Identifying chromosomal regions associated with glucocorticoid-regulated gene transcription

by
Delsy Martinez

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Natalia Jura
E438674A382B42F...

Natalia Jura
Chair

DocuSigned by:
Keith Yamamoto
DocuSigned by:
Danica Fujimori
82967A5AE18B4D6...

Keith Yamamoto

Danica Fujimori

Committee Members

**Acknowledgements**

I was fortunate to have two incredible mentors during my time at UCSF who were committed to my growth as a scientist and as a human being. Thank you to JJ Miranda for starting me on my graduate school journey by patiently teaching me all the way back to basics on how to be a better scientist. The time I spent in your lab was not only incredibly valuable for me as a student but always so much fun alongside Sam, Kristin, Amanda and Stephanie. You all helped prepare me to be a working member of Keith Yamamoto's lab. I appreciate the expertise and guidance I received from everyone in the Yamamoto lab, especially Kirk and Matthew. Thank you to Keith, who saw me through to the end of my time here at UCSF. My admiration and respect for you as a scientist and as a person knows no bounds. On behalf of everyone in your lab, at UCSF, and the scientific community, we thank you Keith for your unrelenting service to science. Thank you to my thesis committee Danica and Natalia for your understanding and willingness to help. Thank you to my colleagues here at UCSF for creating a welcoming and helpful atmosphere. Finally, thank you to the Tetrad Graduate program for the opportunity to learn and grow as a scientist.

Para mis padres, Francisco y Marina, gracias por esta opportunidad. Todo lo que logro en esta vida es por ustedes y sus sacrificios. Esta licenciatura es tanto sus logros como la mia porque por ustedes es quien hago la lucha. Being a first-generation Mexican-American is my greatest struggle and my proudest title because I am laying a foundation for our family. The love and support I received from my parents, my brother Alan, my friends and my extended family have powered me through this difficult but rewarding experience. To my high school sweetheart now husband, thank you for letting me pursue my dreams by holding me up and supporting me all the way. In this life, my unconditional love for you is one of the few things I am certain of Vincent.

This degree is just as much yours as it is mine and I am eternally grateful for all of you.

**Contributions**

**The contents of Chapter 2 are modified and reproduced from the following manuscript, which is in preparation:**

Martinez D, Pack L, Miranda J, Yamamoto K. Multiple RNA-seq comparison identifies a robust set of metrics for differential gene expression. *Unpublished, in preparation*.

**The contents of Chapter 3 are modified and reproduced from the following manuscript, which is in preparation:**

Ehmsen K, Martinez D, Wissink E, Knuesel M, Bernal T, Chan M, Cooper S, Miranda J, Lis J, Yamamoto K. Multiple glucocorticoid receptor-occupied loci collectively impact a glucocorticoid responsive gene. *Unpublished, in preparation*.

**Identifying chromosomal regions associated with glucocorticoid-regulated gene transcription**

**Delsy Marina Martinez**

## Abstract

Glucocorticoid (GC) response elements (GREs) are genomic segments that confer GC-regulated transcription in by recruiting hormone-bound glucocorticoid receptor (GR) and nucleating assembly of transcriptional regulatory complexes (TRCs). The locations of GR binding, the functionality of those GR occupied regions (GORs) as GREs, and the molecular features and spatial organization that characterize active GREs are gene-, cell- and physiological-context specific, and poorly understood. Moreover, identification of the gene(s) targeted for regulation by a given GRE has been inferred by proximity, or examined outside the normal chromosomal context, rather than rigorously validated. We approached these two issues in two human cell lines with distinct tissue origins, treated or not with a hormonal ligand that activates GR. First, we took a systems approach to examine the GC response, cataloging GORs by ChIP-seq, comparing RNA-seq defined transcriptome datasets from three different laboratories, mapping short bidirectional transcripts by Pro-seq, and assessing higher order genome structure by *in situ* Hi-C. To identify a functional GRE, we focused on a single 1.4 Mb topological domain bearing a GC-regulated gene and multiple GORs, and used Cas9 mutagenesis for in-genome GOR editing, coupled with transcriptional analysis to assess GRE activity and identify target gene(s). Our work established an experimental and analytic workflow for identification of robust sets of GC-regulated genes, and for unequivocal determination and validation of GRE activity. We found some but not all of the GORs dispersed across the topological domain contributed to GRE activity, the GRE directly regulated only one or two of the seven genes within the domain, and that features such as bidirectional transcripts or chromosome looping were seen at some but not all functional GORs. These results are consistent with context-specific combinatorial assembly of TRCs into functional GREs, which together enable GCs to orchestrate organismal developmental and physiological actions comprised of gene- and cell-specific transcriptional regulatory events.

**Table of Contents**

**List of Figures**

**List of Tables**

## Chapter 1: Introduction

Specific patterns of gene expression are required for developmental and physiological processes. Key players in this arrangement are genomic response elements, DNA segments that bind transcriptional regulatory factors (TFs) and nucleate assembly of multifactor transcriptional regulatory complexes (TRCs) to activate or repress target gene transcription. Advances in genomic technologies enable description of chromatin structure and putative response elements both at specific loci and genome-wide, but defining molecular determinants of their regulatory activities is greatly complicated by variation dependent on gene-, cell- and physiologic-context. Moreover, it appears that response element regulatory activities must be assessed in their normal chromosomal environments, suggesting that genome editing methodologies provide the only viable strategy for functional dissection, and for identification of cognate target genes. It seems likely that response element activities in each specific context may derive from unique combinations of bound factors and other molecular features, thus generating context-specific regulation. Thus, very few response elements and target genes have been unequivocally identified, despite commonplace assignment of genomic regions as enhancers based on proxy datasets (Halfon et al., 2019). By extension, the functions of bound factors and molecular features at a *bona fide* putative response element have not been established, and the gene(s) inferred to be regulated by putative response elements remain unknown.

Several molecular features have been claimed to correlate with response element activities. Open chromatin, genomic regions highly accessible to nucleases, and presumably to TFs, have been described, commonly together with particular histone modifications, such as H3K27Ac and H3K4me1 (Shlyueva et al., 2014). Other investigations cataloged TF binding

sequence (TFBS) motifs, or motifs for multiple different TFs tightly clustered in genomic segments (Yanez-Cuna et al., 2012; Meireles-Filho et al., 2009), and monitored their occupancy by corresponding TFs or by coregulatory factors recruited by them (Rogatsky et al., 2003; Weikum et al., 2017). More recently, short bidirectional transcripts, so-called eRNAs, have been mapped at candidate response elements (Halfon et al., 2019).

In addition to these molecular features, at least two classes of higher-order genome structure have been suggested to be relevant. First, chromatin loops, up to 450 kb, that appear to bring into physical proximity certain promoters of regulated genes and putative response elements (Fraser et al. 2009, Bonev et al. 2016; Kaduake et al. 2009, Pombo et al. 2015); while spatial proximity is a tempting determinant of activity and target gene identity, direct studies have shown that it is not sufficient as an indicator of regulatory function (Shlyueva et al., 2014). Second, topologically associating domains (TADs), which are typically demarcated by bound Cohesin complex and CCCTC-binding protein (CTCF), and contain characteristic chromatin modifications and histone marks between these boundaries (Rao et al. 2014). Disruption of TAD boundaries can affect expression of nearby genes and promote disease states (Matharu et al., 2015), and a common speculation is that TADs define interaction zones that constrain the range over which response elements can act. TAD boundaries appear to be generally conserved across several cell types and species (Rao et al., 2014). The domains are classified as euchromatin-like, compartment A, or heterochromatin-like, compartment B (Lieberman-Aiden et al. 2009), and are thought to enable intra- but not inter-domain looping interactions (Dixon et al., 2012). However, due to the lack of standard criteria for defining TAD boundaries, different researchers have assigned domains across a range from 40 kb to 3 Mb.

Although the molecular features and higher order structures described above comprise a provocative roster of correlates to putative response element function, the overarching problem is that response elements and their target genes have themselves not been functionally validated. Therefore, we have set out in this work to begin to define functional response elements and their target genes. Our approach is to focus on the actions of a single TF, the human glucocorticoid receptor (GR), the founding member of the nuclear hormone receptor family, and likely the best characterized metazoan TF. GR is constitutively expressed in virtually all vertebrate cells (Weikum et al., 2017), residing inactive in the cytoplasm until it binds a glucocorticoid (GC) ligand (such as cortisol, the natural human hormone or dexamethasone (dex), a synthetic GC drug), whereupon it translocates into the nucleus, binds to context-specific genomic GC response elements (GREs) (Chandler et al., 1983) and confers gene-, cell- and physiologic-context specific transcriptional regulation (Yamamoto et al., 1985, Yamamoto et al., 1998).

Ligand-gating of GR activity allows candidate GREs and target genes to be inferred in comparative experiments carried out in the presence and absence of dex. For example: (i) GR ChIP-seq reveals thousands of genomic GR occupied regions (GORs) in ligand-treated cells (Reddy et al., 2009; Encode Consortium 2012); (ii) microarray analyses implies hundreds of genes either induced or repressed upon dex treatment, many of them cell-type specifically, in lung carcinoma (A549) versus osteosarcoma (U2OS) cells (Rogatsky et al. 2013).

The context specificity of GR action has been discussed as a paradox in which this single TF controls a precise transcription program in a given setting, but displays facile plasticity, dramatically changing binding sites and target genes when the setting is altered (Weikum et al.,

2017). A consequence is that GR action must be analyzed both at a systems level, e.g., to identify all genes regulated in a given context, and locus specifically, e.g., to describe the determinants and mechanism of action of an individual GRE. In the current work, we employ both approaches, describing whole genome approaches to defining TADs, chromosome loops, protein-coding and non-coding transcripts in presence and absence of dex, with an emphasis on definitive target gene identification; at the single locus level, we use genome editing procedures to unequivocally identify a functional GRE and to characterize its activities.

## Chapter 2: Genome-wide analysis of GC response

### Introduction

GR is expressed in virtually all vertebrate cells, but its actions are highly context specific, *e.g.,* mediating immunosuppression and anti-inflammation in immune cells, modulating glucose and lipid metabolism in liver, reducing bone and muscle mass, driving lung maturation and surfactant biosynthesis, promoting cell proliferation in the dentate gyrus of the hippocampus. The implication is that a single DNA-binding TF, potentiated by a single hormonal ligand (cortisol in humans, or a synthetic homolog such as dex), is somehow controlling transcription of distinct batteries of genes in different cell contexts (Weikum et al., 2017). Thus, visualizing the spectrum of candidate GR-target genes and GREs in a given context could be achieved by various systems approaches that compare, for example, full transcriptomes and genomic structure from hormone treated and control cells.

RNA-seq provides a sensitive quantitative strategy for monitoring transcription at the whole genome level. A typical experimental workflow involves RNA extraction, RNA fragmentation and reverse transcription, library construction and sequencing (Han et al., 2015). The computational and systems biology that follows depends upon the end goal, be it identifying new transcripts or alternative splicing analysis. RNA-seq has commonly been employed to measure differential expression, in which statistically significant differences in read counts are detected between two experimental conditions (Anjum et al., 2016). Unfortunately, however, agreement has not been achieved on a standard protocol, analysis pipeline and statistical metrics to identify differentially expressed genes (DEGs) or to infer biological relevance. The ENCODE Consortium established standards, guidelines and experimental practices for RNA-seq, *e.g.,*

information to report for each sample, number of replicates and sequencing depth, but these guidelines have not been widely adopted, nor do they address the computational and system biology aspects.

While procedural and computational differences are known to affect RNA-seq results (Li et al., 2014, T'Hoen et al., 2013, Khanin et al., 2013), it has not been generally considered whether statistical metrics for defining DEGs (*e.g.*, log2FoldChange > 1) may exclude biologically relevant transcripts, or whether such tools should or should not be deployed prior to a systems biology step. Currently, the False Discovery Rate adjusted p (q) value<0.05 is a commonly used cutoff value for differential expression tests but is not guaranteed to be the first metric used to filter datasets. Instead, emphasis is placed on log2FoldChange values being greater than a user-specified threshold, but there is no *a priori* reason that a large log2FoldChange is more biologically relevant than a small log2FoldChange (Zarse et al., 2012).

With these concerns in mind, we set out to compare RNA-seq data from dex-treated and control human A549 cells, collected in three different laboratories, but analyzed through the same computational pipeline, with the gene lists subjected to pathway analysis. We sought to establish and justify the use of metrics in a specified order that represents the biology of the glucocorticoid response in those cells.

In addition to this transcriptome determination, we carried out two further systems analyses. First, we sought to map short, labile bidirectional transcripts, enhancer RNAs (eRNAs) or distal transcribed elements (dTREs), suggested to be selectively expressed at putative

response elements. Because eRNAs are not detected by standard RNA-seq, we collaborated with John Lis (Cornell University, Ithaca, NY) to perform Precision nuclear Run-On Sequencing (Pro-seq) in A549 and U2OS cells, with and without dex treatment. Pro-seq enables quantitative tracking of nascent transcripts genome-wide at nucleotide resolution (Wissink et al., 2019), which in turn identify distal transcribed elements.

Finally, to provide a rational metric for defining the genomic segment searched for GREs that regulate a given target gene, we used *in situ* Hi-C to visualize the three-dimensional genome architecture in intact A549 and U2OS nuclei, inferring higher order chromatin structure, for which we suggest a standard criterion for demarcating TADs.

## Results & Discussion

### Identification of GR-regulated genes in one cell and physiologic context

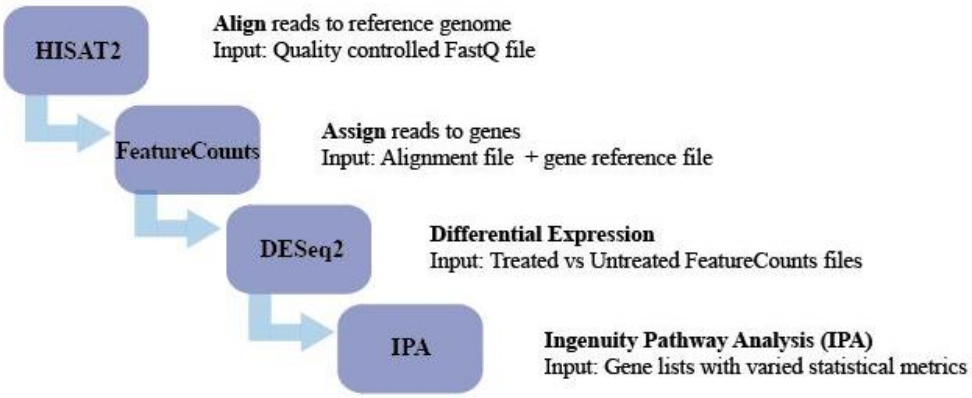
*Description of RNA-seq datasets analyzed*

We examined three RNA-seq datasets produced by three laboratories from different research institutions (Figure 2.1; denoted as D1, D2, and D3). Each dataset consists of three biological replicates A549 cells treated with 100 nM dex or vehicle (EtOH) for 4 hrs. Each dataset sought to identify GC regulated genes—but acquisition of each entailed various pre-analysis (wet lab) differences (See Materials and Methods). Experimental factors that can affect differential expression analysis range from RNA extraction methods to sequencing depth, and can even be as seemingly trivial as the serum source, due to FBS-associated RNA contaminants (Wei et al., 2016).

A

| Dataset (Institute) | # of treated vs un-treated replicates | Illumina Sequencing Platform System | Single vs paired end, read length | Avg. # of sequencing reads | Avg. # of mapped reads |
|---|---|---|---|---|---|
| D1 (Gladstone) | 3 vs 3 | HiSeq4000 | SE, 50 | 36 x$10^6$ | 25 x$10^6$ |
| D2 (UCSF) | 3 vs 3 | HiSeq2500 | SE, 100 | 30 x$10^6$ | 30 x$10^6$ |
| D3 (Duke) | 3 vs 3 | HiSeq2000/ 2500 | PE, 51 | 78 x$10^6$ | 61 x$10^6$ |

B




C

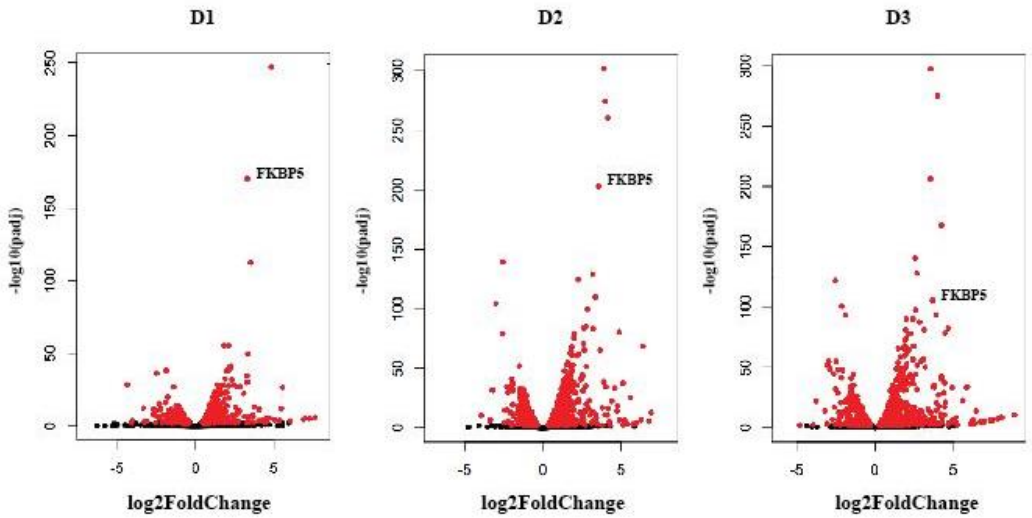



**Figure 2.1: Overview of experimental design of RNA-seq datasets from multiple laboratories.** (A) Description of RNA-seq datasets of A549 cells treated with and without 100 nM dex for 4 hours from 3 different laboratories. (B) Computational and systems biology pipeline devised for determination of differentially expressed genes that are glucocorticoid regulated. (C) Volcano plots. Red dots signify differentially expressed genes with q <0.01. *FKBP5*, a canonical glucocorticoid regulated gene, is present in all 3 datasets.

Several groups have probed sources of experimental variation extensively, such as mRNA (polyA+) enrichment (Zhao et al., 2018) and sequencing platform (Li et al. 2014). D1 used total RNA for cDNA synthesis but with a kit whose proprietary combination of enzymes allows for preferential priming of non-rRNA sequences and therefore a reduced number of reads from rRNA, whereas D2 and D3 isolated mRNA from total RNA. rRNA and tRNA make up >95% of total RNA and does not allow for efficient transcript/gene detection if not removed or biased against by selective priming. Poly(A) selection provides good recovery of mRNAs but biologically relevant RNA species lacking poly(A) go undetected. Hence, each RNA selection approach has advantages and disadvantages. All three datasets used the Illumina sequencing system, HiSeq. D3 doubled the amount of reads due to paired end (PE) sequencing, and D2, though single end (SE), used 100 bp read length; both approaches provided better alignment accuracy.

*Computational analysis pipeline*

We identified differentially expressed genes (DEGs) in each individual dataset with the pipeline outlined in Figure 2.1. We used HISAT2, a fast, efficient pipeline that employs splice junctions and hierarchical indexing for fast alignment to the hg38 human reference genome (Kim et al., 2015) rather than *de novo* transcriptome assembly, which was unnecessary in our case. Sequencing depth would further limit our ability to differentiate between transcripts/isoforms of genes for some of the datasets, therefore, we employed a gene-level summation using Featurecounts (Liao et al., 2013). This read summation process maps the number of reads to genomic features (*e.g.*, exons) with a hg38 refFLAT GTF file containing chromosomal coordinates of exons and coding regions we provided along with the alignment file from HISAT2. Finally, we used DESeq2 (Love et al., 2014) to identify DEGs between vehicle- and dex-treated samples.

As a quality control measure, principal component analyses (PCA) of individual datasets revealed strong separation of vehicle and dex-treated samples across PC1, which provided the most variance. While each dataset identified GC-regulated genes, a PCA with the combined datasets displayed clustering based on dataset instead of treatment conditions as the primary source of variances (Figure 2.2). Sequencing read type and length conditions contributed to the variance as two of the three datasets displayed greater alignment due to PE sequencing or longer read lengths, which increased accuracy in alignment and therefore allocation of counts. Thus, combining samples from different datasets without accounting for variance across datasets is uninformative. When accounting for batch effects, the effect of treatment conditions by combining treated and vehicle samples from all of the datasets displayed 98% of all DEGs from the individual datasets. We developed an approach to infer "robust DEGs", which are GC-responsive despite variance across the three datasets, using statistical metrics that probe biologically relevant pathways.

*Metrics for assessing robustness*

1. The False Discovery Rate adjusted q value is more appropriate than p value, given the need for multiple correction testing. A p value $< 0.05$ gives evidence against the null hypothesis, measuring the likelihood that a gene is significantly differentially expressed in the vehicle versus dex-treated samples. The q value is superior when dealing with thousands of genes, because it measures the likelihood of false positives; a $q<0.05$ threshold is a strong determinant of statistical significance. With a $q<0.05$, we identified 733, 1450, and 2115 differentially expressed genes in D1, D2, and D3, respectively (Figure 2.3). With a more stringent threshold of $q<0.01$, we lose 28% of regulated genes across all the datasets.

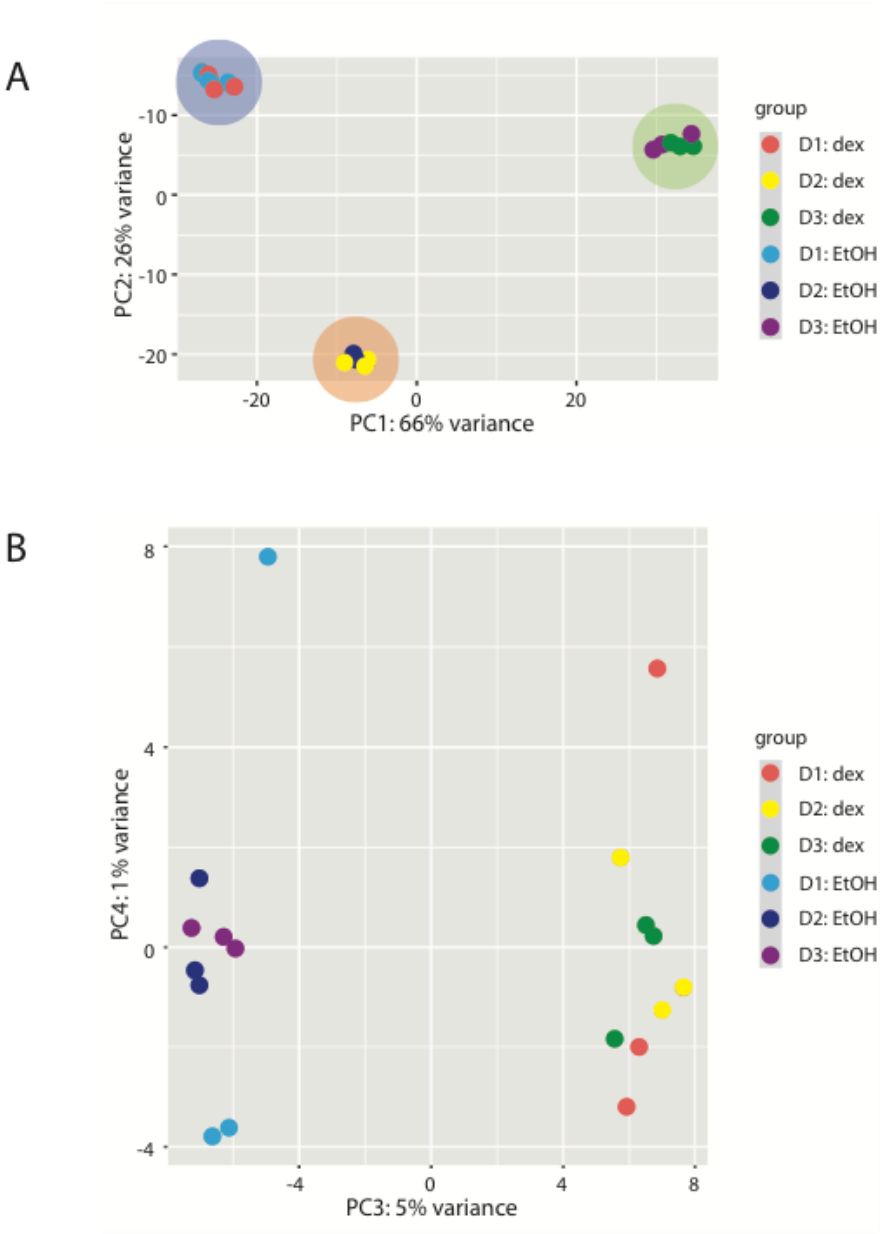
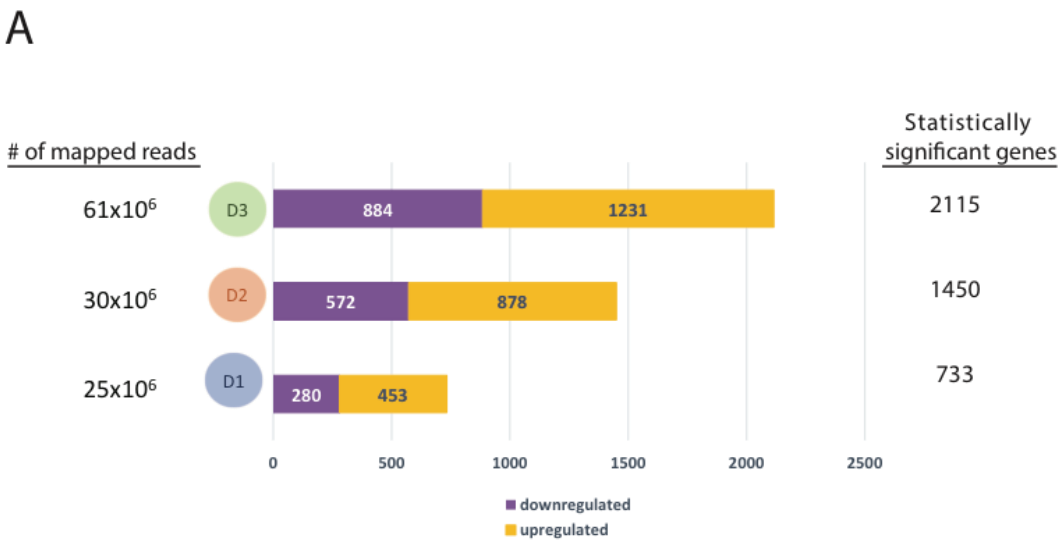



**Figure 2.2: Combined PCA of RNA-seq datasets primarily differentiates individual datasets instead of treated versus untreated samples.** (A) PC1 and PC2, which account for the majority of the variance, differentiate datasets whereas (B) PC3 and PC4 differentiate treatment.



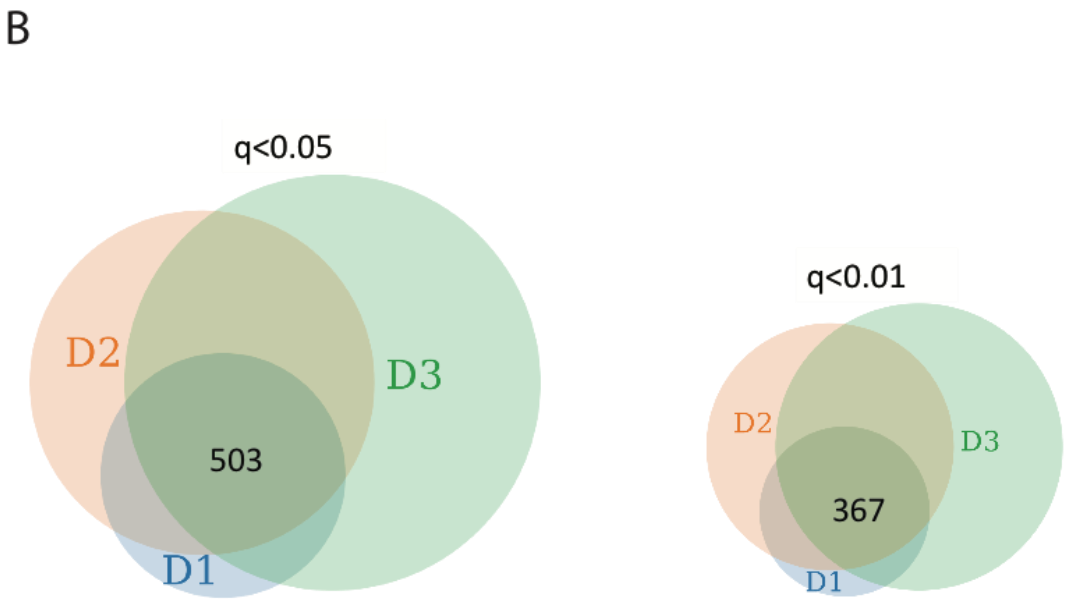




**Figure 2.3: Filtering list of differentially expressed genes for statistical significance and consistency.** (A) Number of differentially expressed upregulated (gold) and downregulated (purple) genes with q<0.05. (B) Venn diagrams showcasing consistent genes from all 3 datasets with 503 genes with q<0.05 (left) and 367 genes with q<0.01 (right).

*Consistent biological behavior across datasets and computational tools*

2a) An additional criterion for robustness was consistent biological behavior. A gene met this metric if it was statistically significant in all three datasets and was consistently upregulated or downregulated in all three datasets. With a $q<0.05$ we identified 503 genes versus 367 genes with $q<0.01$ that were consistently upregulated or downregulated in every dataset (Figure 2.3) (Table 2.1); no genes switched between dex-activation and -repression at $q<0.05$. The mean log2foldchange value of these robust DEGs was $\pm 0.3$ at $q<0.05$, and $\pm 0.4$ at $q<0.01$, which is equivalent to a greater than 1.23-fold change for differentially expressed genes in all datasets. The range of the mean log2FoldChange for these 503 genes is +6.87 to -3.48, which is equivalent to 112-fold increase and 11-fold decrease in expression, respectively. DESeq2 by default finds an optimal value at which to filter low count genes. Unsurprisingly, genes that met the default low count filter but had low mean normalized counts in one or two of the datasets commonly had large deviations in mean log2FoldChange values. Examples of upregulated genes in all three datasets were *TFCP2L1* and *ACSL1,* and consistently down-regulated genes included *PLK2* and *IER5,* all of which align with known effects of GCs on lipid metabolism and stress response (D'Ippolito et al., 2018).

2b) To determine biological relevance of thresholds, and justify cutoffs, we used QIAGEN's web-based software application, Ingenuity Pathway Analysis (IPA, https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/). The manually curated content of the Ingenuity Knowledge Base efficiently determined the biological context of specific gene lists, and assessed effects of altering q values or log2foldchange values on pathway ranking and presence. IPA determines pathway significance with $p<0.05$ values reflecting the

likelihood of non-random overlap between inputs and pathways. D1 had 206 significant IPA pathways, with statistically significant DEGs at $q<0.05$. Interestingly, though D2 and D3 had more statistically significant DEGs at $q<0.05$ compared to D1, D2 and D3 had 194 and 197 pathways, respectively, and they were neither complete subsets of each other nor of D1. Table 2.2 compares the z-scores of pathways generated from genes with $q<0.05$ in individual datasets.

Analyses of pathways generated from the robust gene list (Table 2.1) under different thresholds, $q<0.05$ and $q<0.01$, altered the p value ranking of pathways, but few pathways were gained or lost. The number of statistically significant pathways overlapped by greater than 80% regardless of q value threshold (Figure 2.4). The top ranking pathways were Glucocorticoid receptor signaling (p value=2.76E-07), Colorectal cancer metastasis signaling (p value=1.17E-06), NRF2-mediated oxidative stress response (p value=2.86E-06), IL-7 signaling (p value=6.69E-06) and p53 signaling (p value=7.96E-06 (Table 2.3). Acknowledging that IPA pathways containing more well-documented molecules, such as cancer pathways that are highly studied and reported, can skew the significance of dataset gene lists and pathway rankings, we were nevertheless pleased that GR signaling was top-ranked.

2c) Another criterion for robustness involved using another parametric differential expression tool, EdgeR (Robinson et al., 2010, Liu et al., 2015), to compare the number of DEGs produced in the individual datasets with our set metric of $q<0.05$. EdgeR yielded >80% of the DEGs produced from DESeq2, which is a common range between the two different tools (Schurch et al., 2016).

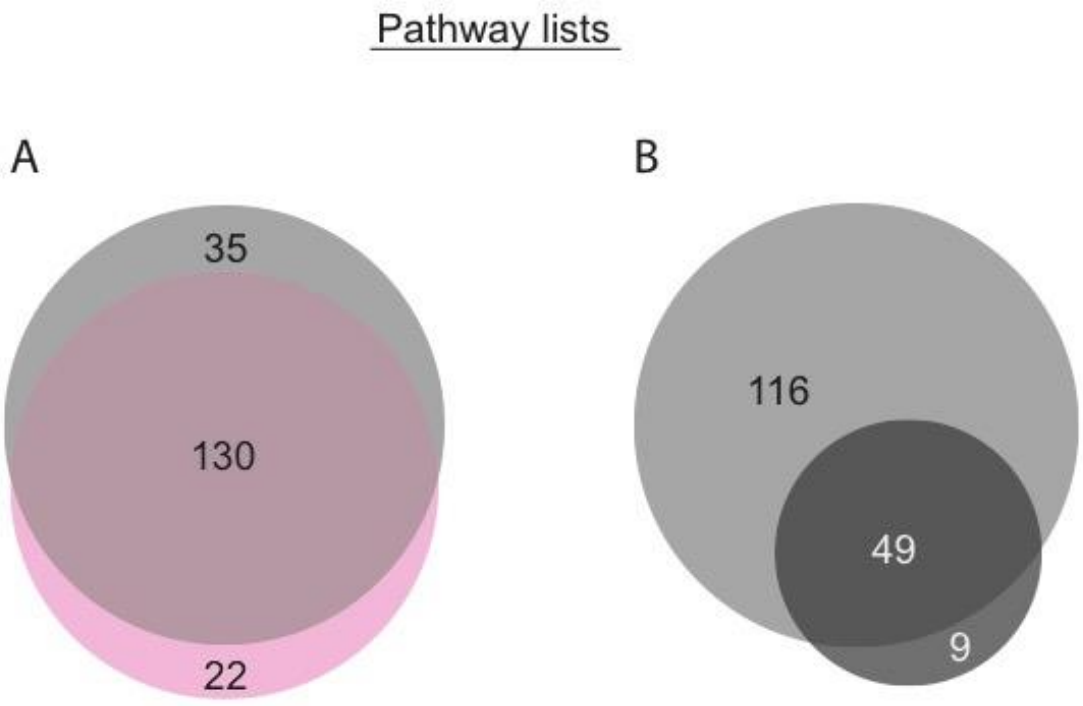



**Figure 2.4: Significant IPA pathways consistent with varying q values but massive differences observed with imposed log2FoldChange (L2FC) threshold.** (A) Overlap of significant* pathways generated from robust differentially expressed genes with a q<0.05 (Grey) and q<0.01 (Pink). (B) Overlap of significant* pathways generated from robust differentially expressed genes with a q<0.05 (Grey) manually imposed log2FoldChange (L2FC) $\pm$1 (Dark Grey). *IPA determined p value in which p<0.05 reflects likelihood overlap between inputs and pathway are not random.

Importantly, the commonly used log2FoldChange >1 threshold is not a sound filter for biological relevance. Large fold changes do not guarantee biological relevance, while genes with small fold changes may be biologically relevant yet discarded. With our robust gene list at the $q<0.05$ threshold, we determined that a log2FoldChange$\pm$1 cutoff excluded 50% of genes and >60% of pathways (Figure 2.4). We could not confidently manually impose log2FoldChange filters because the biological processes, molecular functions and components generally remained consistent while varying q value thresholds but varied drastically when filtering with a stringent log2foldchange threshold. These biologically relevant pathways are derived from what is currently available in the literature by IPA and would take an additional effort to distinguish if the remaining pathways, when imposing log2foldchange thresholds, selected for essential pathways. The likelihood that a large log2foldchange selects for core pathways is contrary to published work showing genes not highly differentially expressed are biologically relevant (Zarse et al., 2012). Finally, it should be noted that RNA levels may not predict protein levels, as processes downstream of transcription may also be regulated (Vogel et al., 2012).

*Guidelines for DEG identification*

Based on our analysis of three RNA-seq datasets, we propose guidelines for metrics to identify robust DEGs: (1) filter by q value; (2) filter by consistent behavior; (3) allow the q value to determine the log2foldchange. We opted for a more inclusive $q<0.05$ value, as a 5% chance of a false positive seemed acceptable against the risk losing information at a more stringent threshold; notably, the pathway analysis using a $q<0.01$ setting overlapped strongly with that at the $q<0.05$ setting. It is strongly advised to have at least 3 replicates for statistical validation of genes whose differential expression is reliable within a dataset. Our analysis of three datasets, which allowed

comparisons within and across datasets, captured >500 DEGs, with some genes unique to each dataset and some differences in pathway analysis; nevertheless, we found 503 consistent DEGs. Of course, it is not customary to compare RNA-seq datasets from different labs, but our analysis confirms that nonidentical results can emerge from slight differences in experimental and analytic approaches.

In summary, we identified genes that were statistically significant based on a q value threshold of less than 0.05 or 0.01, which set a log2FoldChange cutoff to $\pm$ 0.3-0.4 in each dataset. We further distinguished genes that were robust to experimental variation (library preparation, sequencing system, PE vs SE, and read length), as assessed by consistency in expression in all 3 datasets. We ran these robust gene lists with varying q values and found that the IPA pathways were generally conserved. In contrast, imposing a stringent log2FoldChange threshold severely constrained the number of pathways with published findings linking them to the genes on our list. Therefore, the log2foldchange threshold is set by q value, rather than arbitrarily assigned.

**Bidirectional transcripts in response to GC exposure in two cell lineages**

We also looked at non-protein coding transcripts, as it has been claimed that short, bidirectional, noncoding transcripts from so-called distal transcribed regions (dTREs) are characteristic of functional response elements (hence, have been denoted as enhancer RNAs, eRNAs). We treated U2OS and A549 cells with ethanol, 1 nM, or 100 nM dex for 45 min, then performed PRO-seq (Erin Wissink, Cornell University). dTREs were identified using dREG (Wang et al., 2016) and differential expression was performed with DESeq2 (Love et al., 2014). We found that the overall dTRE landscape was A549- and U2OS-specific. PRO-seq identified

~63,175 constitutive dTREs in A549 cells and ~ 51,234 in U2OS cells (Figure 2.5). Intriguingly, a fraction of these dTREs (593 (0.9% of detected dTRE loci) in A549, 2,055 in (4% of detected dTRE loci) in U2OS are differentially responsive to GC signaling (100 nM dex relative to EtOH) at a p value ≤0.05 (in A549, log2FoldChange max = 8.1 [chrX:86,147,390-86,147,800], min = -5.5 [chr10:100,696,430-100,696,959], median = 2.89, mean = 2.31 ± 2.88; in U2OS, log2FoldChange max = 7.2 [chrX:43,655,190-43,655,640], min = -5.1 [chr10:100,347,760-100,348,300], median = 1.82, mean = 1.47 ± 2.14). Further study is required to assess whether eRNAs play a role in GRE activity (See Chapter 3).

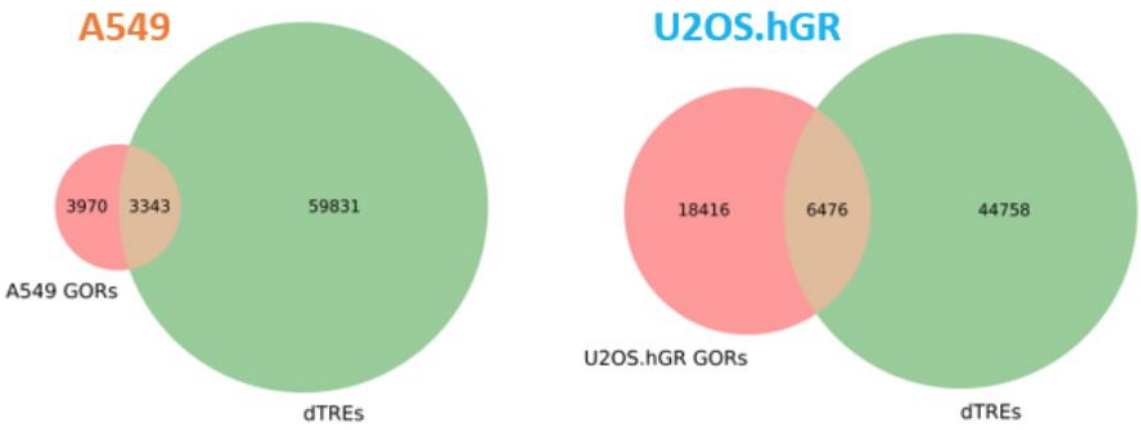



**Fig 5: Distal biredictional transcripts (dTREs) overlap with GORs distinctly in A549 and U2OS cell lines.** 46% of A549 GORs overlap with dTREs whereas 26% of GORs overlap with dTREs in U2OS cells.

**Human genome structure in response to GC exposure in two cell lineages**

To assess whether chromosome topological interactions (large-scale topological domains and/or smaller-scale intrachromosomal looping) may be related to or functional in GC regulation, we performed *in situ* Hi-C in both A549 and U2OS cell lines treated with 100 nM dex or EtOH vehicle for 1.5 hr. Hi-C relies on DNA proximity to produce genome-wide DNA-DNA contact maps.

We first examined interchromosomal interactions between whole chromosomes in the human genome for either cell line. Heatmaps display the observed interactions between chromosomes relative to random expectations (Figure 2.6; clusters of red indicate preferential association between chromosomes whereas blue clusters indicate avoidance). As expected from prior reports (Lieberman-Aiden et al., 2009), gene-rich chromosomes preferentially associated with each other and to a lesser degree, gene-poor chromosomes also associated. We subtracted the dex- and vehicle-treated samples' interaction frequencies for each chromosome after normalization by sequencing depth. The differences in the interaction frequencies did not favor association or avoidance between chromosomes, and dex treatment did not detectably alter chromosomal interactions.

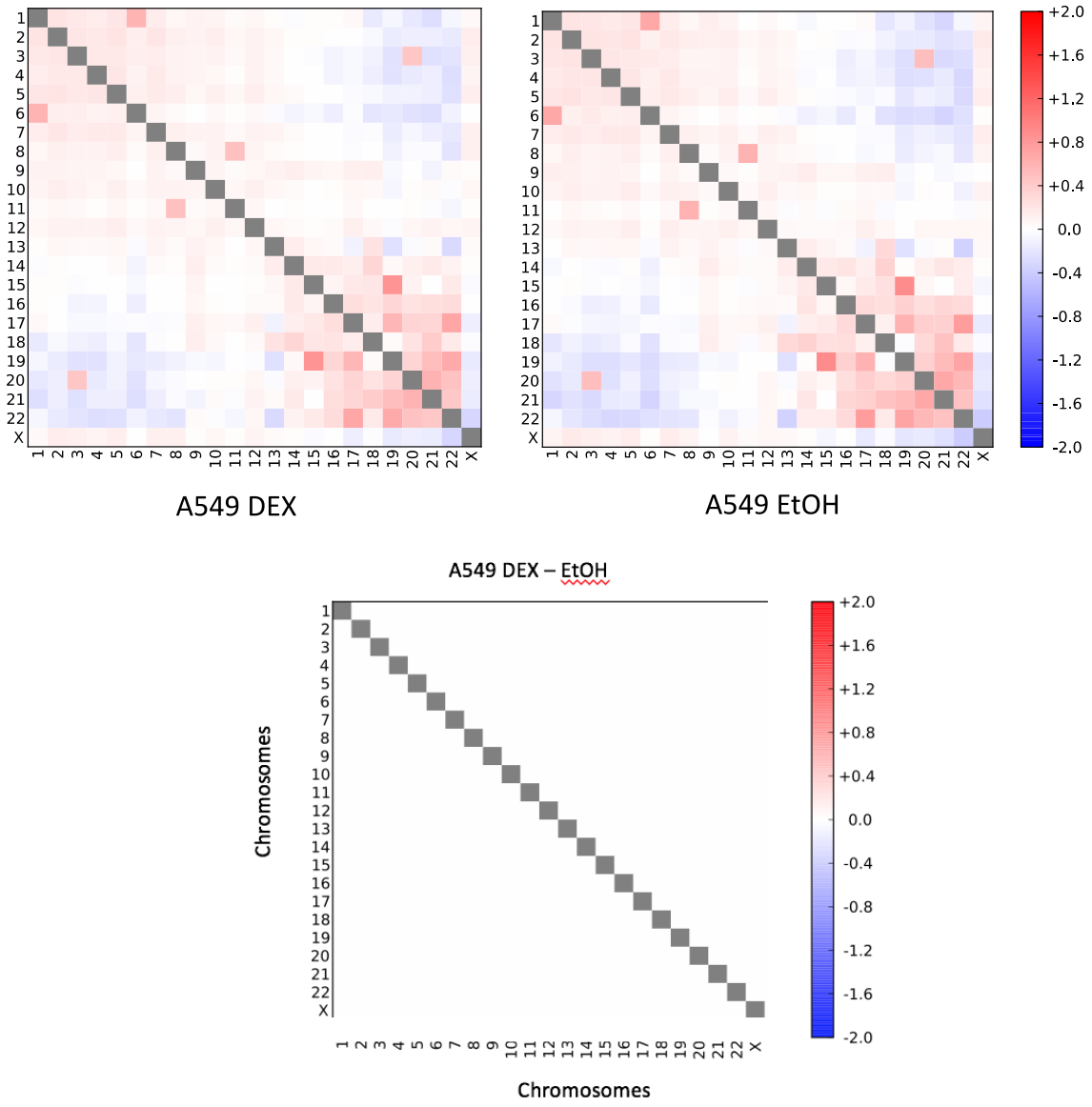



**Figure 2.6: Interchromosomal contacts do not appear or disappear with glucocorticoid treatment in A549 cells**. Interchromosomal interactions in A549 cells treated with 100nM dex (left) or vehicle (right). Heatmaps of chromosome association where observed counts are normalized against random expectation and shown on log2 scale. Red indicates enrichment and blue indicates depletion of interactions. A549 dex and EtOH interchromosomal heatmaps were normalized by valid read pairs and subtracted (bottom).

Intrachromosomal heatmaps display the same chromosomal region mapped to itself. Interactions on the diagonal are enriched, because these regions are close in three-dimensional space, whereas off-diagonal interactions represent long range interactions. At 100 kb resolution, dex did not provoke appearance or disappearance of putative TAD structures larger than 2 Mb encompassing GC-regulated genes in either A549 or U2OS, consistent with the previous observation of conservation of TAD boundaries and loops across several cell types and species (Rao et al., 2014). Figure 2.7, upper panels, display intrachromosomal Hi-C data in dex-treated and control A549 cells for a 10 Mb region of chromosome 10 in which each pixel is a 100-kb bin of the genome and a putative TAD is visible with the GC-regulated gene, *ANKRD1* promoter at the TAD border. Figure 2.7, bottom panel, examines a 2 Mb segment of this region in the dex treated sample.

We sought to use *in situ* Hi-C to assess in A549 and U2OS cells how a GC regulated gene may be influenced by DNA elements potentially within a TAD via chromatin loops. While we secured reproducible results, resolution was limited to 100kb, so we could not be confident that we were capturing chromatin loops that might bridge candidate GREs and cognate promoters, or even relatively small topological structures that might limit the search space for candidate response elements. For that reason, we relied on intrachromosomal Hi-C contacts at higher resolution (~5kb), in the A549 cell line provided by T. Reddy (Duke University). In the case of GC response, higher order genome structure appears to be 'pre-wired' in that chromatin loops detected before and after GC exposure are similar, albeit with some changes in interaction frequency (D'Ippolito et al., 2018).

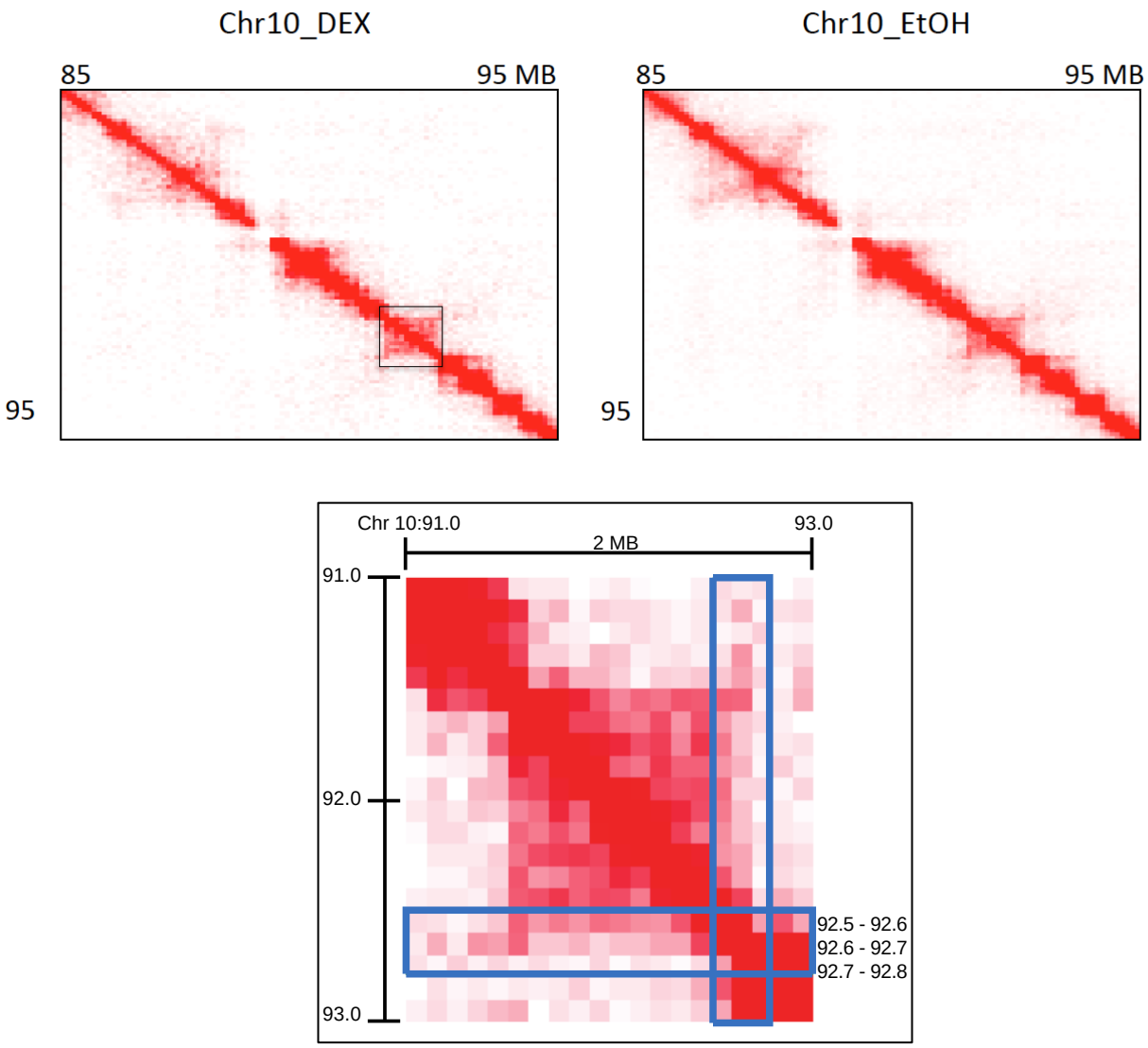



**Figure 2.7**: **Putative TAD in A549 cells containing glucocorticoid regulated gene**. Intrachromosomal interactions within 10 Mb region of chromosome 10 in A549 cells treated with 100nM dex (left) or vehicle (right) normalized by valid read pairs. Red indicates enrichment of interactions and every pixel corresponds to 100 kb region. Bottom box: Zoom in of putative ~2 Mb TAD containing GC regulated gene, *ANKRD1*, in blue box at potential TAD border.

## Material and Methods

### RNA-seq

*Experimental Biology*

For D1:

Two T-225 flasks of A549 cell lines were maintained at 37 °C with 5% $CO_2$ (v/v) in DMEM H-16 low-glucose media supplemented with 5% (v/v) fetal bovine serum. Confluent cells were treated with 100nM dex (D4902-25MG) or EtOH-vehicle for 4 hr (old media was switched out for dex- or vehicle-containing media). RNA from $3x10^6$ cells was purified using Qiagen RNeasy mini kit with QIAshredder columns and optional DNAse step with final elution in 50 μL Ambion RNase-free water; eluants were snap frozen in liquid $N_2$ and stored at -80°C. 15-25 ng total RNA was used to prepare amplified cDNA using Nugen's Ovation RNAseq system V2 kit. 3 μg cDNA was sheared with an S2 Focused-ultrasonicator (Covaris) set at intensity 5, duty cycle 10%, cycles/burst 200, and time = 60s for 2 cycles for fragment sizes ~100-400 bp. Libraries were constructed with 100ng in the Ovation Ultralow System V2 #1-16 (Part #0344) with 8 cycles amplification and quantified on a 2100 Bioanalyzer System (Agilent) with High Sensitivity DNA Kit. Each library was sequenced on a HiSeq 4000 (Illumina) using single reads of 50 bp in length.

For D2:

Refer to Pack, L.R., 2017:

A549 cells were grown in 15 cm dishes using DMEM H-16 low glucose media supplemented with 5% (v/v) fetal bovine serum. Two 15 cm dishes were used for each condition analyzed by RNA-seq: i) Control siRNA/ethanol, ii) Control siRNA/dex. Pool of siRNA acting as a non-targeting negative control (Darmacon, D-0018190- 10-20) were used for the reverse

transfections. 200 pmol of siRNA was diluted in 3.5 mL of Optimem-I media, followed by the addition of 42 µL of RNAiMAX (ThermoFisher) for each 15 cm plate. The mixtures were added to the 15 cm plates and incubated for 20 min at room temperature (rt). Following incubation, $1.3x10^6$ cells were plated in 16.5 mL of standard growth media. The cells were incubated with their respective siRNA pools for 72 hr after which the media was replaced with fresh media containing either 100 nM dex in 0.2% ethanol or 0.2 % ethanol for 4 hr. Following incubation with dex or ethanol, the media was removed and cells were collected by scraping with 1 mL of RLT buffer from the RNAeasy kit (Qiagen). RNA was isolated using Qiashredder and RNAeasy mini columns. RNA quantity was measured using Nanodrop spectroscopy, quality using the Bioanalyzer, and knockdown efficiency using qPCR. mRNA was isolated from total RNA using Oligotex mRNA isolation (Qiagen) as described in the protocol with two modifications: i) following the removal of supernatant from the Oligotex beads, water and OBB buffer were added and the heating, cooling and pelleting steps repeated; ii) beads were treated twice with 70 µL of elution buffer. Following mRNA isolation, rRNA contamination was assessed using the Bioanalyzer . mRNA was precipitated with sodium acetate, isopropanol, and glycoblue and resuspended in 9 µL of 10 mM Tris pH7.0. To fragment the RNA, samples were heated to 95 °C for 2 min followed by 1 µL of fragmentation buffer (Ambion) and incubation at 95°C for 2 min; 1 µL of stop solution was then added. Samples were run on a 10% TBU gel (Invitrogen) at 200 V for 50 min. Gels were visualized by Sybr Gold and 80-120 bp RNA was cut from the gel. The RNA was gel extracted by first pulverizing the gel pieces and then incubating in 300 µL of DEPC water at 70 °C in a Thermomixer. Supernatant was collected through a SpinX column and precipitated with sodium acetate, ethanol, and glycoblue. Following precipitation, the RNA was resuspended in 7 µL of 10 mM Tris pH=7. 1 µL of 10x PNK buffer, 1 µL of Superase Inhibitor,

and 2 µL of PNK were added to each sample, which were then incubated at 37°C for 1 hr. Following incubation with PNK, linker ligation was achieved by adding 6 µL of PEG, 1 µL of linker-1, 1 µL of DTT, 1.1 µL of ligation buffer, and 1.5 µL of Truncated T4 RNA ligase 2 (NEB M0242L). Linker ligation was performed for 2 hr at 37°C. Following ligations samples were precipitated and pellets were resuspended in 8.5 µL 10 mM Tris pH7.0 and run on a 10% TBU gel for 50 min at 200V. The ligated samples were cut from the gel and gel extracted as described. Following precipitation, samples were resuspended in 11 µL of 10 mM Tris pH7.0. 0.8 µL of oCJ200 reverse transcription (RT) buffer was added to the ligated RNA and was incubated at 65°C for 5 min and 35°C for 5 min. Following incubation, 4 µL of 5x RT buffer, 1 µL of DTT, 1 µL of dNTPs, 1 µL of Superase Inhibitor, and 1 µL of reverse transcriptase was added to the RNA primer mixture. Samples were incubated at 52°C for 12 min. Samples were then incubated with 2 µL NaOH for fifteen minutes at 95°C. After boiling, 2 µL of HCl was added to each sample to neutralize the pH. Samples were precipitated in Tris pH8.0, glycoblue, and ethanol, and run on a 10% TBU gel for 1 hr and 20 min at 200V. DNA was imaged and extracted. Samples were resuspended in 15 µL of 10 mM Tris pH8.0 and were circularized through the addition of 2 µL of 10x circ ligase buffer (epiBio), 1 µL of 20 mM ATP, 1 µL of 50 mM $MgCl_2$, and 1 µL of circ ligase (epiBio). Samples were incubated at 60°C for 1 hr and circ ligase was heat inactivated at 80°C for 10 min. After circularization, 3 µL of circ product was PCR amplified using Phusion. The PCR primers were primer 0231 and the indexed primers of interest. PCR conditions included an initial denaturation at 98°C for 30 sec followed by 10 or 12 cycles of denaturation at 98°C for 10 sec, annealing at 60°C for 10 sec, and extension at 72°C for 5 sec. Samples were run on a 8% TBE gel at 180V for 47 min. Amplified products were gel extracted. Following PCR amplification, the quality of the libraries was determined by

Bioanalyzer. Libraries were quantified by qPCR using the KAPA library quantification standards. We generated 5 nM solutions of each library and combined 2.5 μL of each sample for LRPA and LRPB. Samples were sequenced by the UCSF center for advanced technology on an Illumina Hi-seq using Rapid Run single reads of 100 bp in length. The sequencing primer used was oNTI202. Unpublished raw fastq files were provided by Lindsey Pack.

For D3:

Refer to links below for experimental procedures. Raw fastq files were downloaded from Encode.

https://www.encodeproject.org/experiments/ENCSR632DQP/

https://www.encodeproject.org/experiments/ENCSR326PTW/

*Computational Biology*

RNA-seq profiling for three biological replicates from each of three datasets were performed and yielded ~25–60 x$10^6$ mapped sequences. For quality control, mapping and read quantification, we employed the web-based platform Galaxy (usegalaxy.org). FastQC was used to evaluate the quality of reads (Andrew 2010). Raw Fastqs from each dataset were uniformly processed and analyzed, with the exception that Cutadapt (v1.16.3) was used for D2 to remove the 3' sequence (CTGTAGGCACCATCAATATCTCGTATGCCGTCTTCTGCTTG) (Marcel 2011). Reads were mapped using HISAT2 (v2.1.0+galaxy3) to the hg38 genome with default settings. We employed a gene-level summarization using Featurecounts (v1.6.3+galaxy2) with a hg38 refFLAT GTF file (UCSC Main on Human: refFLAT(genome)) containing chromosomal coordinates of exons and coding regions we provided along with the alignment file from

HISAT2 and default settings. Finally, we used R-DESeq2 (v1.22.2) to identify DEGs between vehicle (EtOH) and dex-treated samples in individual datasets.

*Systems Biology*

We submitted gene lists of individual datasets and robust gene list with q value thresholds <0.05 or <0.01 through the Ingenuity Pathway Analysis (IPA) tool. We used the Core Analyses feature to obtain relevant relationships, mechanisms, functions and pathways for a given gene list.

**Pro-seq**

Cells were maintained in DMEM supplemented with FBS and pen/strep. 5 $x10^6$ cells were plated per experiment 24 hr prior to treatment. Media was supplemented with ethanol or the appropriate concentration of dex for 45 min and kept in a 37° C incubator with 5% $CO_2$ during the incubation. Cells were kept on ice during the extraction protocol. Cells were washed with PBS, then incubated in PBS supplemented with 10 mM PBS for 5 min. Cells were scraped and placed in 15 mL Falcon tubes, then washed twice with PBS. Cells were then incubated in permeabilization buffer for 5 min, pelleted, and washed twice with permeabilization buffer. Cells were resuspended in freezing buffer and flash frozen. Run-on reactions with biotin-11-CTP and biotin-11-UTP were performed for 5 min. RNA was isolated, fragmented on ice with 0.2 N NaOH, and underwent buffer exchange with a P-30 column. Three biotin enrichments were performed with Dynabead streptavidin beads, and between enrichments, the 3' adapter was ligated, the 5' end was repaired with RppH and PNK enzymatic treatments, and the 5' adapter was ligated. RT was performed with SuperScript III, and 13 cycles of PCR were used to amplify the library, followed by clean-up with Ampure beads (1.6x ratio).

### *In Situ* Hi-C

Passage modified from Moquin et al., 2017:

*In situ* Hi-C was performed with 5 x$10^6$ cells per experiment as described (Rao et al., 2014), with slight modifications. After end repair and washes, Dynabeads (Thermo Fisher Scientific) with bound DNA were resuspended in 10 mM Tris, 0.1 mM EDTA, pH 8.0, and transferred to new tubes. Sequencing libraries were created from bound DNA by using an Ovation Ultralow library system V2 kit (NuGEN), with one modification. After adapter ligation, because DNA was still attached to the beads, water instead of SPRI beads was added to the reaction mixture. Beads with bound DNA were purified by use of a magnet, washed, and resuspended in 10 mM Tris, 0.1 mM EDTA, pH 8.0. After library amplification, SPRI beads were added as directed to purify the amplified DNA. Quantitation and size distribution of libraries were performed using a Bioanalyzer High Sensitivity DNA kit (Agilent). Fifty-base PE reads were sequenced on a HiSeq instrument (Illumina). Once sequenced, PE reads were aligned to human reference genome by use of the Hi-C User Pipeline (HiCUP), version 0.5.0, using default parameters to generate a set of interactions. We used the human hg19 sequence. The HiCUP processing steps remove PCR duplicates as well as invalid read pairs, including those that are self-ligated or map to identical or adjacent fragments. Only alignments with mapq scores of ≥30 were retained. Data sets contained on average 25 x$10^6$ valid PE Hi-C contacts after quality control filtering.

## Chapter 3: Identification of a glucocorticoid response element and a cognate target gene

### Introduction

GR regulates gene networks that are precisely determined in a given context, yet displays remarkable plasticity as a function of cell type and physiological state. It accomplishes this feat by binding at context-specific genomic sites and provoking assembly of context-specific TRCs, which in turn modulate context-specific processes in mRNA production, such as initiation, release of stalled RNA polymerase II, elongation, splicing, etc. (Weikum et al., 2017). This extreme context specificity enables global regulators like GR to control organismal processes as aggregate outcomes of distinct effects in different cells and tissues, developmental stages and physiologic states. Context specificity also greatly complicates characterization and mechanistic analysis of response element activities, as there is no single set of molecular characteristics, no simple genomic map of functional GREs, no single mechanistic action that can be ascribed to functional GR. Rather, GREs are comprised of context-specific combinations of molecular features, higher order genomic arrangements and TRC components, which together modulate different steps in the transcription of cognate target genes.

A consequence of this complexity is that systems approaches cannot identify functional response elements such as GREs. Most investigations have failed to appreciate this important point, and have used molecular features, higher order genomic organization and spatial proximity as surrogate criteria of response element function and activity. Candidate response elements have also been transferred onto plasmids and their actions measured on linked minimal promoter-reporter gene constructs, despite clear evidence that native chromosomal context and target gene promoter context are critical determinants of response element function.

We conclude that at our present level of knowledge, functional response elements must be defined individually, and must be validated genetically in their normal chromosomal environment. Therefore, in this work, we interrogated our genome-wide datasets to select a gene that is GR-regulated in A549 and U2OS gene, and that resides in a well-resolved TAD bearing multiple GORs. We used CRISPR/Cas9-directed GOR ablation to define the functional GRE, combined with transcript analyses to identify the target gene(s).

**Results & Discussion**

*GR occupancy on the human genome varies across cell lineages*

We mapped by ChIP-seq GR occupied regions (GORs) genome-wide in A549 and U2OS cells, treated with 100 nM dex or EtOH-vehicle for 1.5 hr. We observed 7, 313 GORs in A549, of which 67.4% are shared with U2OS and 24, 891 GORs in U2OS of which 19.8% are shared with A549; > 40 and >70% of the GORs summits encompassed a canonical GBS motif in A549 and U2OS, respectively (Figure 3.1). Clearly, GORs are in substantial excess of GC regulated genes (Figure 3.2).

We focused on a TAD defined by Hi-C contacts occurring within a 1.4 Mb region of chromosome 10 that is conserved across several cell types, including A549 (Figure 3.3). The TAD encompasses 7 coding genes. Only *ANKRD1* is GC responsive in A549, with the nearest GC responsive gene is more than 2 Mb away and outside of the TAD. In contrast, in U2OS, *ANKRD1*and *HECTD2* are dex-responsive within the TAD. *ANKRD1* resides near one boundary, and *HECTD2* is near the center of the TAD. Interestingly, *ANKRD1* is upregulated in A549 cells at 100 nM dex, whereas in U2OS, *ANKRD1* is activated at 1 nM but repressed at 100 nM dex;

*HECTD2* is activated at both dex concentrations in U2OS. The TAD includes 4 and 15 GORs in A549 and U2OS, respectively; all 4 A549 GORs coincide with U2OS GORs.

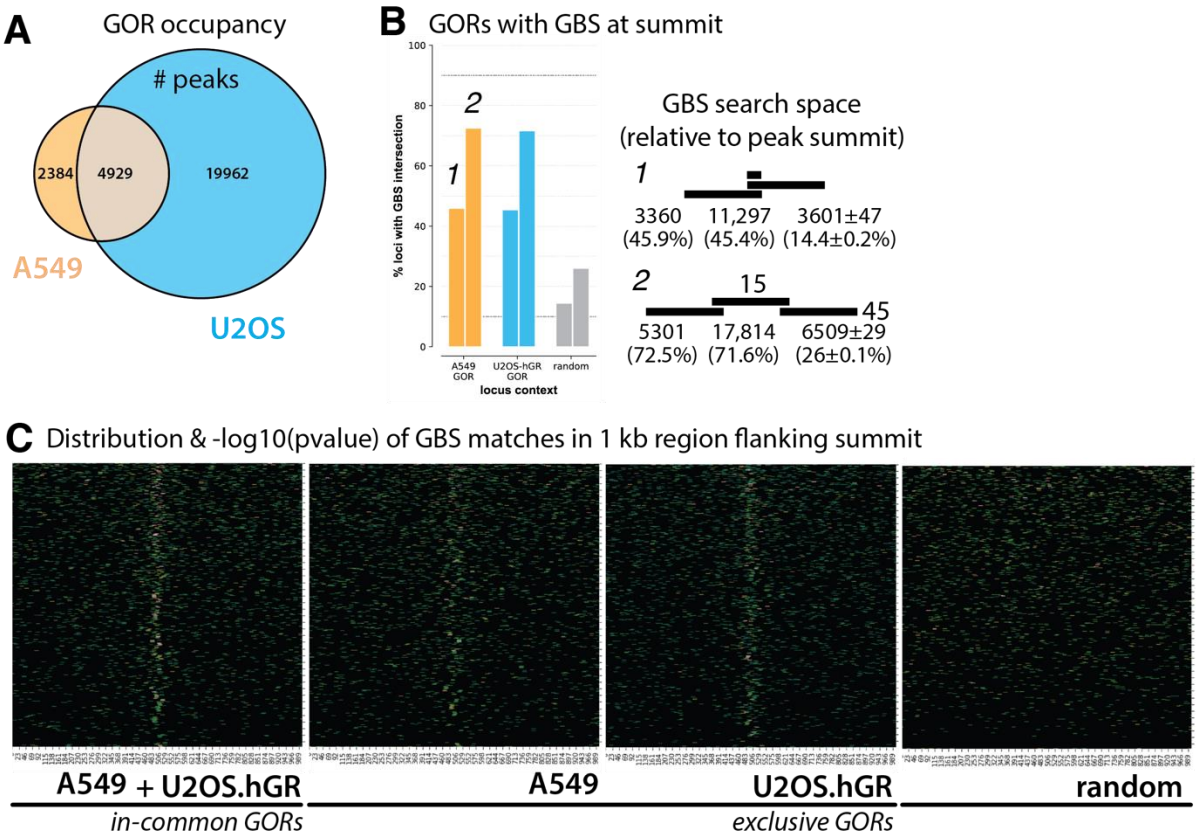



**Figure 3.1: GOR overview in A549 and U2OS cells.** (A) Overlapping and distinct GR occupancy in A549 (orange) and U2OS (blue) cell lines. (B) Number of loci with GBS intersecting with one (*1*) or fifteen (*2*) bp of GOR summit versus random 1 or 15 bp region (grey). (C) Enrichment of GBS motif matches across 1 kb zones centered on GOR peak summit (Heatmap color scale ranges from green to blue to pink (most significant)).

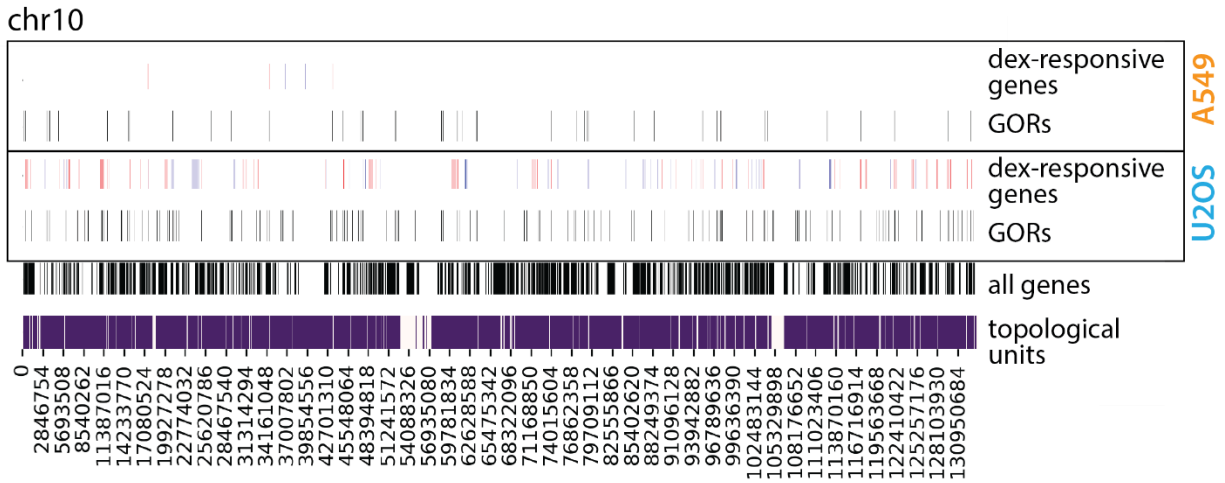



**Figure 3.2: Gene, GOR and TAD overview in chromosome 10.** Example overview of topological units (purple), genes (all vs dex-responsive (blue = downregulated, red= upregulated), GORs in A549 and U2OS cells.

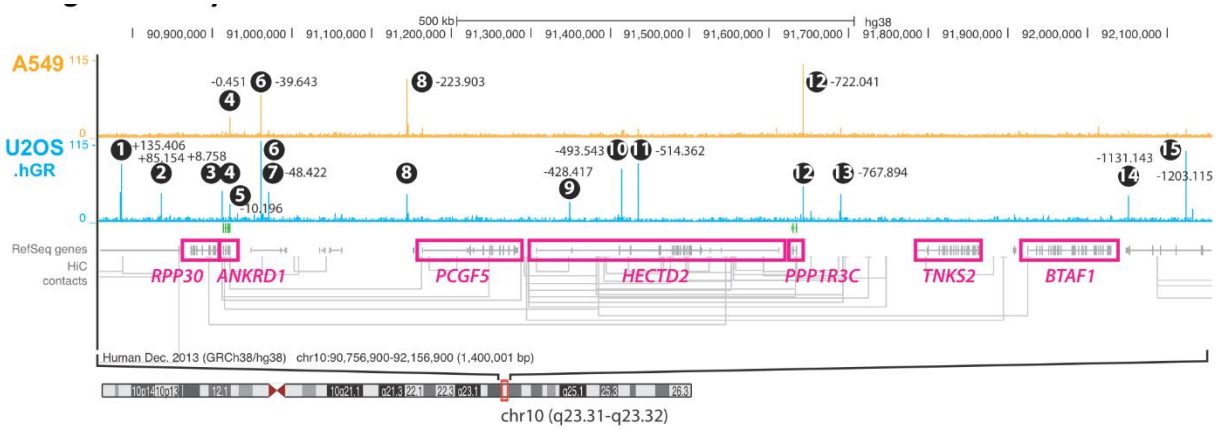



**Figure 3.3: 1.4 Mb topological unit with distinct GOR occupancy and GC gene regulation in A549 and U2OS cells**. GORs numbered left to right in 1.4 Mb region on chromosome 10 in A549 (orange) and U2OS (blue) cells with some overlapping (4, 6, 8, 12) and unique ChIP-seq peaks. Genes within TAD are boxed in red. *ANKRD1* is GC responsive in both A549 and U2OS cells and *HECTD2* is responsive only in U2OS cells. Hi-C contacts in grey not extending past housekeeping genes, *RPP30* and *BTAF1*, serve as the demarcation of TAD.

*Cas9 mutagenesis of GR occupancy at single loci*

We used directed Cas9 genome editing to generate chromosomal deletions or insertions (indels) that disrupt GBS motifs, or GOR peak summits for GORs lacking a GBS. A GOR regulating *ANKRD1* was discovered by a homozygous, ~120 bp deletion of GOR4. GOR4 is slightly upstream of the *ANKRD1* promoter (-0.451 kb) and contains a single GBS. In A549, a single H3K27Ac site was disrupted by deletion of GOR4 (Figure 3.4), and dex induction of *ANKRD1* declined by 75% (Figure 3.5). In U2OS, *ANKRD1* is no longer upregulated at 1 nM and further downregulated at 100 nM with loss of GOR4 (Figure 3.6).

We examined the GOR4 region in the A549 wild type and GOR4 mutant by ChIP-seq, and confirmed the loss of the GOR4 peak (Figure 3.4). In A549, *ANKRD1* expression in the absence of dex was elevated by 2.5-fold in the GOR4 mutant, implying that non-GR TF or chromatin remodeling binding sites may have been deleted or created by the ~120bp GOR4 deletion. RNA-seq established that only *ANKRD1* transcription levels were changed significantly by the GOR4 mutant, examining the surrounding ±1 Mb region, either in the absence (Table 3.1) or the presence (Table 3.2) of dex. Genome-wide, dex-regulated expression of four additional genes, none residing on chromosome 10, appeared to be altered in the GOR4 mutant (Table 3.3). Further testing is required to validate these findings, and to test whether dex regulation of those genes is primary, *i.e.*, controlled directly by GR, or secondary. Our provisional conclusion is that GOR4 displays GRE activity that is nearly fully specific to *ANKRD1* in A549. Interestingly, this appears to contrast with a report in which a Cas9-driven deletion of a single H3K27Ac locus in *HCT116* colon cancer cells produced large scale changes in gene expression of several genes ±1 Mb from the deletion (Tak et al., 2016).


**Figure 3.4: Validation of GOR4 deletion in A549 cells.** A549 wildtype (WT, purple) sequence containing one GBS highlighted in red. A549 GRE mutant (MT) is a homozygous 120 basepair deletion which ablates GBS and GOR4 upstream of *ANKRD*1 that overlaps with H3K27Ac mark (green).

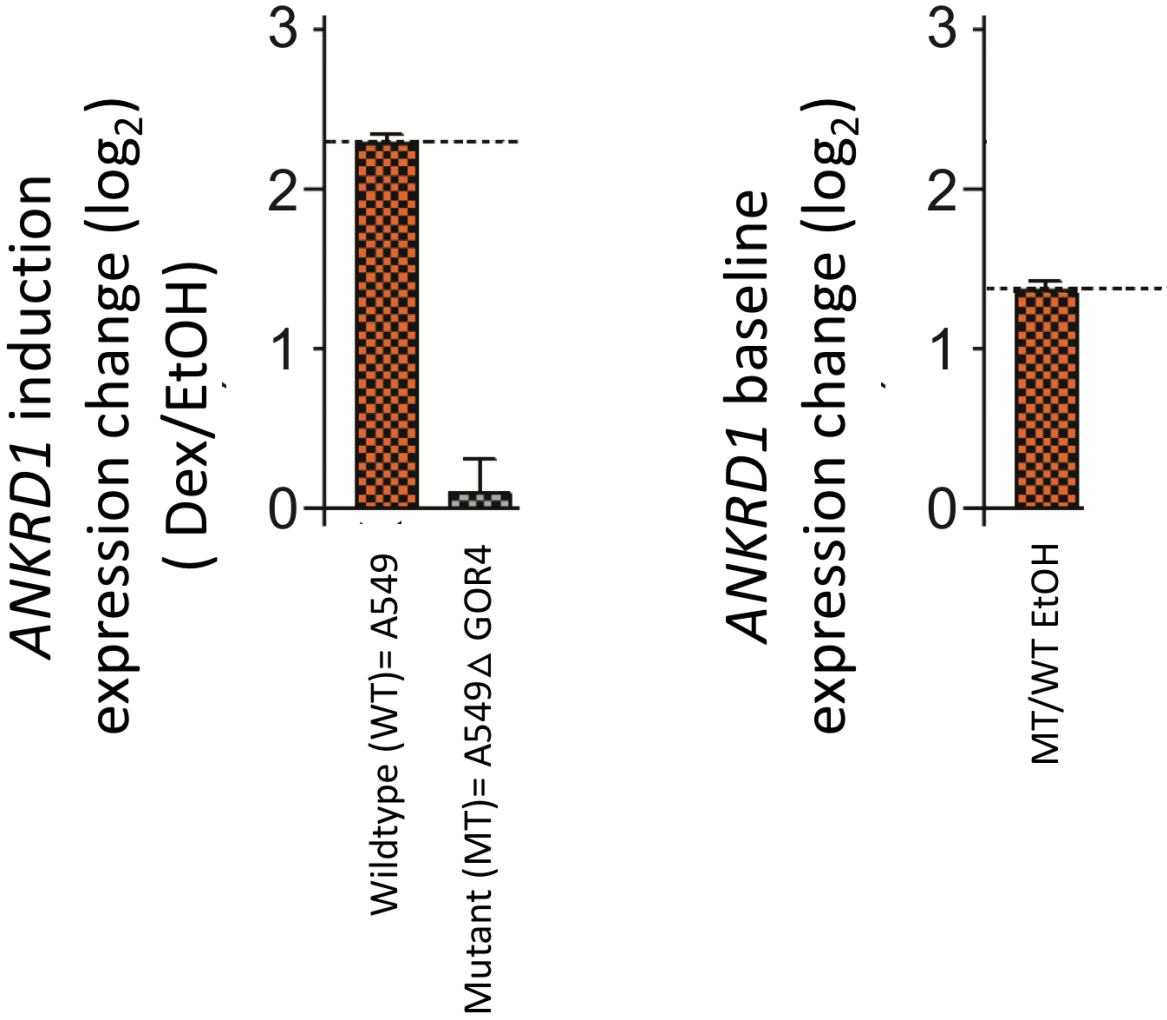



**Figure 3.5: *ANKRD1* induced and basal expression is affected by deletion of GOR4 in A549 cells.** RNA-seq and qPCR show a 4-fold decrease in *ANKRD1* expression when comparing wildtype (WT) and mutant (MT) levels (left) and 2.5-fold increase in *ANKRD1* expression in mutant when compared to wildtype basal levels (right).

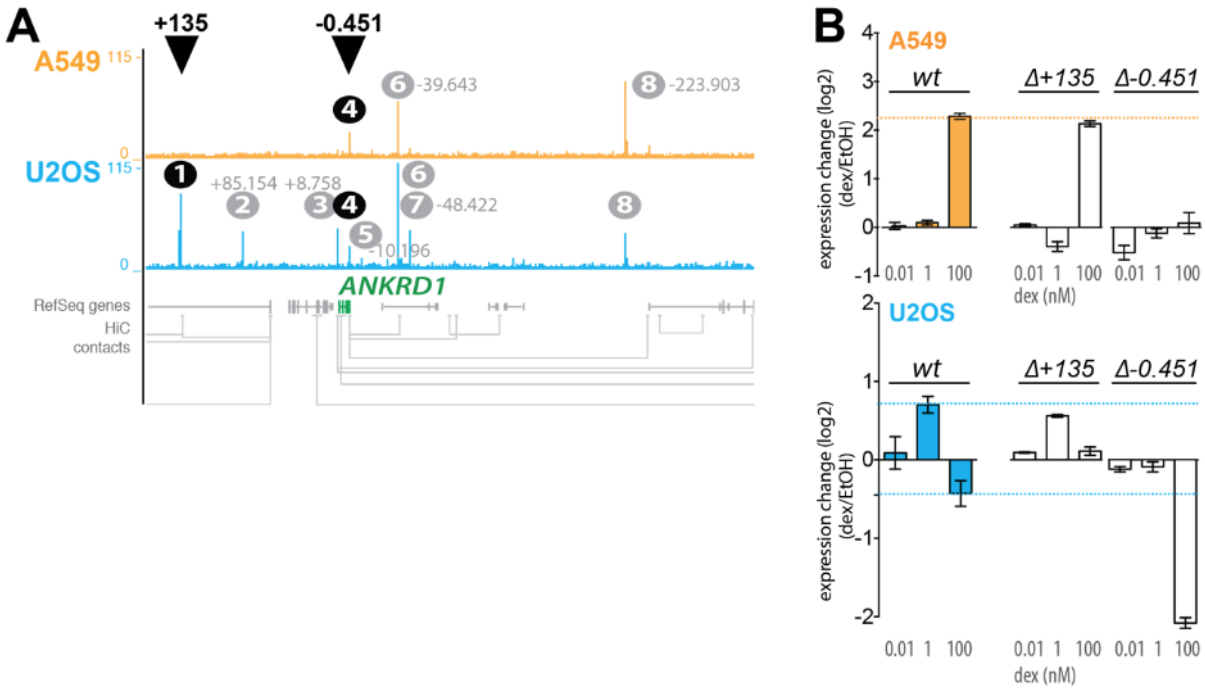



**Figure 3.6: Multiple GORs affect expression of single GC regulated gene, *ANKRD1*.** (A) Zoom in on chromosome 10 topological unit focusing on GOR1 (+135kb) downstream of *ANKRD1* and GOR4 (-0.451kb) upstream of *ANKRD1*. (B) *ANKRD1* regulatory analysis of GOR1 (+135kb) and GOR4 (-0.451kb) ablations across 0.01, 1, and 100 nM dex dose.

*GRE composition*

We have found that multiple GORs can influence transcription of a single GC-responsive gene (Figure 3.6). Interestingly, mutation of GOR4 (-0.451kb) in U2OS abrogated *ANKRD1* upregulation at 1 nM, but strengthened its downregulation at 100 nM, whereas mutation of GOR1 (+135kb) abrogated its downregulation at 100 nM. Hence, it appears that GR regulates expression in *ANKRD1*, at least in U2OS, from a composite GRE that includes at least two GORs separated by >135kb. It will be very interesting to interrogate the other GORs within this TAD. We speculate that control by dispersed composite GREs will prove to be common, whereas some GORs in the TAD will lack detectable activity in the two contexts assessed here, but may well be highly functional in other contexts.

Finally, we note that thousands of GORs overlap with intergenic or intronic regions that produce dTREs in each cell line (Figure 2.5). GOR1 overlaps with 1 dTRE in U2OS cells whereas GOR4 does not display dTREs, but genome-wide, 26% and 46% of GORs overlap with dTREs in U2OS and A549 cells, respectively. We detected 37 dTREs within the TAD; 1 overlapped with GOR6 in A549 versus 5 overlapping with GORs 1, 5, 6, 7, and 9 in U2OS cells. Of the 37 dTREs, only 1 dTRE had expression affected significantly ($p \leq 0.05$) by dex in A549 cells and it did not overlap with a GOR. In U2OS, only 1 dTRE had expression affected significantly by dex and it overlapped with GOR6. While a role for dTREs/eRNAs in GRE activity remains to be investigated, it would be consistent with the context specificity of response element composition and function to assume that dTREs will be found at some but not all GORs, and at some but not all GREs.

## Materials and Methods

### GR ChIP-seq

A549 and U2OS.hGR cells were grown in ~3 T-225 cm flasks to 90% confluency. After 90 min treatment with dex (Sigma) at 100 nM or ethanol (Koptec), cells were harvested by trypsinization, counted on a hemocytometer, and distributed to 50-mL Falcon tubes in volumes corresponding to ~1.8 x $10^7$ cells/tube (anticipating ~1.8 x $10^7$ cells/ChIP). 36.5% formaldehyde was added to suspended cells to a final concentration of 1% v/v; after incubating 3-10 min at RT, formaldehyde was quenched by adding 2.5 M glycine to 0.3 M, RT 5 min, followed by transfer to ice. Cells were recovered by centrifugation at 450g, 5 min, 4 ˚C, then washed twice by resuspension and pelleting in 20 mL ice-cold TBS (100 mM Tris-HCl, pH 7.5 @ 4 ˚C/150 mM NaCl) on ice. Cells were then washed 3x at RT in 1 mL MC lysis buffer (10 mM Tris-HCl, pH 7.5 @ RT, 10 mM NaCl, 3 mM MgCl2, 0.5% (v/v) Tergitol type NP-40), resuspended in 1 mL RT MC lysis buffer and transferred to a 1.5 mL Eppendorf tube. Cells were pelleted at 200g, 5 min; residual buffer was removed and cells were frozen in liquid N2 for storage at -80 ˚C.

Frozen nuclear pellets were thawed in cool water, resuspended in 180 µL MNase reaction buffer (10 mM Tris-HCl, pH 7.5 at RT, 10 mM NaCl, 3 mM MgCl2, 1 mM CaCl2, 4% (v/v) Tergitol type NP-40) supplemented with PMSF to 1:100, and the volume was taken to 270 µL with MNase reaction buffer. MNase (New England Biolabs) was diluted 1:10 in MNase reaction buffer, and 1.35 µL was added to the resuspended chromatin and incubated at 37 ˚C, 5 min. A cOmplete Mini, EDTA-free Protease Inhibitor Cocktail tablet (Roche) was dissolved in 500 µL MNase buffer (PIn cocktail); MNase reaction was stopped by adding 5.4 µL 0.2 M EGTA (pH 8), 7.2 µL 100 mM PMSF, 14.65 µL PIn cocktail, 14.65 µL 20% SDS, and 14.4 µL 5 M NaCl,

with gentle tube inversion to mix. 164 µL volume was transferred to each of two 1.5 mL Bioruptor®Plus TPX microtubes (Diagenode, Denville, NJ) and sonicated in a Bioruptor® Plus (UCD-300) Sonication System using intensity setting 'H' (320 W) and sonication parameters "CYCLE Num:30, Time ON:30sec, Time OFF:30sec". During a 15-min rest interval, samples were vortexed for 5 sec, spun down in a microfuge, and transferred to new TPX tubes on ice, followed by a second round of sonication (amounting to 60 cycles total). 150 µL Dynabeads™ Protein G slurry (Invitrogen) was mixed with 20 mg N499 (rabbit α-human GR IgG) antibody plus 450 µL Lysis Buffer 2 (10 mM Tris-HCl, 1 mM EDTA, 150 mM NaCl, 5% (v/v) glycerol, 0.1% (w/v) sodium deoxycholate, 0.1% (w/v) SDS, 1% (v/v) Triton X-100, pH 8 at 4 ˚C; no PIn added) in a 1.5 mL Eppendorf tube, incubated 1 hr with rolling in 4 ˚C cold room; tubes were placed in magnetic rack and supernatant was removed. Sonicated chromatin samples were pelleted at maximum speed, 10 min, 4 ˚C, and white pellet and cloudy suspension above pellet were recovered by transferring from 1.5 µL TPX tubes to new 1.5 mL tubes. 10 µL aliquot of combined input was set aside for later processing. 25 µL 100X Halt™ Protease Inhibitor Cocktail (Thermo Scientific) was added to a 5 mL tube with 2.5 mL Dilution Buffer (identical to Lysis Buffer 2, except without SDS). ~275 µL chromatin (from ~1.5 x 107 cells) was added to the Dilution Buffer+beads in the 5 mL tube, effectively diluting the chromatin 1:10. The 5 mL tube was sealed with parafilm and incubated 4 h on a roller in a 4 ˚C cold room. During this time, input material was reverse-crosslinked by adding TE (10 mM Tris-HCl, 1 mM EDTA, pH 8; pH 7.5 at RT) to a total volume of 80 µL, followed by addition of 100 µL ChIP Elution Buffer (50 mM Tris-HCl, pH 7.5 @ RT, 10 mM EDTA, 1% SDS) and 20 µL Pronase (Roche, 20 mg/mL) for incubation at 42 ˚C for 2 hr, then 65 ˚C overnight. 100X Halt PIn was warmed to RT, and 15 µL was added to 1.5 mL of each of three Wash Buffers (A-C; Wash Buffer A: Buffer A

containing 10 mM Tris-HCl, 1 mM EDTA, 150 mMNaCl, 5% (v/v) glycerol, 0.1% (w/v) sodium deoxycholate, 0.1% (w/v) SDS, 1% (v/v) Triton X-100, pH 8.0; Wash Buffer B: Buffer A with 500 mM NaCl and Halt protease inhibitor mixture; Wash Buffer C: 20 mM Tris-HCl, 1 mM EDTA, 250 mM LiCl, 0.5% (v/v) Nonidet P-40, 0.5% (w/v) sodium deoxycholate, Halt protease inhibitor mixture, pH 8.0 ). Beads were washed consecutively in 1.5 mL of each of the Wash Buffers A-C (1X Halt PIn), by gently adding buffer to resuspend the beads, then placing the tube on a magnetic rack and removing the supernatant after beads settled. Chromatin (from 1.5 x $10^7$ cells on beads from 150 µL slurry) was eluted from beads in 300 µL Elution Buffer/Reverse-Crosslinking Buffer (10 mM Tris-HCl, 1 mM EDTA, 0.7% (w/v) SDS, pH 8 at RT) by incubating beads in buffer for 5 min, RT with gentle pipetting to occasionally mix, then allowing beads to settle on magnetic rack and transferring eluant volume to a new 1.5 mL Eppendorf tube. To reverse crosslinks, 450 µL Adjustment Buffer (50 mM Tris, 10 mM EDTA, 0.45% SDS pH 7.0 at RT) was added with 82.5 µL Pronase (20 mg/mL) to 300 µL eluted chromatin, followed by incubation at 42 ˚C for 2 hr, then 65 ˚C overnight. DNA was subsequently cleaned from "input/MNase only" samples using a Qiagen PCR Purification Kit, and MinElute PCR Purification Kit (Qiagen) columns were used to purify ChIPs (one column/each ChIP, using 2.5 mL ERC buffer or 4.16 mL PB buffer), eluted in 15 µL EB. Recovered DNA was stored at -20 ˚C. Libraries were generated using an Ovation® Ultralow System V2-32 (NuGEN Technologies, Redwood City, CA), quantified on a 2100 Bioanalyzer System (Agilent, Santa Clara, CA) with High Sensitivity DNA Kit. Each library was sequenced on a HiSeq (Illumina) using single reads of 50 bp in length. Bigwig files were generated using the MACS2 callpeak algorithm in Galaxy (usegalaxy.org), and displayed as a custom track in the

UCSC Genome Browser (http://genome.ucsc.edu) with files hosted at Cyverse Discovery Environment (https://de.cyverse.org/de/).

**GBS selection**

*ANKRD1* ENSG00000148677 TSS was defined as CAGE peak at chr10:90,921,087 in hg38 (GRCh38/hg38 human genome assembly, accession GCA_000001405.15), FANTOM5 CAGE phase 1&2 pooled human tracks (fantom.org). A549 and U2OS.hGR GR occupied regions called by MACS2 in a 1.4 Mb vicinity of *ANKRD1* TSS (GRCh38/hg38 chr10:90,756,900-92,156,900) were selected for functional analysis, with GOR identifier designated based on approximate below-summit GBS position relative to TSS. 500-1000 bp regions were recovered from the UCSC Genome Browser using 'Get DNA' function and populated into SnapGene.dna files for archiving and analysis. Putative direct DNA-binding motifs recognized by GR at peak summits were identified by scanning DNA files for a degenerate GBS 'match' ('extremely generic GBS motif' 5'-NNNACANNNNGTNCNN-3') and by analysis in rsat matrix-scan (http://rsat01.biologie.ens.fr/rsat/matrix-scan_form.cgi) using the TRANSFAC NR3C1 positional weight matrix with markov order 1 and p value upper threshold 5e-2 (0.05).

**GBS/GOR editing**

Single guide (sg) RNAs that deliver *S. pyogenes* Cas9 to genomic target loci were identified using two publicly available SpCas9 sgRNA design tools, sgRNA Designer (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design) and CRISPR-MIT (crispr.mit.edu): for 250-bp regions flanking a GR ChIP-seq peak-centered GBS or GR ChIP-seq peak summit (i.e., GOR), guide selections identified in sgRNA Designer were cross-referenced

to guides identified in CRISPR-MIT in MySQL. sgRNA sequences were populated as 'primers' in SnapGene sequence files and prioritized for selection based on targeting within or near GBS motifs, high on-target efficacy scores and high efficiency scores. For Cas9 RNPs, sgRNAs were synthesized as T7 RNA Pol in vitro transcription (IVT) products from a double-stranded DNA template. dsDNA template was synthesized by PCR using 4 primers in multiplex: 18-mer ML557 (TAA TAC GAC TCA CTA TAG), 22-mer ML558 (AAA AGC ACC GAC TCG GTG C), 93-mer ML611 (AAA AGC ACC GAC TCG GTG CCA CTT TTT CAA GTT GAT AAC GGA CTA GCC TTA TTT AAA CTT GCT ATG CTG TTT CCA GCA TAG CTC TTA AAC) and target-specific 58-mer comprising 5' 18-mer (5'-TAATACGACTCACTATAG-3') and 3' 20-mer (5'-GTTTAAGAGCTATGCTGGAA-3'). 100-µL PCRs were performed as follows: 20-µL 5X Phusion buffer (125 mM TAPS-HCL, 250 mM KCl, 10 mM MgCl2, 5 mM βmercaptoethanol), 2 µL 10 mM dNTPs, 8 µL ML557+558 at 12.5 µM each, 0.5 µL ML711 at 4 µM, 0.5 µL target-specific oligo at 4 µM, 1 µL Phusion pol, cycled using thermocycler program IVT_TMPL (95 ˚C 30 sec, 95 ˚C 15 sec, 57 ˚C 15 sec, 72 ˚C 15 sec (cycle to step 2 30x), 72 ˚C 30 sec, 10 ˚C indefinitely. PCR products was isolated using DNA Clean & Concentrator-5 kit (Zymo) and eluted in 12 µL nuclease-free H2O. IVTs were performed in 100 µL reaction volumes with 5X reaction buffer (components), 2 µL each NTP (each at 25 mM), 5 µL 100 mM DTT, 600-700 ng template DNA (10 µL PCR product), incubated 4 h–o/n, 37 ˚C. RNA was isolated using RNA Clean & Concentrator-5 columns (Zymo), eluted in 15 µL nuclease-free H2O; RNA concentration was estimated in a NanoDrop™ Microvolume Spectrophotometer (ThermoFisher) and diluted as appropriate for sgRNA stock ~100 µM (estimating that sgRNA MW ~37 kDa, 3700 ng/µL ~100 µM). sgRNAs were assembled with Cas9-NLS protein (QB3 MacroLab, Berkeley, CA; qb3.berkeley.edu/macrolab/) as follows: for single RNP

nucleofections, sgRNA volumes corresponding to 40-100 pmol sgRNA were distributed to 500 µL non-stick, nuclease-free Eppendorf tubes (Ambion) on ice; 40 pmol Cas9-NLS (1 µL Cas9-NLS at 40 µM/6.4 mg/mL in 20 mM HEPES-KOH, pH 7.5, 150 mM KCl, 10% glycerol, 1 mM DTT; Mw 160.95 kDa) was added, then 50 pmol carrier DNA (0.5 µL Alt-R Cas9 electroporation enhancer nucleic acid at 100 µM, IDT); the resulting ~2.5 µL volume was incubated at 37 C for 15 min, transferred to ice; 10- 11 µL cells suspended in buffer R at a concentration of 2.5 x 106 cells/mL were then added directly to the RNP volume, for a final concentration of ~1.6 pmol/µL RNP (~1.6 µM RNP in cell suspension, ratio of ~108 RNP molecules per cell). 10 µL RNP+cell suspension was transferred into 10 µL Neon tip and nucleofected into A549 and U2OS.GR populations using the Neon™ Transfection System (ThermoScientific) with 10 µL Neon tips, and the following electroporation settings (pulse voltage (V), pulse width (ms), pulse #: A549: 1200, 30, 2; U2OS.GR: 1200, 10, 4). Nucleofected cells were delivered to 12-well or 6-well dishes containing pre-warmed DMEM/high glucose (HyClone)/10% FBS (GemCell) for recovery, and incubated for 24 – 72 h before FACS isolation of individual cells.

**Clonal isolation by FACS**

Single cells were delivered to 100 µL HAM'S F-12 media (Lonza, Basel, Switzerland), (A549) or DMEM/5% FBS mixed 1:1 with conditioned media (U2OS.hGR) in individual wells of 96-well plates (Corning, Kennebunk, ME) by FACSAria2 (BD Biosciences, San Jose, CA) in the UCSF Center for Advanced Technology, and grown for 3-4 weeks with regular media replacement after 2 weeks. Media in 96-well plates was changed by aspiration using 8-channel adapter (Argos Technologies, Elgin, IL; EV503) attached directly to aspirator tubing and fitted

with sterile, disposable pipet tips (Rainin SS-L10); all processing for media addition was performed using sterile filter tips.

Allele description (genotyping)—Cells were prepared for genotyping by removing media, washing with 100 µL DPBS/Modified (–calcium/–magnesium) (HyClone Laboratories, Logan, UT), and trypsinization with 30 µL 0.5% trypsin-EDTA (Gibco/ThermoScientific, Waltham, MA). 15-20 µL volume from each trypsinized well was transferred to corresponding well of 96-well, 0.2 mL/well TempPlate semi-skirted polypropylene PCR plate (USA Scientific, Ocala, FL) for lysis, and 100 µL fresh media was added to remaining cells in 96-well plate for return to culture. Cells in polypropylene plates were sealed with cold storage foil (USA Scientific), lysed by adding 15 µL 2X lysis buffer+1:100 Recombinant PCR Grade Proteinase K (Roche, Basel, Switzerland), with thermocycler incubation at 65 °C 30 min., 95 °C 15 min. Amplicons were prepared for massively parallel sequencing in two PCR reactions, performed in Hard-Shell® PCR plates, 384-well, thin-wall (Bio-Rad Laboratories, Hercules, CA): PCR1 (~219 bp amplicons)—4 µL lysate in 20 µL PCR volume, TCHDWN: cycled at 98 °C 2 min 30 sec, [98 °C 30 sec, 57-62 °C 20 sec, 72 °C 30 sec (30x)], 72 °C 8 min. PCR2 (~302-bp amplicons)—0.5 µL PCR1 template in 20 µL volume, with i5 and i7 indexed primers at 200 nM. SampleSheet preparation for Illumina MiSeq (Illumina, San Diego, CA) was automated for barcoded amplicons using SampleSheet.py (Ehmsen et al., in preparation, https://github.com/YamamotoLabUCSF). Following PCR2, 5 µL from each well was pooled, 100 µL pooled amplicons were column-cleaned (Zymo DNA Clean & Concentrator-5, Genesee Scientific, San Diego, CA), concentration was estimated by NanoDrop (ThermoFisher Scientific), and library was quantified using the KAPA Library Quantification Kit for Illumina

platforms at 2-3 concentrations (serially diluted to 1,000,000 – 100,000,000-fold dilutions to reach within commercial standards, according to manufacturer's instructions and quantification template (www.kapabiosystems.com). Sequencing was performed using MiSeq Reagent Kit v2 (300-cycles) or MiSeq Reagent Nano Kit v2 (300-cycles) according to manufacturer's instructions, at 8 – 12 pM (typically 10 pM) library with φX DNA (Illumina PhiX Control v3) at 5-30% (typically 5%). Data were monitored in Illumina BaseSpace (basespace.illumina.com), fastq files were directly transferred from the MiSeq instrument to an external drive for processing, and processed in bash for Mac OS to identify top reads per well. Top reads were aligned to a reference sequence in SnapGene 4.0.8 (GSL Biotech LLC, Chicago, IL; www.snapgene.com) using Tools⟶Align Multiple Sequences, to assess alleles and genotypes in clones. Target clones were expanded from 96-well plates to 48- or 12-well plates, 6-well plates, and finally to 100-mm plates, by trypsinization, etc., from which three vials were frozen in DMEM/5% FBS/5% DMSO in styrofoam blocks or Mr. Frosty freezing containers (Thermo Scientific) and archived for long-term storage in liquid N2.

**Regulatory analysis by qPCR**

3 mL cells at 1.2 – 1.5 x $10^5$ cells/mL were plated in 6-well dishes. 24 – 36 h later, media was removed and replaced with 2.4 mL media with charcoal-stripped FBS (Omega Scientific, Tarzana, CA). 3 h later, 600 µL media containing 5X dex (Sigma Scientific, St. Louis, MO) was added. 4 hr later, media was aspirated, cells were rinsed with 3 mL PBS, and following aspiration, cells were lysed in situ with 350 µL RLT buffer (1:100 β-mercaptoethanol). Cell lysate was transferred to 1.5 mL Eppendorf tubes, flash-frozen in liquid N2, and stored at -80 °C. RNA was isolated by RNEasy (Qiagen, Hilden, Germany) according to manufacturer

instructions, with DNase treatment (50 µL RNase-free DNase (Qiagen)), eluted in 30 µL nucleasefree H2O. RNA concentration was determined by NanoDrop. cDNA was synthesized in 20-µL volumes with 1 µg RNA template, 4 µL 5X iScript reaction mix, 1 µL iScript reverse transcriptase, incubated under ISCRIPT thermocycler program in RNase-free 0.5-mL microfuge tubes (Ambion) (25 °C 5 min, 42 °C 30 min, 85 °C 5 min, 4 °C indefinitely). RNA stocks were stored at -80 °C after flash-freezing; cDNA reaction products were stored at -20 °C. cDNA reactions were diluted 4-fold, with 4 µL/reaction (50 ng/reaction). 6 µL primers at 0.83 µM (each) were added; 20 uL final qPCR reaction volume, the final working concentration of each oligo will be 250 nM. Add 10 µL SsoAdvanced Supermix low-retention tips and multichannel pipet if possible, 95 °C 30 sec, 95 °C 5 sec, 57 °C 30 sec, (cycle to step 2 39x). Note plate types, Microseal 'B' optically clear adhesive seals (Bio-Rad). Note primer IDs. qPCR primer pairs were designed using IDT PrimerSelect tool and selected for assay use based on certification for between-cycle 2-fold amplification efficiency across a 7-sample, 10-fold serial dilution of cDNA (maximum 50 ng cDNA tested, 2-fold amplification efficiency accepted up to empirical cutoff of Cq = 33, beyond which linearity abruptly declines) and target specificity as monitored by qPCR melt curve analysis. Primers mixed as pairs (667 nM in H2O) and cDNA (50 ng/ µL) were prepared in separate wells of 384- well source plates (Labcyte) and delivered to 384-well white PCR plates (Bio-Rad) at 3000 nL and 1000 nL/well, respectively, using an Echo® 525 Acoustic Liquid Handler (Labcyte Inc., San Jose, CA). 4000 nL SsoAdvanced Supermix (BioRad) was then added from 6-well reservoir source plate (final assay concentrations: 50 ng cDNA/reaction, 250 nM/primer, 8 µL reaction volume). Plates were sealed with Microseal 'B' optically clear adhesive seals (Bio-Rad), centrifuged 5 min. at 1500g, and processed for qPCR at 95 °C 30 sec, 95 °C 5 sec, 57 °C 30 sec, (cycle to step 2 39x), with endpoint melt curve analysis. RNA only

(no RT) controls were processed in parallel for every primer pair and cDNA sample, certifying qPCR signal attributable to amplification from cDNA template.

**Locus evaluation (topological unit/TAD designation)**

Metazoan genomes are increasingly recognized to exhibit intrachromosomal looping or nested sets of heightened interaction/proximity frequencies at distances ranging from several kb to hundreds of kb; although many long-range proximities may be incidental (without evolutionarily selected function), others may participate in regulatory control of specific genes or gene hubs. It is presently difficult to define stable borders between sub-chromosomal regions in which unexpectedly high proximity interactions can be detected by chromosomal conformation assays; we chose to bin TADs (topological units) based on publicly available 5-kb resolution HiC data (Reddy lab, Duke University, Chapel Hill, NC). Topological domains were defined based on publicly available HiC datasets (D'Ippolito et al. 2018), rendered as tabix files hosted at Cyverse Discovery Environment for viewing in the UCSC Genome Browser. Briefly, we converted HICCUPS file contact data for 5-kb genomic units to tabix files (Li et al., 2011); then using custom Python code, we populated chromosome-length lists with binary ('0' vs. '1') definitions for each bp, with '0' denoting no evidence of that bp partaking in long-range interaction with another bp block, and '1' denoting HiC evidence for that bp partaking in a long-range interaction with another bp block. The resulting Python lists were processed in pandas data frames to mark chromosomal units that either comprised topological interactions or were void of topological interactions, thereby fractionating the genome into topological units. We mapped genes, GORs, and dTREs into these units based on bedtools intersects. HiC contact heat maps were additionally visualized in cloud-based Juicebox (www.aidenlab.org/juicebox/).

## Appendix to Chapter 3

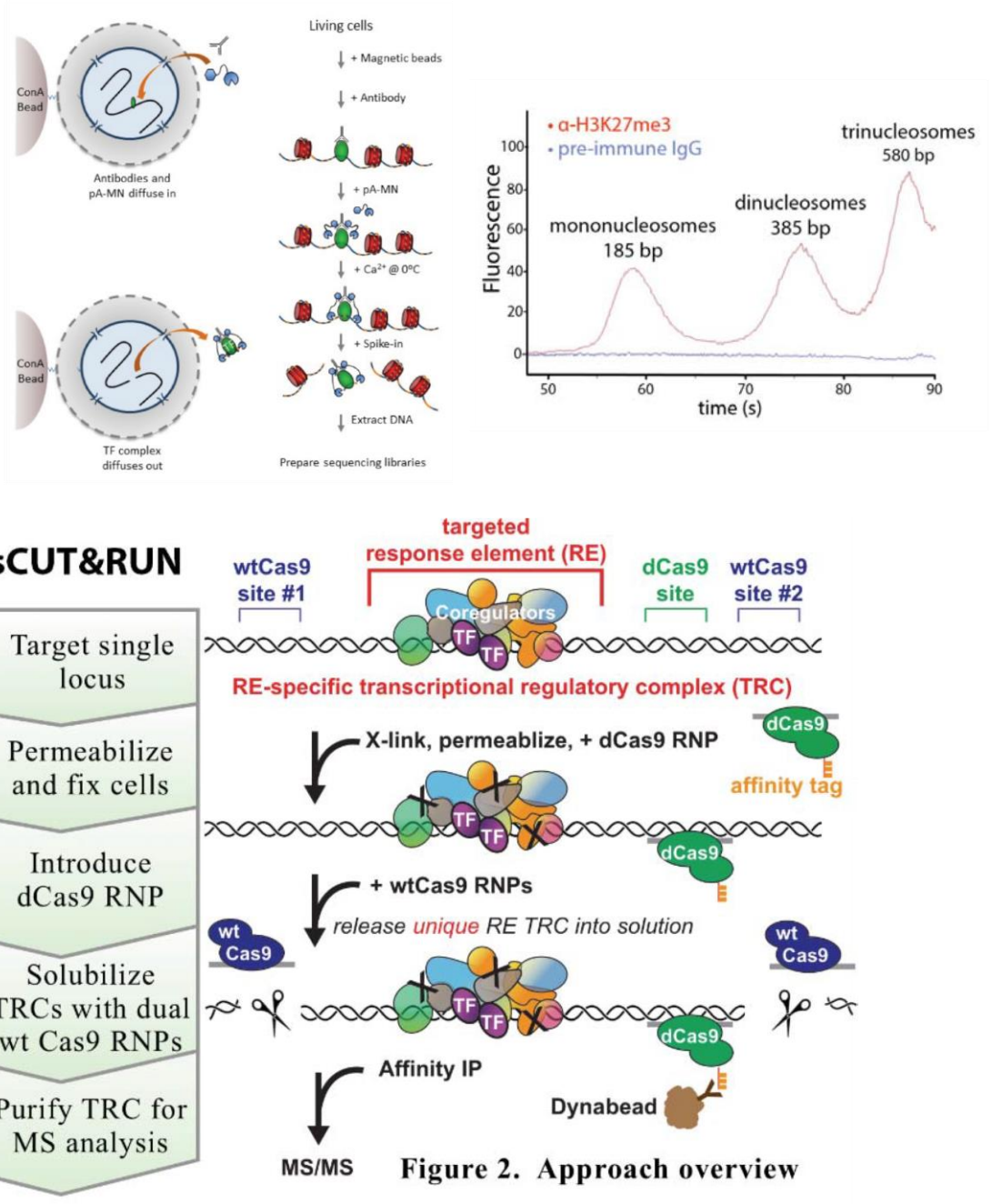
Once GREs are functionally identified, we aspire to identify the protein components of the TRCs at those loci. We have begun to develop and validate a new technology, CasCUT&RUN (Matthew Knuesel, UCSF), which will enable for the first time isolation and purification of *in vivo* assembled, GRE-specific TRCs. Ultra-high resolution mass spectrometry will then be used to identify resident proteins and their post-translational modifications.

### Results

*Chromatin fragment recovery from CUT&RUN*

In preliminary studies, we have validated CUT&RUN (Skene et al., 2017) as the basis for modification to CasCUT&RUN. DNA fragments of the appropriate size were solubilized from digitonin-permeabilized, Concanavalin-A bead immobilized cells, and incubated with control or H3K27Me3-antibody followed by ProteinA-MNase (Appendix Figure 3.1). These validated procedures serve as the basis for development of CasCUT&RUN.

CUT&RUN




**Appendix Figure 3.1: Validation of CUT&RUN for use in CasCUT&RUN.** CUT&RUN schematic from Skene et al., 2017 (left) and Bioanalyzer trace of ProteinA-MNase cleaved and solubilized DNA from 42,000 cells incubated with anti-H3K27Me3 or pre-immune IgG control antibody (right). CasCUT&RUN approach overview (bottom).

## Chapter 4: Discussion

Jacob and Monod (1961) established that prokaryotic transcription is regulated by a TF bound adjacent to a target gene promoter and affecting RNA polymerase function. However, this concept seems insufficient to account for transcriptional regulation in metazoans, where genes are expressed with remarkable cell- and physiological-context specificity. Britten and Davidson introduced the idea of combinatorial regulation (1969), which in principle could enable context specificity if metazoan TFs were dynamic multifactor TRCs assembled as context-specific combinations of broadly expressed TFs and co-regulators.

Our lab showed that GR, which regulates distinct gene networks in different cell-and physiological contexts, receives and integrates multiple signals (*e.g.*, hormonal ligands, DNA binding sequences, post-translational modifications, interacting non-GR TFs) as allosteric effectors that together drive distinct GR conformations bearing specific patterns of interaction surfaces for association with particular co-regulator factors, such as histone modification enzymes and chromatin remodeling machines, etc. Hence, context-specific signaling to GR results in context-specific TRC assembly, in turn conferring context-specific regulatory functions on GR, which alone is merely a DNA-binding scaffold protein lacking intrinsic transcriptional regulatory activity (Weikum et al., 2017).

Context-specific combinatoriality introduces enormous complexity into identification and validation of GRE or any response element, as there is no simple "GRE code" that specifies DNA binding sites, TRC components, genome features or higher order organization, or even a regulatory action or mechanism, that corresponds to a functional GRE. Thus, systems analyses

cannot identify GREs, and instead, they must be validated individually using genome editing procedures that allow genetic analysis in normal chromosomal context. Similarly, the capacity of metazoan transcriptional regulation to operate over long range, vastly greater and more flexible than the base-pair positional specificity of prokaryotic response elements relative to their target gene promotors, complicates identification of cognate target gene(s) for a given GRE, from amongst all candidate GC-regulated genes, identified by whole genome systems approaches such as RNA-seq.

Decades ago, our lab demonstrated sequence-specific binding by GR, showed that a DNA fragment bearing GR-binding sequences could confer GC regulation on a remotely positioned heterologous promoter, and denoted that first functional response element as a GRE (Chandler et al.,1983). Since that time, progress in securing a full understanding of the defining properties and mechanisms for response element actions, and in unequivocal identification of target genes, has been severely hampered by failure to acknowledge and address context specificity. Instead, numerous reports have appeared describing systems analyses that catalog features whose relationships to response element activity rely on untested assumptions and/or flawed assays, together with inference of target gene identity based virtually solely on linear or topologic proximity.

Clearly, neither systems nor reductionist approaches alone can predict response elements or provide insight into allostery-determined combinatorial regulation of transcription. In the present work, we have established for the first time a standard for unequivocal identification of functional response elements and cognate target genes, and a path forward for determination of

TRC composition and mechanism. Context specificity demands that GREs be characterized and validated individually, using first genetic, and eventually molecular and biochemical approaches. Target genes can then be imputed by assessing the effects of GRE mutations on genome-wide GC-regulated transcription.

In time, with CRISPR mutagenesis of candidate GREs and genetic screens like Perturb-seq, it should be possible to probe functionality of GORs and apply machine learning algorithms to high throughput studies such as RNA-seq, HiC, and Pro-seq, to divulge classes of response elements. With validated response elements and cognate target genes in hand, we will be able to determine whether chromatin loops are essential GRE properties, linking GORs to each other or to a target gene promoter; we could assess whether topological unit demarcation functions to constrain GRE activity to that domain; we could test whether eRNAs are functional components of GREs. With successful development of CasCut&Run, we could define the outcome of context specificity, identifying the composition of TRCs that are the products of signal-driven allostery. In summary, functional validation of response elements and their cognate target genes, as we have described here, is an essential first step to derive mechanistic insights into context-specific metazoan transcriptional regulation; our work with GR can be generalized to other eukaryotic TFs and response elements.

When numerous GRE-target gene combinations have been defined, we predict that subsets of GREs that control a given physiologic property, *e.g.,* GC-mediated immunosuppression, will be found to bear nonidentical but overlapping features and characteristics. This potential to assign GREs to functional and compositional sub-classes could

open a pathway to design and screen new therapeutics for treatment of diseases and pathologic conditions influenced by glucocorticoids.

## References

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A., and Rai, A. (2016). Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. Journal of Computational Biology 23:4, 239-247.

Bonev, B., Cavalli, G. (2016). Organization and function of the 3D genome. Nature Reviews Genetics, 17, 661-678.

Britten, R.J., Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science, 165(3891):349-57

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

Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74.

D'Ippolito, A. M., McDowell, I. C., Barrera, A., Hong, L. K., Leichter, S. M., Bartelt, L. C., Vockley, C.M.,… and Reddy, T. E. (2018). Pre-established Chromatin Interactions Mediate the Genomic Response to Glucocorticoids. Cell Systems, 7(2), 146-160.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.H., Ye, Z., Kim, A.,… and Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature, 518, 331-336.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376.

ENCODE consortium. (2009). Standards, Guidelines and Best Practices for RNA-Seq. Vasa, 0(June), 1–7.

Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M.A., Hayashizaki, Y., Blanchette, M., Dostie, J. (2009). Chromatin conformation signatures of cellular differentiation. Genome biology, 10, R37.

Halfon, M. (2019). Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. Trends Genet. 35, 93–103.

Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B (2015). Advanced applications of RNA sequencing and challenges. Bioinformatics and Biology Insights, 9, 29–46.

Jacob, F., Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol., 3, 318-56.

Kaduake, S., Blobel, G.A. (2009). Chromatin loops in gene regulation. Biochim Biophys Acta, 1789, 17–25.

Khanin, R., Rapaport, F., Betel, D., Liang, Y., Krek, A., Mason, C. E., … Zumbo, P. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biology, 14(9), R95.

Kim, D., and Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12 (4).

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files, *Bioinformatics*, 27, 5, 718–719.

Li, S., Tighe, S. W., Nicolet, C. M., Grove, D., Levy, S., Farmerie, W., Mason, C. E. (2014). Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. Nature Biotechnology, 32(9), 915–925.

Liao, Y. and Smyth, G. K. and Shi, W (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30 (7).

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science, 326(5950), 289 LP – 293.

Liu, R., Holik, A. Z., Su, S., Jansz, N., Chen, K., Leong, H. S., … Ritchie, M. E. (2015). Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic acids research*, *43*(15), e97.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 1–21.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10-12.

Matharu, N., Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. PLOS Genetics 11(12): e1005640.

Meireles-Filho, A. C. A., and Stark, A. (2009). Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. Curr. Opin. Genet. Dev. 19, 565–570.

Moquin, S. A. (2017). Novel molecular insights inot Epstein-Barr virus. University of California.

Pack, L. R. (2016). Regulatory mechanisms that govern the activity of human histone demethylase KDM4C.University of California.

Pombo, A., Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nature Reviews Molecular Cell Biology, 16, 245–257.

QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159(7), 1665–1680.

Reddy, T. E., Pauli, F., Sprouse, R. O., Neff, N. F., Newberry, K. M., Garabedian, M. J., & Myers, R. M. (2009). Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome research, 19(12), 2163–2171.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England), 26(1), 139–140.

Rogatsky, I., Wang, J. C., Derynck, M. K., Nonaka, D. F., Khodabakhsh, D. B., Haqq, C. M., … Yamamoto, K. R. (2003). Target-specific utilization of transcriptional regulatory surfaces by the

glucocorticoid receptor. Proceedings of the National Academy of Sciences of the United States of America, 100(24), 13845–13850.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A.,… and Barton, G.J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA, 22:839-851.

Shlyueva, D., Stampfel, G., Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nature Reviews Genetics 15, 272-286.

Skene, P. J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife: 6:e21856.

T'Hoen, P. A. C., Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., … Lappalainen, T. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nature Biotechnology, 31(11), 1015–1022.

Tak, Y.G., Hung, Y., Yao L., Grimmer, M.R., Do, A., Bhakta, M.S., … Farnham, P.J. (2016). Effects on the transcriptome upon deletion of a distal element cannot be predicted by the size of the H3K27Ac peak in human cells. Nucleic Acids Res. 44(9):4123–4133.

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics, 13, 227.

Wang, Z., Martins, A.L., Danko, C.G. (2016). RTFBSDB: an integrated framework for transcription factor binding site analysis. Bioinformatics 32:3024-6.

Wei, Z., Batagov, A.O., Carter, D.R.F., and Krichevksy, A.M. (2016). Fetal bovine serum RNA interferes with the cell culture derived extracellular RNA. Scientific reports, 6, 31175.

Weikum, E.R., Knuesel, M.T., Ortlund, E.A., Yamamoto, K.R. (2017). Glucocorticoid receptor control of transcription: precision and plasticity via allostery. Nature Reviews Molecular Cell Biology, 18, 159–174.

Wingett, S., Ewels, P., Furlan-Magaril, M., (2015). HiCUP: pipeline for mapping and processing Hi-C data. F1000Res. 4:1310.

Wissink, E.M., Vihervaara, A., Tippens, N.D., Lis, J.T. (2019). Nascent RNA analyses: tracking transcription and its regulation. Nat Rev Genet, epub ahead of print.

Yamamoto, K. R. (1985). Steroid receptor regulated transcription of specific genes and gene networks. Annu. Rev. Genet. 19, 209–252.

Yamamoto, K. R., Darimont, B. D., Wagner, R. L. & Iñiguez-Lluhí, J. A. (1998). Building transcriptional regulatory complexes: signals and surfaces. Cold Spring Harb. Symp. Quant. Biol. 63, 587–598.

Yáñez-Cuna, JO., Dinh, HQ., Kvon, EZ., Shlyueva, D., Stark, A. (2012). Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. Genome Res. 22(10):2018-30

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., von Schack, D. (2018) Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: poly A+ selection versus rRNA depletion. Scientific reports, 8, 4871.

Zarse, K., Schmeisser, S., Groth, M., (2012). Impaired insulin/IGF1 signaling extends life span by promoting mitochondrial L-proline catabolism to induce a transient ROS signal. Cell Metab.15(4):451–465.

**Table 2.1 Robust differentially expressed genes that are glucocorticoid regulated derived from 3 different RNA-seq datasets filtered by q<0.05 and consistency.**

| Geneid | D1_MNC | D2_MNC | D3_MNC | ML2FC | PSE |
|---|---|---|---|---|---|
| ZBTB16 | 11 | 18 | 20 | 6.87 | 2.27 |
| IP6K3 | 10 | 55 | 57 | 6.75 | 1.81 |
| KLF15 | 4 | 76 | 47 | 6.36 | 2.15 |
| BEST2 | 4 | 18 | 86 | 6.16 | 2.26 |
| EDN3 | 17 | 23 | 343 | 6.11 | 1.84 |
| STAC2 | 42 | 273 | 329 | 5.31 | 0.92 |
| TGFBR1 | 2403 | 4088 | 27848 | 1.19 | 0.23 |
| ITGA5 | 576 | 2511 | 4722 | 1.18 | 0.26 |
| TNS4 | 7983 | 8041 | 9636 | 1.18 | 0.21 |
| FAM222B | 1301 | 1802 | 2481 | 1.17 | 0.21 |
| NAV2 | 1065 | 1614 | 2554 | 1.16 | 0.29 |
| MTSS1L | 637 | 5209 | 2317 | 1.16 | 0.23 |
| NOL3 | 109 | 1131 | 695 | 1.16 | 0.32 |
| CHST7 | 144 | 1141 | 1780 | 1.13 | 0.30 |
| RAP1GAP2 | 124 | 594 | 1896 | 1.12 | 0.37 |
| SNX8 | 212 | 3816 | 2355 | 1.12 | 0.27 |

MNC= Mean Normalized Counts
ML2FC= Mean log2FoldChange
PSE= Propagated Standard Error

For full table refer to Supplemental material.

**Table 2.2. Significant IPA pathways with z-scores* generated when comparing significant genes (q<0.05) from individual RNA-seq datasets.**

| Canonical Pathways | D1 | D2 | D3 |
|---|---|---|---|
| 1D-myo-inositol Hexakisphosphate Biosynthesis II (Mammalian) | 2.643906 | 1.73539 | 1.619185 |
| Actin Cytoskeleton Signaling | 2.462857 | 1.388326 | 2.490645 |
| Adipogenesis pathway | 2.389909 | 4.292143 | 3.97629 |
| Aldosterone Signaling in Epithelial Cells | 3.280751 | 4.532621 | 3.182586 |
| AMPK Signaling | 3.562997 | 1.883844 | 1.919056 |
| Apelin Cardiac Fibroblast Signaling Pathway | 2.309694 | 1.377331 | 1.966831 |
| Aryl Hydrocarbon Receptor Signaling | 1.777732 | 1.806659 | 3.614807 |
| Axonal Guidance Signaling | 2.212286 | 5.505336 | 4.364587 |
| B Cell Receptor Signaling | 6.486534 | 3.255869 | 4.843051 |
| cAMP-mediated signaling | 3.374472 | 3.561877 | 2.012879 |
| Cardiac Hypertrophy Signaling | 2.268857 | 3.021608 | 2.223682 |
| CD27 Signaling in Lymphocytes | 2.160571 | 3.754794 | 5.436833 |
| CD40 Signaling | 2.934406 | 2.281777 | 5.615934 |
| Cholecystokinin/Gastrin-mediated Signaling | 3.767673 | 3.744966 | 4.40559 |
| Coagulation System | 2.304078 | 1.965551 | 2.890629 |
| Colorectal Cancer Metastasis Signaling | 6.071404 | 7.828613 | 5.269203 |
| D-myo-inositol (1,3,4)-trisphosphate Biosynthesis | 2.643906 | 1.73539 | 1.619185 |
| Death Receptor Signaling | 3.705599 | 4.21389 | 2.242668 |
| Dopamine-DARPP32 Feedback in cAMP Signaling | 2.615848 | 3.561038 | 1.527505 |
| Endocannabinoid Cancer Inhibition Pathway | 3.732522 | 5.906265 | 2.617165 |
| Ephrin A Signaling | 1.309416 | 2.523677 | 1.385245 |
| ERK/MAPK Signaling | 5.500968 | 3.146976 | 2.251855 |
| Erythropoietin Signaling | 2.129856 | 2.316979 | 2.026099 |
| FAT10 Cancer Signaling Pathway | 3.205523 | 2.173376 | 4.68403 |
| G-Protein Coupled Receptor Signaling | 6.271612 | 4.242412 | 4.272729 |
| Germ Cell-Sertoli Cell Junction Signaling | 2.249039 | 5.426497 | 5.492191 |
| Glioblastoma Multiforme Signaling | 1.65853 | 4.34853 | 1.938122 |
| Glucocorticoid Receptor Signaling | 6.975216 | 2.997772 | 3.726176 |
| GNRH Signaling | 3.897797 | 3.598599 | 3.725523 |
| Gα12/13 Signaling | 2.638167 | 2.693414 | 2.757247 |
| Gαq Signaling | 4.1459 | 2.909962 | 2.849026 |
| Hepatic Cholestasis | 3.672366 | 1.789902 | 4.591475 |
| HER-2 Signaling in Breast Cancer | 1.911651 | 2.306768 | 2.126638 |
| HGF Signaling | 2.240166 | 5.133435 | 3.667105 |
| HIPPO signaling | 2.785495 | 5.395457 | 1.740107 |
| HMGB1 Signaling | 4.891232 | 1.568172 | 2.82961 |
| Human Embryonic Stem Cell Pluripotency | 2.587956 | 3.309917 | 1.981469 |
| IGF-1 Signaling | 2.977284 | 2.613001 | 2.290066 |
| IL-15 Signaling | 2.821795 | 1.441507 | 2.185325 |
| IL-17 Signaling | 2.515274 | 2.452659 | 4.169711 |
| IL-17A Signaling in Fibroblasts | 1.597324 | 2.516619 | 6.070165 |
| IL-6 Signaling | 5.092213 | 3.372676 | 4.838371 |
| IL-7 Signaling Pathway | 5.062841 | 2.852634 | 3.640857 |

| Canonical Pathways | D1 | D2 | D3 |
|---|---|---|---|
| IL-8 Signaling | 5.534844 | 2.878552 | 5.281046 |
| ILK Signaling | 3.17568 | 2.536524 | 4.294798 |
| Integrin Signaling | 3.98307 | 3.168003 | 1.875243 |
| JAK/Stat Signaling | 3.790389 | 2.965972 | 3.753987 |
| Leukocyte Extravasation Signaling | 2.03031 | 2.301522 | 2.348736 |
| Molecular Mechanisms of Cancer | 5.307066 | 5.63485 | 4.325399 |
| Mouse Embryonic Stem Cell Pluripotency | 1.508663 | 2.940608 | 1.885225 |
| Neuregulin Signaling | 5.062841 | 2.452659 | 2.676043 |
| NF-κB Signaling | 5.555803 | 2.589104 | 3.873892 |
| NRF2-mediated Oxidative Stress Response | 5.603326 | 5.03428 | 4.930995 |
| Osteoarthritis Pathway | 7.043288 | 9.387797 | 8.868343 |
| p53 Signaling | 4.788103 | 4.264839 | 5.68214 |
| p70S6K Signaling | 2.247157 | 2.821382 | 2.866489 |
| Pancreatic Adenocarcinoma Signaling | 4.409603 | 1.905684 | 2.749329 |
| PEDF Signaling | 2.452856 | 3.602853 | 3.53153 |
| Phospholipase C Signaling | 3.430408 | 1.510878 | 1.662007 |
| PI3K Signaling in B Lymphocytes | 2.365433 | 3.422913 | 3.058173 |
| PI3K/AKT Signaling | 4.750667 | 4.597864 | 3.305299 |
| PPARα/RXRα Activation | 5.591886 | 2.06237 | 3.518141 |
| Production of Nitric Oxide and Reactive Oxygen Species in Macrophages | 2.754767 | 2.865988 | 3.542982 |
| Prolactin Signaling | 2.044714 | 2.17237 | 2.767532 |
| Protein Kinase A Signaling | 6.552419 | 6.728815 | 4.192154 |
| PTEN Signaling | 5.666282 | 2.182808 | 2.710541 |
| Pyridoxal 5'-phosphate Salvage Pathway | 2.955853 | 3.812655 | 3.728923 |
| RANK Signaling in Osteoclasts | 3.280535 | 3.489395 | 4.584293 |
| Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes | 2.749788 | 2.418637 | 2.540031 |
| Regulation of the Epithelial-Mesenchymal Transition Pathway | 4.142112 | 4.153821 | 3.892539 |
| Role of IL-17A in Arthritis | 2.142947 | 2.475979 | 5.228995 |
| Role of IL-17F in Allergic Inflammatory Airway Diseases | 2.51678 | 1.347999 | 2.74813 |
| Role of JAK2 in Hormone-like Cytokine Signaling | 3.169118 | 2.610244 | 2.319303 |
| Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis | 7.704222 | 10.22705 | 7.500913 |
| Role of NFAT in Cardiac Hypertrophy | 2.964131 | 4.263755 | 2.621085 |
| Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis | 4.586137 | 6.058212 | 3.692843 |
| Role of Tissue Factor in Cancer | 3.466248 | 1.316639 | 1.696787 |
| Salvage Pathways of Pyrimidine Ribonucleotides | 2.392311 | 2.260049 | 2.502588 |
| Semaphorin Signaling in Neurons | 2.160571 | 1.34905 | 1.751605 |
| Signaling by Rho Family GTPases | 2.464832 | 1.7461 | 3.506808 |
| STAT3 Pathway | 3.466248 | 1.316639 | 5.058376 |
| Superpathway of D-myo-inositol (1,4,5)-trisphosphate Metabolism | 2.237757 | 2.503949 | 1.88316 |
| Superpathway of Inositol Phosphate Compounds | 3.589532 | 2.905062 | 2.061907 |
| Tec Kinase Signaling | 4.403678 | 1.727516 | 1.358031 |
| Thrombin Signaling | 2.447693 | 1.857448 | 1.825301 |
| TNFR2 Signaling | 2.604299 | 3.731476 | 5.946274 |
| Type II Diabetes Mellitus Signaling | 5.569767 | 4.350193 | 3.499235 |
| VDR/RXR Activation | 3.708659 | 4.896869 | 4.007937 |
| VEGF Signaling | 4.229428 | 3.095633 | 1.98179 |

| Canonical Pathways | D1 | D2 | D3 |
|---|---|---|---|
| Wnt/β-catenin Signaling | 3.333721 | 6.831619 | 4.49353 |
| Xenobiotic Metabolism Signaling | 1.761764 | 1.937908 | 2.578716 |
| Adipogenesis pathway | 2.389909 | 4.292143 | 3.97629 |
| Induction of Apoptosis by HIV1 | 1.870271 | 1.726929 | 1.849297 |
| Adrenomedullin signaling pathway | 2.525959 | 1.965462 | 1.914548 |
| Apoptosis Signaling | 2.98057 | 4.48977 | 1.742702 |
| April Mediated Signaling | 2.843392 | 1.686666 | 3.927846 |
| B Cell Activating Factor Signaling | 2.7277 | 1.564348 | 3.731838 |
| Circadian Rhythm Signaling | 3.241586 | 1.605227 | 2.395433 |
| D-myo-inositol (1,4,5,6)-Tetrakisphosphate Biosynthesis | 2.157214 | 1.482325 | 1.400982 |
| D-myo-inositol (3,4,5,6)-tetrakisphosphate Biosynthesis | 2.157214 | 1.482325 | 1.400982 |
| Factors Promoting Cardiogenesis in Vertebrates | 1.539124 | 2.503278 | 1.511383 |
| GADD45 Signaling | 1.659524 | 2.297122 | 1.538295 |
| Hepatic Fibrosis / Hepatic Stellate Cell Activation | 3.946922 | 1.346507 | 5.203837 |
| IL-1 Signaling | 2.579651 | 1.819164 | 2.323672 |
| iNOS Signaling | 3.26515 | 1.347999 | 4.059563 |
| NF-κB Activation by Viruses | 3.015909 | 1.358075 | 1.422867 |
| LPS-stimulated MAPK Signaling | 2.452856 | 1.358075 | 2.164607 |
| p38 MAPK Signaling | 2.823452 | 1.462723 | 2.116711 |
| Phagosome Formation | 4.032202 | 1.316639 | 1.696787 |
| Phosphatidylethanolamine Biosynthesis II | 1.494239 | 1.396366 | 1.796522 |
| PPAR Signaling | 2.811401 | 2.082144 | 2.340891 |
| RhoA Signaling | 1.744632 | 2.182808 | 1.622267 |
| Role of PKR in Interferon Induction and Antiviral Response | 2.009216 | 1.564348 | 3.0521 |
| Small Cell Lung Cancer Signaling | 3.31868 | 1.335007 | 3.992418 |
| T Cell Exhaustion Signaling Pathway | 2.628443 | 1.440375 | 1.652109 |
| Toll-like Receptor Signaling | 5.276365 | 2.144276 | 4.143018 |
| TWEAK Signaling | 3.099294 | 1.965551 | 5.19109 |
| Wnt/Ca+ pathway | 1.837685 | 2.523677 | 1.385245 |
| Type I Diabetes Mellitus Signaling | 2.014714 | 1.693589 | 2.774934 |

*z-score is a statistical measure of the match between expected relationship direction and observed gene expression. A z-score > 2 or < -2 is considered significant. Note that the actual z-score is weighted by the underlying findings, the relationship bias, and dataset bias.

**Table 2.3 Canonical IPA pathways from robust gene list with log(p-values).**

| Ingenuity Canonical Pathways | -log(p-value) |
|---|---|
| Glucocorticoid Receptor Signaling | 6.56E+00 |
| Colorectal Cancer Metastasis Signaling | 5.93E+00 |
| NRF2-mediated Oxidative Stress Response | 5.54E+00 |
| IL-7 Signaling Pathway | 5.17E+00 |
| p53 Signaling | 5.10E+00 |
| Molecular Mechanisms of Cancer | 5.05E+00 |
| Protein Kinase A Signaling | 4.96E+00 |
| G-Protein Coupled Receptor Signaling | 4.70E+00 |
| Insulin Receptor Signaling | 4.61E+00 |
| Aldosterone Signaling in Epithelial Cells | 4.45E+00 |
| Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis | 4.41E+00 |
| Neuregulin Signaling | 4.39E+00 |
| B Cell Receptor Signaling | 3.94E+00 |
| ErbB Signaling | 3.91E+00 |
| Wnt/β-catenin Signaling | 3.90E+00 |
| Integrin Signaling | 3.88E+00 |
| Cholecystokinin/Gastrin-mediated Signaling | 3.87E+00 |
| PI3K/AKT Signaling | 3.84E+00 |
| Phagosome Formation | 3.76E+00 |
| VEGF Signaling | 3.74E+00 |
| ERK/MAPK Signaling | 3.69E+00 |
| Osteoarthritis Pathway | 3.56E+00 |
| p70S6K Signaling | 3.53E+00 |
| HMGB1 Signaling | 3.51E+00 |
| VDR/RXR Activation | 3.48E+00 |
| Paxillin Signaling | 3.44E+00 |
| Human Embryonic Stem Cell Pluripotency | 3.40E+00 |
| PTEN Signaling | 3.38E+00 |
| Tec Kinase Signaling | 3.35E+00 |
| Regulation of the Epithelial-Mesenchymal Transition Pathway | 3.33E+00 |
| Role of NFAT in Cardiac Hypertrophy | 3.26E+00 |
| cAMP-mediated signaling | 3.24E+00 |
| Pyridoxal 5'-phosphate Salvage Pathway | 3.24E+00 |
| Type II Diabetes Mellitus Signaling | 3.16E+00 |
| Circadian Rhythm Signaling | 3.13E+00 |
| T Cell Exhaustion Signaling Pathway | 3.11E+00 |

| Ingenuity Canonical Pathways | -log(p-value) |
|---|---|
| Macropinocytosis Signaling | 3.09E+00 |
| Endocannabinoid Cancer Inhibition Pathway | 3.09E+00 |
| Adipogenesis pathway | 3.09E+00 |
| Role of JAK2 in Hormone-like Cytokine Signaling | 3.07E+00 |
| Superpathway of Inositol Phosphate Compounds | 3.06E+00 |
| IGF-1 Signaling | 3.03E+00 |
| JAK/Stat Signaling | 3.03E+00 |
| PPARα/RXRα Activation | 3.03E+00 |
| Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses | 2.97E+00 |
| HER-2 Signaling in Breast Cancer | 2.88E+00 |
| Salvage Pathways of Pyrimidine Ribonucleotides | 2.85E+00 |
| Glioblastoma Multiforme Signaling | 2.83E+00 |
| ILK Signaling | 2.81E+00 |
| Neuroinflammation Signaling Pathway | 2.75E+00 |
| IL-8 Signaling | 2.69E+00 |
| SPINK1 General Cancer Pathway | 2.64E+00 |
| Germ Cell-Sertoli Cell Junction Signaling | 2.63E+00 |
| Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis | 2.63E+00 |
| HIPPO signaling | 2.55E+00 |
| NF-κB Signaling | 2.51E+00 |
| Gαq Signaling | 2.47E+00 |
| FAT10 Cancer Signaling Pathway | 2.47E+00 |
| Cell Cycle: G1/S Checkpoint Regulation | 2.44E+00 |
| IL-3 Signaling | 2.41E+00 |
| Prolactin Signaling | 2.39E+00 |
| Nitric Oxide Signaling in the Cardiovascular System | 2.39E+00 |
| Factors Promoting Cardiogenesis in Vertebrates | 2.36E+00 |
| Production of Nitric Oxide and Reactive Oxygen Species in Macrophages | 2.35E+00 |
| IL-17 Signaling | 2.34E+00 |
| Gα12/13 Signaling | 2.33E+00 |
| Virus Entry via Endocytic Pathways | 2.32E+00 |
| Caveolar-mediated Endocytosis Signaling | 2.31E+00 |
| IL-4 Signaling | 2.26E+00 |
| HGF Signaling | 2.25E+00 |
| Actin Cytoskeleton Signaling | 2.25E+00 |
| Pancreatic Adenocarcinoma Signaling | 2.23E+00 |
| PDGF Signaling | 2.21E+00 |
| Toll-like Receptor Signaling | 2.17E+00 |

| Ingenuity Canonical Pathways | -log(p-value) |
|---|---|
| 1D-myo-inositol Hexakisphosphate Biosynthesis II (Mammalian) | 2.16E+00 |
| D-myo-inositol (1,3,4)-trisphosphate Biosynthesis | 2.16E+00 |
| Coagulation System | 2.14E+00 |
| Apelin Endothelial Signaling Pathway | 2.13E+00 |
| Thrombin Signaling | 2.12E+00 |
| Glioma Invasiveness Signaling | 2.11E+00 |
| GADD45 Signaling | 2.09E+00 |
| Axonal Guidance Signaling | 2.09E+00 |
| Role of NANOG in Mammalian Embryonic Stem Cell Pluripotency | 2.07E+00 |
| Phospholipase C Signaling | 2.07E+00 |
| ErbB4 Signaling | 2.06E+00 |
| FAK Signaling | 2.01E+00 |
| PI3K Signaling in B Lymphocytes | 1.98E+00 |
| Dopamine-DARPP32 Feedback in cAMP Signaling | 1.97E+00 |
| Huntington's Disease Signaling | 1.97E+00 |
| Signaling by Rho Family GTPases | 1.97E+00 |
| IL-15 Signaling | 1.96E+00 |
| IL-6 Signaling | 1.96E+00 |
| Ephrin A Signaling | 1.92E+00 |
| Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes | 1.92E+00 |
| Apelin Cardiac Fibroblast Signaling Pathway | 1.91E+00 |
| 14-3-3-mediated Signaling | 1.91E+00 |
| Inositol Pyrophosphates Biosynthesis | 1.90E+00 |
| Small Cell Lung Cancer Signaling | 1.89E+00 |
| Regulation of Cellular Mechanics by Calpain Protease | 1.89E+00 |
| Chronic Myeloid Leukemia Signaling | 1.89E+00 |
| Erythropoietin Signaling | 1.87E+00 |
| Mouse Embryonic Stem Cell Pluripotency | 1.87E+00 |
| Superpathway of D-myo-inositol (1,4,5)-trisphosphate Metabolism | 1.86E+00 |
| Sirtuin Signaling Pathway | 1.86E+00 |
| CXCR4 Signaling | 1.83E+00 |
| eNOS Signaling | 1.80E+00 |
| Fcγ Receptor-mediated Phagocytosis in Macrophages and Monocytes | 1.78E+00 |
| TGF-β Signaling | 1.76E+00 |
| Ephrin Receptor Signaling | 1.76E+00 |
| mTOR Signaling | 1.74E+00 |
| Epithelial Adherens Junction Signaling | 1.74E+00 |
| Lymphotoxin β Receptor Signaling | 1.74E+00 |

| Ingenuity Canonical Pathways | -log(p-value) |
|---|---|
| ERK5 Signaling | 1.74E+00 |
| Clathrin-mediated Endocytosis Signaling | 1.73E+00 |
| LPS-stimulated MAPK Signaling | 1.72E+00 |
| NF-κB Activation by Viruses | 1.72E+00 |
| Antiproliferative Role of TOB in T Cell Signaling | 1.72E+00 |
| Ovarian Cancer Signaling | 1.71E+00 |
| EGF Signaling | 1.71E+00 |
| Cardiac Hypertrophy Signaling | 1.71E+00 |
| VEGF Family Ligand-Receptor Interactions | 1.70E+00 |
| Leukocyte Extravasation Signaling | 1.68E+00 |
| Thrombopoietin Signaling | 1.66E+00 |
| Glutamine Biosynthesis I | 1.65E+00 |
| Sphingosine-1-phosphate Signaling | 1.65E+00 |
| Fc Epsilon RI Signaling | 1.64E+00 |
| TR/RXR Activation | 1.62E+00 |
| Th1 and Th2 Activation Pathway | 1.62E+00 |
| AMPK Signaling | 1.61E+00 |
| Natural Killer Cell Signaling | 1.59E+00 |
| Renin-Angiotensin Signaling | 1.59E+00 |
| RAR Activation | 1.58E+00 |
| Hepatic Cholestasis | 1.58E+00 |
| Role of Tissue Factor in Cancer | 1.56E+00 |
| STAT3 Pathway | 1.56E+00 |
| Pregnenolone Biosynthesis | 1.55E+00 |
| ErbB2-ErbB3 Signaling | 1.55E+00 |
| RANK Signaling in Osteoclasts | 1.55E+00 |
| TNFR2 Signaling | 1.55E+00 |
| Semaphorin Signaling in Neurons | 1.53E+00 |
| Docosahexaenoic Acid (DHA) Signaling | 1.50E+00 |
| Androgen Signaling | 1.49E+00 |
| Th1 Pathway | 1.48E+00 |
| CD40 Signaling | 1.47E+00 |
| Xenobiotic Metabolism Signaling | 1.45E+00 |
| GNRH Signaling | 1.44E+00 |
| 3-phosphoinositide Biosynthesis | 1.44E+00 |
| SAPK/JNK Signaling | 1.41E+00 |
| Amyotrophic Lateral Sclerosis Signaling | 1.40E+00 |
| D-myo-inositol (1,4,5,6)-Tetrakisphosphate Biosynthesis | 1.38E+00 |

| Ingenuity Canonical Pathways | -log(p-value) |
|---|---|
| D-myo-inositol (3,4,5,6)-tetrakisphosphate Biosynthesis | 1.38E+00 |
| Apelin Pancreas Signaling Pathway | 1.38E+00 |
| TWEAK Signaling | 1.38E+00 |
| Angiopoietin Signaling | 1.37E+00 |
| Histidine Degradation VI | 1.37E+00 |
| Acute Phase Response Signaling | 1.37E+00 |
| Gap Junction Signaling | 1.37E+00 |
| Glutamine Degradation I | 1.36E+00 |
| Neuropathic Pain Signaling In Dorsal Horn Neurons | 1.34E+00 |
| Induction of Apoptosis by HIV1 | 1.34E+00 |
| IL-12 Signaling and Production in Macrophages | 1.33E+00 |
| p38 MAPK Signaling | 1.32E+00 |
| Actin Nucleation by ARP-WASP Complex | 1.31E+00 |
| G Beta Gamma Signaling | 1.31E+00 |
| Telomerase Signaling | 1.31E+00 |

TABLE 3.1

**Basal expression changes of genes +/- 1 Mb from GRE deletion on chromosome 10.**

| Gene | Coordinates | $log_2$Fold Change | q-value |
|---|---|---|---|
| *LOC101926942* | 92,162,278-92,300,562 | ND | ND |
| *HTR7* | 92,500,576-92,617,671 | ND | ND |
| *RPP30* | 92,631,474-92,668,312 | 0.05 | 0.99 |
| *ANKRD1* | 92,671,857-92,681,032 | -1.52 | 0.01 |
| *XLOC 008559* | 92,707,057-92,751,889 | ND | ND |
| *LOC105378430* | 92,792,923-92,801,012 | ND | ND |
| *LINC00502* | 92,805,565-92,821,916 | ND | ND |
| *NUDT9P1* | 92,911,761-92,912,837 | -0.06 | 0.99 |
| *PCGF5* | 92,922,769-93,044,088 | 0.003 | 0.99 |
| *HECTD2-AS1* | 93,066,719-93,371,217 | 2.84 | 0.85 |
| *HECTD2* | 93,170,096-93,274,520 | 0.28 | 0.96 |
| *PPP1R3C* | 93,388,197-93,392,858 | 0.28 | 0.86 |
| *TNKS2-AS1* | 93,542,596-93,558,048 | ND | ND |
| *TNKS2* | 93,558,151-93,625,232 | 0.25 | 0.96 |
| *FGFBP3* | 93,666,345-93,669,258 | 0.65 | 0.40 |
| *BTAF1* | 93,683,736-93,790,080 | -0.06 | 0.99 |

[a] Row shading indicates differentially regulated genes with a q-value < 0.05.

TABLE 3.2

**Genes +/- 1Mb of GRE deletion on chromosome 10 whose drug effects are different in the presence of mutation.**

| Gene | Coordinates | $log_2$Fold Change | q-value |
|---|---|---|---|
| *LOC101926942* | 92,162,278-92,300,562 | ND | ND |
| *HTR7* | 92,500,576-92,617,671 | 2.1 | 1 |
| *RPP30* | 92,631,474-92,668,312 | -0.03 | 1 |
| *ANKRD1* | 92,671,857-92,681,032 | 2 | 7.39E-10 |
| *XLOC 008559* | 92,707,057-92,751,889 | -0.6 | 1 |
| *LOC105378430* | 92,792,923-92,801,012 | ND | ND |
| *LINC00502* | 92,805,565-92,821,916 | ND | ND |
| *NUDT9P1* | 92,911,761-92,912,837 | 1.5 | 1 |
| *PCGF5* | 92,922,769-93,044,088 | 0.38 | 1 |
| *HECTD2-AS1* | 93,066,719-93,371,217 | -2.95 | 1 |
| *HECTD2* | 93,170,096-93,274,520 | -0.21 | 1 |
| *PPP1R3C* | 93,388,197-93,392,858 | 1 | 1 |
| *TNKS2-AS1* | 93,542,596-93,558,048 | 0.89 | 1 |
| *TNKS2* | 93,558,151-93,625,232 | 0.26 | 1 |
| *FGFBP3* | 93,666,345-93,669,258 | 0.08 | 1 |
| *BTAF1* | 93,683,736-93,790,080 | 0.34 | 1 |

[a] Row shading indicates differentially regulated genes with a q-value < 0.05.

TABLE 3.3

**Candidate genes whose expression is primarily affected by GRE deletion.**

| Gene | Coordinates | $log_2$Fold Change | q-value |
|---|---|---|---|
| *GPR153* | Chr1: 6,247,346 - 6,260,990 | 2.30 | 0.02 |
| *ETNK2* | Chr1: 204,131,061 - 204,152,182 | 1.80 | 0.01 |
| *ZBTB18* | Chr1: 244,048,939 - 244,057,476 | -1.13 | 0.03 |
| *ANKRD1* | Chr10: 90,912,096 - 90,921,276 | -2.01 | 0.19 |
| *SALL1* | Chr16: 51,135,975 - 51,152,316 | 1.84 | 0.02 |

Differentially regulated genes with a q-value < 0.2.

**Publishing Agreement**

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

***Please sign the following statement:***

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

Daley Martinez
Author Signature

8/31/19
Date