differential analysis of sequence count was done again with DESeq2 for two biological replicates transfected with the GBS-HS library in GR18 after 4 hours hormone treatment (Section 4.1.3). The data was of high quality (sequencing depth and number of variants) and there was little deviation between the two replicates (Figure 4.7 on page 54).

After merging the count data of the two replicates, I received count information for a total 65305 of variants, and found 2261 differential regulated sequences (adj. p-value$<$0.01, Table 4.1). I found 3.5 times more differential expressed sequences at 4 hours than at 15 hours treatment. Finding less differential regulated sequences after 15 hours was probably due to variability between the 3 replicates, resulting in contradictory count information for some sequences, which were removed from analysis by DESeq2. Detected expression changes in the 4h STARR-seq were in a similar range (-6.6 to 7.4 log_2foldchange) to the 15h STARR-seq. These findings indicated, that 4 hours of hormone induction are sufficient to detect expression changes as a result of GR action.

Next, I wanted to analyze if sequences differ in activity between 4 hours and 15 hours hormone treatment. Therefore, I compared the two time points with DESeq2 and chose as threshold an adjusted p-value$<$0.1 and mean(count)$>$100. I found from 65503 tested sequences 66 to be differently regulated between the two

Table 4.1.: Overview of number of differentially regulated sequences of GBS-HS library after 4 h hormone treatment

adj. p-value	total DE	LFC	# variants
<0.1	7069	up	4407, 6.7%
		down	2662, 4.1%
<0.01	2261	up	1813, 2.8%
		down	448, 0.7%

time points. 30 sequences appeared to be regulated only after 4 hours (up and down) and 36 sequences appeared to be regulated only after 15 hours (up and down). Most of the identified sequences did not possess a large mean count, so there is a high risk of being false-positive, unless sequences are validated, which I could not do for time reasons. Examples of identified sequences can be seen in figure 4.12. AGAACATGAAGCCAA appeared to be down regulated only after 15 hours, whereas AGAACATTCGTACAA was not differentially regulated anymore after 15 hours. Interestingly, AGAACATCGTGGATT seemed to be up-regulated after 4 hours, but was down-regulated after 15 hours. Altogether, the comparison of the two time points showed that they were to a large degree in agreement for sequence activity and I could not detect a clear example of a GBS that follows different kinetic. Also secondary effects seemed not to play a major role in the 15 h incubation, since I found only a few sequences that might be differential regulated between 4h and 15h hormone treatment.

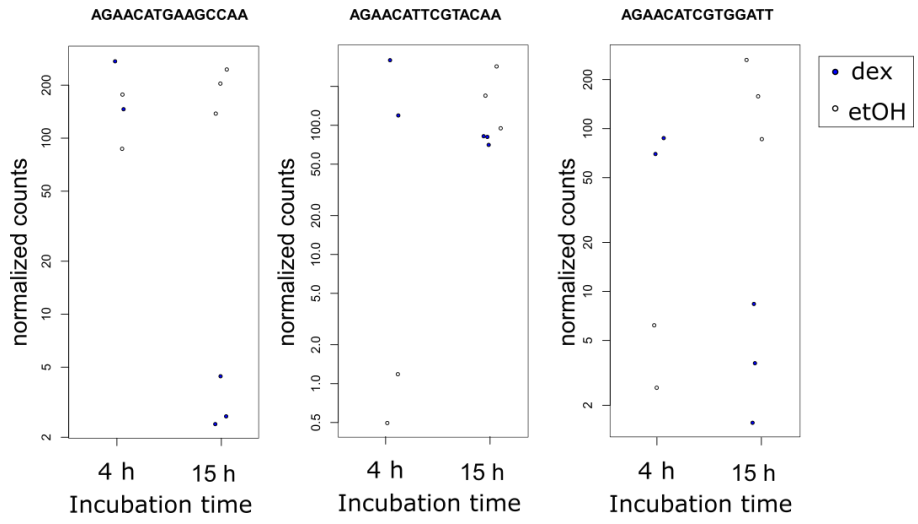


Figure 4.12.: Comparison between time and normalized gene expression counts.. Three example were chosen that showed differential expression between 4 h and 15 h. Counts are plotted for dex sample and etoh control of each replicate.

Diversity between GBS Flanks and Sequence Activity

To analyze if not only the GBS, but also flanking region sequence influence GR's activity, I tested the Cgt and Sgk flank library after 4h hormone treatment. For the two libraries, I was able to extract information for all possible sequence variants with at least 1000 counts at each condition. The range of foldchange was smaller (-0.84<LFC<1.13) for STARR-seq experiment with flank libraries compared to STARR-seq experiment with GBS-HS library at 4h hormone treatment. Cgt and Sgk are two validated active GBS variants [1, 50] and they were active in the STARR-seq GBS-HS library experiment with a $\log_2$foldchange of 1.8 and 1.6 (p-value<0.01). On the contrary, Cgt and Sgk in the flank library with very same flanks as in the GBS-HS library showed no differential expression, $\log_2$foldchange -0.1 and 0.1, respectively. From 1024 tested flanks of Cgt and Sgk only 20% showed significant up-regulation and 17% showed significant down-regulation (see table 4.2). However, this interpretation of the DESeq2 analysis might be misleading, since an analysis like this is based on the assumption that only a minority of sequences actually change gene expression and that the majority does not change expression. In this case, however, we expect that most candidate sequences are active and lead to up-regulation. Further, for etoh control and dex samples a similar number of reads were sequenced and we sampled only part of the population of reads and not the total population of expressed reads. Extremely strong expressed sequences may dominate sequencing and may therefore distort sequencing results. Therefore, flanks marked as significantly up-regulated might belong to a group of sequences exhibiting expression stronger than average after hormone treatment. These flanks mark a group of flanks that enhance activity of GBS. The flanks with negative foldchange might actually be flanks that have less than average expression and are reduced in the dex sample, because of the large number of strongly activated sequences. These flanks mark a group of flanks that blunt activity of GBS. These circumstances did not mean I could not analyze the data and I was still able to draw conclusions from flanking region usage in the different groups.

Table 4.2.: Overview of differentially regulated flanking sites of Cgt and Sgk

GBS	LFC	adj. p-value<0.1	adj. p-value<0.01	Differing activity Cgt vs. Sgk
Cgt	up	208, 20%	143, 14%	-
	down	178, 17%	80, 8%	1
Sgk	up	196, 19%	126, 12%	1
	down	173, 17%	70, 7%	1

A color chart of the flanks enhancing the activity (adj. p-value<0.01) of Cgt and Sgk, ranked by foldchange, showed the enrichment of TACNN in the flanks for both GBSs (Figure 4.13). By eye, the color chart for the enhancing flanks looked similar for Cgt and Sgk. Yet, Cgt and Sgk differed, when studying the flanks with blunted activation. Cgt showed enrichment of ATNNN flanks, whereas Sgk showed enrichment of NGNTN flanks in the blunted activation group. Interestingly, after

close analysis of the TACnn motif in the enhancing flanks, the sequence seemingly formed a third GR-halfsite or a second GBS: AGAACAtttTGTACGTACNNCTAGATCG. The existence of an additional binding sequence for GR might explain quite well the strong enrichment of this sequence in the top hits by recruitment of more GR.

By comparing Cgt and Sgk flank library, I was expecting to see only little differences in activity between flanks, because the two GBS are nearly identical and differ in only one base position (Figure 4.4). A direct comparison of flanks of Cgt with Sgk in DESeq2 revealed only 3 differentially regulated flanking regions (adj. p-value<0.1, Figure 4.14 and Table 4.2). The differentially regulated flanks, CTATG and CATGT, showed some variation in count number between replicates and need to be validated to make any assumptions. The third differentially regulated flank, GGAAA, showed reasonable differences between Cgt and Sgk and looked like a good example for GBS-specific effect of flanks.

In summary, flanks can modulate GR activity both ways and can enhance and blunt activity of GBS. For TACNN flanks, the discovery of a second GBS might well explain the strong enhancing effect. It is also possible that flanks might contain TFBS that interfere with GR action.

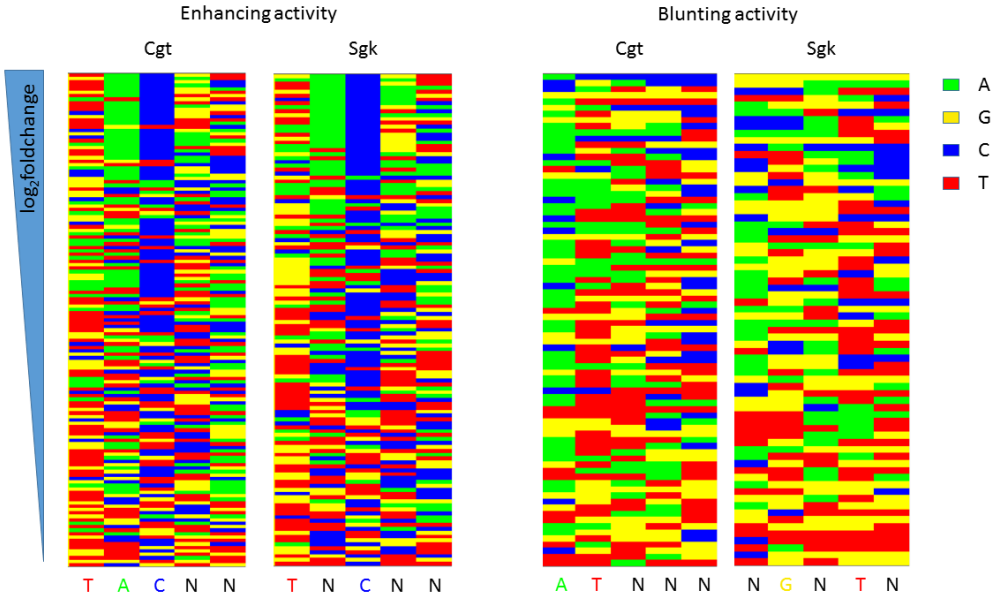


Figure 4.13.: Flanks correlate with activity. Color charts of the differentially up- (enhancing activity) and down-regulated (blunting activity) flanking regions for Cgt and Sgk were ranked to log_2foldchange (adj. p-value<0.01, top to bottom). Threshold for consensus sequence at the bottom of color chart was 40%.

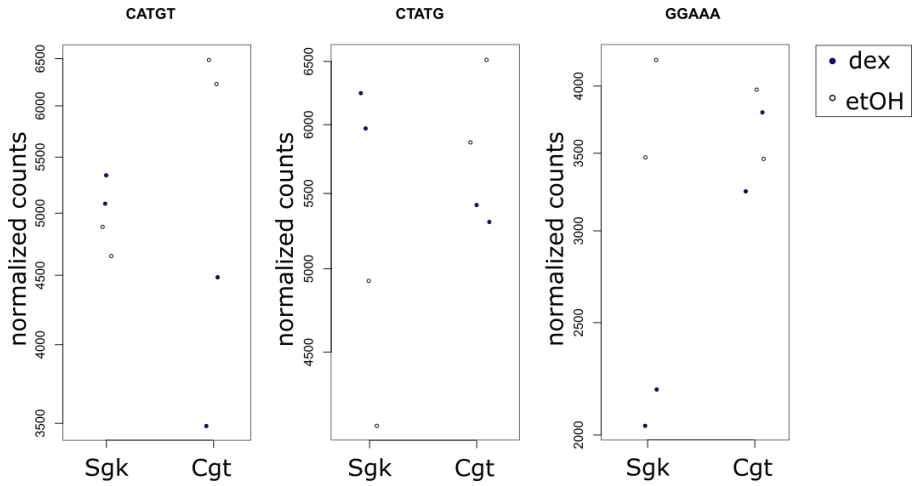


Figure 4.14.: Three flanks were significantly differently regulated between Cgt and Sgk. CATGT showed blunting effects, when placed next to Cgt. GGAAA showed blunting effects, when placed next to Sgk. CTATG showed dual effects: enhancing activity for Sgk and slight blunting effect for Cgt.

No Activation of STARR-seq Vector in A549 cells

As a next step, I wanted to test if there exists a cell-type specific activity of GBS between different cell lines. Glucocorticoid signaling has tissue-specific effects and leads to differential activation of genes that vary between tissues. The GR binding sequence might be one component in determining tissue-specific effects. Therefore, I repeated the STARR-seq experiment with the GBS-HS library in A549, a lung cell line, and I wanted to compare it with U2OS cells, derived from bone. Unfortunately, there was little to no induction after 15 h hormone treatment in both biological replicates in A549 and no differentially expressed sequences could be extracted. Despite the fact that the transfection in A549 worked well (>90% transfected A549 cells, Amaxa kit T) and that I found a large diversity of identified sequence variants (64037) after transfection, supported this assumption (Section 4.1.3). Also the amount of GR protein in A549 is similar to GR18 cells (Meijsing Lab, unpublished data). All this might indicate that the promoter-reporter gene complex of the STARR-seq screening vectors was not active in A549, which might be A549 specific. Therefore, A549 cells could not be used for my STARR-seq experiment at these conditions.

4.1.6. In Depth Analysis of STARR-seq Data

After demonstrating that the STARR-seq experiment with synthetic GBS library worked successfully in GR18 cells, I turned to conduct an in depth analysis of STARR-seq data, by focusing on the role of the spacer and flanking regions in modulating GR activity.

A Study of GBS Spacer

The STARR-seq approach, I developed, allowed us to study a large variation of spacer sequences together with GR halfsite sequences and to associate these with GR activity. For figure 4.15a I selected two GBSs, one forming a perfect palindromic sequence TGTTCT and one forming an imperfect palindromic sequence TGTACG and plotted the activity of all 16 spacer variants of these GBSs in the STARR-seq experiment (GBS-HS library, 4h). In general, the imperfect GBS showed on average higher activation rates than the perfect GBS. Strikingly, for both GBS the "green" spacer group showed only weak activation (mean($\log_2$foldchange)=1.5 and 1.4), whereas the "red" spacer group showed strong activation (mean($\log_2$foldchange)=2.7 and 2.2). The green color plate corresponds to the spacer sequences with an A at position 1 and the red color plate corresponds to the spacer sequences with a C at position 1. The other two groups, yellow (G1) and blue (T1), show intermediate activation. To test if in general C1 leads to higher activity than A1 independent of GBS-halfsite composition, I extracted information from the STARR-seq experiment (GBS-HS library, 4h) of all sequences, resembling a GBS (NNNGTNCN) and plotted the activity of GBS with A1 and C1 spacer (Figure 4.15b). Indeed, C1 spacer led to significantly stronger activation than A1 spacer (p-value$<$0.008, Wilcoxon-rank-sum-test). Research by Watson et al. revealed that GR prefers pyrimidine (T and C) at spacer positions and demonstrated that by changing a spacer of Fkbp5 from GGG to AAA reduced GR activation rates in luciferase assays [50]. This is consistent with my observation that spacer composition affects GR activity. Moreover, they showed that changes in spacer composition affect the D-loop structure of the dimerisation interface, although the spacer is not directly contacted by the GR protein. In summary, I showed that a single spacer position can significantly affect gene expression and that my finding is in line with previous research. Yet, with the STARR-seq method I am able to test and compare sequence features on a much larger scale compared to previous research.

Luciferase Assay Validates Flanks Activity

To test if the results are also transferable to other experiments, I compared STARR-seq activity of selected flanks with their activity in luciferase reporter assays (Figure 4.16). So far, STARR-seq experiments were validated by repeating STARR-seq experiments with single STARR-seq variants by qPCR. The sequence environment of the GBS in the two assays differ. In STARR-seq, the GR binding sequences is separated from the promoter and acts as an enhancer. In the luciferase reporter vectors is the GBS part of the promoter. The TCTCG flank was differentially regulated in the STARR-seq Cgt flank experiment (foldchange=1.26, adj. p-value$<$0.0006), whereas slight variations of the sequence, (CCTCG, CATCG,TATCG), showed no modulation of activity. A similar result for these flanks of Cgt was obtained in luciferase reporter assay, again only the TCTCG flanks showed strong activation of 4.3 and CCTCG was slightly activated with 1.35 fold(dex/etOH). CATCG and TATCG were not active in

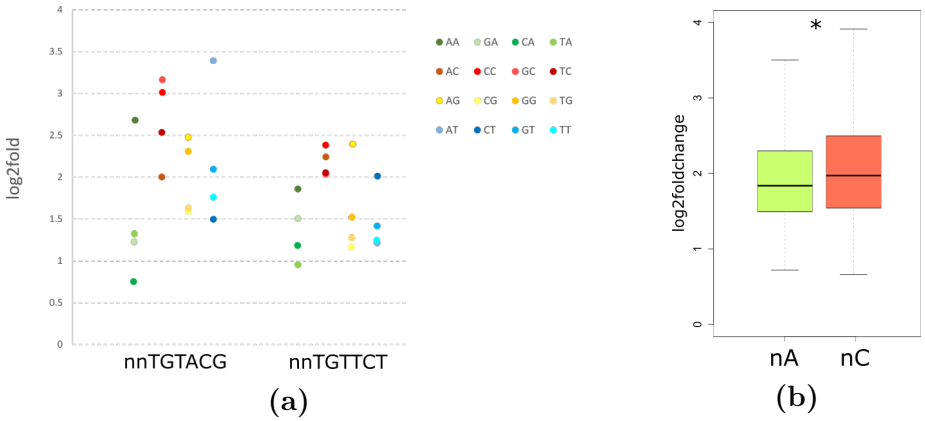


Figure 4.15.: Nucleotide composition in spacer correlates with gene activity. (a) STARR-seq activity of 16 spacer variants of a palindromic (TGTTCT) and a non-palindromic (TGTACG) GBS halfsite in GR18 after 4 h hormone treatment were compared. C at position 1 in spacer showed on average higher activity than the other 3 nucleotide. A at position 1 appeared to lead to weaker activation, which is recapitulated in panel (b). Comparing all differential up-regulated GBS with A or C at spacer position 1. C1 spacer led to significant higher activation rates than A1 (Wilcoxon-rank-sum-test, p-value < 0.008).

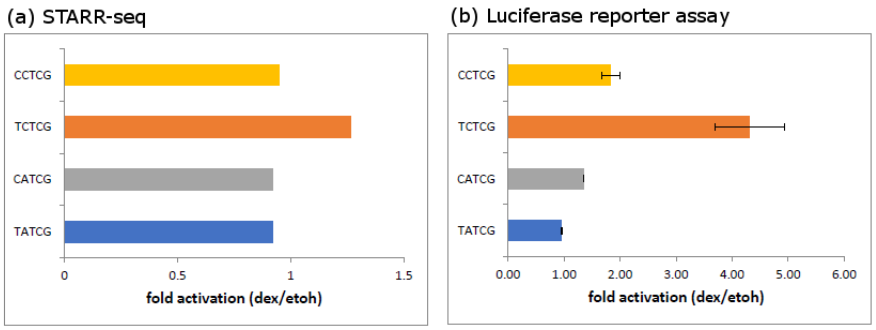


Figure 4.16.: Transient luciferase reporter assay validated activity of flanks in STARR-seq. (a) STARR-seq activity for tested flanks was in agreement with (b) luciferase activity. Average fold induction upon 1 μM dexamethasone (dex) treatment relative to ethanol (etoh) vehicle ±SD (n=2) is shown.

both experiments. In this way, the modulation of activity of the flanks from the STARR-seq experiment could be recapitulated in the luciferase reporter assay and shows that GR action might be similar in promoters and enhancers despite their structural differences.

4.2. Flanking Sites of GBS Modulate GR's Activity and DNA Shape

So far, I could show that in transient reporter assays, GR shows sequence dependent differences in gene activation. However, until now I did not show that this may also happen in an endogenous context, and what could be causing these differences in activity. Hence, I set out to search for activity-associated GBS variants in the genomic context. I found that the direct flanking sites of a GBS correlate with target gene activity and that structural changes of DNA and GR affect target gene expression downstream of GR binding. The following section will be part of a publication (Schöne et al., submitted) and was done in collaboration with the following researchers: Dr. Marcel Jurk, the group of Dr. Bruno Kieffer (IGBMC, Illkirch dedex), Petra Imhof (FU Berlin) and Remo Rohs (USC, Los Angeles).

4.2.1. GBS Variants Correlate with GR Activity in a Genomic Context

To assess whether GBS variants may indeed play a role in fine-tuning the activity of GR towards individual endogenous target genes, I analyzed genomic data to see if the level of GR activity correlates with the presence of specific GBS variants near these regulated genes. Therefore, I first grouped genes regulated by GR in U2OS cells into strong responders (top 20% with greatest fold induction upon dexamethasone treatment, 290 genes) and a control group of weak responders ($\log_2$(absolute foldchange)$\leq$|0.72|) (Figure 4.17). Next, I associated GR-bound regions (ChIP-seq data) to regulated genes when a ChIP-seq peak was located within a window of 40 kb centered on the transcriptional start site (TSS) of the gene. In this manner, I analyzed the 20% strongest GR responsive genes and associated 543 peaks with these genes. For comparison, a control group with similar peak number was generated consisting of 532 peaks associated with weak GR responsive genes. For each group of peaks, I conducted a de-novo motif search with *peak-motifs* [72]. For both groups, I identified the GR motif and motifs for AP1 and SP1, which are known cofactors of nuclear receptors [77, 78]. The core GR motif was similar for both groups (Figure 4.17) and closely matches the GR consensus sequence (Figure 2.3a). However, there are some subtle differences in the preferred nucleotide at individual positions. For instance, the spacer for GBSs associated with weak responders preferentially contains a G or C at position -1, whereas no such preference is observed for GBSs associated with strong responders. This is consistent with previous studies showing that the sequence of the spacer can modulate GR activity [1, 50]. Furthermore, I found that the nucleotide flanking each half-site (position -8 and +8)

exhibited high information content, with sequence preferences that were different for peaks associated with strong and weak responder genes (Figure 4.17). For GBSs associated with strong GR responsive genes the flanking nucleotide was preferentially A or T, whereas for the GBS associated with weak GR responsive genes the flanking nucleotide was preferentially G or C. Because the motifs uncovered by the de-novo motif search are not necessarily present at different frequencies for the two groups, I quantitatively compared the occurrence of A/T and G/C flanked motif matches between the "strong" and "weak"-associated peaks. Consistent with the outcome of the de-novo motif search, this analysis showed more matches for the A/T flanked motif for strong-responder-associated peaks than for weak-responder-associated peaks, whereas the opposite was found when I scanned with the G/C flanked motif (Figure 4.18). Together, this suggests that GBS variants may indeed play a role in modulating GR activity towards endogenous target genes, and hints at a possible role for the bases directly flanking the half-sites in this process.

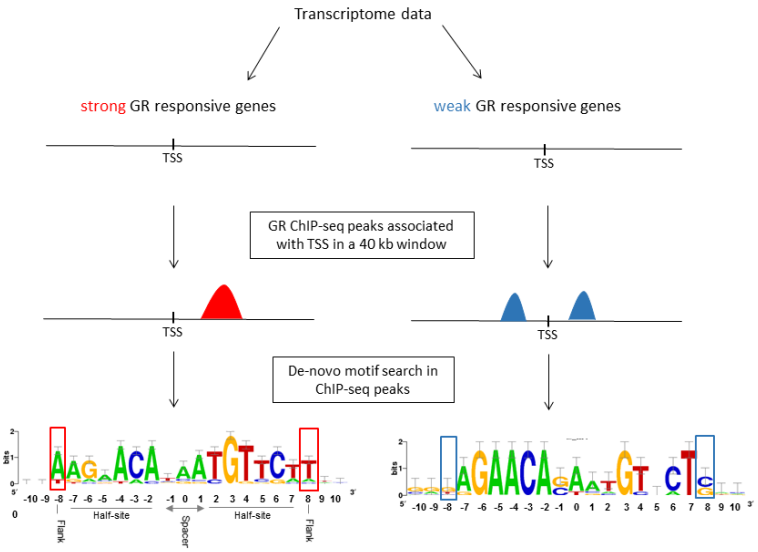


Figure 4.17.: Analysis of GR transcriptome and ChIP-seq data revealed association of GBS direct flanking sites with gene regulation. For this analysis, GR regulated genes were binned into strong and weak responsive genes and GR ChIP-seq peaks were associated with these genes by distance association. De-novo motif analysis led to identification of flanking sites (red and blue rectangle) as factor for GR gene regulation.

4.2.2. GBS Flanking Sites Modulate GR Activity

To test the role of bases flanking the half-site (position -8 and +8) in modulating GR activity, I generated transient luciferase reporters for five GBS variants (Cgt, FKBP5-1, FKBP5-2, Pal and Sgk) where I flanked each by either A/T or by G/C (Figure 4.19a). These reporters displayed comparable basal activities (not shown),

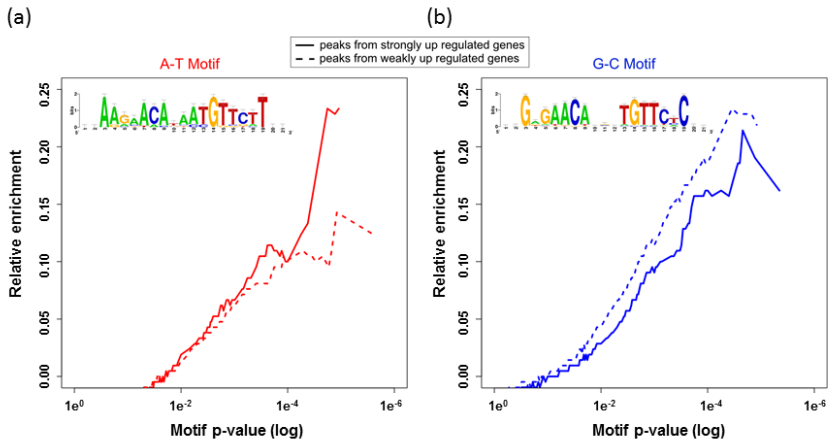


Figure 4.18.: Comparison of motif p-value distribution confirms motif association with regulation group. (a) The A-T motif was found more frequently in the strong responder group (solid line) than in the weak responder group (dashed line). (b) G-C motif was found slightly more frequently in the weak responder group than in the strong responder group.

whereas the level of induction upon hormone treatment varied between the sequence variants (Figure 4.19a). Consistent with the observations for endogenous GR target genes, the A/T flanked GBSs showed higher reporter gene activity than the G/C flanked GBSs for four out of five tested GBS variants, whereas little to no effect of changing the flanks was observed for the Pal sequence (Figure 4.19a). For example, the activity of A/T flanked Cgt was twice that of the G/C flanked version of this GBS. Together, these experiments indicate that the flanking sites can indeed modulate GR activity and from now on, I use the term 'flanking site effect' to refer to the dependency of GR target gene expression on GBS flanking site sequence. Notably, the Sgk and Cgt GBSs showed the greatest flanking site effect whereas the effect for the Pal and FKBP5-1 GBSs was small. When comparing the sequences of these GBS variants, I observed that the second half site (position 2-7) forms an "imperfect" palindromic sequence (not matching TGTTCT) for the GBSs with the greatest flanking site effect (Cgt and Sgk) whereas this sequence is palindromic for Pal and FKBP5-1. To test whether the "imperfect" half site of Cgt and Sgk is responsible for the flanking site effect, I generated new luciferase reporter constructs with mixed flanking sites (A/C and G/T) (Figure 4.19b). These experiments showed that the imperfect half site is indeed mainly responsible for the flanking site effect. On average there is a 98% increase in activity when I changed the flank at the imperfect site, whereas this increase is lower (18%) when I changed the flank at the "perfect" half site.

To find out what causes the flanking site effect, I focused on the Cgt and Sgk GBS in the next experiments, because they showed the strongest influence of the flanking sites. To study the role of flanking sites in the chromosomal context, I stably integrated a Sgk-GBS luciferase reporter in U2OS cells at the *AAVS1*

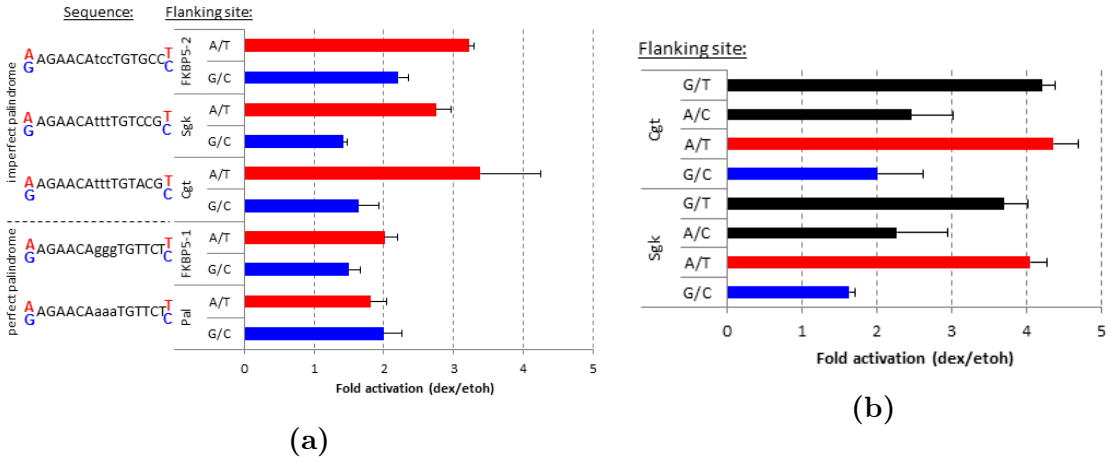
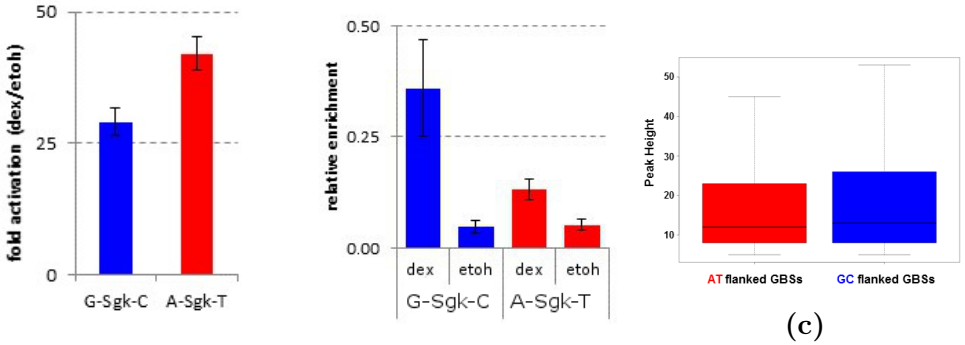


Figure 4.19.: A/T flanking sites correlate with stronger reporter gene activation in transient luciferase reporter assay. (a) For four out of five tested GBS (Cgt, Fkbp5-1, Fkbp5-2, Pal and Sgk) show A/T flanking sites stronger reporter gene activation than G/C flanking sites. (b) Mixed flanking sites reveal that flanking site effect originates from imperfect halfsite in Cgt and Sgk. Average fold induction upon 1 μM dexamethasone (dex) treatment relative to ethanol (etoh) vehicle ±S.E.M. (n≥3) is shown.

locus to simulate an endogenous gene environment. Matching what I observed with the transiently transfected reporters, I again found that the integrated A/T-flanked Sgk showed a 1.5 times greater reporter activity than the G/C flanked GBS (Figure 4.20a). At this point, I started to ask in which way the flanking sites influence GR's activity. To determine whether the flanking site effect might be caused by a change in the intrinsic affinity of the DNA binding domain (DBD) for GBSs, I conducted electrophoretic mobility shift assays (EMSAs). However, arguing against a role for changes in the intrinsic affinity, I found similar K_D values for both A/T and G/C flanked Cgt and Sgk GBSs (Table 4.3). In a second approach, I also studied GR binding *in vivo* to A/T flanked and G/C flanked Sgk versions of the stably integrated reporter constructs from the previous experiment by ChIP experiments (Figure 4.20a). Remarkably, the GR occupancy of G/C flanked Sgk was twice that of the A/T flanked Sgk (Figure 4.20b), despite the fact that A/T flanked Sgk leads to higher gene activation. Similarly, GR binding was essentially the same when comparing the peak height of all GR ChIP-seq peaks in U2OS containing an A/T flanked GBS with those flanked by G/C (Figure 4.20c), showing that peak height and flanking sites are seemingly independent. Together, I therefore conclude that the flanking site effect appears not to be a consequence of changes in DNA binding affinity or occupancy *in vivo*.

Table 4.3.: *In vitro* binding affinity of GR for GBS-flank variants

GBS	Flank	K_D [μM]	SD
Sgk	G/C	0.55	0.07
	A/T	0.70	0.3
Cgt	G/C	0.98	0.09
	A/T	1.09	0.12



(a) (b) (c)

Figure 4.20.: Effect of flanking site in a chromosomal context. (a) Stably integrated A/T flanked Sgk-GBS lead to higher luciferase reporter gene activation than integrated G/C flanked reporter. (b) GR occupancy levels for integrated Sgk-GBS reporters with flanks as indicated were analyzed by chromatin immunoprecipitation followed by qPCR for cells treated with either 1 μM dexamethasone (dex) or ethanol (etoh). Average relative enrichment for 3 clonal lines from 3 biological replicates, ±SD, are shown. (c) Peak height of all ChIP-seq peaks in U2OS containing A/T or G/C flanked GBS show no significant difference in peak height between A/T and G/C flanked GBS (p-val=0.2208, Wilcoxon-rank-sum-test).

4.2.3. Flanking Sites Modulate DNA Structure

A previous study has shown that the sequence of the spacer influences DNA shape and GR activity [50]. Therefore, I investigated if flanking-site-induced changes in DNA shape might play a role in mediating the flanking site effect. To test whether the local structure of the DNA molecule is affected by the flanking site of the GBS, I collaborated with Iris Dror and Remo Rohs (University of Southern California, Los Angeles). We compared predicted DNA shape features between G/C (75 GBSs) and A/T (83 GBSs) flanked GBSs from peaks associated with strong and weak responder genes (as used in figure 4.17), respectively (Figure 4.21). Their prediction for the minor groove width showed only a slight difference between GBS flanked by G/C and A/T at positions -8 and +8 (matching the flanking sites). More strikingly, at position -7, +7, -6 and +6 the predicted minor groove width in A/T flanked GBS is not only narrower than the rest of the GBS but also narrower than the

corresponding position in G/C flanked GBS (Figure 4.22). Importantly, the overall nucleotide composition of the GBS and its surrounding was comparable for the two groups of sequences (Figure 4.21), indicating that the effect on the two neighboring nucleotides is a consequence of changing the sequence of the flanking site.

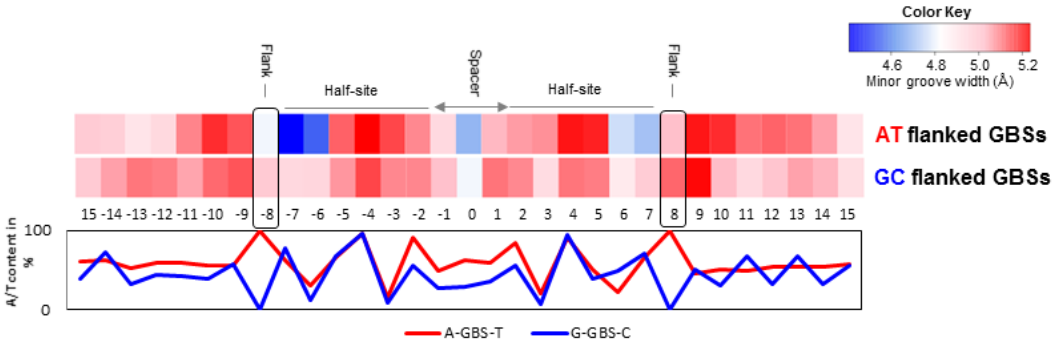


Figure 4.21.: Flanking sites affect DNA structure. At the top, the mean minor groove width for single nucleotide position was predicted for a group of 83 A/T flanked GBSs associated with strong GR responders and for a group of 75 G/C flanked GBSs associated with weak GR responders. At the bottom the AT content (in %) was plotted for each group.

Next, we repeated the DNA shape prediction for individual GBSs, tested in the luciferase reporter constructs (Figure 4.23). Since the first half site (-7 to -2) is identical in all tested GBS, it is not surprising that all GBS have the same minor groove width at these positions. Here, I focused on the flanking site and second half site. For both Cgt and Sgk GBS, the minor groove width at the flanking position 8 is slightly narrower in the G/C flanked version than in the A/T flanked version. In contrast, the neighboring positions 6 and 7 exhibit a narrower minor groove width in A/T flanked versions. This result suggests that the crucial structural DNA change occurs at position 6, 7 and 8. For the G/C flanked Pal and Fkbp5-1 GBS variants (which do not exhibit a flanking site effect) the minor groove width is already quite narrow at these positions and the A/T flanked GBSs show a further narrowing at position 7. Whereas in contrast to what we observed for Sgk and Cgt, the minor groove is wider at position 6 perhaps explaining why Pal and Fkbp5-1 do not exhibit a flanking site effect. All in all, the structural effects might be subtle, but the effect on gene activation, I observed, was strong.

4.2.4. NMR and MD Simulations Revealed that GBS Flanking Site Influence GR Structure

Overall, the predicted changes in DNA structure induced by the flanking site suggest that DNA shape may serve as an input that regulates GR activity. To determine if the flanking site influences GR structure, my colleague Dr. Marcel Jurk probed the DBD of GR in complex with flank-site Cgt variants by 2D (1H, 15N) heteronu-

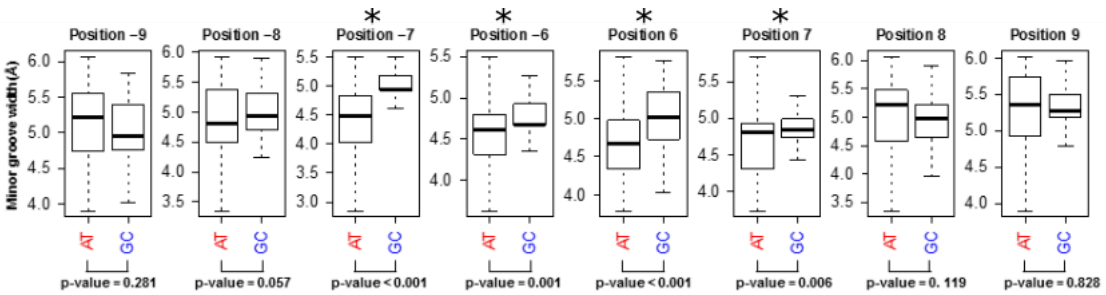


Figure 4.22.: Boxplot of minor groove width for individual positions comparing A/T and G/C flanked GBS. Boxplots are shown for predicted MGW only for the positions around the flanking site. Position -7,-6, 6, and 7 show significant difference between AT flanked GBS and GC flanked GBS groups (Wilcoxon-rank-sum-test).

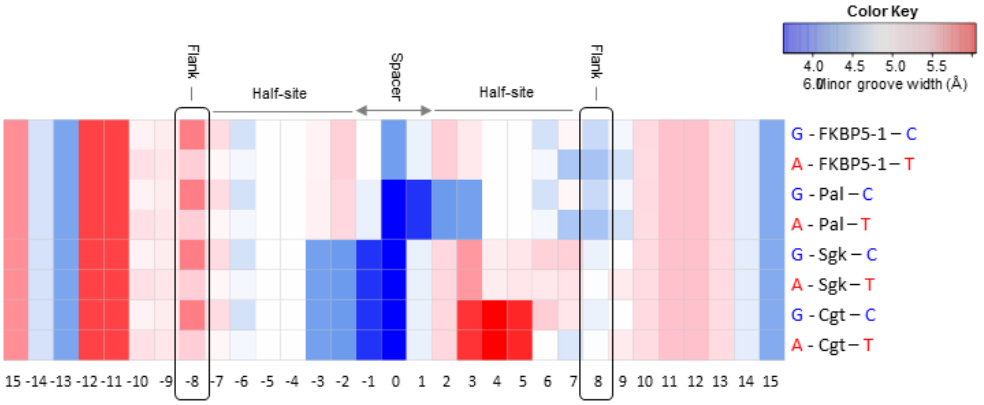


Figure 4.23.: Minor groove width prediction for individual GBS variants tested in luciferase assay. Minor groove width analysis reveals small changes in DNA structure around flanking sites for individual GBS at position 6, 7, 8 and 9. Indicating structural differences of DNA between perfect and imperfect GBS and between G/C and A/T flanked GBS.

clear single quantum coherence spectroscopy (HSQC) (Figure 4.24). In this way, the spectrum of a recorded protein forms a fingerprint-like picture, where peaks can be assigned to amino acid residues. In (1H, 15N) HSQC, the chemical environment of the N-H group of the protein backbone is measured and a shift in the spectra corresponds to a change in the chemical environment of the corresponding residue, usually induced by a structural change. When we compared the spectra between G/C and A/T flanked Cgt, we found a striking number of residues with significant chemical shift perturbations (Figure 4.25a). Next, we selectively changed the flanking site at either the "perfect" half site (Chain A) or at the "imperfect" half site (Chain B), which is mainly responsible for the flanking site effect. These experiments showed that changing the flanking site on the imperfect half-site (AT/AC; Figure 4.25b), resulted in significant chemical shift perturbations for several residues (Ser448, Thr456, Arg488, Tyr497, Asn506, Lys511). Similarly, changing the flank for the perfect half site (GC/AC, Figure 4.25a) induced peaks shifts for multiple residues. Interestingly however, the residues affected overlapped for some residues (T456 and Y497), whereas they were flank-specific for others.

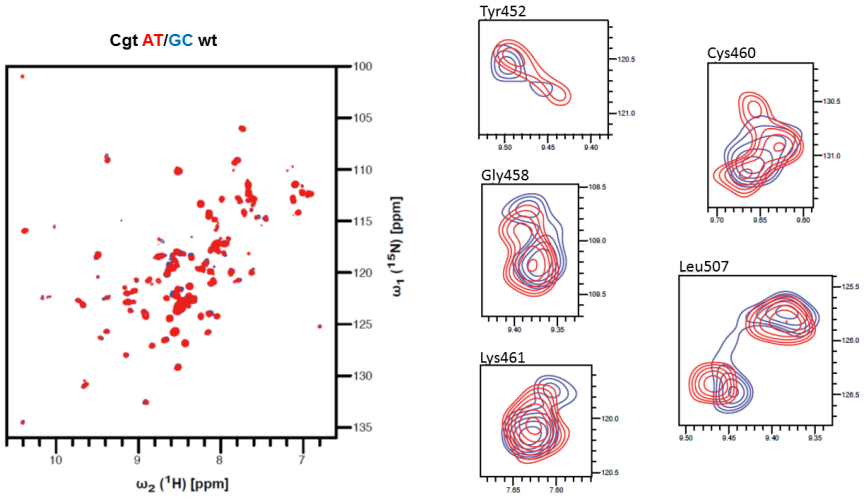


Figure 4.24.: 2D (1H,15N)-HSQC of GR bound to Cgt-AT vs. Cgt-GC. Left panel shows recorded NMR 2D spectra for Cgt-AT (red) vs. GC (blue). Right panel shows zoom-in of 5 individual residues in the vicinity of the helix 1 in GR exhibiting peak-splitting and shifting.

Several residues that map to the DNA-recognition helix 1 (Gly458, Cys460 and Lys461) showed different peak characteristics. Specifically, we observed shifting and splitting of peaks (Figure 4.24, right panel). Splitted peak patterns are characteristic of either conformational exchange within each monomer or different chemical environments (i.e. conformations) of the individual monomers. Helix 1 sits in the major groove opposite to the positions (-6, -7 / +6, +7) where the flank-site induces a narrowing of the minor groove in A/T flanked GBS. Consequently, the DBD of GR might contact DNA differently for example by contacting other nucleotide positions

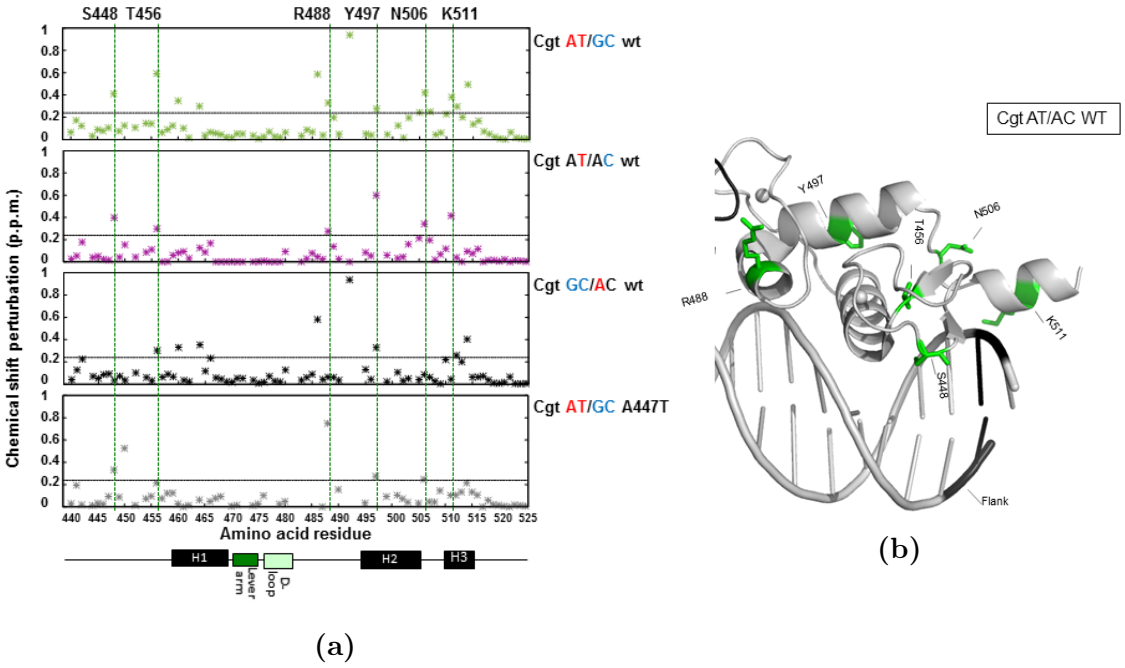


Figure 4.25.: NMR experiments reveal global structural changes between GC and AT flanks. (a) Chemical shift perturbation (CSP) was calculated for each recorded amino acid residue between GC and AT flanked Cgt. Black line marks significance cut-off (Average +1 standard deviation). Green dashed lines mark amino acid residues with significant shifts when comparing the A/T and A/C sequences. (b) Overlay of crystal structure with shifted amino acid residue (green) from AC/AT flank.

when the sequence of the flank is changed. To test this, our collaborator Isabelle Lebars (group of Bruno Kieffer, IGBMC, Illkirch dedex) analyzed the protein:DNA complex again by NMR spectroscopy but this time not by observing the resonances of the protein but those of the DNA itself. She assigned the imino protons in the 1D spectra for Cgt flanked by either A/T or G/C and increasing amounts of protein were titrated to both DNAs to determine if the flank influences DNA positions that are contacted (Figure 4.26). Consistent with the crystal structure for GR, these experiments indicate that the DBD contacts the half-sites of the GBS, for example at positions -6 (G6), -4 (T41), -3 (G40) or +2 (T14), +4 (T16). Furthermore, we find evidence for contacts outside the 15 bp consensus sequence (G46, position -9) again in agreement with contacts by helix 3 in the crystal structure. However, the base G46 next to the flanking site was contacted differently between flanking variants. When we compare the bases contacted between A/T and G/C flanked DNAs, however, most bases show evidence of being similarly contacted by the DBD of GR indicating that the flanking site does not influence which nucleotides are contacted by the DBD of GR. This data is in good agreement with the K_D, derived from EMSAs, and indicates that binding of GR to DNA is not changed upon changing the sequence of the flanks.

The NMR experiments indicate that the flanking site induces several conformational changes in the DBD of GR. One limitation of the NMR experiments is that we do not know if the observed conformational changes occur on chain A, chain B or both monomers of GR. Therefore, we turned to molecular dynamics (MD) simulations, conducted by Mahdi Bagherpoor Helabad (group of Petra Imhof, FU Berlin), to simulate how changing the flanks influences the dimer and the individual monomers. When we compared the overall trajectories of the molecule, however, we did not observe a significant structural difference for either chain A or chain B, comparing the RMSD (root-mean-square deviation) values for individual residues between the A/T and the G/C-flanked Cgt GBS. Similarly, we only observed subtle changes when we compared the RMSF (root-mean-square fluctuation) (Figure 4.27a), a measure of flexibility of the DBD, between the two Cgt flank variants. The changes that do occur however, predominantly map to residues at the dimerization interface (D-loop) of both chain A and chain B. In addition, the RMSF values for chain B when bound to the G/C flanked GBS showed higher values indicating that chain B's interaction with the DNA for this sequence is more dynamic (Figure 4.27a). Finally, we compared the last 50 ns of median GR-DBD structures when bound to A/T- or G/C-flanked Cgt and again found little deviation between these structures except for the lever arm, which connects the dimerization interface with the DNA recognition helix. Interestingly however, changing the flanks appears to result in a change of the relative positioning of the dimer-halves when we aligned the median structures for both flank-variants on chain A (Figure 4.27b).

Together, the structural approaches indicate that flanking site induces several changes in the DBD of GR, which include conformational changes, changes in flexibility and a relative repositioning of the two dimer-halves. These structural changes may bring GR bound at A/T flanks into a more favorable position to interact with

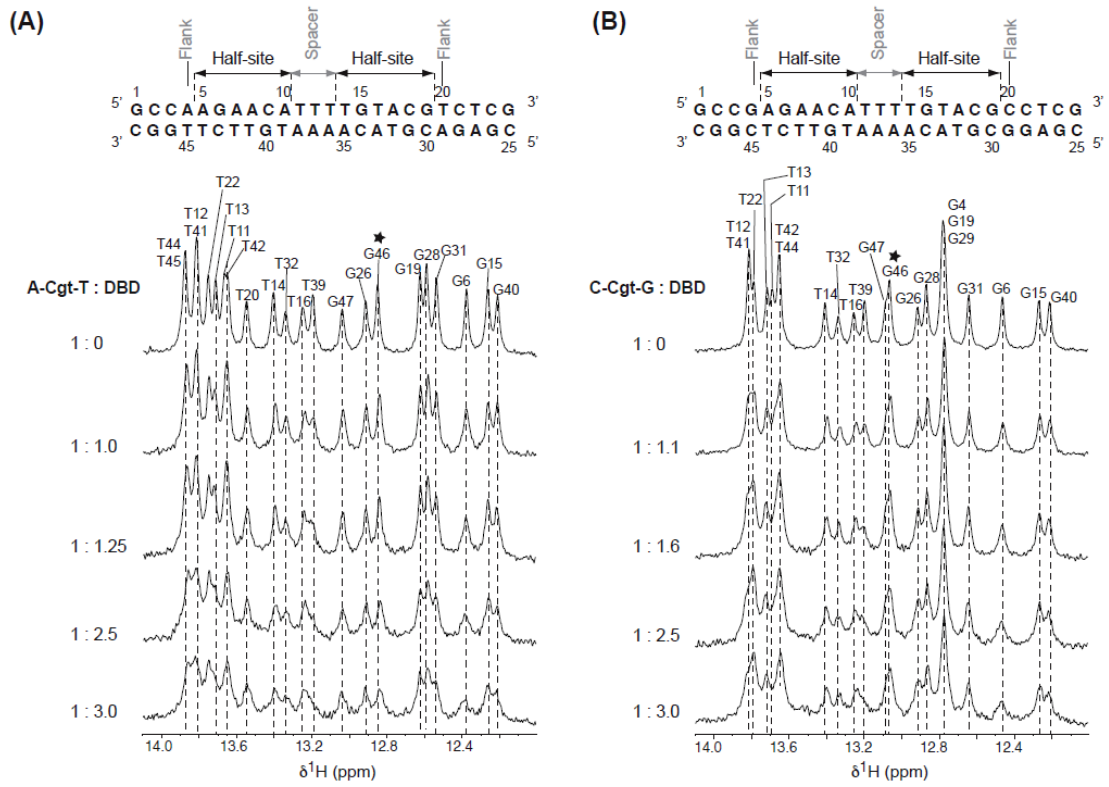


Figure 4.26.: (1H)-NMR of DNA:GR-DBD titration experiments reveal DNA-Protein contacts. (A) Titration of GR DBD with A/T-flanked Cgt GBS and (B) with G/C-flanked Cgt GBS reveals contacted DNA bases. The secondary structure of DNAs are indicated on top. The imino proton regions of DNAs are shown upon increasing amount of protein. The DNA:DBD ratio is indicated on the left. The star indicates the residue that exhibits the most important broadening, which is different between A/T-flanked Cgt and G/C-flanked Cgt.

coregulators compared to G/C flanks and could explain the differences in GR activity.

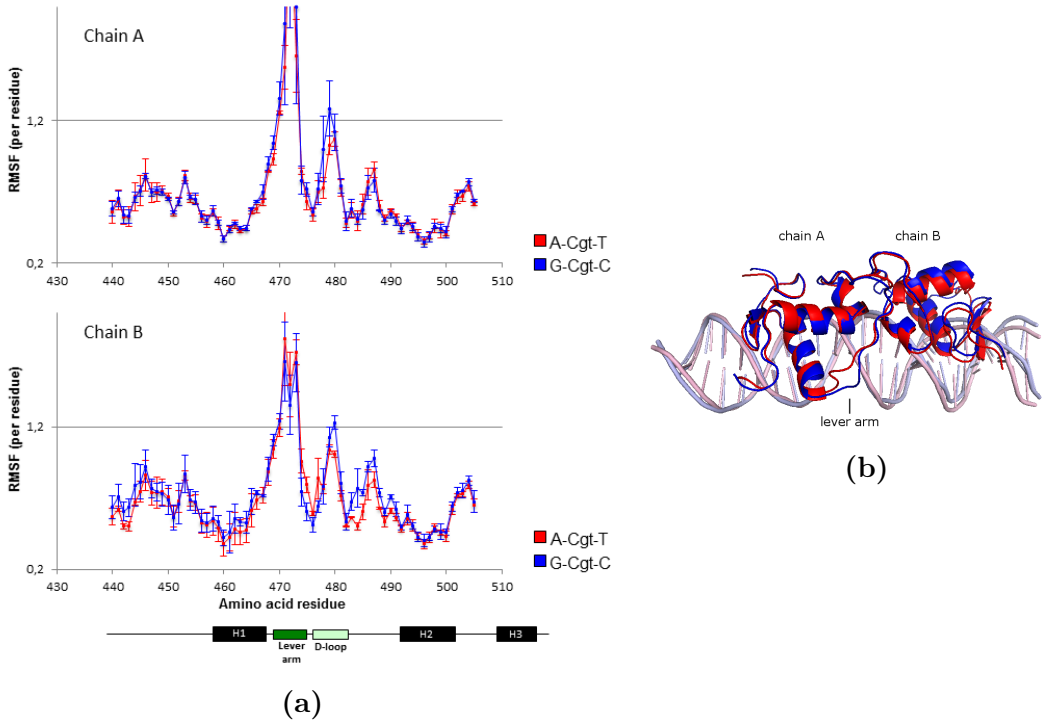


Figure 4.27.: Molecular dynamics simulations reveal repositioning of dimer-halves. (a) RMSF of backbone atoms for each amino acid as indicated for chain A (top) and chain B (bottom) of GR over the first 100 ns MD simulation for GR bound to A/T flanked Cgt (red) and GR bound to G/C flanked Cgt (blue). Shown are averages and standard deviation from 3 different MD runs. (b) MD simulations over last the 50 ns aligned on chain A reveal relative repositioning of GR dimer-halves. The lever arm is different in chain A and B for GC (blue) and AT (red) flanked Cgt.

4.2.5. Intact Dimer Interface is Required for the Flanking Sites Effect

To investigate how the DBD of GR might "read" the shape of DNA to modulate GR activity, I tested the role of several candidate amino acid residues of the DBD that contact the DNA. As candidates I chose R510, which is part of helix 3 and contacts the flanking base directly according to the crystal structure. Similarly, K511 might contact the flanking base and it shows a significant chemical shift in our NMR experiments upon changing the flanking site sequence (Figure 4.25a). In addition, I tested K461 and K465, which reside in the DNA recognition helix 1. Based on the crystal structure, K461 makes a base-specific contact with the G at position -6/+6 in the major groove opposite to the position where the flank induces a change in minor groove width, whereas K465 contacts the phosphate backbone.

When we mutated R510, K511 or K465 from an arginine to an alanine (R510A, K511A or K465A), the flanking site effect was still observed in transient luciferase reporter assays arguing against a role of these residues in "reading" the DNA to modulate GR activity (Figure 4.28a). Mutating K461 to an alanine resulted in a marked decrease in GR-dependent activation for the A/T flanked GBS and only a small loss of activation for the G/C flanked GBS, consistent with previous studies of this mutation [79]. Interestingly however, there was still some residual activity for the G/C flanked GBS, whereas no activation was seen for the A/T flanked variant, which is more active for wildtype (Figure 4.28a). This suggests that the K461 residue might play a role in interpreting the flank-site encoded instructions and is in line with previous studies [79] that uncovered a role of this residue in interpreting the signaling information provided by response elements at which GR binds.

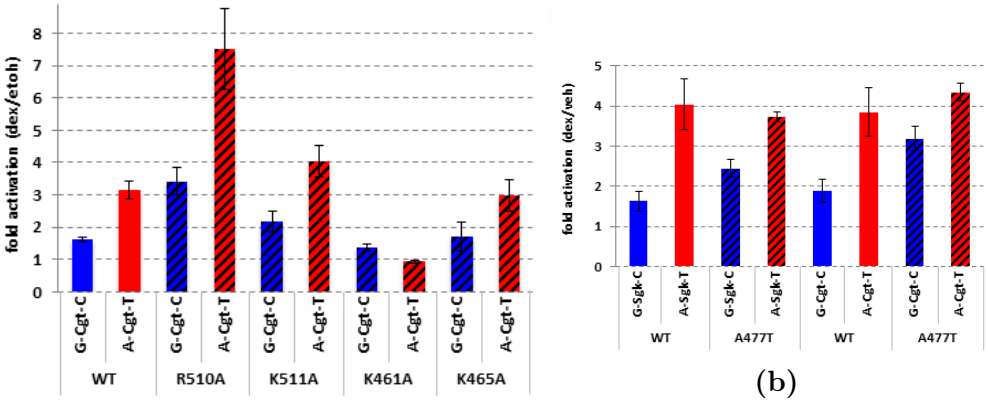


Figure 4.28.: Global GR structure involved in flanking site effect not a direct "reader". Transient luciferase reporter assays were used to find (a) "reader" of the flanking sites by mutating K461, K465, R510 and K511 in GR to alanine. (b) Dimerisation mutant A477T shows that a intact dimerisation interface is needed for the flanking site effect. Average induction upon 1 μM dexamethasone (dex) treatment relative to ethanol (etoh) vehicle $\pm$ S.E.M. (n$\geq$3) is shown.

Previous studies have shown that an intact dimer interface is required to read DNA shape and to direct sequence-specific GR activity when changing nucleotides of either the spacer or of GR half sites [50]. To test if the dimer interface also plays a role in mediating the flanking site effect, I tested the effect of disrupting the dimerization interface on flank-site-induced modulation of GR activity. As reported previously, mutating Alanine 477 of the dimer interface resulted in GBS-specific effects. For the A/T flanked GBSs Cgt and Sgk, the difference in activity between wild type and the A477T mutant GR was small (Sgk: 8% decrease; Cgt 13% increase, Figure 4.28b). In contrast, for the flanks with the lower activity, G/C, the A477T mutation resulted in a more pronounced increase in activity for both

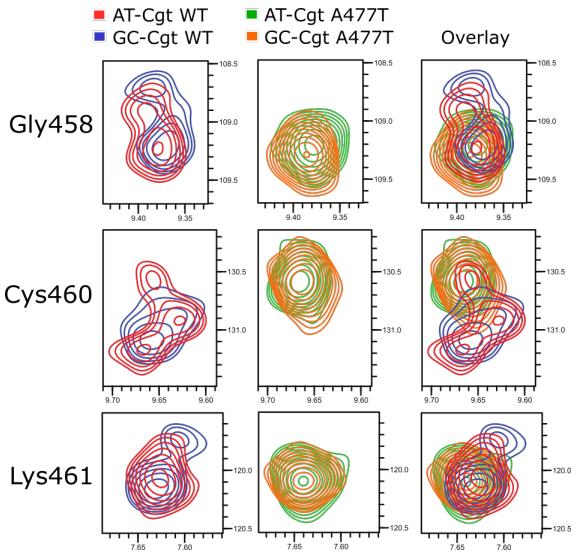


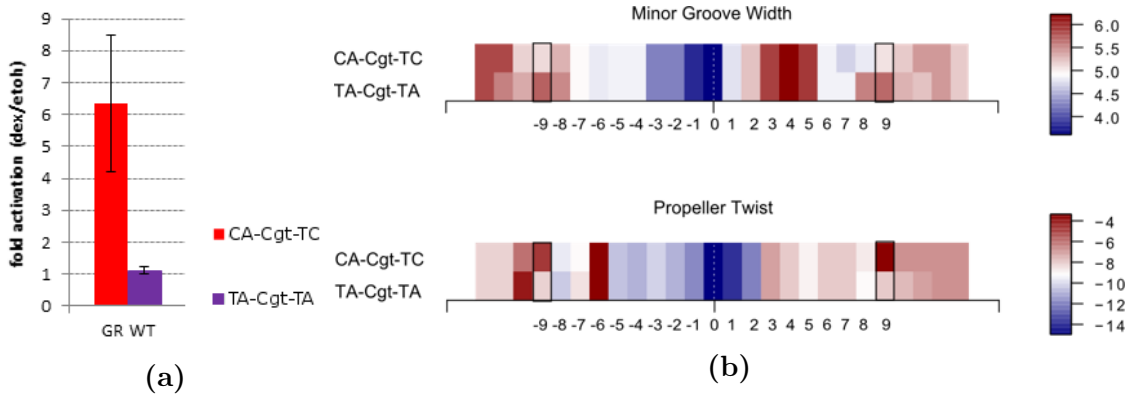
Figure 4.29.: Intact dimerisation interface is needed for alternative conformation in the flanking site effect. 1H,15N-HSQC Zoom-ins on selected peaks show flank-specific peak splitting of residues in wildtype GR and a single peak overlapping in both Cgt flank constructs for GR dimerisation mutant A477T.

GBSs tested (Sgk: 50% increase; Cgt: 69% increase, Figure 4.28b). Consequently, the difference in activity between the A/T and G/C versions of Cgt and Sgk is smaller for the dimer mutant than for wild type GR (Figure 4.28b) indicating that the dimerization domain is involved in transmitting the flanking site effect. Yet to make it clear, the dimerization interface lies on the opposite side of the GR monomer relative to the flanking nucleotide position suggesting that a more global change in GR's structure may be induced by the flanking sites.

To further elucidate the role of the dimer interface in transmitting the flanking site effect, my colleague Dr. Marcel Jurk and me studied the effect on the A477T mutation of flanking-site-induced conformational changes of GR by 1H,15N-HSQC. This analysis uncovered two things. First, several of the residues with significant chemical shift perturbations for wild type (C460, F464, M505, L507, R511, T512, K514) no longer show a significant shift when we compare the G/C and A/T flanked Cgt for the A477T mutant (Figure 4.25a). Second, several peaks that show flank-specific patterns of peak splitting (e.g. C460) for wild type GR, show an overlapping single peak for the mutated A477T DBD (Figure 4.29). This indicates that flanking site can only induce alternative conformations of the DBD when the dimerization interface is intact. Together, the functional and structural analyses of the consequences of disrupting the dimer interface, argue for a prominent role of the dimer interface in facilitating flanking site induced changes in GR conformation and activity.

4.2.6. Second Flanking Site Affects GR Activity

I could show that the structure of the DNA at the GBS is influenced by the base of the first flanking site and that this tunes the activity of GR. The NMR analysis of the DNA (Figure 4.26) showed that GR also contacts the second next flanking base and that this is done differently depending on the sequence of the first flanking base, which lead to differences in GR activity. In consequence, I wanted to test if the second next flanking base in addition to the first flanking base has an effect on GR action. Therefore, I exchanged the C at position -9 and 9 to T and A, respectively. This had an extreme effect on the activity of GR turning a GBS (CA-Cgt-TC) with high activity to a non-active GBS (TA-Cgt-TA, Figure 4.30a). Similar to the first flanking site, there was no apparent difference in intrinsic binding affinity for the two second flanking site constructs, which cannot explain the extreme difference activity. CA-Cgt-TC was bound by GR with an K_D of 0.41 μM (from 4 independent replicates, ±SEM=0.07) and TA-Cgt-TA with an K_D of 0.47 μM (from 4 independent replicates, ±SEM=0.14). Apparently, TA-Cgt-TA is bound by GR *in vitro*, but seems to be inactivated. To explain this, I conducted a structural analysis of the DNA with the *DNAshapeR* tool [55] and plotted predicted minor groove width and predicted propeller twist for each base position (Figure 4.30b). The difference in propeller twist is mostly between position -10 and -7, whereas differences in minor groove width were seen between 7 and 9 at the opposite end. Interestingly, CA-Cgt-TC shows a long stretch (from position 6 to 9) of narrow minor groove which is broken by a wide minor groove in the TA-Cgt-TA at position 8 and 9. At position 8 and 9 the minor groove is directly contacted by GR. Maybe the structural changes at this position lead to this extreme effect on GR activity, but are still speculative. Yet, the example of the second flanking site represents another case, where a highly active core GBS is present and can be bound by GR, but turned inactive through a still unknown mechanism, which might be connected to changes in DNA structure and GR structure.



(a) (b)

Figure 4.30.: Effect of the second flanking sites on GR activity and DNA structure. (a) Comparison of transcriptional activation by the Cgt GBS sequence flanked by either CA/TC or TA/TA for wildtype GR in transient luciferase reporter assay. Average fold induction upon 1 μM dexamethasone (dex) treatment relative to ethanol vehicle (etoh) ±S.E.M. (n≥3) is shown. (b) Top: predicted minor groove width and bottom: predicted propeller twist for individual bases for the Cgt GBS flanked by either CA/TC or TA/TA; prediction based on *DNAshapeR*.

4.3. Recruitment of Coregulators by GR in a Sequence-dependent Manner

4.3.1. DNA pull-down of GR and Coregulators

We have shown that the DNA structure of GBSs affect the structure of the bound GR dimer. But one open question still remains. Does the GR structure affect interaction with coregulators downstream of GR binding? My hypothesis is that GR bound at A/T flanked Cgt might have a more favorable structure for recruitment of coregulators, whereas the structure of GR bound at G/C flanked Cgt is less favorable for recruitment of coregulators. GR can recruit coregulators (e.g. SRC1-3, p300, TBP, PGC1, STAT, TRAPs, SHARP, ...) from a large pool of around 300 coregulators [80] to form complexes and recruit the basal transcriptional machinery, including RNA polymerase PolII (reviewed in [81]). Many coregulators are shared between steroid receptors, since the coregulator interaction domains (AF1 and AF2) are partially conserved in the steroid receptor family. To quantitatively test the recruitment of coregulators to GR in an unbiased manner, I designed and conducted DNA pull-down experiments to capture directly and indirectly DNA bound proteins. Since it is nearly impossible to test all possible coregulators individually, I chose mass-spectrometry (MS) to identify potential GBS-specific GR coregulators and to quantify the recruitment to GR by label-free quantification.

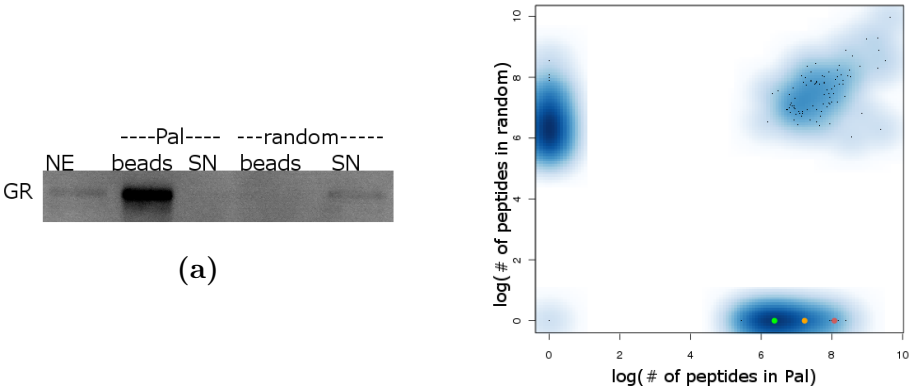
To test if the pull-down assay coupled with MS is able to identify GR coregulators, I started by comparing two bait-DNA sequences. One bait-DNA contained a Pal

GBS (AGAACAaaaTGTTCT) to specifically pull-down GR protein and other bound nuclear proteins. The second bait-DNA, a control bait-DNA, contained a random sequence (AGAAACaaaGTTTCT), retaining a similar nucleotide composition for both bait DNA. A wester-blot analysis of the pulled-down proteins showed a specific pull-down of GR by the Pal-GBS bait whereas little to no binding was observed for the random bait (Figure 4.31a). GR was specifically enriched on the DNA containing Pal and depleted from the nuclear extract. As a next step, the pull-down samples were submitted to MS, which confirmed the specific enrichment of GR on the Pal-GBS bait (Figure 4.31b). No GR peptides were identified on the random bait, whereas they were identified for the Pal bait by MS. This showed that the assay worked as expected for directly DNA bound GR.

Next, I looked at the possible enrichment of coregulators. Multiple proteins could be found to be specifically enriched on Pal-DNA compared to random sequence. In total, around 600 proteins were identified for both DNA baits. The MS data, however, does not reveal any of the known coregulators of GR, like p300 or SRC1. Most identified peptides matched ribonucleoproteins and RNA binding proteins, which are in general highly abundant in affinity purification coupled with MS, and can be considered as background contaminants [82]. However, I identified two other interesting candidates, Sam68 and NCOA5, which are known to interact with other nuclear receptors (Figure 4.31b). NCOA5 was found to interact with two other nuclear receptors, ER and AR [83, 84]. Sam68, was found to interact with NCOA5 in yeast-two-hybrid screens and is a ligand-dependent coactivator of AR [85, 86]. Unfortunately, western blot analysis of DNA pull-down samples for Sam68 showed no specific enrichment as I also detected signal for Sam68 at the random sequence at similar intensity as at that seen for the Pal sequence. However in MS, only peptides matching to the N-terminal part of Sam68 where identified and the antibody I used was targeted the C-terminal part of Sam68, which might be an alternative splice form of Sam68. A repetition with a new antibody raised against the N-terminal part of Sam68 might show specific enrichment of Sam68 at GR, but was not done.

In the end, I may have found two new coregulators (Sam69 and NCOA5) of GR, but I did not find known GR coregulators due to several reasons. First, the interaction between GR and coregulators could be too unstable to pull-down sufficient amounts to be detected by MS. Second, the context of the short bait-DNA may not be sufficient to assemble coregulator complex at GR bound to the DNA. As a next step, I repeated the pull-down experiment by cross-linking bound proteins on the DNA-bait with formaldehyde or BS3, two protein:protein cross-linking agents. Yet, these experiments did not identify known or new coregulators of GR. I could only find Sam68 and NCOA5 again. Additionally, I tested a different bait-DNA with a GilZ-GBS, AT-flanked Sgk and GC-flanked Sgk, but I could not identify known GR coregulators or GBS-specific coregulator recruitment.

Consequently, the DNA pull-down experiment failed to detect GBS-specific recruitment of coregulators to GR. Similar experiments were attempted by other researchers and to my knowledge were also unsuccessful, except for Foulds et al. who have successfully pulled-down coregulator complexes for ER [87]. However,



(a)

(b)

Figure 4.31.: Analysis of GR and other peptides pulled by DNA-baits. (a) GR was detected in nuclear extract (NE) before application to beads. Pal bait beads showed a clear enrichment of GR compared to beads with random bait. Supernatant (SN) after random bait beads incubation showed similar GR levels to NE, whereas GR was depleted in SN after Pal bait beads incubation (b) MS analysis of pulled-down protein peptides revealed an enrichment of GR (red dot), Sam68 (orange dot) and NCOA5 (green dot) on Pal baits compared to random baits.

in contrast to my approach they used a long (<1kb) bait DNA , which contained 4xERE sites together with a E4 promoter, which simulated a more endogenous environment. Adapting their method could be difficult, because I would introduce too many variables (multiple binding sites and promoter), which might obscure the effect introduced by GBS variants.

In summary, the pull-down of DNA bound nuclear proteins efficiently identified direct interaction partners. Therefore, the method could be adapted for the project of my colleague Jonas Telorac for pulling down proteins that bind to specific sequences that prevent GR binding. This assay resulted in the identification of paraspeckle components, presented in a manuscript in which I am co-author [19].

4.3.2. Activity of BATF3 Varies between A/T and G/C Flanked GBS

I chose to directly test coregulator interaction of GR with flanked GBS. Since, the DNA pull-down assay failed to identify and quantify GBS-specific coregulators. I chose the GR coregulator BATF3, a transcription factor that belongs to the AP-1 family, and tested its effect on GR-dependent activation, comparing GBS variants with different flank sites. Increasing amounts of BATF3 expression plasmid were co-transfected together with GR and G/C or A/T flanked Cgt luciferase reporter in U2OS (Figure 4.32). The overall activation of expression upon hormone treatment was very weak, which complicates interpretation of the data. Yet, it can be clearly

seen that GR activity rises continuously with increasing amount of BATF3 for the A/T flanked Cgt reporter, while the G/C flanked Cgt reporter remained inactive until the highest BATF3 concentration (30 ng). A/T flanked Cgt exhibited higher reporter activity in the previous experiments, but it is also seemingly more sensitive to BATF3. This might indicate that the structure of GR at the A/T flanked Cgt is indeed more favorable for interaction with coregulators like BATF3. GR at G/C flanked Cgt might interact with BATF3 only at high concentrations. Notably, this difference in coregulator recruitment might explain why structurally different GR dimers lead to different activation rates of target genes.

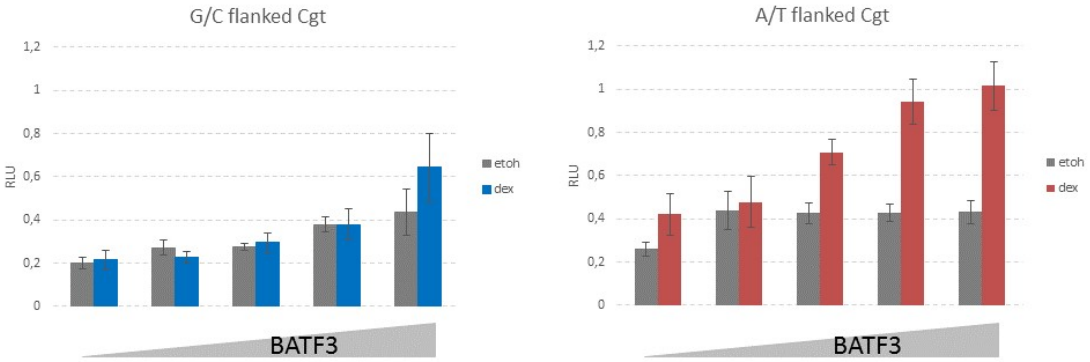


Figure 4.32.: A/T flanked Cgt shows higher activity upon increasing amounts of BATF3. Increasing amounts of BATF3-expressing plasmid were transfected (0 ng - 30 ng) together with GRα expression plasmid and G/C or A/T flanked Cgt luciferase reporter in U2OS. Average relative light units (RLU) upon 1 μM dexamethasone (dex) treatment relative to ethanol vehicle (etoh) ±S.E.M. (n=3) is shown.

4.4. Study of Expression Noise during Gene Activation by GR

As discussed in the introduction, several mechanisms exist to define how much gene product is produced. I could show that GR can modulate the quantity of target gene expression. GR can potentially generate the same average expression in a cell population but with high or low cell-to-cell variability. This variability may be an additional feature of how GR fine-tunes target gene expression. High expression noise results in a cell population with large variations in gene expression level with some cells producing extremely low amounts of gene product whereas other cells produce high amounts of gene product. In contrast, low noise results in a more similar content of gene product for individual cells within a cell population. Here, I set out to analyze the variability of gene levels induced by GR in individual cells and how this may depend on the sequence composition of the GR response

element. Therefore, I analyzed expression noise of single GBSs, of GBS triplicates and of GR composite sites containing binding sites for GR and other transcription factors. I analyzed fluorescent protein content to measure expression noise under the assumption that translation and mRNA and protein degradation rates are similar for the fluorescent proteins GFP and mCherry within the U2OS cell line.

4.4.1. Expression Noise of GR Reporter Constructs

Expression Noise Scales with Mean Expression for Single GBS Variants

To measure expression noise, I tested 5 different single GBS variants (1xCgt, 1xFkbp 5-1, 1xFkbp5-2, 1xGilz, 1xPal) inserted into the STARR-seq screening vector and measured the GFP fluorescence for ≈1000 cells. To normalize and control for extrinsic noise and transfection efficiencies, I co-transfected cells with pSV40::mCherry plasmid. I observed that single GBS variants showed slight differences in noise (Figure 4.33a). As a general rule, I observed that GBSs with a high expression mean exhibited a lower noise level than GBS with a low expression mean. This finding is consistent with previously published data by Sharon et al [31]. It is probably due to the fact that cells with higher expression express more copies of a gene and are more robust to expression deviations. Missing one copy in a hundred has a small effect on the individual GFP signal compared to mean GFP signal (1% reduction). On the contrary, if a cell with lower expression misses one copy from an average of 5 copies within a cell population, the signal deviates already by 20% from the mean GFP signal and leads to higher noise. In summary, GBSs with high expression, exhibit less noise than GBSs with low expression.

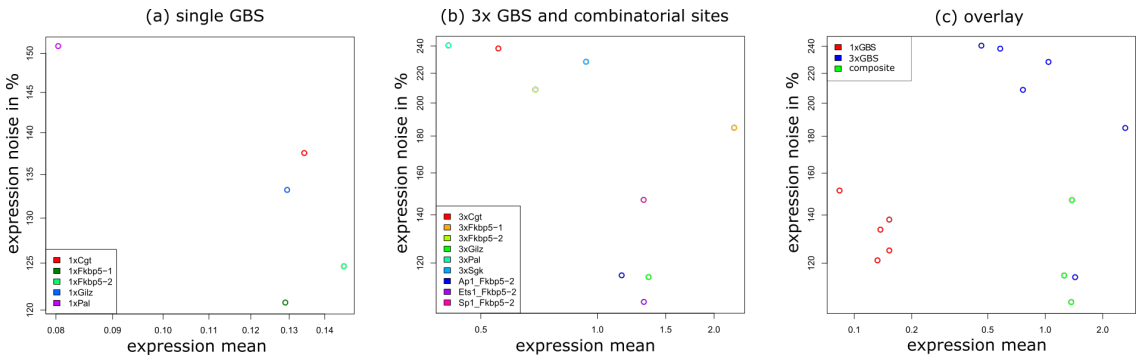


Figure 4.33.: Expression noise depends on the composition of the GR response element.(a) Expression noise and expression mean of single GBS are negatively correlated. (b) Combinatorial site for GR and Ap1, Sp1 and Ets1 exhibit lower noise but similar expression mean compared to 3xGBS. (c) Overlay of all analyzed GBS constructs. Transfected cell were treated overnight with 1 μM dexamethasone before fluorescence analysis. Only GFP+ and mCherry+ cells were analyzed (≈1000 cells per construct) and the geometric coefficient of variation was plotted as expression noise.

Higher Noise for multiple GBS

As a next step, I repeated the experiment with reporters with 3 copies of the same GBS (3xCgt, 3xFkbp5-1, 3xFkbp5-2, 3xGilz, 3xPal, 3xSgk), which were separated by 4 bp each (Figure 4.33b). Again, in general for the 3xGBS reporters, transcriptional noise was negatively correlated with mean expression. Interestingly, the noise was higher for 3xGBS compared to single GBS constructs, yet mean expression was increased compared to single GBS constructs (Figure 4.33c). Again this finding is coherent with findings by Sharon et al. [31], who showed that multiple TFBSs for GCN4 increase expression mean but also expression noise. Sharon et al claimed that 1-dimensional sliding along the genome is a major determinant of TF binding. The authors explain that adjacent TFBSs result in lower binding rates to each site and that the promoter switches slower between transcriptionally active and inactive states, which results in higher noise.

In this way, I could show that in general GR functions as GCN4. But there is one exception: the 3xGilz construct expresses extremely low noise with high mean expression. However, the single Gilz-GBS showed similar characteristics in line with the other single GBSs, confirming that low expression noise is not a general characteristic of Gilz (Figure 4.33a). 3xGilz showed no GFP expression without hormone treatment (2%GFP positive cells, similar to untransfected cells), showing that expression is GR dependent and may not be a consequence of binding of an unknown transcription factor. Why 3xGilZ is so different to the other tested GBS variants is still an open question and needs to be further analyzed. Yet, it might represent a case of cooperative binding of GR proteins increasing GR binding rates and lowering noise.

Low Noise at Composite Sites

In the genome, we observe not only GBSs driving expression of target genes. In general, GR binds the genome together with other TFs to act on gene expression. These sites containing a GBS and an additional TFBS are called composite sites. To test the effect of such composite sites on expression noise, I constructed reporters with a single Fkbp5-2 GBS and placed 10 bp upstream a binding site for a GR cofactor (Ap1, Sp1 and Ets1), identified by enrichment in GR ChIP-seq peaks in U2OS (Figure 4.33b). The expression mean of these composite sites was high and the expression noise was low compared to 1xFkbp5-2 and 3xFkbp5-2 (Figure 4.33c). Notably, expression mean was similar for the three different composite sites and only the noise was variable. This effect on expression was hormone dependent, since the action of Ap1, Sp1 and Ets1 alone showed only weak expression mean and high noise in the ethanol control (Figure 4.34). Comparing 1xFkbp5-2 reporter and not hormone activated 1xFkbp5-2+Ap1 reporter alone showed that the effect on expression mean may be multiplicative and not additive compared to induced 1xFkbp5-2+Ap1 reporter (more than 10 times increase). The reduction in noise and increase in mean expression might again be the consequence of cooperative binding

at these sites. It was shown that due to cooperative binding with Ap1 or FoxA1, GR binds with higher rates to GR composite sites [88, 89].

Together, my findings uncovered different ways by which GR can affect gene expression and expression variability. Consequently, GR might employ different mechanisms on genes in a cell population, showing similar expression mean but a different degrees of cell-to-cell variability. Gene A could be under the control of a composite site and gene B could be controlled by multiple GBSs. According to my findings, the expression of gene A may show lower cell-to-cell variability than the expression of gene B in a cell population.

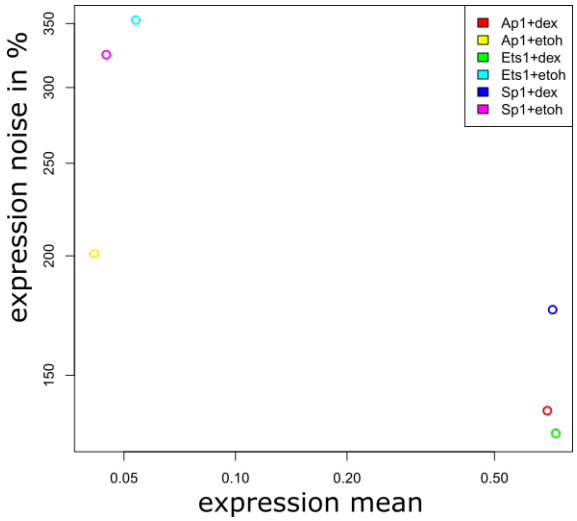


Figure 4.34.: High expression of composite sites is GR dependent. Repeat of noise measurement for not hormone treated cells reveals very weak activation by Ap1, Sp1 and Ets1 alone. Transfected cells were treated over-night with 1 μM dexamethasone and 0.1% ethanol before fluorescence analysis. Only GFP+ and mCherry+ cells were analyzed.

5. Discussion

In my thesis, I have shown that the DNA sequence variants of a GBS and its flanking sequences interfere with gene expression by GR. By using STARR-seq with a synthetic GBS library I could show that GBS variants can induce different levels of gene expression. In the endogenous genomic context, I found that especially the direct flanking sites play a role in fine-tuning GR action. My work indicated that DNA acts as an allosteric factor and induces conformational changes in GR downstream of binding, which may change interactions with cofactors and coregulators. Furthermore, I have shown that expression noise of reporter genes can be affected by sequence variants and composition of the GR response element. A detailed discussion of my findings and its implications are presented below.

5.1. STARR-seq with Synthetic GBS Library

5.1.1. Design of Library Insert is a Key Step

The STARR-seq method allowed me to assay a large number of GBS variants in parallel for their GR activity, something that has not been done before. I was able to find that GBS variants and the flanking sequence of GBS modulate GR activity. The advantage of this method is that the read-out of TF action is activity and not affinity. In this way, the method is unique to other well-established methods for studying TFBS variants like SELEX-seq or protein binding microarray [76, 90]. In these methods, researchers correlate sequence variants with TF binding strength, but receive no downstream information for their biological activity. Moreover, the STARR-seq method with synthetic TFBS library could be applicable for all kinds of DNA binding proteins that lead to a change in gene expression. Studying even small variations of binding sequence may help us to understand the sequence code of TF action.

A single GBS alone showed high activation rates of the STARR-seq reporter and integration of complete ChIP-peaks was unnecessary (Figure 4.2). This made the design of the GBS library easier, since inserts are short (<200bp). Moreover, integration by In-Fusion reaction (Takara clontech) of small inserts worked with higher efficiency compared to longer inserts with more than 750 bp.

Nonetheless, the biological question one can answer with this method highly depends on the design of the test library. The test library depends on the length and complexity of the regulatory sequence one is interested in. In my case, I was not interested in studying variations of a complete enhancer with multiple TFBSs, which

is rather complex. I was interested in studying variations of a single GR binding sequence. The number of variants of a regulatory sequence one can possibly test is limited, because there are several bottle-necks in the STARR-seq method. The most limiting bottle-neck I encountered was the transformation of *E. coli*, because in general only one or two plasmids are transferred in a single bacterium during transformation [91]. Thus to increase library complexity, one needs to increase the amount of transformants. In my experience in *E. coli* transformation, 1 million different plasmid variants can be transformed without too many adaptations, but testing already more than 10 randomized base pairs (4^{10}) exceeds this number. The transfection in human cells was not rate limiting in terms of how many sequence variants can be assayed, since multiple plasmids are transfected in a single cell, which was also shown by the high number of variants after plasmid library isolation from transfected cells. However, one question needs to be asked: Is it really necessary to cover all possible sequence variants to answer a biological question? For GR (GBS-HS library) most tested sequences did not show a GR dependent activation, since these sequence did not form a functional GBS. One way to answer this could be via a selection step of potential GBSs and to completely design a synthetic library [12]. Additionally, one can analyze only sequences really existing in the genome, that are bound by a TF (DNA accessibility) like done in the CapStarr-seq approach [42]. This might be close to the true nature, but it also means limiting research and introducing biases. Firstly, it is limited to known mechanisms (the selection criteria), hiding potential unknown mechanisms. Secondly, it is also of interest to investigate sequences that lead to no or only weak activation to understand transcription factor action. Following this concept, a different approach would be to "neglect" sequence variant coverage. Farley et al. developed a method similar to STARR-seq based on MPRA and transfected synthesized enhancer sequences with up to 49 bp randomized [43], so in principle billions of sequence variants could be in the library. However, only a subset actually is present in the library and thus a subset is subsequently assayed, yet from those sequences they could draw important conclusions regarding activity of enhancer variants. Altogether, to answer a biological question a complete coverage of sequence variants is not absolutely necessary, but has to be taken into account when designing a test library.

Independent of the number of sequence variants, sequencing depth influences the number of sequence variants than can be analyzed in parallel and affects the downstream analysis of differential expression. As I have seen in my data, for some sequences no reads were detected in the etoh control sample, but were well expressed in the dexamethasone treated sample. Because no information exists in the control sample, I had to assume these sequences have a high activation rate. DESeq2 runs into the problem of overestimating these sequences as done for the nonGBS. Activity of such low count GBS variants proved not to be reproducible when I tested them in isolation and therefore, I thus excluded them (Figure 4.9c). Still, it is valuable to keep information of sequences with low etoh counts, because some might be biological meaningful.

One drawback of the STARR-seq method is that it is an episomal method not

chromosomal. Therefore, we might overlook mechanisms involved in gene regulation that appear only in the genomic context. For example, histone occupancy differs between chromosomes and plasmids [92]. Also 3D-structural arrangements are limited on plasmids due to space, whereas it plays a role in the genome for formation of gene regulation units [93]. To study gene regulation also in chromosomal context, integration of the STARR-seq construct into the genome would be advantageous for future applications. Therefore, I am developing a version of the STARR-seq method where I use ZFN-driven targeted integration to study gene regulation in a defined chromosomal context.

5.1.2. Investigation of Technical Problems

One problem occurred that concerned sequences with low counts, but significant foldchanges, and no clear GR binding sequence (nonGBS sequences). These sequences might have been interesting, since they may represent cases of cooperative binding of GR with different cofactors. Yet, after validation these sequences revealed to be false-positive sequences. To circumvent misleading interpretation of the data, I chose a stringent cut-off of mean counts to remove these sequences. But there are also other ways to treat this problem. One could train a model that would detect these sequences, and thus remove false-positives in a more sophisticated way. Further, one could increase sequencing depth to reduce the number sequences with very low or not existing counts.

Another problem that occurred was the variability in activity of sequences between replicates, which can be of biological or technical origin. There is a large variability in the replicates of the 15h GBS-HS library compared to the 4h treatment (Figure 4.7). This could be explained by the routine and experience I gained while conducting the experiments, since time-wise I conducted the 15h experiment before the 4h experiment.

Additionally, the length of hormone incubation can potentially lead to secondary effects in expression pattern, meaning GR induced expression of proteins that could possibly interfere with GR action. I compared 4h and 15h hormone treatment times and I could not detect major differences in sequence expression between the two time points. This indicated that secondary effects did not play a major role after 4h. However, 4h could also be too long and secondary effects took already place, since GR action takes place within minutes after hormone application [75]. One could therefore also test shorter hormone incubation time. Additionally, after hormone treatment cells could be treated with cycloheximide to prevent protein biosynthesis and therefore secondary effects.

One general problem of the STARR-seq method is the lack of independent replicates for individual sequences. In the alternative method MPRA, one enhancer can be tested multiple times by usage of different barcodes and an average activity can be calculated from all incidents (barcode variants) detected during an experiment [39]. This has the advantage that technical errors introduced by library constitution, RNA stability, low coverage and PCR amplification can be reduced by

having replicates. However, this also reduced the number of sequences that can be tested, because the sequencing depth needs to be increased in order to compensate for the increase in barcode variants. My goal was to test as many variants of a GBS as possible for the trade-off of replicates. To assure that activities of individual binding sites were reproducible, I performed multiple biological replicates of each experiment.

A different problem was that some sequences may have different STARR-seq activities in individual context and in library context. One of the tested sequences of the STARR-seq medium group, produced an exceptionally high foldchange (Figure 4.10b), when tested in individually but only a medium signal in library context. This may originate from library constitution, but could be also caused by different unknown factors. This variant may be an exception, but has to be considered and is a reason why interesting sequences need to be validated.

Unexpectedly, in contrast to what we found in U2OS cells, we found that single GBSs were unable to activate the STARR-seq reporter in A549 cells. Apparently, gene expression could not be activated in this cell line, although from other experiments (luciferase reporter assay) GR shows good activation rates in the A549 cell line. Possibly, the SCP1 promoter cannot be activated in A549 cells. Promoter activity can be cell-type specific and may depend on recruitment of cell-type specific general transcription factors [94]. One way to get around this problem is to use a different promoter. Introducing a new promoter that is active in A549 may help, but it makes past experiments conducted with the SCP1 promoter in U2OS cell line difficult to compare with. Since I am not by design bound to the A549 cell line, it would be easier to test different cell lines, that express GR, for STARR-seq activity. The MCF7 cell line could be an interesting candidate cell line. It not only expresses GR but also ER, PR and AR [95]. So almost all members of the steroid receptor family could be tested in one cell line just by addition of the different steroid ligands, helping us to understand the gene regulation code in the binding sequences of the different steroid receptors.

5.1.3. Effect of Flanking Sequences on GBS Activity

I was able to show that the DNA sequence around a GBS plays a role in enhancing or reducing GR dependent gene expression through activity analysis of flanking sequence variants. One prominent sequence feature I identified, was the -tACNN- sequence flanking directly the Cgt and Sgk GBS, which correlated with strong GR activation. In broader context with the following upstream sequence, this sequence feature most likely generated a second GBS next to the designed GBS. As a consequence, this may lead to binding of more GR molecules and can explain the increase in activation strength. I identified, however, other sequences leading to an enhancing or blunting of gene expression. Several reasons may explain the effect of these flanking sequences. Structural differences of the DNA may affect the structure of nearby GBS and may have an allosteric effect on GR as will be discussed for the direct flanking sequence in the following section. Further, target binding sites for

different TFs may be generated in some sequence variants, that could interfere with GR action and can lead to a change in STARR-seq activation. So far, I did not scan the flanking sequence for the occurrence of new TFBS. This could be one potential way to analyze the found sequences. For newly identified sequences, interacting proteins could be identified using the DNA pull-down assays described in section 4.3.1 to confirm specific recruitment of other TFs to the DNA. In the end, several mechanisms may influence gene activity through the flanking sites but need thorough investigations to test them.

5.1.4. Insights into the Regulatory Code of GBSs

The next step will be to use the STARR-seq data to study the regulatory code in the sequence of a GBS. I am now able to associate a large number of GBS variants with a specific GR activity. I would like to study how sequences encode the GR activity. First, I could cluster sequences according to sequence and GR activity, for example using self-organizing maps (SOM). Similarly as I have done for the in depth analysis (without SOM) of the spacer (Figure 4.15), where I showed that the nucleotide composition of spacer position could be associated with strong or weak GR activity. Maybe further nucleotide positions of a GBS can be associated with a specific GR activity. This may reveal also interdependencies between nucleotide positions. For example, the C-rich GBS variants, that were found to be associated with strong GR activity, showed an enrichment of T0, C2 and C7 for this GBS variant. There seems to be a strong dependence between these nucleotides for inducing a strong GR activity.

A more complex application of the STARR-seq data would be to correlate activity with DNA shape prediction of individual sequences. One could cluster sequences according to GR activity and DNA shape (instead of sequence). This may help to deepen the understanding of the regulatory code of GBSs, because different sequences may have a similar DNA shape, but do not cluster in the same group following sequence composition. According to my findings, DNA shape plays an important role in GR action and similar shapes might induce similar responses in GR.

5.2. Flanking Sites Modulate GR's Action

Transcription factors find binding sites through DNA base and shape readout [45]. Both mechanisms add to the binding specificity; their influence on the gene expression level has not been well studied. It is well accepted that binding affinity correlates with activation. Here, in my case I showed that DNA shape and sequence also influence gene expression, that this influence is independent of binding affinity, and that they affect downstream targets of GR binding. I could show that the sequence of the first and second flanking site of a GBS strongly impair or elevate GR-dependent gene regulation without affecting *in vitro* GR binding (Table 4.3 and

section 4.2.6). The binding study was not only limited to *in vitro* binding experiments, I also used ChIP experiments as a proxy for *in vivo* binding. Although ChIP results can be distorted by cross-linking efficiencies, we change only two bases outside of the core DNA recognition site and the result of the ChIP assays were in-line with *in vitro* binding assays.

The interesting question remained, what happens to GR downstream of binding that results in such strong changes in gene expression? Previous studies showed that GBSs have the potential to induce conformational changes in the GR protein [1, 50] and that GBS halfsites and spacer induce this. In my work, I could show that flanking sites are also capable of inducing conformational changes. I saw that this effect is not caused by a direct "reader" that contacts the flanking bases. Residue R510 and K511 form contacts with the flanking positions, but however, upon mutation of these residues the flanking site effect persisted showing the cause is not the local readout of base contacts. The flanking sites were predicted to affect the DNA shape, also of the surrounding nucleotides, which varied depending on the flanking sequences. In summary, the different structural experiments revealed that the conformation of the dimer partners, the dynamics and the relative position of the dimer partners are affected by flanking sites.

According to my findings, the dimerization interface plays an important role, because a functional dimerization interface was necessary for the flanking site effect (Figure 4.28). Further, the conformational differences induced by flanks were reduced for GR A477T, lacking a functional dimer interface. The dimerization interface seems to place a structural stress on the two dimer partners, because they are limited by the partner in positioning. This might explain why we see different molecule dynamical behavior in the RMSF data at the flanking sequences (Figure 4.27a). For G/C flanks, the GR protein seems to move more in search of a favorable position under constrains by the dimer interface. As soon as this is relieved by the A477T mutation, each GR molecule might settle in a favorable position (Figure 4.29). The DNA structure of a base pair also affects surrounding sequences explaining why K461, a residue opposite to the predicted minor groove width changes, shows an interesting effect when mutated to an alanine. Now, the flanking site effect seems to be inverted for A/T and G/C flanks (Figure 4.28). K461 was shown to be important for GBS recognition by GR [79]. It was suggested that K461 is a device for GR to sense and interpret DNA sequence and may also drive GR conformation to an active form, which is consistent with my findings. Additional to sensing the DNA sequence, K461 seems to sense the DNA structure and drives GR conformation. Moreover, in our NMR experiment, K461 and two other neighboring residues (C460, G458) show peak splitting and shifting dependent on flanking sequence (Figure 4.29), indicating conformational differences of the helix 1 of the DBD between DNA constructs but also between dimer partners.

Altogether, my findings show that DNA sequence and shape both play an important role in defining GR action. DNA shape brings GR in an activation-competent form. In more detail, this means GR may bind to two sequences, but only one can induce the activation-competent form, which opens up the activation domains and

leads to higher quality recruitment of coregulators (Figure 5.1). This mechanism is likely not limited to GR. Since the steroid receptor family shares many structural features like DNA binding domain, ligand binding domain and activation domain. It is highly likely that they share this activation mechanism and might explain, despite the high similarity of binding site preferences, why steroid receptors have distinct target sites. Steroid receptors may react differently to DNA shape and may induce different structural conformations alias activation-competent forms. Also for other TFs (Hox-family, bHLH TF) flanking sites and DNA structure play an important role for binding site specificity [36, 76]. So the core motif seems to provide the molecular basis of TF binding, but it is the shape of the DNA and of the surrounding sequences that may make a sequence active and functional.

DNA shape may help GR to induce different levels of transcription at different target genes and explains in part the diversity of GR action. I identified G/C and A/T flanks in correlation with endogenous gene expression level. In this way, GR supports the cell to produce gene products at different levels.

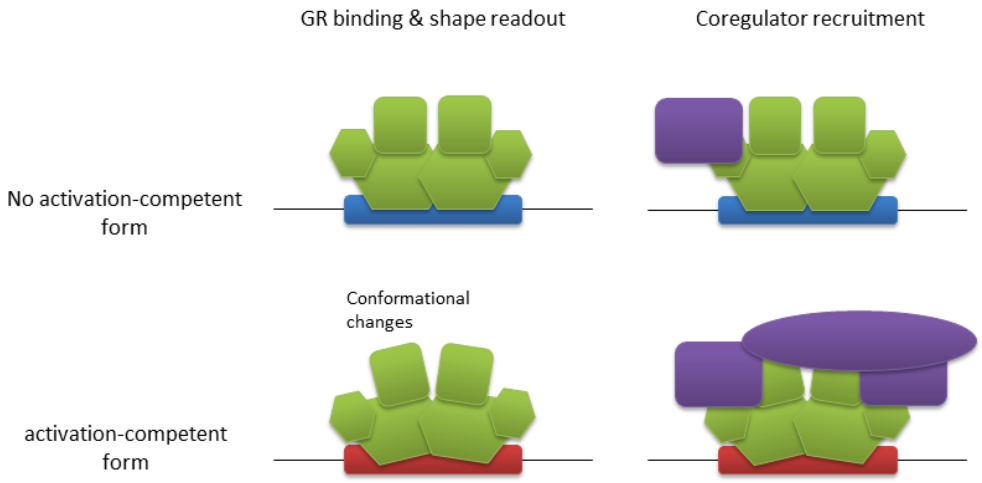


Figure 5.1.: Cartoon depicting GR action at different GBS with differing DNA shape. The first GBS results in a GR form that can not activate gene expression after binding. The second GBS induces conformational changes transforming GR in an activation-competent form to recruit coregulators and activate gene expression.

5.3. Towards New Insights into the Steroid Receptor Family with STARR-seq

Previously, I brought up the case of the high similarity in binding preferences between steroid receptor family members. Despite the similar core binding sequence each member of the steroid receptor family controls different cellular processes, even

when present in the same tissue. One way to explain the final specificity of receptor binding could be coregulator interaction with the receptor. Coregulators may introduce conformational changes in the steroid receptor, which change DNA binding specificity. However, a different scenario based on my findings regarding the GBS flanking sites could be true as well. The DNA itself introduces conformational changes of the receptor that opens up specific interaction sites for coregulators, which stabilize receptor binding at the DNA (Figure 5.1). The retention time of the receptor-coregulator complex is increased to last long enough to recruit the transcriptional machinery [96]. A different steroid receptor member might recognize the same site, but upon binding is not transformed in an activation-competent form and binding is unstable, because it is not supported by other coregulators and cofactors and finally falls off the DNA again. In a slightly different scenario, binding of the steroid receptor may lead to interaction with different coregulators, since steroid receptors differ in the N-terminal activation domain which only shows limited sequence conservation between steroid receptors. Studies of AR and GR activity at the same site revealed that GR and AR binding result in different transcriptional outcomes and that this is linked to differences in interaction with HDACs downstream of steroid receptor binding [97]. What comes first, conformational change induced by DNA or induced by coregulator? Probably, both mechanism play an equally important role and both pieces have to fit together to produce a functional output.

My adapted STARR-seq approach may help to solve part of the steroid receptor puzzle. In the future, one could test steroid receptors in parallel with the same library to identify sequence preferences and potential differences. For example, one could test AR and GR on the GBS-HS library and read out the activation profile of each sequence for each receptor. This might give a better understanding of the sequence code for receptor action. In a different experiment, one could knock-out or overexpress one specific coregulator of the steroid receptor family and test how the sequence preference changes. These are all experiments that can be performed easily with the STARR-seq method with synthetic binding site library. These types of experiments might advance our understanding of steroid receptors action and may help to produce better designed drugs for targeting receptor action in the future. Maybe this leads to targeting of specific coregulators that are involved in a certain cellular process of the receptor and might help to reduce off-target effects of treatments.

5.4. Expression Noise

Next to expression strength, my studies uncovered that expression noise is influenced by variations in GR binding regions. I discovered that the expression noise of a reporter gene is influenced by the composition of the GR binding region regulating it. Additionally, I could show that enhancer and not only promoter affect expression noise. In which way expression noise is affected by GR, can be linked to mean

expression strength of a population, to multiple GBSs and to interaction of GR with cofactors, which might have negative and positive effects on noise. So far, studies including mine, have mostly depended on reporter plasmids to measure expression noise. However, in this way, we lack the information of what happens at endogenous genes. Analysis of endogenous genes is far more difficult than the simplicity of reporter genes. Therefore, the next step will be to learn more about expression noise in the chromosomal context. Recent advances in single cell technology allow us to analyze gene expression in single cells, like single cell RNA-seq (scRNA-seq). Mantsoki et al. analyzed expression variability in mammalian embryonic stem cells with scRNA-seq [59]. They observed that many highly expressed genes have low noise and were enriched for cell cycle genes, whereas some genes with high noise (also highly expressed) form co-expression clusters and are enriched for response genes of the DNA repair and DNA damage system. Additionally, they showed that technical noise can be a problem in scRNA-seq, but mostly for lowly expressed genes. Thus, Mantsoki et al. selected an expression threshold and analyzed only genes with abundant expression. In the future, scRNA-seq will probably be improved and technical noise reduced, allowing us to study expression noise of most genes.

Why is studying the noise of gene expression of importance? Because noisy expression is one natural way of generating diversity in a cell population without affecting the genome sequence [98]. On the contrary, precise expression of a gene might be important for some genes, for example master regulators that control important developmental processes. Furthermore, miss-regulation of noise might be linked to diseases. It was shown that a change in expression variability is linked to neurological disorders [99]. In this study, Mar et al. did not focus on difference in gene expression between schizophrenia and Parkinson's disease, but they focused on differences in expression variability. Expression variability was constrained in schizophrenia patient and high in Parkinson's disease compared to a healthy control group. Therefore, expression noise is important to study in diseases and it may affect cellular networks and disease process. These findings may have implications for GR and downstream targets. GR is an important widely-used therapeutical target and understanding the effect of GR on gene expression and expression noise might help us to understand and therefore, reduce side effects and improve disease outcomes.

5.5. Insight into Transcriptional Regulation

Transcriptional regulation is a complex concert of many factors. In my thesis, I could show that the DNA shape, plays an important role in this concert. The DNA is able to actively affect the structure and activity of its interacting transcription factor. Not only the sequence of the TFBS but also the flanking base have a similar effect. In addition to GR, for two other TF families, Hox and bHLH, it was previously described that DNA shape played a role in TF binding [36, 76]. Thus, DNA shape may be a general mechanism for many transcription factors, which affects binding specificity and gene expression. This also brings mutations found in the flanking

regions of TFs into the focus. Most of these mutations may not have a dramatic effect on gene expression, but some mutations might lead to formation of a DNA-TF complex with enhanced or reduced transcriptional activity. These changes might not have a predicted effect on the affinity of a TF for the binding site, yet might influence the expression of associated genes.

Researchers mostly measured the effect of a sequence variation on TF action by affinity-driven methods (Selex-seq, PBM) [76,90]. High binding affinity is considered to be a reliable readout for activity, however this assumption seems to be not true for all TFs. In the case of GR, the strongest measured binding affinity for a GBS, called Pal, exhibited very low GR activation [1]. So far we have only low-throughput data for GR showing that GR activity is not always correlated with binding affinity [1,97]. With the STARR-seq method we have a tool at hand to also directly measure GR gene activation and correlate this information with binding affinity data, for example from protein binding microarrays (PBM), in a high-throughput manner, shedding light on this paradox. Notably, also in nature high affinity sites are not necessarily responsible for conferring biological responses. The contrary seems to be true as well. Low binding affinity TFBS with suboptimized sequences seem to confer higher binding specificity and lead to less ectopic expression of target genes than sites with high affinity [43,48]. Additionally, multiple low affinity binding sites also lead to robustness in expression. So for enhancer function, less affinity seems to be advantageous in at least some cases. Following this idea, purely affinity-driven methods bear the danger of producing misleading sequence variants, which show higher binding affinity but are not improving binding site identification in the genomes, because they are not naturally occurring [48]. Further findings showed that transcriptional regulation of TFs is strongly linked with binding dynamics rather than occupancy [96, 100], yet also binding dynamics can be misleading in some cases [97]. In the end, only measuring expression strength seems to be a reliable functional readout of TF action. This may be a result of binding of TF and interaction with downstream factors. Focusing downstream of TF binding on cofactor and coregulator interaction may reveal how TFBS sequence variants influence gene-specific transcriptional outputs.

For gene regulation, the promoter likely integrates information coming from several enhancers bound by multiple TFs. So the effect of a single TFBS might be compensated by other TFBSs in enhancer or promoter, showing a certain robustness of gene expression. Nonetheless, this information may help us to improve the predictive models and quantitative prediction of gene expression from sequence information.

5.6. Recent Developments of Enhancer Studies

Despite the recent development of the assays, enhancer identification and characterization in mammals is still challenging because of the complexity of the genome. Preselection steps are so far necessary to reduce the complexity and size of the

genomes, but they may introduce biases and blind spots for unknown factors. However, with the fast advances of the technologies, we might be able to also overcome the remaining obstacles.

The recently developed CRISPR/Cas system is one of these methods revolutionizing our current way of research [101, 102]. CRISPR/Cas allows to directly edit the genome in a fast, easy, cheap and highly accurate way at nucleotide resolution in almost any biological system. With CRISPR/Cas, we can study gene regulation in an endogenous context and change the way we study gene regulation. It reduces the importance of classical transient reporter studies, which were controversial in the scientific community because of their artificiality. Now, we can "easily" modify regulatory sequences directly at almost any gene of interest to study the effect on transcriptional regulation. Research is currently ongoing to apply the CRISP/Cas system in combination with other high-throughput methods.

Using the CRISPR/Cas system, we could provide the ultimate proof that GBS variants tune the expression of individual GR target genes by changing the GBS sequence in the endogenous genomic context. Assaying how this changes the GR-dependent regulation of associated genes, might provide insights in the endogenous regulation of expression by GR.

5.7. Conclusion

With my work, I could show that GBS variants may influence the structure of the bound GR, which could affect further downstream the recruitment of coregulators. In this way, a GBS may fine-tune the expression of GR target genes, seemingly disconnected from the binding affinity of GR. Further, I showed that not only the core sequence of a GBS may influence GR action, also the sequence flanking a GBS is involved in this process. Therefore, the sequence context of a GBS is also important for GR function. Additionally, the composition of GR bound regulatory regions play a role in the variability of cell-to-cell expression of GR target genes. This may have consequences for the diverse regulatory networks in which GR is involved, but further experiments in the endogenous genomic context are necessary to unravel this.

GR is a seemingly sophisticated transcription factor with equally sophisticated GR binding sites. The large variation of GBSs in the genome might therefore be a consequence of the diverse functions GR has to exert in mammals.

A. Abbreviations

bp	basepair
ChIP	chromatin-immunoprecipitation
DBD	DNA binding domain
DNA	deoxyribonucleic acid
GBS	GR binding sequence
GR	glucocorticoid receptor
lfc	log foldchange
MD	molecular dynamic
MS	mass-spectrometry
N	IUPAC nucleotide code for any base
nt	nucleotide
PBM	protein binding microarray
PMW	Position weight matrix
RNA	ribonucleic acid
RNA PolII	RNA polymerase 2
RFU	relative fluorescence units
RLU	relative light units
RT	room temperature
scRNA-seq	single cell RNA-sequencing
SD	standard deviation
TF	transcription factor
TFBS	transcription factor binding sites
TSS	Transcription start site
ZFN	Zinc-finger nuclease

B. List of DNA Oligomers

Table B.1.: Cloning Primers for PGL3 promoter vector

GBS	flanks	forward	reverse
Cgt	G-C	CGAGAACATTTTGTACGCC	TCGAGGCGTACAAAATGTTCTCGGTAC
	A-T	CAAGAACATTTTGTACGTC	TCGAGACGTACAAAATGTTCTTGGTAC
	G-T	CGAGAACATTTTGTACGTC	TCGAGACGTACAAAATGTTCTCGGTAC
	A-C	CAAGAACATTTTGTACGCC	TCGAGGCGTACAAAATGTTCTTGGTAC
Sgk	G-C	CGAGAACATTTTGTCCGCC	TCGAGGCGGACAAAATGTTCTCGGTAC
	A-T	CAAGAACATTTTGTCCGTC	TCGAGACGGACAAAATGTTCTTGGTAC
	G-T	CGAGAACATTTTGTCCGTC	TCGAGACGGACAAAATGTTCTCGGTAC
	A-C	CAAGAACATTTTGTCCGCC	TCGAGGCGGACAAAATGTTCTTGGTAC
FKBP5	G-C	CGAGAACAGGGTGTTCTCC	TCGAGGAGAACACCCTGTTCTCGGTAC
	A-T	CAAGAACAGGGTGTTCTTC	TCGAGAAGAACACCCTGTTCTTGGTAC
FKBP5-2	G-C	CGAGAACATCCTGTGCCCC	TCGAGGGGCACAGGATGTTCTCGGTAC
	A-T	CAAGAACATCCTGTGCCTC	TCGAGAGGCACAGGATGTTCTTGGTAC
Pal	G-C	CGAGAACAAAATGTTCTCC	TCGAGGAGAACATTTTGTTCTCGGTAC
	A-T	CAAGAACAAAATGTTCTTC	TCGAGAAGAACATTTTGTTCTTGGTAC

Table B.2.: Primers for site-directed Mutagenesis

rGR K465A	fw	AGCTGCAAAGTATTCTTTGCAAGAGCAGTGGAAGGAC
	rev	GTCCTTCCACTGCTCTTGCAAAGAATACTTTGCAGCT
rGR K511A	fw	GAACCTTGAAGCTCGAGCAACAAAGAAAAAAATC
	rev	GATTTTTTTCTTTGTTGCTCGAGCTTCAAGGTTC
TA-Sgk-TA	fw	CTCTATCGATAGGTACTAAGAACATTTTGTCCGTATCGAGATCTGCGATCTGCATC
	rev	GATGCAGATCGCAGATCTCGATACGGACAAAATGTTCTTAGTACCTATCGATAGAG
TA-Cgt-TA	fw	CTCTATCGATAGGTACTAAGAACATTTTGTACGTATCGAGATCTGCGATCTGCATC
	rev	GATGCAGATCGCAGATCTCGATACGTACAAAATGTTCTTAGTACCTATCGATAGAG

Table B.3.: List of DNA Oligomer

Name	description	sequence
SS063	fw, SAA-GFP	ATACGGTACCGTGCCAGAACATTTCTCTATCGATA
SS064	rev, SAA-GFP	TCAAGTCGACGGATCCTTATCGATTTTACC
R5	AVVS1 Integration	CTGGGATACCCCGAAGAGTG
LucNested	AVVS1 Integration	TCAAAGAGGCGAACTGTGTG
G/C-Sgk	EMSA-fw	Cy5/ACCGAGAACATTTTGTCCGCCTC
G/C-Sgk	EMSA-rev	GAGGCGGACAAAATGTTCTCGGT
A/T-Sgk	EMSA-fw	Cy5/ACCAAGAACATTTTGTCCGTCTC
A/T-Sgk	EMSA-rev	GAGACGGACAAAATGTTCTTGGT
A/T-Cgt	EMSA-fw	Cy5/ACCAAGAACATTTTGTACGTCTC
A/T-Cgt	EMSA-rev	GAGACGTACAAAATGTTCTTGGT
G/C-Cgt	EMSA-fw	Cy5/ACCGAGAACATTTTGTACGCCTC
G/C-Cgt	EMSA-rev	GAGGCGTACAAAATGTTCTCGGT
CA-CGT-TC	EMSA-fw	Cy5/TACCAAGAACATTTTGTACGTCTCG
CA-CGT-TC	EMSA-rev	CGAGACGTACAAAATGTTCTTGGTA
TA-CGT-TA	EMSA-fw	Cy5/TACTAAGAACATTTTGTACGTATCG
TA-CGT-TA	EMSA-rev	CGATACGTACAAAATGTTCTTAGTA
SS186	mRNA Starr-seq	CAAACTCATCAATGTATCTTATCATG
SS189	mRNA RPL19	GAGGCCAGTATGTACAGACAAAGTGG
hRPL19	qPCR-fw	ATGTATCACAGCCTGTACCTG
hRPL19	qPCR-rev	TTCTTGGTCTCTTCCTCCTTG
GFP-mRNA	qPCR-fw	GGCCAGCTGTTGGGGTGTC
GFP-mRNA	qPCR-rev	TTGGGACAACTCCAGTGAAGA
pgl3promoter-luc	qPCR-fw	GATGCGGTGGGCTCTATG
pgl3promoter-luc	qPCR-rev	GAGTTAGGGGCGGGACTATG
hFKBP5	qPCR-fw	GCATGGTTTAGGGGTTCTTG
hFKBP5	qPCR-rev	TAACCACATCAAGCGAGCTG
hZBTB16	qPCR-fw	CTCCTTGAGGGAAAGAACACAC
hZBTB16	qPCR-rv	ACAGACGCAGGGCATTTTAC
A-Cgt-T	NMR, fw	GCCAAGAACATTTTGTACGTCTCG
A-Cgt-T	NMR, rev	CGAGACGTACAAAATGTTCTTGGC
G-Cgt-C	NMR, fw	GCCGAGAACATTTTGTACGCCTCG
G-Cgt-C	NMR, rev	CGAGGCGTACAAAATGTTCTCGGC
A-Cgt-C	NMR, fw	GCCAAGAACATTTTGTACGCCTCG
A-Cgt-C	NMR, rev	CGAGGCGTACAAAATGTTCTTGGC
G-Cgt-T	NMR, fw	GCCGAGAACATTTTGTACGTCTCG
G-Cgt-T	NMR, rev	CGAGACGTACAAAATGTTCTCGGC
SS194	library generation	GGCCGAATTCGTCGAGTGAC
SS192	cDNA amplification	GGGCCAGCTGTTGGGGTG*T*C*C*A*C
SS193	cDNA amplification	CTTATCATGTCTGCTCGA*A*G*C
SS195	input amplification	TATCATGTCTGCTCGAAGCGG
SS196	input amplification	GGATTTGATATTCACCTGGC

Table B.4.: DNA Oligomer for DNA pull-down

Pal-fw	Btn-teg-CAAAAGATCGCTGCAGACTTGAACAGAACAAAATGTTCTACTTTGTC
Pal-rev	GACAAAGTAGAACATTTTGTTCTGTTCAAGTCTGCAGCGATCTTTTG
scrambled-fw	Btn-teg-CAAAAGATCGCTGCAGACTTGAACAGAAACAAAGTTTCTACTTTGTC
scrambled-rev	GACAAAGTAGAAACTTTGTTTCTGTTCAAGTCTGCAGCGATCTTTTG
GilZ-fw	Btn-teg-CAAAAGATCGCTGCAGACTTGAACAGAACATTGGGTTCCACTTTGTC
GilZ-rev	GACAAAGTGGAACCCAATGTTCTGTTCAAGTCTGCAGCGATCTTTTG
A/T-Sgk, fw	Btn-teg-CAAAAGATCGCTGCAGACTTGAAAAGAACATTTTGTCCGTCTTTGTC
A/T-Sgk, rev	GACAAAGACGGACAAAATGTTCTTTTCAAGTCTGCAGCGATCTTTTG
G/C-Sgk, fw	Btn-teg-CAAAAGATCGCTGCAGACTTGAAGAGAACATTTTGTCCGCCTTTGTC
G/C-Sgk, rev	GACAAAGGCGGACAAAATGTTCTCTTCAAGTCTGCAGCGATCTTTTG

Table B.5.: gBlocks for STARR-seq screening vector

1xPal	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAaa** **aTGTTCT**CGTCGCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
1xFKBP5-1	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAggg** **TGTTCT**CGTCGCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
1xGilz	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAttg** **GGTTCC**CGTCGCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
1xFkbp5-2	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAtcc** **TGTGCC**CGTCGCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
1xCgt	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGCG**AGAACAttt** **TGTACG**CCTCGCTAGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
3xGilz	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAACAttgGGTTCC** GAAC**AGAACAttgGGTTCC**TCGA**AGAACAttgGGTTCC**AGATCGGGAGCACACGTCTGAAC TCCAGTCACTCGACGAATTCGGCC
3xCgt	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAACAtttTGTACG** GAAC**AGAACAtttTGTACG**TCGA**AGAACAtttTGTACG**AGATCGGGAGCACACGTCTGAAC TCCAGTCACTCGACGAATTCGGCC
3xSgk	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAACAtttTGTCCG** GAAC**AGAACAtttTGTCCG**TCGA**AGAACAtttTGTCC**GAGATCGGGAGCACACGTCTGAAC TCCAGTCACTCGACGAATTCGGCC
3xPal	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAACAaaaTGTTCT** GAAC**AGAACAaaaTGTTCT**TCGA**AGAACAaaaTGTTCT**AGATCGGGAGCACACGTCTGAAC TCCAGTCACTCGACGAATTCGGCC
3xFkbp5-1	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAACAgggTGTTCT** GAAC**AGAACAgggTGTTCT**TCGA**AGAACAgggTGTTCT**AGATCGGGAGCACACGTCTGA ACTCCAGTCACTCGACGAATTCGGCC
3xFkbp5-2	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**AGAACATCCTGTGC** **C**GAAC**AGAACATCCTGTGCC**TCGA**AGAACATCCTGTGCC**AGATCGGGAGCACACGTCTG AACTCCAGTCACTCGACGAATTCGGCC
Ap1–Fkbp5-2	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAT** **CCTGTGCC**CGTCGCTAAG**TGAGTCA**CCTAGTTAGATCGGAAGAGCACACGTCTGAACTC CAGTCACTCGACGAATTCGGCC
Sp1–Fkbp5-2	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAT** **CCTGTGCC**CGTCGCTAAG**CCCTCCCC**CCTAGTTAGATCGGAAGAGCACACGTCTGAACTC CAGTCACTCGACGAATTCGGCC
Ets1–Fkbp5-2	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGCGAA**AGAACAT** **CCTGTGCC**CGTCGCTAAG**ACAGGA**CCTAGTTAGATCGGAAGAGCACACGTCTGAACTC CAGTCACTCGACGAATTCGGCC

Table B.6.: Oligomer (29 bp) for STARR-seq Input Library

Cgt-flank	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**CAGCGCAAGAACAttt** **TGTACGNNNNNNCT**AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
Sgk-flank	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**CAGCGCAAGAACAttt** **TGTCCGNNNNNNCT**AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC
GBS-HS	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT**CAGCGAAAGAACAtnn** **NNNNNNCGTCGCT**AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGACGAATTCGGCC

C. Vector Maps

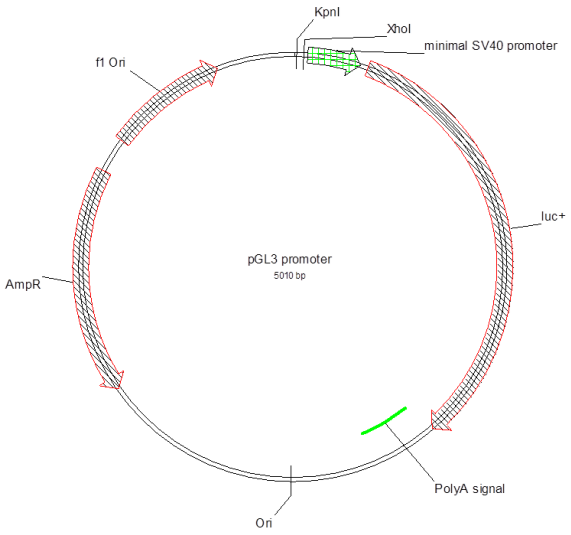


Figure C.1.: Vector map of pGL3 promoter (Promega)

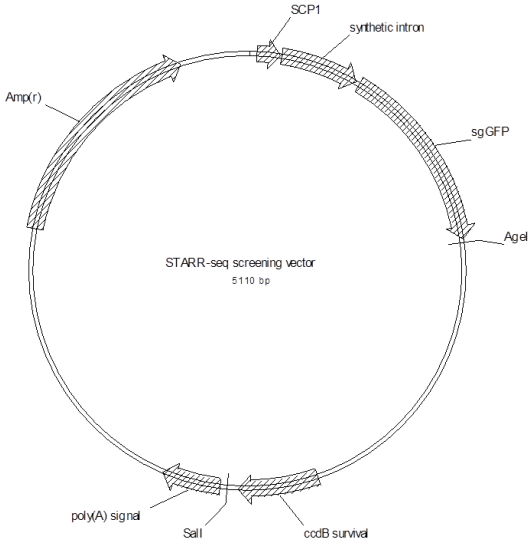


Figure C.2.: Vector map of human STARR-seq screening vector

References

[1] Meijsing, S. H., Pufall, M. A., So, A. Y., Bates, D. L., Chen, L., and Yamamoto, K. R. (2009) DNA Binding Site Sequence Directs Glucocorticoid Receptor Structure and Activity. *Science,* **324**(5925), 407–410.

[2] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (apr, 2009) A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics,* **10**(4), 252–63.

[3] Solomon, M. J., Larsen, P. L., and Varshavsky, A. (jun, 1988) Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell,* **53**(6), 937–947.

[4] Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (feb, 2002) Capturing chromosome conformation. *Science (New York, N.Y.),* **295**(5558), 1306–11.

[5] WATSON, J. D. and CRICK, F. H. (apr, 1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature,* **171**(4356), 737–8.

[6] Lawrence, M., Daujat, S., and Schneider, R. (jan, 2016) Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics,* **32**(1), 42–56.

[7] Washietl, S., Pedersen, J. S., Korbel, J. O., Stocsits, C., Gruber, A. R., Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., Ucla, C., Wyss, C., Antonarakis, S. E., Denoeud, F., Lagarde, J., Drenkow, J., Kapranov, P., Gingeras, T. R., Guigó, R., Snyder, M., Gerstein, M. B., Reymond, A., Hofacker, I. L., and Stadler, P. F. (jun, 2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome research,* **17**(6), 852–64.

[8] Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M. (sep, 2012) An integrated encyclopedia of DNA elements in the human genome. *Nature,* **489**(7414), 57–74.

[9] Ohno, S. (1972) So Much 'Junk DNA' in our Genome.. *Evolution of genetics systems. Brookhaven Symposium on Biology,* **23**(23), 366–370.

[10] Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. (mar, 2012) Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell research,* **22**(3), 490–503.

[11] Banerji, J., Rusconi, S., and Schaffner, W. (dec, 1981) Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell,* **27**(2), 299–308.

[12] Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (may, 2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology,* **30**(6), 521–530.

[13] Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., Berns, A., Wessels, L. F., van Lohuizen, M., and van Steensel, B. (aug, 2013) Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell,* **154**(4), 914–927.

[14] Link, N., Kurtz, P., O'Neal, M., Garcia-Hughes, G., and Abrams, J. M. (nov, 2013) A p53 enhancer region regulates target genes through chromatin conformations in cis and in trans.. *Genes & development,* **27**(22), 2433–8.

[15] Dean, A. (jan, 2006) On a chromosome far, far away: LCRs and gene expression. *Trends in genetics : TIG,* **22**(1), 38–45.

[16] Babbitt, C. C., Markstein, M., and Gray, J. M. (jun, 2015) Recent advances in functional assays of transcriptional enhancers. *Genomics,*.

[17] Kornberg, R. D. (may, 2005) Mediator and the mechanism of transcriptional activation. *Trends in biochemical sciences,* **30**(5), 235–9.

[18] Berger, S. L. (may, 2007) The complex language of chromatin regulation during transcription. *Nature,* **447**(7143), 407–12.

[19] Telorac, J., Prykhozhij, S. V., Schöne, S., Meierhofer, D., Sauer, S., Thomas-Chollier, M., and Meijsing, S. H. (mar, 2016) Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements. *Nucleic acids research,* p. gkw203.

[20] Gao, F., Foat, B. C., and Bussemaker, H. J. (mar, 2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC bioinformatics,* **5**(1), 31.

[21] Shlyueva, D., Stampfel, G., and Stark, A. (apr, 2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics,* **15**(4), 272–86.

[22] Levo, M. and Segal, E. (jul, 2014) In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics,* **15**(7), 453–68.

[23] Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E., and Weng, Z. (jan, 2012) Modeling gene expression using chromatin features in various cellular contexts.. *Genome biology,* **13**(9), R53.

[24] Ligr, M., Siddharthan, R., Cross, F. R., and Siggia, E. D. (apr, 2006) Gene expression from random libraries of yeast promoters. *Genetics,* **172**(4), 2113–22.

[25] Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (may, 2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences of the United States of America,* **107**(20), 9158–63.

[26] Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R., and Willard, H. F. (jan, 1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature,* **349**(6304), 38–44.

[27] Straub, T., Gilfillan, G. D., Maier, V. K., and Becker, P. B. (oct, 2005) The Drosophila MSL complex activates the transcription of target genes. *Genes & development,* **19**(19), 2284–8.

[28] Harvey, M., McArthur, M. J., Montgomery, C. A., Butel, J. S., Bradley, A., and Donehower, L. A. (nov, 1993) Spontaneous and carcinogen-induced tumorigenesis in p53-deficient mice. *Nature genetics,* **5**(3), 225–9.

[29] Kim, H. D. and O'Shea, E. K. (nov, 2008) A quantitative model of transcription factor-activated gene expression. *Nature structural & molecular biology,* **15**(11), 1192–8.

[30] Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (dec, 2014) Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation. *Nature,* **518**(7540), 556–559.

[31] Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z., and Segal, E. (oct, 2014) Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome research,* **24**(10), 1698–706.

[32] Oakley, R. H. and Cidlowski, J. A. (feb, 2011) Cellular processing of the glucocorticoid receptor gene and protein: new mechanisms for generating tissue-specific actions of glucocorticoids. *The Journal of biological chemistry,* **286**(5), 3177–84.

[33] Thomas-Chollier, M., Watson, L. C., Cooper, S. B., Pufall, M. a., Liu, J. S., Borzym, K., Vingron, M., Yamamoto, K. R., and Meijsing, S. H. (oct, 2013) A naturally occuring insertion of a single amino acid rewires transcriptional regulation by glucocorticoid receptor isoforms. *Proceedings of the National Academy of Sciences of the United States of America,*.

[34] Thornton, J. W. (may, 2001) Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proceedings of the National Academy of Sciences of the United States of America,* **98**(10), 5671–6.

[35] Thornton, J. W. and Desalle, R. (jun, 2000) A New Method to Localize and Test the Significance of Incongruence: Detecting Domain Shuffling in the Nuclear Receptor Superfamily. *Systematic Biology,* **49**(2), 183–201.

[36] Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. (nov, 2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell,* **131**(3), 530–43.

[37] Mann, R. S., Lelli, K. M., and Joshi, R. (jan, 2009) Hox specificity unique roles for cofactors and collaborators. *Current topics in developmental biology,* **88**, 63–101.

[38] Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (mar, 2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.),* **339**(6123), 1074–7.

[39] Birnbaum, R. Y., Patwardhan, R. P., Kim, M. J., Findlay, G. M., Martin, B., Zhao, J., Bell, R. J. A., Smith, R. P., Ku, A. A., Shendure, J., and Ahituv, N. (oct, 2014) Systematic Dissection of Coding Exons at Single Nucleotide Resolution Supports an Additional Role in Cell-Specific Transcriptional Regulation. *PLoS Genetics,* **10**(10), e1004592.

[40] Inoue, F. and Ahituv, N. (sep, 2015) Decoding enhancers using massively parallel reporter assays. *Genomics,* **106**(3), 159–64.

[41] Murtha, M., Tokcaer-Keskin, Z., Tang, Z., Strino, F., Chen, X., Wang, Y., Xi, X., Basilico, C., Brown, S., Bonneau, R., Kluger, Y., and Dailey, L. (may, 2014) FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nature methods,* **11**(5), 559–65.

[42] Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T., Fernandez, N., Ballester, B., Andrau, J. C., and Spicuglia, S. (2015) High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications,* **6**, 6905.

[43] Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., and Levine, M. S. (oct, 2015) Suboptimization of developmental enhancers. *Science,* **350**(6258), 325–328.

[44] Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (oct, 2009) The role of DNA shape in protein-DNA recognition. *Nature,* **461**(7268), 1248–53.

[45] Rohs, R. (2010) Origins of specificity in protein-DNA recognition. *Annual review of biochemistry,* pp. 233–269.

[46] Siggers, T. and Gordân, R. (feb, 2014) Protein-DNA binding: complexities and multi-protein codes. *Nucleic acids research,* **42**(4), 2099–111.

[47] Stormo, G. D. (jan, 2000) DNA binding sites: representation and discovery. *Bioinformatics,* **16**(1), 16–23.

[48] Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. S., and Stern, D. L. (jan, 2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell,* **160**(1-2), 191–203.

[49] Kim, S., Broströmer, E., Xing, D., Jin, J., Chong, S., Ge, H., Wang, S., Gu, C., Yang, L., Gao, Y. Q., Su, X.-d., Sun, Y., and Xie, X. S. (feb, 2013) Probing allostery through DNA. *Science (New York, N.Y.),* **339**(6121), 816–9.

[50] Watson, L. C., Kuchenbecker, K. M., Schiller, B. J., Gross, J. D., Pufall, M. A., and Yamamoto, K. R. (jul, 2013) The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nature structural & molecular biology,* **20**(7), 876–83.

[51] Adcock, S. A. and McCammon, J. A. (may, 2006) Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews,* **106**(5), 1589–615.

[52] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., and Mackerell, A. D. (mar, 2010) CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of computational chemistry,* **31**(4), 671–90.

[53] Sklenar, H., Wüstner, D., and Rohs, R. (feb, 2006) Using internal and collective variables in Monte Carlo simulations of nucleic acid structures: chain breakage/closure algorithm and associated Jacobians. *Journal of computational chemistry,* **27**(3), 309–15.

[54] Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., Di Felice, R., and Rohs, R. (jul, 2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic acids research,* **41**(Web Server issue), W56–62.

[55] Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (dec, 2015) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics (Oxford, England),*.

[56] Zoller, B., Nicolas, D., Molina, N., and Naef, F. (jul, 2015) Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular systems biology,* **11**(7), 823.

[57] Arias, A. M. and Hayward, P. (jan, 2006) Filtering transcriptional noise during development: concepts and mechanisms. *Nature reviews. Genetics,* **7**(1), 34–44.

[58] Chalancon, G., Ravarani, C. N. J., Balaji, S., Martinez-Arias, A., Aravind, L., Jothi, R., and Babu, M. M. (may, 2012) Interplay between gene expression noise and regulatory network architecture. *Trends in genetics : TIG,* **28**(5), 221–32.

[59] Mantsoki, A., Devailly, G., and Joshi, A. (feb, 2016) Gene expression variability in mammalian embryonic stem cells using single cell RNA-seq data. *Computational biology and chemistry,*.

[60] Rogatsky, I. and Others, A. (nov, 2003) Target-specific utilization of transcriptional regulatory surfaces by the glucocorticoid receptor. *Proceedings of the National Academy of Sciences,* **100**(24), 13845–13850.

[61] Pearce, D. and Yamamoto, K. R. (feb, 1993) Mineralocorticoid and glucocorticoid receptor activities distinguished by nonreceptor factors at a composite response element. *Science (New York, N.Y.),* **259**(5098), 1161–5.

[62] DeKelver, R. C. and Others, A. (aug, 2010) Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome research,* **20**(8), 1133–42.

[63] Rogatsky, I., Trowbridge, J. M., and Garabedian, M. J. (jun, 1997) Glucocorticoid receptor-mediated cell cycle arrest is achieved through distinct cell-specific transcriptional regulatory mechanisms. *Molecular and cellular biology,* **17**(6), 3181–93.

[64] Kotin, R. M., Linden, R. M., and Berns, K. I. (dec, 1992) Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *The EMBO journal,* **11**(13), 5071–8.

[65] Wu, K. K. (2006) Analysis of Protein-DNA Binding by Streptavidin–Agarose Pulldown. In *Methods in Molecular Biology* Vol. 338(1), pp. 281–290.

[66] Mittler, G., Butter, F., and Mann, M. (2009) A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Research,* pp. 284–293.

[67] Cox, J. and Mann, M. (dec, 2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology,* **26**(12), 1367–72.

[68] Gal, M., Schanda, P., Brutscher, B., and Frydman, L. (feb, 2007) UltraSOFAST HMQC NMR and the repetitive acquisition of 2D protein spectra at Hz rates. *Journal of the American Chemical Society,* **129**(5), 1372–7.

[69] Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D. M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (jul, 2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic acids research,* **43**(W1), W50–6.

[70] Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (jan, 2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature protocols,* **3**(10), 1578–88.

[71] Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J., and van Helden, J. (feb, 2011) Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic acids research,* **39**(3), 808–24.

[72] Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (dec, 2011) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research,* **32**(0), 1–9.

[73] Quinlan, A. R. and Hall, I. M. (mar, 2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England),* **26**(6), 841–2.

[74] Love, M. I., Huber, W., and Anders, S. (dec, 2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology,* **15**(12), 550.

[75] John, S., Johnson, T. A., Sung, M.-H., Biddie, S. C., Trump, S., Koch-Paiz, C. A., Davis, S. R., Walker, R., Meltzer, P. S., and Hager, G. L. (apr, 2009) Kinetic complexity of the global response to glucocorticoid receptor action. *Endocrinology,* **150**(4), 1766–74.

[76] Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L. (apr, 2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports,* **3**(4), 1093–104.

[77] Diamond, M., Miner, J., Yoshinaga, S., and Yamamoto, K. (sep, 1990) Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. *Science,* **249**(4974), 1266–1272.

[78] Safe, S., Kim, K., and Kim, K. (nov, 2008) Non-classical genomic estrogen receptor (ER)/specificity protein and ER/activating protein-1 signaling pathways. *Journal of molecular endocrinology,* **41**(5), 263–75.

[79] Starr, D. B., Matsui, W., Thomas, J. R., and Yamamoto, K. R. (may, 1996) Intracellular receptors use a common mechanism to interpret signaling information at response elements. *Genes & development,* **10**(10), 1271–83.

[80] Becnel, L. B., Darlington, Y. F., Ochsner, S. A., Easton-Marks, J. R., Watkins, C. M., McOwiti, A., Kankanamge, W. H., Wise, M. W., DeHart, M., Margolis, R. N., and McKenna, N. J. (jan, 2015) Nuclear Receptor Signaling Atlas: Opening Access to the Biology of Nuclear Receptor Signaling Pathways. *PloS one,* **10**(9), e0135615.

[81] McKenna, N. J. and O'Malley, B. W. (feb, 2002) Combinatorial Control of Gene Expression by Nuclear Receptors and Coregulators. *Cell,* **108**(4), 465–474.

[82] Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J.-P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z.-Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J. R., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A.-C., and Nesvizhskii, A. I. (aug, 2013) The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature methods,* **10**(8), 730–6.

[83] Sauvé, F., McBroom, L. D., Gallant, J., Moraitis, A. N., Labrie, F., and Giguère, V. (jan, 2001) CIA, a novel estrogen receptor coactivator with a bifunctional nuclear receptor interacting determinant. *Molecular and cellular biology,* **21**(1), 343–53.

[84] Gao, S., Li, A., Liu, F., Chen, F., Williams, M., Zhang, C., Kelley, Z., Wu, C.-L., Luo, R., and Xiao, H. (dec, 2013) NCOA5 haploinsufficiency results in glucose intolerance and subsequent hepatocellular carcinoma. *Cancer cell,* **24**(6), 725–37.

[85] Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J.-H.,

Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T., and Hayashizaki, Y. (mar, 2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell,* **140**(5), 744–52.

[86] Rajan, P., Gaughan, L., Dalgliesh, C., Robson, C. N., Leung, H. Y., and Elliott, D. J. (2008) The RNA-binding and adaptor protein Sam68 modulates signal-dependent splicing and transcriptional activity of the androgen receptor. *Journal of Pathology,* **215(1)**, 67–77.

[87] Foulds, C. E., Feng, Q., Ding, C., Bailey, S., Hunsaker, T. L., Malovannaya, A., Hamilton, R. a., Gates, L. a., Zhang, Z., Li, C., Chan, D., Bajaj, A., Callaway, C. G., Edwards, D. P., Lonard, D. M., Tsai, S. Y., Tsai, M.-J., Qin, J., and O'Malley, B. W. (jul, 2013) Proteomic Analysis of Coregulators Bound to ERα on DNA and Nucleosomes Reveals Coregulator Dynamics. *Molecular cell,* **51**(2), 185–99.

[88] Biddie, S. C., John, S., Sabo, P. J., Thurman, R. E., Johnson, T. a., Schiltz, R. L., Miranda, T. B., Sung, M.-H., Trump, S., Lightman, S. L., Vinson, C., Stamatoyannopoulos, J. a., and Hager, G. L. (jul, 2011) Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular cell,* **43**(1), 145–55.

[89] Belikov, S., Berg, O. G., and Wrange, Ö. (dec, 2015) Quantification of transcription factor-DNA binding affinity in a living cell. *Nucleic acids research,* pp. gkv1350–.

[90] Riley, T. R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R. S., and Bussemaker, H. J. (jan, 2014) SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods in molecular biology (Clifton, N.J.),* **1196**, 255–78.

[91] Goldsmith, M., Kiss, C., Bradbury, A. R. M., and Tawfik, D. S. (jul, 2007) Avoiding and controlling double transformation artifacts. *Protein engineering, design & selection : PEDS,* **20**(7), 315–8.

[92] Mladenova, V., Mladenov, E., and Russev, G. (apr, 2014) Organization of Plasmid DNA into Nucleosome-Like Structures after Transfection in Eukaryotic Cells. *Biotechnology & Biotechnological Equipment,* **23**(1), 1044–1047.

[93] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot,

E., Blüthgen, N., Dekker, J., and Heard, E. (may, 2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature,* **485**(7398), 381–5.

[94] Deato, M. D. E. and Tjian, R. (sep, 2007) Switching of the core transcription machinery during myogenesis. *Genes & development,* **21**(17), 2137–49.

[95] Horwitz, K. B., Zava, D. T., Thilagar, A. K., Jensen, E. M., and McGuire, W. L. (aug, 1978) Steroid Receptor Analyses of Nine Human Breast Cancer Cell Lines. *Cancer Res.,* **38**(8), 2434–2437.

[96] Hager, G. L., McNally, J. G., and Misteli, T. (sep, 2009) Transcription dynamics. *Molecular cell,* **35**(6), 741–53.

[97] Tesikova, M., Dezitter, X., Nenseth, H. Z., Klokk, T. I., Mueller, F., Hager, G. L., and Saatcioglu, F. (apr, 2016) Divergent Binding and Transactivation by Two Related Steroid Receptors at the Same Response Element. *The Journal of biological chemistry,.*

[98] Raj, A., Rifkin, S. A., Andersen, E., and van Oudenaarden, A. (feb, 2010) Variability in gene expression underlies incomplete penetrance. *Nature,* **463**(7283), 913–8.

[99] Mar, J. C., Matigian, N. A., Mackay-Sim, A., Mellick, G. D., Sue, C. M., Silburn, P. A., McGrath, J. J., Quackenbush, J., and Wells, C. A. (aug, 2011) Variance of gene expression identifies altered network constraints in neurological disease. *PLoS genetics,* **7**(8), e1002207.

[100] Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G., and Lieb, J. D. (apr, 2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature,* **484**(7393), 251–5.

[101] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (aug, 2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.),* **337**(6096), 816–21.

[102] Haeussler, M. and Concordet, J.-P. (apr, 2016) Genome Editing with CRISPR-Cas9: Can It Get Any Better?. *Journal of genetics and genomics,* p. Epub ahead of print.