On the evolution of gene transcription regulation in metazoans

by

D. A. Ramirez Hernandez

B.S., Universidad Nacional Autónoma de México, 2015

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Molecular, Cellular, and Developmental Biology

2023

Committee Members:

Sara Sawyer, Chair

Robin Dowell

Edward Chuong

Shelley Copley

Nolan Kane

Ramirez Hernandez, D. A. (Ph.D., Molecular, Cellular, and Developmental Biology)

On the evolution of gene transcription regulation in metazoans

Thesis directed by Prof. Robin Dowell

The regulation of gene expression is essential for organisms to respond and adapt to their ever changing environments. They do so by sensing stimuli, transducing the signals, and then mounting specific, appropriate transcription programs. Through time, DNA is shaped by both natural selection and drift, ultimately acquiring changes that rewire gene transcription regulatory networks.

Here, I present work on deciphering how the regulation of gene transcription, upon two signaling responses, has evolved across animals. First, I describe the primary and downstream transcription response across the primate phylogeny with the activation of the transcription factor p53, which orchestrates the cellular response to stressors such as DNA damage. Second, I present a dataset on the primary and downstream transcriptional responses across animals, and within diverse human ethnic populations, of the cellular response to pathogens as directed by the cytokine interferon, specifically IFN-β. Finally, I show preliminary results on chromatin condensation changes arising upon abrupt mechanical stressors on human and pig cells. The investigation presented here showcases the dynamic ways in which different species evolved to sense and react to their unique surroundings in order to survive.

Dedication

To my family and friends.

Acknowledgements

I thank the Dowell and Allen Labs (past and current members), the BioFrontiers Next Generation Sequencing Core Facilities, the Biochemistry Cell Culture Core, the Molecular Cellular and Developmental Biology department, the Interdisciplinary Quantitative Biology program, el Consejo Nacional de Ciencia y Tecnología de México (CONACYT), the National Science Foundation (NSF) for ABI grant number 1759949, the National Institutes of Health (NIH) for grant R01 GM12587, the SMART and STEM Routes student groups, and Angélica Ramírez for the primate silhouettes.

And the following labs for sharing their primate cell lines: Sara Sawyer's Lab at the University of Colorado, Yoav Gilad's Lab at the University of Chicago, Evan Eichler's Lab at the University of Washington, and Lucia Carbone's Lab at Oregon Health & Science University.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Background

1.1 Introduction

On an unremarkable rocky planet in the Orion Arm of the Milky Way galaxy, a long time ago a series of chemical reactions slowly began the formation of self-replicating entities that human scientists may define as life [213]. Countless iterations and prototypes must have been tried until evolution chose a sufficiently good chemical substrate in which to encode the instructions for perpetuating those primeval life forms. The last common ancestor for the extant life on Earth is estimated to have appeared around 3.8 billion years ago [201, 6]. Over time, these DNA- and RNA-based lifeforms diversified by means of natural selection, colonizing the surface of the planet Earth, bringing about a tremendous variety of shapes and forms, and filling all possible environments [177]. Populations are altered by evolution to better survive and pass on their genetic material to their offspring [52]. Understanding these evolutionary processes is paramount to better recognizing the place of species in the tree of DNA-based life. Surely, it is a worthy goal for us to understand how sufficiently intelligent primates evolved on Earth and then developed the scientific tools necessary to ask themselves: What are we, and what is this universe?

The terrestrial tree of life is divided into three broad domains: the Archaea, Bacteria, and Eukarya [202]. In this text, however, I limited my study to multicellular eukaryotes: the domain that arose when, according to the endosymbiotic theory [162], some free-living prokaryotes were taken inside other free-living prokaryotes, giving rise to nucleated cells. In particular, I focused my investigation on the evolution of the molecular mechanisms that control how, when, and where

genes are transcribed; the first step for gene expression [26].

My research follows my desire to better understand how gene transcriptional regulation evolves, allowing organisms to explore their fitness landscape by fine-tuning the expression of their genetic information. I set out to improve knowledge on how terrestrial life has diversified throughout the surface of our planet, giving rise to endless forms, but also shaping complex genetic networks that allow organisms to respond and adapt to abrupt perturbations in their environments. The perturbations of interest were in the form of stressors that influence the integrity of their genomes, the response to infection, and mechanical forces that disrupt cell growth.

In this chapter, I outline the background material necessary to understand the work presented in this thesis. Specifically I will first describe the mechanisms of eukaryotic gene transcription regulation (Section 1.2), as the systems I studied are all within this domain. Next, I will cover the molecular approaches employed in studies on gene transcription regulation (Section 1.3). Then, a discussion of the mechanisms by which gene transcription regulation evolves across species (Section 1.4), a widespread phenomenon harnessed to acquire new phenotypes. I will then provide background into three regulatory pathways: TP53 (Section 1.5), interferon (Section 1.6), and cellular mechanosensing (Section 1.7) as these are these systems employed in my studies. Finally, I will end with a preview of the remainder of the thesis (Section 1.8).

1.2 Mechanisms of eukaryotic gene transcription regulation

1.2.1 On genes and their structure

In order to understand the evolution of the regulatory mechanisms that control eukaryotic gene expression, I need to start with defining what a gene is. Historically, the concept of a gene was coined before we even understood what DNA was. A gene was an abstract concept that simply denoted an "unit of trait inheritance".

The structure of DNA was elucidated in 1953, [202, 100, 121], and it was described as being composed by two opposite oriented strands forming a double helix. By convention, each strand is

referred to from the 5' to the 3' direction, referring to the position of the phosphate group linked to the 5' carbon of the ribose sugar ring, and the position of the next nucleotide attached to the 3' carbon of the same ribose molecule, respectively.

After the discovery of the DNA structure, genes were subsequently defined as discrete points in chromosomes, and later on as concrete intervals of DNA within those chromosomes [150]. Here, I will sidestep the troublesome, albeit interesting, philosophical debate of what the above definition really means, and will align with the Encyclopedia of DNA Elements Consortium (ENCODE) definition: A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products [67].

Eukaryotic genes are composed of many DNA elements, and come in two loosely defined categories: genes that code for proteins, and genes that do not code for proteins. In either case, they have an initiation region at their 5' end, referred to as the promoter, where the transcription machinery is assembled to transcribe the gene, converting the DNA information into RNA transcripts. In protein-coding genes, these transcripts have two regions at both of their 5' and 3' ends, called untranslated regions (UTR), that contain many regulatory elements that control translation initiation, transcript stability and degradation, sometimes by forming secondary structures [13]. Internally, protein-coding genes are composed of stretches of DNA called exons and introns. As genes are transcribed into RNA, the introns are removed and the two adjacent exons are spliced together in a process called splicing. Exons contain the genetic information that dictates the amino acid sequence of the proteins that the gene codes for, following a degenerate genetic code where triplets of nucleotides code for single amino acids [47]. Introns, on the other hand, do not code for amino acid sequence and their function remains debated, as they have poor conservation across species. Introns, however, are known to affect transcription initiation and splicing, and sometimes contain smaller regulatory non-coding genes [36].

The number of genes, including protein-coding and non-coding genes, varies significantly in eukaryotes. This number ranges from as few as ∼12,000 genes in the budding yeast Saccharomyces cerevisiae [16], to ∼63,000 genes in humans [138], to as many as ∼110,000 genes in the common

wheat [38]. However, not all genes are expressed at the same time. Instead, their expression is restricted to specific times (e.g. during development or as a response to certain stressors),or to specific cells within tissues (e.g. blood immune cells, neurons, adipocytes).

1.2.2 On transcription factors, promoters, enhancers, and eRNAs

The transcription of a gene is regulated by the concerted activity of diffusible transcription factors (also referred as trans-regulatory elements), and of DNA sequences (also referred to as cis-regulatory elements) that are either proximal (i.e. promoters), or distal (i.e. enhancers, silencers, and insulators) to the genes they regulate.

With regards to the trans-acting regulatory elements, transcription factors (TFs) are proteins that bind with high affinity to specific DNA short sequences (i.e. motif instances) and help with the recruitment of the transcription machinery to transcribe their target genes. TFs can sometimes work as single proteins, or as mutimeric TF protein complexes [174]. In humans, there are currently around 800 annotated TFs (i.e. with direct or indirect evidence for gene transcription regulation activity), though it is estimated that there are around twice as many TFs total in our genome [106]. They generally are composed of at least two protein elements, a DNA-binding domain (DBD), and a docking site for other proteins such as a subunit of the RNA polymerase II (Pol II), a chromatin remodeler, a cofactor, or another TF. However, they can be as simple as a single DBD that influences gene expression by solely precluding the binding of other proteins. Some TFs contain multiple DBDs, as in the case of the zinc-finger TF family, which is posited to undergo recurrent gene duplication events while switching between different DBDs modules to acquire new unique DNA-binding specificities. Other important DBDs are C2H2 arrays, homeodomains, T-box, AT hook, Forkhead, bZIP, bHLH, and Ets. Some TFs have a very narrow set of genes they regulate, whereas others are general TFs (e.g. TBP, TFIIA, TFIIB, and others) that are known to be involved in the recruitment of Pol II for most genes.

Depending on the TF, they either can or cannot bind to DNA while it is coiled around nucleosomes. TFs that can recognize their targets in tightly packaged DNA are known as pioneer

TFs, and they work by recruiting chromatin remodelers to unwound and expose DNA so that other TFs can bind to their otherwise inaccessible sequence motifs [210]. In addition, when multiple TFs bind to a single regulatory DNA sequence, they can act in an additive fashion where the transcriptional output is increased linearly with the number of TF bound to them; or they can act as a tightly controlled binary switch, where the gene is not transcribed until all their TFs have bound to them simultaneously [49].

With regard to the cis-acting regulatory elements, promoters and enhancers are DNAsequences that act as landing pads to specific proteins (e.g. TFs) to activate the transcription of their target gene. The final product of a successful interaction between TFs and promoters and enhancers, is the assembly of a pre-initiation complex, composed by Pol II and its general transcription factors [188]. Historically, promoters have been defined as the locus where transcription of the gene initiates, while enhancers are defined as distal loci that enhance or facilitate transcription. However, recent studies have described promoters with enhancer activity to distal genes[51]. The more the field advances, the blurrier the boundary between promoters and enhancers becomes.

Nonetheless, the field of gene transcription regulation defines promoters as those DNA regions close to the transcription start site (TSS) that have the following features. They contain a core promoter sequence, consisting of a short region of around 100 base pairs (bp) of length centered at the TSS. These core promoters are generally able to initiate transcription by themselves, but at very low levels. They encode the binding sequence to recruit general transcription factors and Pol II itself. In a minority of promoters, there is a small sequence referred to as the TATA-box, found $\sim$30 bp upstream of the TSS, which is widely conserved across eukaryotes. A bigger proportion of core promoters have an initiator motif called Ihr, which is encoded overlapping the TSS, and close to the Ihr there is a downstream promoter element (DPE). Promoters are also characterized by an elevated GC content, and in vertebrates by having high density of CpG dinucleotides. Further, promoter sequences, when active, are depleted from nucleosomes. Nucleosomes downstream of the TSS are precisely positioned, and are generally composed with the histone protein variants such as H3.3 and H2A.Z. These nucleosomes are also modified post translationally by adding up to

three methyl groups in their histone H3 lysine 4 (H3K4me3), and by adding an acetyl group at histone 3 lysine 27 (H3K27ac). The purpose of such chemical modifications is not fully understood [188]. Some studies show that they are deposited before transcription serving as landing pads for chromatin remodelers, such as p300, to facilitate transcription initiation; while other studies suggest that they are deposited after transcription serving as epigenetic memory to facilitate other rounds of transcription [10].

As the ability to identify enhancers improved, the field realized that enhancers, while located distally from the genes they regulate, shared many of the same properties found in promoters [166]. Namely, they have similar motif sequences that allow the assembly of the pre-initiation complex, they experience similar nucleosome positioning dynamics, they have similarly high GC content – although they are not enriched for CpG dinucleotides, they are marked with similar histone variants and post-translational modifications; and have been also observed to be be transcribed when active. Enhancers, it turned out, are transcribed from both DNA strands in a bidirectional manner when they are used to regulate their target genes (i.e. with two Pol II complexes loaded in both strands and transcribing away from the enhancer locus), generating RNA transcripts that are rapidly degraded, named enhancer associated RNAs (eRNAs) [4, 205]. Furthermore, we now know that promoters are also bidirectionally transcribed, but in their case the transcript originating from the strand that contains the gene is successfully elongated and processed, while the other antisense transcript is rapidly degraded just as with eRNAs [188].

Some studies propose that both enhancers and promoters share so many properties because they are essentially the same type of regulatory element observed in different stages in their evolutionary trajectory [82]. They propose that enhancers originate fortuitously, and that over time one of the two strands acquire sequence determinants that stabilize the otherwise unstable eRNA transcripts, forming non-coding genes, and that such genes may then acquire coding potential. Indeed, bidirectional transcription has been observed in all three domains of terrestrial life, including in the promoter regions from bacteria and archaea [197].

The current paradigm posits that promoters and enhancers are in close spatial proximity

to each other when they are actively regulating their target genes [166]. Proteins such as CTCF, cohesin, and YY1, are thought to bring together these loci even though they may be hundreds of thousands of base pairs away in the linear DNA dimension. Once an enhancer and a promoter are in close 3D proximity, the transcription factors that are bound to both of their regulatory DNA sequences, together with cofactors, Pol II, and the multi-subunit mediator complex, serve as an organized protein bridge that stabilizes the contact between both DNA loci [158]. It is thought that promoters become activated by enhancers either by promoting the formation of a fully operational pre-initiation complex itself or by regulating the successful pause-release of Pol II so that it can transition to its transcription elongation phase [166]. Some reports have shown that for signaling-triggered transcriptional responses, the enhancers and their promoters are already primed, in contact with each other but without transcription occurring, even before their their signal has been sensed, possibly as a mechanism to obtain a quick transcriptional response [59].

It is now understood that transcription of a gene does not happen continuously, but rather it occurs in episodes of rapid transcription followed by periods of inactivity that have been defined as transcriptional bursts [108, 185]. These dynamics suggest that the entire regulatory apparatus described above may not be very stable, rendering the whole process able to shift stochastically from inactive to active states, back and forth.

Importantly, the role of eRNAs in gene transcription regulation has not been fully elucidated. eRNAs have been described with variable characteristics, with examples of them being short and long, spliced and unspliced, polyadenylated and non-polyadenylated. They are thought to have many functions, and even to not have a function whatsoever but be mere byproducts of the regulatory process [105]. It makes sense to think of them as a substrate on which evolution has attributed different roles, just as with other types of RNA molecules in cells. Studies have suggested that eRNA can work to stabilize the enhancer and promoter contact loops by working as additional binding platforms for the many proteins that are recruited in these loops. They can recruit additional TFs to strengthen the transcriptional response, and they can recruit chromatin remodelers to make the locus more accessible for the rest of the transcription machinery. They

can also serve by diluting away negative regulators, such as by sequestering (by act of binding) the negative elongation factor NELF to promote successful transcription elongation of the gene they help regulate. The eRNAs can work in cis, by acting on the same locus that they are transcribed from, or to work in trans in other promoter enhancer contacts after they are disloged from the Pol II that synthesized them [7, 163].

1.2.3 On the stages of transcription and RNA processing

All of the above regulatory mechanisms have to occur in order for the transcription of a gene to successfully initiate. However, for a gene to be expressed, many other subsequent molecular activities must take place. The (presumptuously named [48]) central dogma of molecular biology, after all, describes that a gene is transcribed from DNA to RNA, and then it is translated from RNA to protein (which of course only applies for protein-coding genes) [46].In what follows, I will describe in more detail the molecular mechanisms that a nascently transcribed RNA transcript undergoes for it to be fully matured, but will not cover the translation and protein degradation dynamics, as they are beyond the scope of this text.

Pol II does not immediately proceed transcribing through the gene body after it is properly loaded; rather, it stalls in a process termed pausing, about 30 to 60 nucleotides downstream of the TSS [43]. In fact, Pol II is not only paused, but is sometimes dislodged entirely from the DNA as premature termination events. The largest subunit of Pol II, RPB1, has a long unstructured tail at the carboxyl end of its linear sequence, named C-terminal domain (CTD), which in human is composed of 52 repeats of a 7 amino acid long sequence that are heavily modified throughout the transcription cycle in a way that dictates when Pol II can proceed through each step of the transcription process. The kinase CDK7 phosphorylates the serine residues at positions 5 and 7 from the CTD repeats (Ser5 and Ser7), and these modifications are thought to help Pol II escape its tight grip from the mediator complex. These phosphorylated serine residues at the CTD also serve as a landing pad for enzymes that chemically modify the nascent transcripts by adding a 7-methylguanosine at the transcripts 5' end, effectively capping it and protecting it from nucleosomal

degradation [154]. Pol II is stalled at the initiation region by the simultaneous activity of the Negative Elongation Factor (NELF) and Spt5. As long as NELF remains attached to the CTD, Pol II cannot proceed transcribing. The Positive Transcription Elongation Factor (P-TEFb) kinase is then recruited to the promoter region by TFs and the mediator, and phosphorylates Spt5, which in turn dissociates NELF from the complex, releasing Pol II to continue transcribing. The evolution of this pausing mechanism is thought to allow an extra layer of regulation to fine tune the expression levels of genes. Pausing is also thought to help maintain the promoter region devoid of nucleosomes.

After Pol II is released from its pausing, it transcribes through gene bodies at different velocities. The rate depends on the gene, the number of nucleosomes it encounters and histone modifications of these nucleosomes, as well as the number of exons of the gene [85]. In fact, Pol II is the slowest in the first few kilobases, as it still accumulates more Ser2 phosphorylation and dislodges NELF proteins. In general, low complexity DNA sequences and low GC content help Pol II transcribe faster. The nucleosomes associated with active promoters are marked with H3K4me3, and multiple H3 and H4 lysine acetylations. Gene bodies, when transcribed, tend to be marked with mono-ubiquitinated H2B (H2Bub), H3K36me3, H3K79me2 and H3K79me3. The role of each of these modifications, however, has been difficult to ascertain. Further, the kinase P-TEFb associates with other proteins to form the Super Elongation Complex (SEC), and with bromodomain-containing protein 4 (BRD4), which further phosphorylate Ser2 residues on the CTD. As Pol II transcribes, it supercoils the DNA downstream and relaxes the DNA coiling upstream. These two processes, unchecked, severely slow down transcription rates, which is why DNA topoisomerases are known to help alleviate this DNA tension [34].

As eukaryotic genes are transcribed, their RNA sequences are cut and glued back together to remove introns, and to only keep exon sequences in a process called splicing. Human genes have an average of eight introns per gene. But there is also TTN, a gene with up to 362 introns; and intronless genes such as the type I interferon genes IFNA1 and IFNB1 [118]. Splicing occurs simultaneously as transcription, and it is catalyzed by a large ribonucleic complex (consisting of both proteins and RNA molecules) called the spliceosome. In particular, this complex is formed

by 5 distinct ribonucleoproteins (U1, U2, U4, U5, and U6), that are assembled in a combination of monomers, heterodimers, and heterotrimers. Many of these ribonucleoproteins are recruited by the phosphorylated Ser5 of Pol II CTD. This puts the spliceosome in close proximity to the nascent RNA transcript. At the beginning and at the end of an intron of an RNA transcript, there are short RNA sequences known as the 5' splice site (with nucleotide sequence GU) and 3' splice site (with nucleotide sequence AG), respectively. In addition, there are two other sequence elements, called the branch point (a single adenine), and a polypyrimidine tract close to the 3' splice site. The splicing of exons is the result of a complex choreography involving the recognition of these sequence elements by the ribonucleoproteins of the spliceosome, followed by enzymatic cutting and pasting of the two resulting RNA ends [66]. The removal of introns can often be done in different ways for a given gene by skipping introns, resulting in distinct mRNA isoforms for a single gene in a process called alternative splicing, which expands the protein repertoire for a given cell without having to expand its number of genes [11].

Once the gene body has been transcribed, Pol II has to stop transcribing and be released from its template DNA. Not surprisingly, this process of transcription termination is also regulated. This process can be roughly divided into two parts: The first entails finishing transcribing and processing the messenger RNA (mRNA) transcript itself, whereas the second part involves actually stopping the transcription machinery from keeping transcribing. To achieve these processes, several other regulatory protein complexes are recruited as transcription approaches the end of the gene. These include the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulatory factor (CstF), and the cleavage factor 1 and 2 (CFI and CFII); that together form a complex. CPSF directly binds to Pol II and not its CTD, whereas CstF, CFI, and CFII, are recruited by phosphorylated Ser2 at Pol II CTD. At the 3' UTR of a gene sequence, there is a specific motif referred to as the polyadenylation signal (PAS), that once transcribed, is recognized and cleaved by CPSF and CstF, causing the separation of the RNA transcript containing the gene from the rest of the nascent RNA still being synthesized by Pol II. Alongside its cleavage, CPSF also adds a string of adenines to the 3' end of the gene transcript [149]. This polyadenylated string is used by

nuclear export proteins to move the transcript out of the nucleus to be translated, and is also used by ribosomes to stabilize the translation process itself [128]. After the gene transcript cleavage, an exonuclease named XRN2 is recruited to degrade the rest of the nascent transcript coming out of Pol II. It is thought that Pol II is then dislodged from the DNA by the mechanical action of being hit by XRN2 after the later catches up with Pol II (named the torpedo model), or that some allosteric change occurs in Pol II is triggered by contact with XRN2. Transcription termination can occur thousands of kilobases downstream of the PAS signal, and it is known to be another region where Pol II tends to process at a slower rate, causing its pausing [152].

1.2.4 On additional layers of regulation: DNA methylation, nucleosome dynamics, histone modifications, and genome organization

As we have seen, for a protein-coding gene to be transcribed and processed into a matured mRNA, there are many layers of regulation to ensure cells are expressing the correct genes at the correct time and tissue. Yet, there are still additional layers of regulation that have evolved, and that have been already briefly alluded to. Next, I will consider some of them in detail, including the methylation of DNA, the modification of the nucleosomes that compact DNA, and the organization of DNA in the nucleus.

DNA can be methylated by adding a methyl group at the 5th carbon of cytosines, usually occurs only at CpG dinucleotides, and is carried out by DNA methyltransferases (DNMT). This methylation modification has been observed in different contexts that are not only sometimes poorly understood, but outright contradictory. Interestingly, many branches of the DNA-based evolutionary tree have lost this feature, such as the fruit fly Drosophila melanogaster, or the baker's yeast Saccharomyces Cerevisiae, with no evidence that they methylate their DNA [211]. In the context of transcription regulation, promoter regions are enriched by CpG, and can be silenced by their methylation. It is thought that TFs have decreased affinity to methylated DNA. Also, DNMT are known to help recruit chromatin remodelers that then make promoter regions inaccessible by converting them to heterochromatin. Further, in species that reproduce through sex instead of

by clonal expansion, different methylation patterns are sometimes conserved across each parent-derived allele, conferring allele-specific gene expression patterns. DNA methylation is also used to repress repetitive elements in the genome, such as the promoters of transposable elements, to keep them from jumping around with harmful implications for the host. Strangely, DNA methylation is also found in gene bodies, but there it is not associated with silencing, but rather it is rather positively correlated with transcription levels. It is posited DNA methylation at gene bodies helps with transcription elongation [68].

In order for the great majority of TFs and other DNA-binding molecules to attach to DNA, the latter has to be accessible for recognition. In fact, only around 3% of DNA is accessible at any given time. This is no obvious feature, as organisms have evolved to carefully package their DNA into tightly wrapped chromatin, in order to both keep it safe and also to reduce its volume for it to fit inside the nucleus. DNA is coiled around nucleosomes, octamers of distinct histone proteins, and it takes DNA around 147 bps to encircle them. DNA regulatory regions such as strong promoters are known to be almost devoid of nucleosomes, weak promoters display a slight increase of nucleosome occupancy but still low, active enhancers follow with an even greater nucleosome occupancy, inactive enhancers have more nucleosomes; and in the other extreme constitutive chromatin has complete nucleosome occupancy. Although a region such as an active enhancer has little nucleosome occupancy, it has high nucleosome turnover, with their few nucleosomes constantly being repositioned and replaced. Some TFs facilitate the opening of chromatin by passively competing with the introduction of new nucleosomes by binding to their DNA motifs instead of the nucleosomes. Other TFs can actively open chromatin by binding to adjacent regions of DNA and recruiting chromatin remodelers that then displace the nucleosomes nearby [99].

Besides the differential positioning of nucleosomes affecting gene regulation, the nucleosomes themselves are also heavily modified to regulate gene expression. As it was mentioned, nucleosomes are formed by histone proteins, two of each of H2A, H2B, H3 and H4. There is an additional histone referred to as linker H1, which is not part of the nucleosome per se but rather is bound just outside of the core nucleosome helping DNA wound around it. The core nucleosome histones

have unstructured N-terminal tails that can be differently modified to change the behavior of the DNA wrapped around them; these chemical changes include their acetylation, methylation, phosphorylation, ubiquitination, and sumoylation. Some modifications directly change the hydrostatic interactions between DNA and the histones, making their interaction tighter or more relaxed. Other modifications, instead, influence their bound DNA by acting as landing pads for other proteins that then can act on the DNA or on the nucleosome itself. There are dozens to hundreds of modifications that have been observed, and elucidating their function has proven difficult, as it is hard to study them in isolation. Many of them, therefore, are only known to be correlated with specific phenotypes; such as transcriptional activation, transcriptional elongation, transcriptional silencing, DNA repair, silencing of telomeres, heterochromatin regions, mitosis, etc. [109]. If that wasn't convoluted enough, there are many alternative histones known to be also associated to specific genome compartments and functions, such as CENP-A at centromeres, or H2B.W used in testis. Of note, H2A.Z and H3.3, are variants known to be deposited in chromatin that is actively transcribed [123].

On a large scale, eukaryotic genomes are divided into euchromatin (where genes reside and most transcription occurs), and heterochromatin (tightly packaged DNA). With the implementation of techniques that prove the 3D structure of genomes, it was discovered that not all DNA regions are able to interact with all other DNA regions, but that there tends to be clusters of DNA that interact with each other much more frequently. These regions are referred to as topologically associated domains (TADs) and are big, in the order of megabases long. Further still, inside of TADs there are regions that interact with each other much more frequently, and were thus named subTADs, in the order of hundreds to dozen kilobases long [160].

It has been posited that what keeps these DNA regions from interacting outside of their TADs is their physical extrusion by the action of the structural proteins CTCF, cohesin, and condensin [190]. Enhancers generally cannot get in close looping proximity to their target genes unless they both reside within the same subcompartment. In addition, some studies have proposed that some of these domains are not only spatially separated, but that they undergo changes in their solvent and solute concentrations that renders them mutually exclusive from each others, such as changes

in viscosity by modulating the number and types of biomolecules present [95].

Altogether, all of these regulatory mechanisms have evolved through billions of years to facilitate the use of genes at proper times and places by eukaryotic life forms in order to survive in an ever changing environment. They are not perfect, but rather they are sufficient to permit the passing on of the genetic information to the next generation.

1.3 Approaches to study gene transcription regulation

Gene transcription regulation can be studied using a wide variety of techniques. In this section, I will describe some of them in detail. These include genome-wide genomic approaches such as RNA-seq, which measures mRNA steady state levels of all transcribed genes in a genome; precision run-on sequencing (PRO-seq), which directly measures transcription levels by capturing all nascently transcribed RNA molecules; and the Assay for Transposase-Accessible Chromatin (ATAC)-seq, which measures chromatin accessibility. I will also describe approaches that test the transcriptional regulatory activity of DNA sequences, such as with massive parallel reporter assays (MPRAs), and DNA-editing tools such as the variety of implementations of the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 technology.

Because it pertains to both RNA-seq, PRO-seq, and ATAC-seq; I will begin by briefly describing the most common sequencing technologies used in the field to obtain a read out of the nucleotide sequence composition of a given sample.

Sequencing by synthesis is by far the most widely used sequencing technology in the last decade, and lately the company Illumina has dominated in this field. In order to determine the nucleotide sequence of DNA fragments, this technique relies on the fragments having been ligated with specialized DNA-adaptors at both of their 5' and 3' ends. A set of complementary DNA adapters are immobilized on a surface, and the single-stranded DNA fragments to be sequenced are hybridized to the immobilized adapters. The DNA fragments are amplified so that they form a cluster, or spot, of identical DNA sequences. Free-floating nucleotides – modified so that they contain a fluorophore (with each of the four DNA nucleotides having a different fluorophore color)

– are washed through the single-stranded DNA fragments, and they are covalently attached in the first position of the DNA fragment if they are the complement nucleotide. The non-attached floating nucleotides are washed away and a picture of the surface is taken, where the two-dimensional coordinates of the fluorophore signal are tracked. The fluorophores are then chemically quenched, and another cycle of fluorophores is added. The DNA fragments can be sequenced only on one end of the fragment (i.e. single-end reads), or both ends of the fragment can be read (i.e. paired-end reads). At the end of each cycle, the sequencer determines which fluorophore color was added at each position of the surface, and from this information the nucleotide sequence of the fragment that hybridized in that position is obtained. This technique allows for the simultaneous sequencing of up to billions of short DNA fragments, with lengths ranging from 25 to 500 bp [176].

In contrast, in the last few years new technologies allow for the sequencing of long DNA fragments, with a practical limit in the order of dozens to hundreds of kilobases (kb) long. The Pacific Biosciences sequencing technology relies on DNA fragments to be circularized. An immobilized DNA polymerase sits in a nanowell, and it synthesizes the complementary strand of the circular DNA fragment also using fluorescent labeled nucleotides. As the nucleotides are incorporated, their fluorescence is read out. This approach has a high error rate, so each DNA fragment is sequenced multiple times (hence their circularization), and a consensus sequence is obtained [176]. The Oxford Nanopore sequencer is also capable of obtaining long sequencing reads. It relies on distinctive changes in electrical current by each nucleotide that are measured by using molecular motors that push either DNA or RNA fragments through a nanopore [176].

1.3.1 RNA-seq

Short-read RNA-seq technology was developed more than a decade ago. To prepare short-read RNA-seq libraries, total RNA is extracted from cells, and depending on the protocol, an enrichment step can be performed to remove ribosomal transcripts and keep only polyadenylated (i.e. mRNA) transcripts. The RNA is fragmented into smaller fragments, usually 200 nucleotides long, the fragments are reverse transcribed into cDNA, and adaptors ligated to both ends of the

DNA fragments. Usually 25 million reads are obtained per RNA-seq dataset. Paired-end reads are preferred for RNA-seq datasets, as obtaining the sequence from both ends of the cDNA fragments can be used to determine spliced isoforms. In a standard RNA-cell analysis pipeline to determine differentially expressed genes (DEGs), the sequencing reads are mapped to a reference genome (e.g. using HISAT2, STAR, TopHat, etc.), the number of reads per gene are counted (e.g. using featureCounts, HTSeq, etc.), the counts per gene are normalized between samples, and a statistical assessment is performed to test if the number of reads for a given gene (and for all genes) is significantly different between two conditions (e.g. DESeq2, edgeR, limma+voom, etc.) [176].

The above procedure supposes that a bulk cell population, or tissue, was used as input to generate a short-read RNA-seq datasets. Bulk cell assays are easy to use, but have the significant shortcomings of interrogating the average cell population response. If half of the cells are highly expressing a certain gene, and the other half are not expressing that same gene whatsoever, the results will show that the gene is expressed at a moderate level, which is a wrong interpretation of the biomaterial interrogated. There is a variation of RNA-seq called single-cell RNA-seq, and as its name implies, it measures the levels of mRNA coming from individual cells. To achieve this, a population of cells is diluted such that a single cell is placed into tiny wells where their RNA is processed such that the RNA coming from each well is tagged with unique indexes. After all the fragments are sequenced, all reads carrying the same index can be traced back to the same individual cell [176].

Critically, RNA-seq does not measure transcription levels from cells. Instead, it measures steady-state levels of RNA molecules, which is a balance between the production of new RNA transcripts, their accumulation, and their degradation [165].

1.3.2 PRO-seq

In order to directly measure transcription levels from cells, new library preparations were developed that capture nascently transcribed RNA molecules. As described in section 1.2, RNA transcripts are extensively processed in order to become mature mRNA ready to be translated,

in the case of protein-coding genes; or ready to perform their function as non-coding RNAs, for non-coding genes. Also described in section 1.2, both promoters and enhancers are transcribed in a bidirectional manner from a single origin when actively being used for regulation. PRO-seq (and related nascent variants)) is one of a few methods developed that capture unprocessed, nascently transcribed RNA molecules coming off of actively transcribing RNA polymerases, from both gene bodies, as well as from promoter and enhancer regions.

To obtain PRO-seq datasets, a cell population is processed so that their nuclei are extracted and frozen to avoid disrupting the native configuration of actively transcribing RNA polymerases. A mixture of normal and biotinylated-labeled ribonucleotides is added to the thawed nuclei to allow the RNA polymerase to continue transcribing, incorporating both normal and biotinylated-bases to their nascent transcripts. Once a biotinylated-ribonucleotide is incorporated, it blocks the active site of the RNA polymerase and hinders its activity, virtually locking them in place. Total RNA is extracted from the cells, including the biotinylated-nascent transcripts. RNA is fragmented, and streptavidin-coated beads are used to keep only the biotin-labeled transcripts while removing everything else. These nascent transcripts are then ligated with adaptors, reverse complemented into cDNA, amplified, and sequenced [199]. Other variations of nascent transcription datasets rely on a similar logic, with the difference being the way the nascent transcripts are pulled down. For the mammalian native elongating transcript sequencing (mNET-seq), Pol II is pulled down using antibodies and their associated transcripts recovered and processed. For the global run-on sequencing, 5-bromouridine 5'triphosphate nucleotides are used instead of biotinylated-labeled nucleotides, and an antibody that recognizes 5-bromo-2-deoxyuridine is used to pull down the nascent transcripts.

PRO-seq datasets are processed similar to RNA-seq datasets, except in that their aligners do not have to account for spliced reads, and counting reads over gene bodies does not need to consider the boundaries between exons and introns. Loci of bidirectional transcription can be detected by specialized software such as Tfit or dREG [208].

1.3.3 ATAC-seq

Gene transcription regulatory elements are depleted of nucleosomes when they are active [99]. Researchers have relied on this feature to quickly assay what parts of the genome are accessible and inaccessible by exploiting the activity of the hyperactive Tn5 transposase variant. This Tn5 has been engineered to contain adaptor sequences, such that when it introduces itself into open chromatin regions, it also introduces these adaptors to the same locations. After the transposition reaction takes place, the DNA can be amplified using complementary primers to the transposed adaptors, and the library is ready to be sequenced [42].

ATAC-seq datasets are also processed similarly to RNA-seq and PRO-seq datasets. Regions of open chromatin can be detected by using peak callers such as MACS2 [212] or HMMRATAC [183].

1.3.4 Massive parallel reporter assays

Genomic assays such as ChIP-seq, ATAC-seq, and even PRO-seq, are used to annotate loci as putative regulatory elements genome-wide by relying on the correlation of certain chromatin features (e.g. H3K27ac, H3K4me3, TF binding, open chromatin, bidirectional transcription) with DNA sequences known to be regulatory elements. However, to test if those loci really have regulatory potential, correlation is not enough, and direct testing of those DNA sequences are needed.

MPRAs are a powerful tool that simultaneously tests thousands of DNA sequences, with a length of up to a few hundred bp, to see which of them are capable of regulating the transcription of a given reporter gene. They work by taking the DNA sequences to be queried out of their endogenous genomic context, and introducing them into DNA plasmids with a reporter gene. These genes are not able to be transcribed on their own, as they may lack a functional promoter, or they may lack an enhancer to help a basal promoter become active. If testing for promoter potential, the DNA sequence to be tested is introduced immediately upstream of the gene TSS; if testing for enhancer potential, the DNA sequence is introduced somewhere else in the plasmid, often downstream of the

reporter gene. MPRAs can either look for the output of the reporter gene, such as the expression of a fluorescent protein, or they can see the levels of transcription of the reporter gene mRNA. In the case of a sequencing output, the reporter gene can have a specific barcode such that the investigators can match the reporter gene to a specific DNA sequence; or they can design their plasmid such that the downstream DNA sequence is transcribed alongside the reporter gene and they can simply look for the levels of transcription of the inserted DNA sequences themselves [79].

The power of MRPA is not only their capability to test many DNA sequences at once, but that they can test those sequences in a controlled experimental environment where many variables are accounted for. However, the latter is also one of its limitations, as some enhancer sequences may need their endogenous genomic context to properly function (e.g. they lack necessary adjacent TF motifs, or nucleosomes with specific modifications), and are therefore – by MPRA – falsely thought as non-regulatory.

1.3.5 DNA-editing of putative regulatory elements

CRISPR-Cas9 (and its variants) is a powerful DNA-editing technology that allows the editing of DNA in its native genome context [83]. Briefly, the CRISPR-Cas9 activity relies on the Cas9 enzyme that is capable of using a short guide RNA molecule as a template to recognize and cleave DNA sequences. It is now understood that bacteria have evolved a wide variety of CRISPR-like systems as analogs of the eukaryotic multicellular adaptive immune system [102]. Bacteria store short sequences of previous bacteriophage infections into specific regions of their genomes, referred to as clustered regularly interspaced short palindromic repeats (CRISPR). Upon subsequent bacteriophage infections, they transcribe and load these repeats onto their Cas enzymes to recognize and destroy bacteriophages.

Not only has CRISPR-Cas9 been used to directly edit the sequence of putative regulatory elements to test their activity (e.g. by disrupting TF motifs), but the Cas9 enzyme itself has been engineered to deactivate its endonuclease activity. This deactivated Cas9 (dCas9) is only capable of loading guide RNAs and recognizing specific DNA sequences. dCas9 has been fused

to protein domains derived from transcriptional activators or repressors, such as the p300 histone acetyltransferase core domain or the Krüppel-associated box (KRAB) domain, respectively. The resulting dCas9-p300 fusion can be targeted to putative regulatory elements and test their activity by depositing H3K27ac on the adjacent histones. Conversely, repression of those putative elements can be achieved by the dCas9-KRAB fusion, as it then can promote repression by depositing H3K9me3, which in turns transforms the loci into heterochromatin [97].

1.4 The evolution of gene transcription regulation

As we have seen, many layers of regulation have evolved to tightly control how, when, and where genes should be transcribed in the lifespan of a given organism. The rewiring of gene regulatory networks is a double-edge sword, as it is commonly implicated in disease susceptibility [110], but it is also necessary for giving rise to novel phenotypic traits, which are vital for organisms to improve their fitness in ever changing environments [31].

In fact, it has been proposed that the amount of variation observed in extant species cannot be solely accounted for by changes in the number of genes or in changes in the structure of the proteins those genes code for [96]. Instead, morphological evolution seems to be primarily driven by changes in the use of promoters and enhancers, by adding and removing genes from their transcription regulatory networks [156, 117].

Many examples have been described of species acquiring new phenotypes by changing the way they express a few genes. From fish colonizing new ecological niches, such as moving from marine saltwater to freshwater lakes [84]. Butterflies modulating the color patterns of their wings by varying where to express pigment proteins [112]. Peppered moths quickly going from snow-adapted white color to ash-colored pigmentation to survive predation during the industrial revolution in England [74].

There is evidence of rapid turnover of regulatory elements relative to genes [156, 194, 117]. For example, humans surviving deathly diseases by modulating what genes to use in their immune responses [98, 130, 71, 101]. Importantly, species do not choose to acquire these new phenotypes,

but rather they randomly acquire mutations and when these confer beneficial phenotypes, the organism manages to survive and pass on their genes, sometimes fixing the mutations in their entire populations.

With the advent of high-throughput sequencing technologies, researchers have been able to determine changes in whole transcriptomes across species [159, 22], or in the assessment of differences of where TF binding [164] and chromatin marks [194] are located genome-wide between organisms.

Not only are non-coding DNA sequences poorly conserved across species; such as introns, enhancers, and intergenic regions [209], it was found that there is a huge amount of divergence of where TFs bind across species [57, 164]. However, it is important to note that not all TF binding events are functional, with a significant fraction of TF binding not used in gene regulation [175, 9]. Active regulatory elements tend to be associated with the histone mark H3K27ac, so looking for differences across species of this chromatin feature is considered a more reliable indicator of changes in gene regulation than changes in TF binding alone. Still, widespread changes in this and other marks of active regulatory elements have been observed across many species, even at close evolutionary distances – such as between primates [45, 151, 157, 192]. Adding to this, changes in eRNA expression have been observed between humans and chimpanzees [50].

However, it has been proposed that the observed high turnover rate of enhancers may not impact in a drastic way the transcription of their target genes. This is because genes have multiple enhancers, probably having evolved so as to not experience radical expression changes upon losing or acquiring enhancers [64, 29]. This enhancer redundancy is observed at orthologous loci across species, where the loss of an enhancer tends to be compensated by the acquisition of another enhancer nearby in the other species [139, 104].Furthermore, changes in gene transcription across species has been shown to be buffered out as observed by unaltered resulting protein levels [14], or even by opposite changes in the level of those proteins [94].

A gene's transcription can be altered by either changes in the underlying DNA sequence of its promoter or enhancers (referred to as changes in cis), or in changes in the amino acid sequence of its

TF regulators (referred to as changes in trans) [125, 200, 172]. Although both types of changes have been observed to impact gene transcription across species, changes in cis are much more common. Changes in trans, such as changing the protein structure of the DNA-binding domain of a TF, will have an effect in all the genes that such TF regulates; whereas a change in cis, such as a nucleotide substitution at a key TF motif instance, will only affect the single gene that uses that motif for its transcription regulation. Cis and trans labels can be a bit blurry. For example: there can be changes in cis in the promoter of a TF that decrease its transcription, that will then have effects in trans for the genes that the TF can now differentially regulate, based simply on having fewer TF molecules around in the nucleus.

Mutations can arise as DNA substitutions, deletions, insertions, inversions, and depending on where they occur they can affect gene expression in different ways. These changes can directly change the nucleotide composition of TF binding sites (TFBSs), they can alter the chromatin structure by modifying how nucleosomes are positioned or how chromatin can become accessible, they can even disrupt the boundaries of TADs and in doing so precluding the interaction between promoters and their enhancers [73].

Moreover, selfish replicating elements such as transposons [21, 61] and endogenous retroviruses [37] are known to be a common source for the rewiring of existing host gene regulatory networks. When they replicate by integrating into new regions of their host genome, these elements bring with them their promoter sequences. Even though hosts have evolved mechanisms to silence these selfish elements [203], their promoter regions can sometimes become deregulated and affect the transcription of nearby genes.

Taken together, there are many mechanisms by which gene transcription regulation is altered to bring about new phenotypes that species can employ to increase their fitness. Though transcription regulation is a tightly regulated process, it is not impervious to change; and this is a good thing, as otherwise species would not be able to contend with their ever changing environments that they inhabit.

1.5 The p53-triggered transcriptional response

As alluded previously, eukaryotic organisms have evolved mechanisms to ensure the proper storage of their genetic information, by wrapping DNA around nucleosomes [120] inside of the membrane-bound container that is the nucleus. However, DNA can still be damaged throughout the life of individual cells, and in the case of multicellular life forms for the lifespan of the whole organism, which can be up to thousands of years [18]. In response, these organisms have also evolved several mechanisms to quickly fix DNA damage, which can arise in the forms of double-stranded or single-stranded breaks, or abnormal chemical modifications of nucleotides [69].

DNA is damaged by factors that can come from outside or inside of cells. From outside of cells, radiation is one of the main environmental sources of DNA damage. It comes from virtually everything, from rocks and the soil, to cosmic radiation, including the sun. Ionizing radiation has the potential to directly break the covalent bonds in the DNA, or radiolyse the water surrounding DNA, which then has the potential to chemically react and break DNA. Ultraviolet radiation can damage DNA by causing adjacent pyrimidines to covalently attached to each other, causing DNA deformations. Environmental and diet-borne chemicals, such as alkylating agents or aromatic amines, can react with the nitrogens in nucleotides, modifying the normal DNA structure. Extreme heat, extreme cold, and hypoxia have also been shown to damage DNA. From inside of cells, DNA replication errors can damage DNA, as well as spontaneous deamination of nucleotides. In addition, during normal and abnormal metabolism, reactive oxygen species are produced that can also react with and damage DNA [33].

The gene TP53 is found in all multicellular organisms, and it is one of the key genes whose role is to detect DNA damage, among other signals, and mount an appropriate response to safeguard the integrity of the genome [111]. The TF p53, the product of the gene TP53, is constitutively expressed in most cells, but it is kept inactive in the cytoplasm. There, it is continually degraded by the ubiquitin-ligase MDM2. The interaction between MDM2 and p53 is conserved across multicellular animals [214, 19, 88]. After some of the above stressors are sensed by the cell, through poorly

understood mechanisms, the MDM2 interaction with p53 is diminished. One of these mechanisms is the phosphorylation of p53 near or at the binding interface of MDM2. Another described mechanism is the upregulation of another ubiquitin ligase, NEDD4-1, that targets MDM2 for degradation [206]. Once the p53 monomers stop being degraded, they accumulate in the cytoplasm, and are transported to the nucleus where they form homotetramers, bind to p53-responsive elements in the DNA, and upregulate the transcription of p53 target genes. The TF p53 is known to only operate as a transcriptional activator. It directly binds to and upregulates the transcription of dozens of genes within one hour of its activation. Later in time, hundreds of downstream secondary target genes are up and downregulated [3, 5, 181].

One of the main responses controlled by p53 is the arrest of cell cycle progression, so that enough time is given to fix any DNA damage before cell cycle progression can resume. If the DNA damage is severe enough, p53 triggers programmed cell death (i.e. apoptosis), to ensure that the DNA mutations are not propagated through the organism in the form of tumors [8].

In addition, after decades of its study, many other functions have been linked to the cellular role of p53 upon its activation. Besides DNA damage, p53 is known to be activated upon telomere erosion, mitophagy, changes in the cell's redox potential, anoxia, cellular senescence, ribosomal biogenesis, infection by pathogens, inflammation, glucocorticoid-triggered stress, expression of oncogenes, and other stressors. Besides cell cycle arrest and apoptosis, other known cellular responses in which p53 is implicated are: maintenance of pluripotency, cell fate determination, autophagy, changes in epigenetic marks, control of reactive oxygen species, changes in metabolism, the recruitment of immune cells, and the epithelial-mesenchymal transition [89, 70, 111].

Because of its key role in protecting the integrity of the host genome, the transactivating potential of p53 has remained significantly unchanged across evolutionary time, as evidenced by the conservation of its DNA-binding domain across phyla. However, many genes have been observed to have acquired p53-responsive elements in their regulatory sequences [76, 80, 62], even in closely related species [143, 196].

These changes in the p53-responsive regulatory network suggest that p53 has played an

important role in driving the evolution of new traits across species.

1.6 The type I interferon-triggered transcriptional response

The immune system evolved to protect the host against the many pathogens hosts are exposed to throughout their lives [20, 135]. This system can be divided into two subsystems.

The innate immune system is the first line of defense against invading microbes. It includes anatomical barriers such as the skin, or the use of substances such as gastric acid in the stomach; to the complement system, a set of soluble proteins that opsonize pathogens so that they can be phagocytosed by host cells, or that directly disrupts their membrane; to a set of immune cell types, such as natural killer cells, macrophages, neutrophils, and others, with roles ranging from promoting inflammation, to killing infected host cells [140]. Inside cells, there are proteins that are deployed when pathogens are sensed. These defenses are part of the cell-autonomous innate immune system [155].

If the innate immune system does not eradicate the pathogenic threats, the adaptive immune response steps in to finish the job. The adaptive immune system is composed of white blood cells, called lymphocytes, that are produced in two immune organs, the bone marrow and the thymus. These lymphocytes are further divided into many cell subtypes. B-cells are created and matured in bone marrow, whereas T-cells are also made in the bone marrow, but they mature in the thymus. Both of these cells use antibodies to recognize and destroy pathogens. These antibodies have strong affinity for short peptide sequences, called epitopes [63].

Interferons (IFN) are soluble proteins that are used to signal cells throughout the body that pathogens have been detected so that the host can mount the necessary cellular and molecular defenses. IFNs are known to be involved in the control of aspects of both the innate and the adaptive immune systems. In humans, there are multiple IFN genes that are classified into three types, according to the membrane receptors that they have affinity for: type I, type II, and type III IFNs. Type I and type III IFN proteins are known to be used mostly to fight viral infections, whereas type II are thought to be involved in the regulation of inflammation [126].

The type I IFN gene family binds to the receptors IFNAR1 and IFNAR2, and has undergone extensive gene duplication in many parts of the metazoan branch of Earthian life. In humans, for instance, the family is composed of 17 genes: 13 paralogs of IFN-α, and single copies of IFN-ε, IFN-$\varkappa$, IFN-ω, and IFN-β; the latter known to induce a very robust immune response. IFN-γ is the only type II IFN, and it binds to the receptors IFNGR1 and IFNGR2. And for type III there are four IFN-λs which use IL-10R2 and IFNLR1 as their receptors [75].

Type I IFN are deployed when host cells detect pathogen-associated molecular patterns (PAMPs), through the use of pathogen recognition receptors (PRRs). There are PRRs in the plasma membrane, such as TLR4, as well as in endosomes, such as TLR4 and TLR9; and free-floating cytoplasmic PRRs such as RIG-I, that recognize nucleic acid [113]. The use of these PRRs seem to be relatively conserved, though there is some variation of PRR expression, which is posited to be driven by the distinct pathogens each species needs to recognize [12].

Once these PRRs have recognized their PAMPs, the signal is relayed so that a set of immune-related TFs, such as the IFN-regulatory factors (IRF) IRF3 and IRF7, and the Nuclear factor-κB (NF-κB), are recruited to the nucleus to upregulate the transcription of IFN-β and through some feedback mechanisms, the induction of some IFN-α follows soon. IFN-β and IFN-α are synthesized and released from the cell to go and alert neighboring cells. Upon binding with its receptors, IFNAR1 and IFNAR2 undergo conformational changes that in turn induce the kinases JAK1 and TYK2 to activate each other by mutual phosphorylation. These kinases then phosphorylate STAT1 and STAT2. IRF9 forms a trimeric complex, called ISGF3, with the activated STAT1 and STAT2. Other TF complexes, such as a STAT1 homodimer, are also formed. ISGF3 recognizes specific DNA sequences termed interferon stimulated response element (ISRE) located at promoters and enhancers, and regulates the transcription of hundreds of interferon stimulated genes (ISGs) that code for proteins that combat the incoming pathones [126]. Not all ISGs are transcribed simultaneously, but rather, they seem to be expressed in temporal waves, with some ISGs transcribed a few minutes after IFN-β and IFN-α are recognized at the cell membrane, while other ISGs do not appear until a few hours later [131].

Not only have IFN genes diversified through gene duplication across species [103], but the number and types of ISGs have also varied across species [168]. These evolutionary dynamics highlight the strong selective pressure imposed on the immune system to diversify and contend with rapidly evolving pathogens [53]. Indeed, signatures of strong positive selection have been detected in many ISGs [86].

It is not yet clear how new genes are converted into ISGs, but some studies have pinpointed their induction by IFN on the introduction of ISREs close to their promoters, or acting as distal enhancers, as a byproduct of the replication transposable elements and endogenous retroviruses [28].

There are now multiple examples of rewiring of the IFN regulatory network, including between human populations, posited to have been selected for in order for ancient humans to bypass infections with past pathogens [71, 115, 101].

1.7 Cellular mechanosensing and its microenvironment

The spatial organization of chromatin inside cells' nucleus is a highly dynamic process. Not only are chromosomes regularly packaged into pairs of sister chromatids during each replication cycle on dividing cells [182], but the boundaries between the accessible euchromatin and the tightly packaged inaccessible heterochromatin is constantly changing, with actively transcribed genes being moved away from the nuclear lamina [184]. In addition to responding to biological and chemical environmental cues with changes in gene expression, cells are now better understood to also sense mechanical stimuli from their immediate microenvironment, and transduce them with the help of their cytoskeleton towards the nucleus membrane, where they affect genome organization and transcription of genes [187].

In multicellular organisms, cells interact with other cells through direct physical contact, such as through cell-cell junctions; or with a gelatinous extracellular matrix which is composed of secreted proteins such as collagen, glycoproteins, and polysaccharides [133]. Cells use their cytoskeleton, which is composed of structural proteins that form a diverse set of filaments, including

microtubules and actin filaments; to move around (in the case of motile cells), or to withstand the external pressures and avoid being squished. These external forces, such as shearing, compression, and tension; can be directly sensed by cells through the use of membrane proteins (e.g. integrins and cadherins). These mechanical receptors transduce mechanical signals to protein complexes inside the cell, namely actomyosin, that in turn modulate actin filaments through contraction forces. The actin filaments are in direct contact with Nesprin proteins, which in turn contact the linker of nucleoskeleton and cytoskeleton (LINC) protein complex. The LINC is a network of structural proteins that traverse the nuclear lamina or are associated with the inner nuclear membrane.

Upon their stimulation through the mechanical signal transduction coming from outside the cell, proteins of LINC complex can change the level of compaction of the chromatin inside the nucleus through unknown mechanisms. But the effects of this signal transduction are readily observable with microscopy techniques showing nuclear condensation upon changes in the stiffness of the cell's outside environment [187]. In addition, specific TFs have been shown to be imported to the nucleus through nuclear pores after the transduction of mechanical forces outside of the cell, to regulate the expression of genes with functions related to cellular migration and cell differentiation [153, 136, 173].

To sum up, the regulation of gene transcription has evolved, and continues to evolve, to be responsive to a complex interplay of biological signals (e.g. cytokines or hormones), chemical signals (e.g. reactive oxygen species), physical stressors (e.g. radiation), and even mechanical signals (e.g. cell-to-cell contacts, and physical tension). All of these sources impinge upon cells for them to react to their environments and respond accordingly, lest they cease to be alive.

1.8 Preface of following chapters

In the next chapters, I will delve into specific contexts in which gene transcription regulation has been rewired through evolutionary time.

In chapter 2, I will explore the variation across primates in the transcriptional response controlled by the TF p53. The guardian of the genome, as p53 is oftentimes referred to, is a TF

involved in many cellular processes, but its main function is to orchestrate the cellular response upon damage to the genome. Safekeeping the genetic information is crucial for organisms to pass on their genes to the next generation, and as important as it is to keep this role of p53, I describe how evolution has still managed to diversify its regulated transcriptional response.

In chapter 3, I provide insights into how the type I interferon transcriptional response has evolved in metazoans, going as far away in time as the split between mammals and egg-laying birds. As I have mentioned, gene transcription regulation is a highly malleable process that evolution has constantly shaped, a feature that is especially important for the immune system, as it is under a strong selective pressure to diversify. I briefly describe two datasets: one comparing across distinct animal species, and another comparing across different human ethnicities.

In chapter 4, I put forward preliminary evidence that shows how disrupting the immediate microenvironment where eukaryotic cells grow can entail significant consequences for the chromatin dynamics of genomes. I tested how an abrupt removal of human and pig cells from their substrate seems to be mechanically sensed, and in turn alter chromatin accessibility genome-wide.

In chapter 5, I finish by sharing my learned lessons, as well as providing future directions that may help elucidate the molecular underpinnings of my observations.

Chapter 2

The evolution of the p53 transcriptional response in anthropoids

2.1 Introduction

Terrestrial life is characterized by the encoding of genetic information in long chains of nucleic acid in the form of DNA or RNA [2]. One of the hallmarks of these life forms is that they persist through long stretches of time, in the order of billions of years [17], by carefully safekeeping this genetic information and passing it on to their descendants. To withstand an ever hostile Earth environment threatening the integrity of their genomes, life forms have evolved several molecular mechanisms. These include the upgrade from the labile RNA to the sturdy double helix in DNA [39], to the tight packaging of DNA into chromatin by protein complexes in all domains of terrestrial life: such as nucleosomes in eukaryotes [120] and in archaea [144], to chromatin-like structures in bacteria [169].

Notwithstanding the many layers of safeguarding shielding the genome, DNA damage can happen in the forms of breakage of either one or the two DNA strands. There are many sources for these damages, whether originating endogenously or exogenously. From within cells, examples are the production of chemically reactive byproducts of normal or aberrant metabolism. From outside cells, some examples are the exposure to different types of radiation (e.g. ionizing or ultraviolet), to toxic chemical compounds, or to extreme environmental stressors (e.g. hypoxia, heat or cold shocks) [33].

Around 800 million years ago, in the common ancestor to multicellular animals, the TP53 gene evolved to orchestrate the cellular response to many of these DNA damaging threats [111].

This single gene, which codes for the p53 protein, acts as a crucial central node where many sensing signaling pathways converge for the cell to decide how to mount an appropriate response to safekeep the genome of the host organism. Since its discovery by humans 40 years ago, TP53 has been so extensively studied that it is now the most researched gene in human history, which highlights its key role in many corners of molecular biology [56].

p53 molecules exist in the cells' cytoplasm in an inactive state ready to be deployed at a moment's notice. They are sequestered by an ubiquitin ligase called MDM2 that continuously degrades them if p53 activity is not needed. Through many signaling pathways that are triggered by DNA damage and other stressors, the MDM2 proteins are modified in ways that reduce their binding affinity to p53, which permits p53 monomers to migrate to the cell nucleus, form a tetrameric transcription factor complex, and bind to regulatory elements to activate the transcription of a multitude of genes [111]. The canonical nature of the cellular response upon DNA damage ranges from efforts to repair the DNA if the damage is mild, to the elimination of the cell through programmed cellular death if the damage is extensive in order to reduce the risk of the formation of cancer, therefore protecting the multicellular organism [8].

In addition to its role in safeguarding the integrity of the host genome through DNA repair or apoptosis, p53 is now known to regulate many other cellular processes. p53 has been shown to be involved in the sensing of a wide range of stressors such as telomere erosion, cellular senescence, epigenetic changes, changes in redox potential, sudden alterations in the synthesis levels of ribosome complexes; and in the direction of a plethora of cellular responses to these stressors, such as autophagy, the alteration of metabolic pathways, of modifying cellular plasticity and differentiation, in changes in cell cycle progression, and even in the recruitment of immune cells [89, 70, 111].

The role of p53 in evolution cannot be overstated. Ranging from its role in protecting the host from the insidious microevolution process that is cancer progression, and its role in development [35]; to its impact in macroevolution processes by acting as a direct filtering agent that reduces the observed mutation rate in multicellular organisms by correcting DNA mutations that could be otherwise passed on through the germ line [180].

Though the interaction between p53 and MDM2 is quite conserved across metazoans [214, 19, 88], the activity of p53 as a transcription factor (TF) has diversified through evolutionary time [76, 80, 62], with the lost and acquisition of regulatory elements adding the expression of new genes into the p53 responsive network, even at closely related species such as within primates [143, 196].

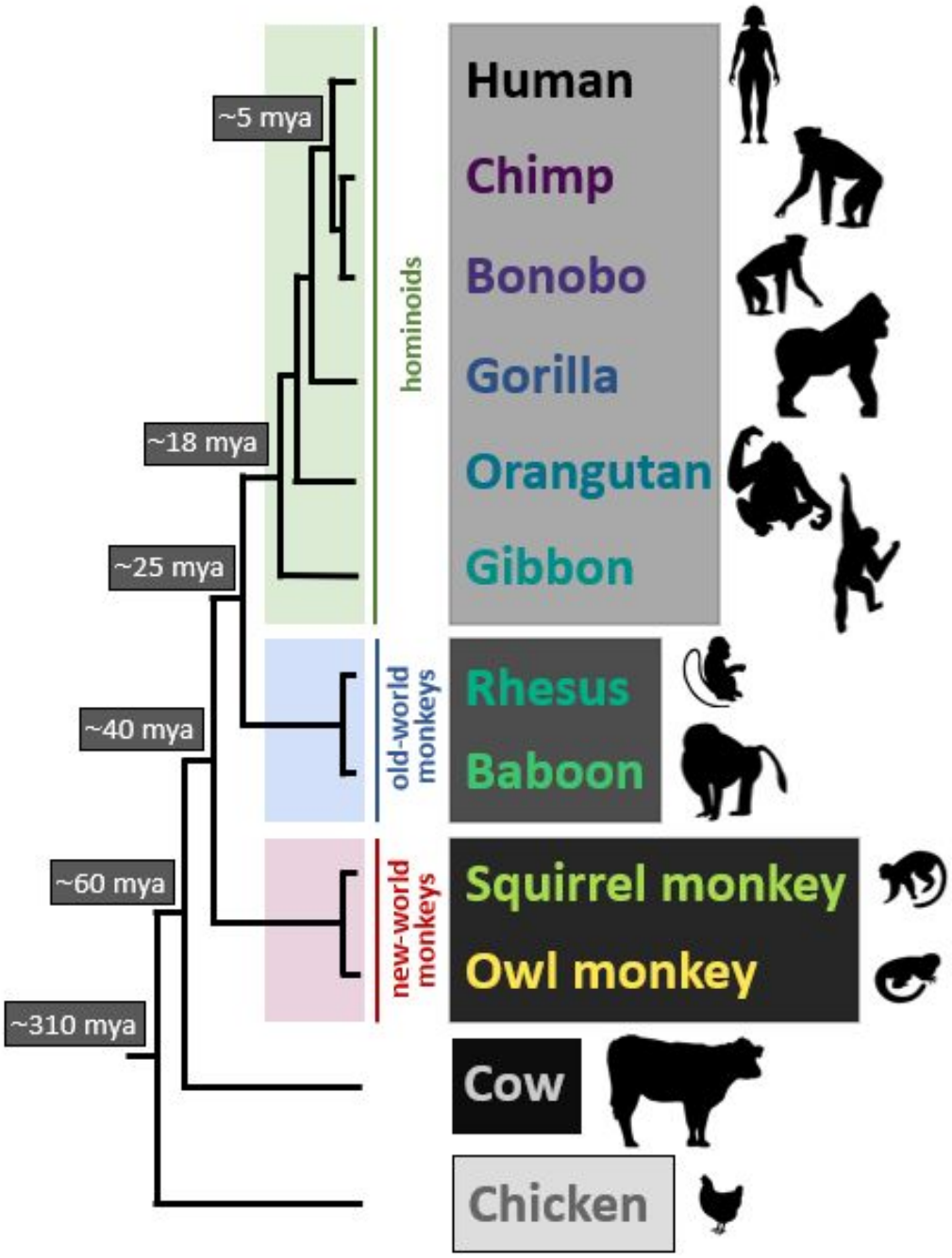
2.2 Experimental system

I decided to examine how the p53 transcriptional regulatory network has evolved in metazoans. To this end, I obtained lymphoblastoid cell lines (LCLs) derived from one individual each from 12 different animal species. From these, I sampled ten anthropoids (also referred to as primates in the text); from which I got six hominoids, Homo sapiens (human), Pan troglodytes (chimp), Pan paniscus (bonobo), Gorilla gorilla (gorilla), Pongo pygmaeus (orangutan), and Nomascus leucogenys (gibbon), whose last common ancestor lived ∼17 million years ago [30]; two old world monkeys, Macaca mulatta (rhesus macaque), and Papio anubis (baboon), whose last common ancestor with hominoids lived ∼25 million years ago [178]; and two new world monkeys, Saimiri boliviensis (squirrel monkey), and Aotus nancymae (owl monkey), whose last common ancestor with hominoids and old world monkeys lived ∼40 million years ago [145]. I also sampled one ungulate: Bos taurus (cow), whose common ancestor with anthropoids lived ∼60 million years ago [27]. And one bird: Gallus gallus (chicken), whose common ancestor with anthropoids and ungulates lived ∼310 million years ago [93] (Figure 2.1).

To study the transcriptional response upon p53 activation I used the drug Nutlin-3a (Nutlin), an imidazoline analog that disrupts the cytoplasmic interaction between MDM2 and p53, and thus activates p53 to do its function in the nucleus as a transcription activator [72]. The activity of Nutlin has been observed in many human and mouse cell lines [132], and I therefore posited it will serve as a p53 activator for my 12 species LCLs.

p53 works as a transcriptional activator by binding to its preferred DNA sequence motifs and upregulates the transcription of dozens of genes as a primary response within one hour of its activation, followed by up and downregulation of hundreds of downstream secondary target

Figure 2.1: Cladogram showing the 12 metazoan species used in the study. The last common ancestor [30, 178, 145, 27, 93] is shown in boxes next to their corresponding inner branch node (mya, million years ago).The color scheme accompanying each species common name font is the same that follows in the subsequent figures in this chapter.



genes [3, 5, 181]. Gene transcription regulation is shaped by evolutionary forces in ways that could potentially change either the cis-acting transcriptional regulatory elements (e.g. promoters and enhancers) or by changing the trans-acting soluble regulators themselves (e.g. TFs) [23]. In the case of p53, though its protein structure has undergone amino acid substitutions throughout the primate phylogeny, the key amino acid sequences in its DNA recognition domain have remained conserved (Figure 2.2). This suggests that changes in the p53-responsive transcription network are due to changes in cis-transcriptional regulatory elements, and not in differences in the way orthologous p53 tetramers recognize their preferred sequence motifs.

I set out to capture both the primary and downstream gene transcriptional responses upon Nutlin-mediated p53 activation, as well as the cis-acting transcriptional regulatory elements that control them. To achieve this, I generated PRO-seq datasets for all 12 species LCLs at 1 hour after the Nutlin stimulation to capture the nascently transcribed primary target genes and their transcribed responsive regulatory elements. To complement the study of these regulatory elements, I made ATAC-seq datasets 1 hour after p53 activation with only the human and bonobo LCLs. Finally, I also prepared RNA-seq datasets for all 12 species LCLs at 6 hours after the addition of Nutlin to capture the immediate downstream secondary transcription of p53-responsive genes.

2.3 Results

2.3.1 Quality check on the p53 activation

First I checked some quality metrics on the obtained datasets. The PRO-seq, ATAC-seq, and RNA-seq were obtained with an appropriate sequencing depth (Figure 2.3) and all passed sufficiently good sequencing quality checks. The first PRO-seq replicates of the human LCL for both DMSO and Nutlin treatments were sequenced at a much higher depth to test the effect of varying sequencing depth in the subsequent analyses. I used Tfit to detect bidirectional transcription loci on the PRO-seq datasets, and I observed that all datasets had a relatively similar number of bidirectionals, except for the chicken sample (Figure 2.4).

Figure 2.2: Top left, DNA-binding domain of the crystal structure of a p53 homotetramer with DNA showing the critical arginines needed for motif recognition, taken from [Baugh2018]. Top right, linear diagram of the p53 amino acid sequence showing the number of mutations found in cancer focused on the DNA-binding domain, taken from [Kato2003]. Bottom, multiple sequence alignment of p53 across the 10 primates used in the study showing amino acids 237 to 281 relative to the human sequence (GenBank: NP_000537.3), generated with the Molecular Evolutionary Genetic Analysis (MEGA) software.

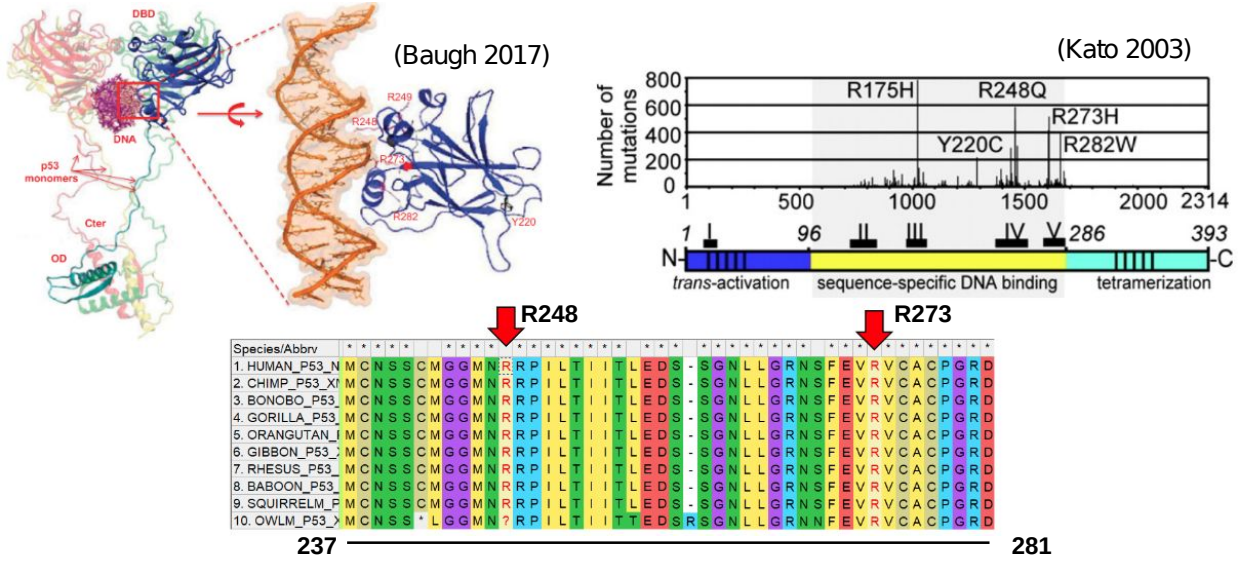


Figure 2.3: Barplot showing the number of short read sequencing reads obtained from the PRO-seq (top) and RNA-seq (bottom) datasets obtained from the 12 species LCLs treated with Nutlin in this study. In light gray are the samples treated with the carrier DMSO, and in dark gray are the samples treated with Nutlin-3a.

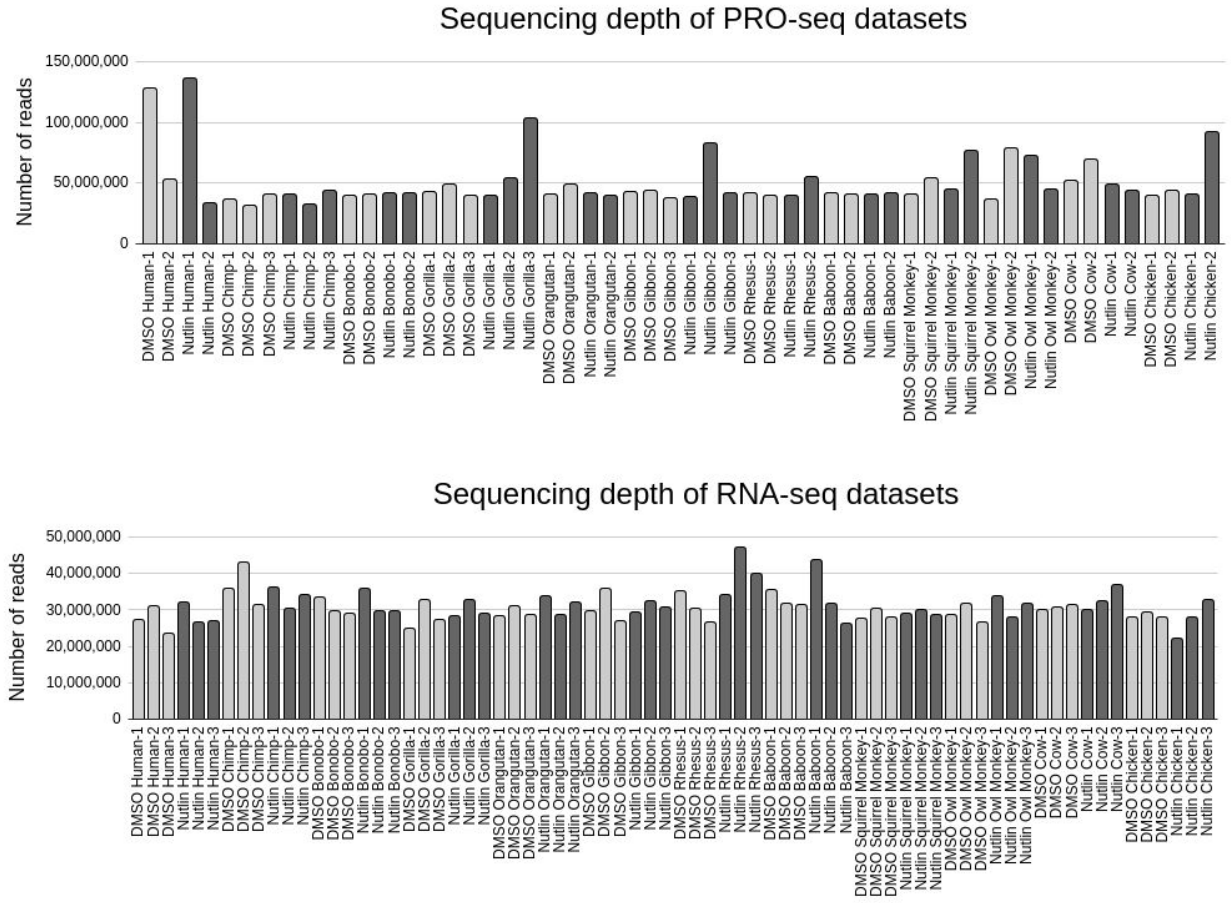
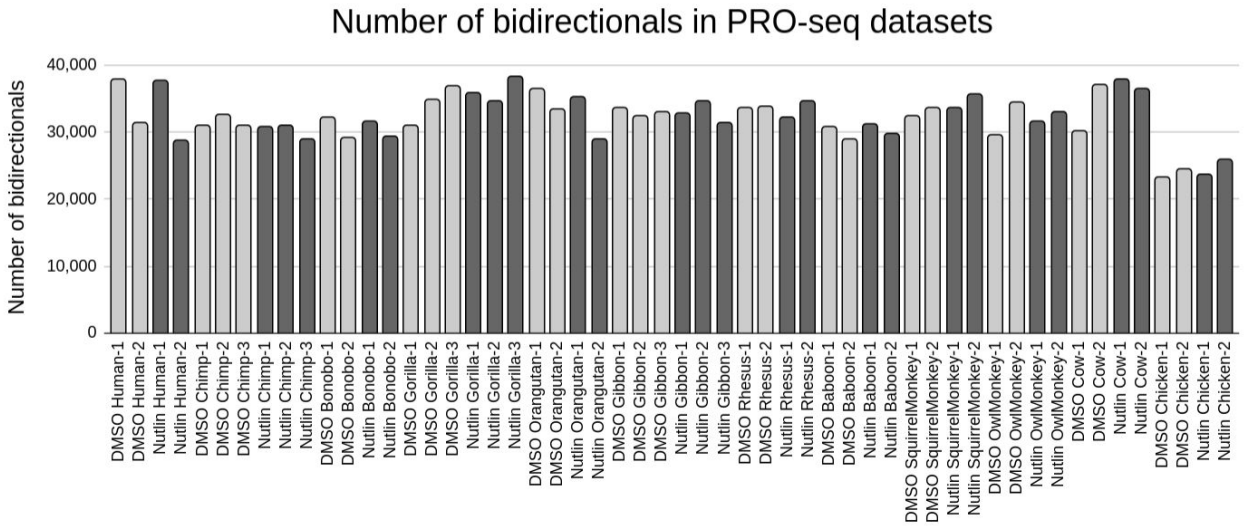


Figure 2.4: Barplots showing the number of bidirectional transcription loci detected by Tfit from the PRO-seq datasets of the 12 species LCLs treated with Nutlin in this study. In light gray are the samples treated with the carrier DMSO, and in dark gray are the samples treated with Nutlin-3a. Most samples have two replicates, except for chimp, gorilla, and gibbon, as these datasets had one replicate with low quality which necessitated a third replicate.



The PRO-seq datasets display nascent transcription signal, (i.e. before RNA is processed, without being capped, spliced, nor polyadenylated). Because I would be comparing PRO-seq and RNA-seq datasets against each other, I wanted to test if there was a significant difference between assessing the fold-change in transcription of genes upon Nutlin induction when considering two different types of gene annotations. The first one was the public gene annotations, where typically are used such that only exons regions are considered and ignores introns. The second was a modified gene annotation where the whole gene body is considered, including the intronic regions. I tested the similarity in fold-change between these two methods using the same RNA-seq dataset with the human LCL treated with Nutlin (Figure 2.5). I observed that although there are a few differences in some genes, the overall fold-changes remain quite similar. I therefore decided to use the whole gene regions to define fold-changes in the rest of the study.

However, there is considerable variation in the quality of the gene annotations across the 12 species used here. For example, the human reference genome hg38 has a vast quantity of annotated genes, whereas the squirrel monkey reference genome saiBol1 lags behind in their number of annotated genes. This difference in the number of defined genes can impact downstream analyses, including the obvious one that one cannot test the induction of a given gene if such a gene is not even being interrogated, to more nuanced impacts such as the number of multiple hypothesis testing corrections that are done to assess statistical significance across species.

To overcome the above hurdle, I set out to define a standard primate gene annotation set across my species of interest, such that each annotation set contains the same number of genes, those that can be assigned unambiguously to orthologs across the species (see Methods section for details). To make these standard gene annotations, I removed genes if they were not present in all species, and also genes that have ambiguous orthology assignments. I refer to this consensus gene set as the "standard annotations" hereafter. I tested the number of differentially expressed genes (DEGs) when using both the full and standard annotations using DESeq2 [119] (Figure 2.6), and observed a similar pattern in the number of DEGs called. Importantly, I observed that neither the cow or the chicken datasets seemed to have been properly stimulated by Nutlin, as observed with

Figure 2.5: Scatterplot showing the RNA-seq FPKM normalized expression values of human genes from the human LCL treated with Nutlin. In the horizontal axis the values were obtained from the unmodified GTF annotation file that has exons and introns, and where the reads were only counted over exons. In the vertical axis the values were obtained from a modified annotation file that has a single genomic interval spanning from the first to the last exon, and reads were counted over the whole region regardless of exons or introns. The dotted diagonal red line represents the identity line.

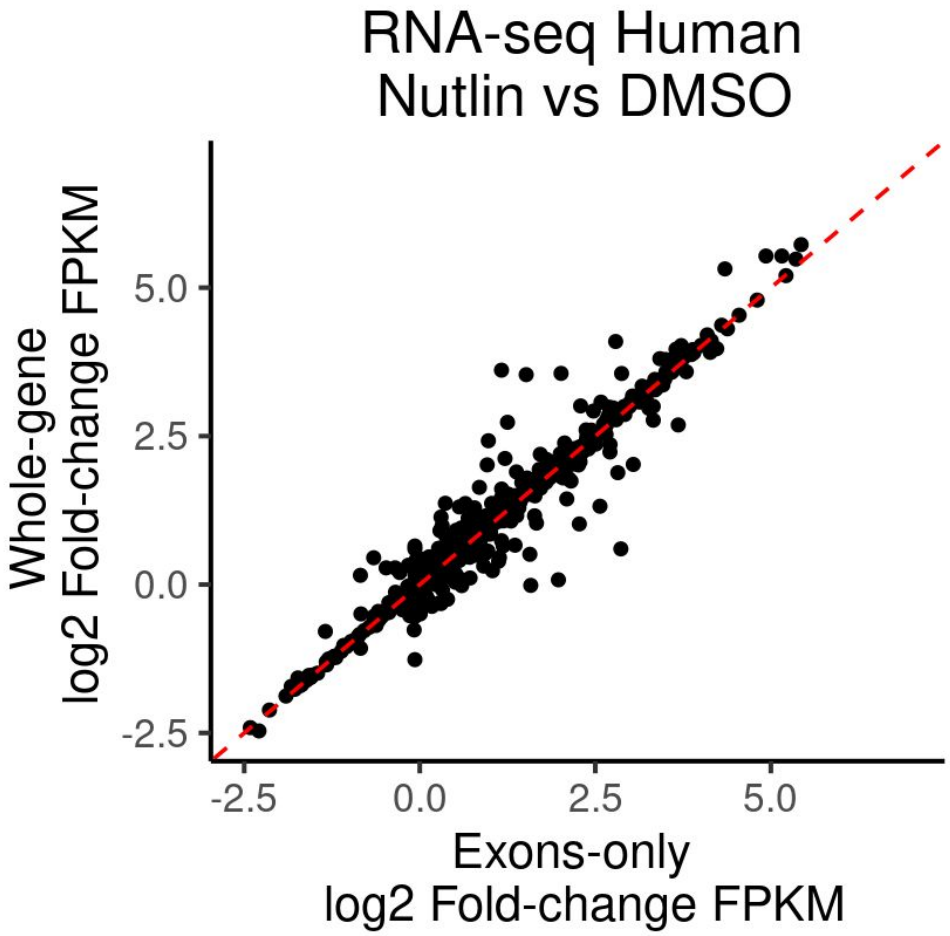
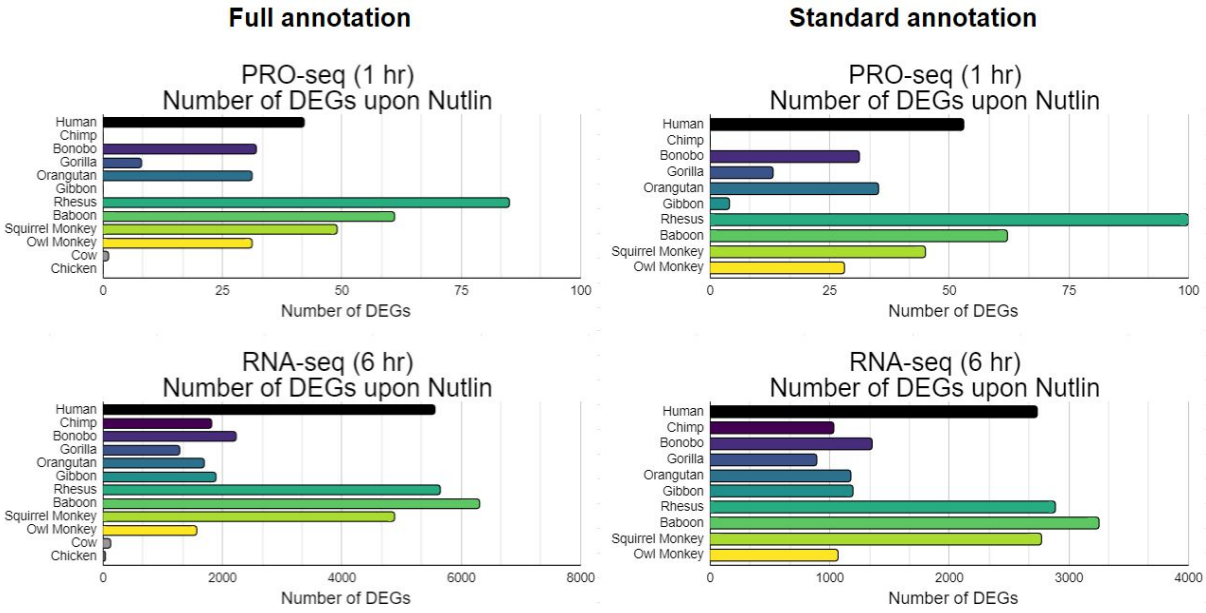


Figure 2.6: Barplots showing the number of differentially expressed genes (DEGs) found in the PRO-seq dataset (top) and in the RNA-seq datasets (bottom) using either each species full public gene annotation set (left) or the standard annotation for the 10 primates (right), on the species LCLs treated with Nutlin. DEGs were obtained with DESeq2 using an alpha level of 0.05. DESeq2 adjusted p-values account for the size of the gene set.



a non-existent or negligible number of DEGs for both their PRO-seq or RNA-seq datasets. Both cow and chicken were dropped out of the subsequent analyses, refocusing the p53 transcriptional evolution study only within anthropoids. The standard annotation, therefore, was used to test DEGs within the primate species in the study.

A closer look at the p53-driven response across primates showed that a few of the better characterized genes upregulated by p53, CDKN1A (also called p21) and MDM2, were significantly upregulated across the 10 primates (Figure 2.7). CDKN1A is a master regulator of the cell cycle, which halts cell cycle progression in damaged cells to prevent cancer. MDM2, is the very protein that negatively regulates p53 activity, and its upregulation is a known negative feedback mechanism that cells have evolved to not let the p53 response get out of control. I observed both (CDKN1A and MDM2) upregulated early on as part of the primary transcriptional response (in PRO-seq).

Using the standard primate gene annotation, I analyzed how closely the overall gene response to Nutlin appeared by clustering the PRO-seq and RNA-seq datasets using principal component analysis using as the gene sets the union of DEGs across the 10 primates (Figure 2.8). The results were very similar in both the PRO-seq and RNA-seq datasets. I saw that the datasets separate first by species, and then by treatment. The datasets also separate by their evolutionary distance, with all hominoids clustering first, while having the old world monkeys further away from the hominids on one side, and the new world monkeys from both the hominoids and the old world monkeys on another side.

To narrow down on the p53-driven gene transcription differences that each primate displays, I performed Gene Set Enrichment Analysis (GSEA) [179] using the Hallmark gene sets from the Molecular Signature Database on the ranked set of DEGs as defined by DESeq2 (Figure 2.9). I observed that both the primary and downstream time points obtained similar enriched gene sets related to cell cycle regulation, and also directly annotated to be part of the p53 regulated gene set.

In addition to observing the overall enriched genic response with GSEA, I tested the enrichment of the differential colocalization of TF motifs with the Nutlin-induced loci as a proxy

Figure 2.7: Volcano plots showing the log2 fold-change in the horizontal axis and the -log10 adjusted p-value on the vertical axis showing the induction of genes in the standard annotation for each of the 10 species LCLs used in the study. Labeled are two canonical genes induced by p53: CDKN1A and MDM2. In the first track the plots are from the PRO-seq datasets, and in the second track the plots are from the RNA-seq datasets.

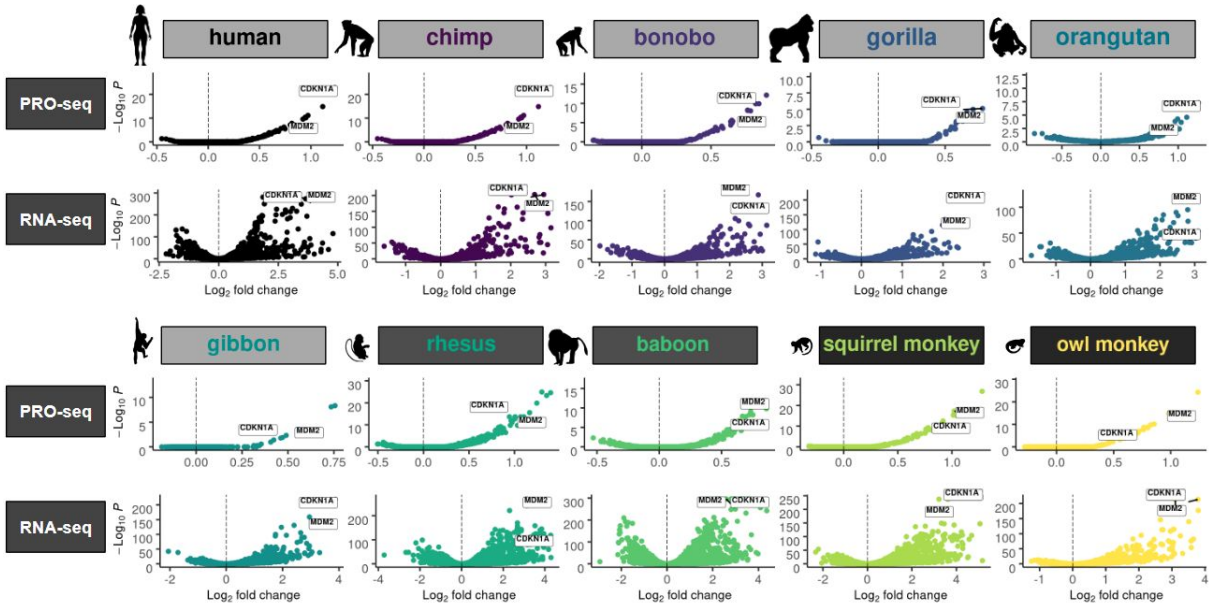


Figure 2.8: Scatterplots showing a principal component analysis (PCA), showing the first principal component in the horizontal axis and the second principal component in the vertical axis. Left, PCA from the PRO-seq datasets showing the 175 genes resulting from the union of all DEGs from the 10 primates. Middle, PCA from the RNA-seq showing the 5412 genes resulting from the union of all DEGs from the 10 primates. Right, cladogram showing the phylogenetic relationship among the 10 primates, displaying the colors used in the PCA dots. Circular shapes denote DMSO-treated samples, and triangles denote Nutlin-treated samples.

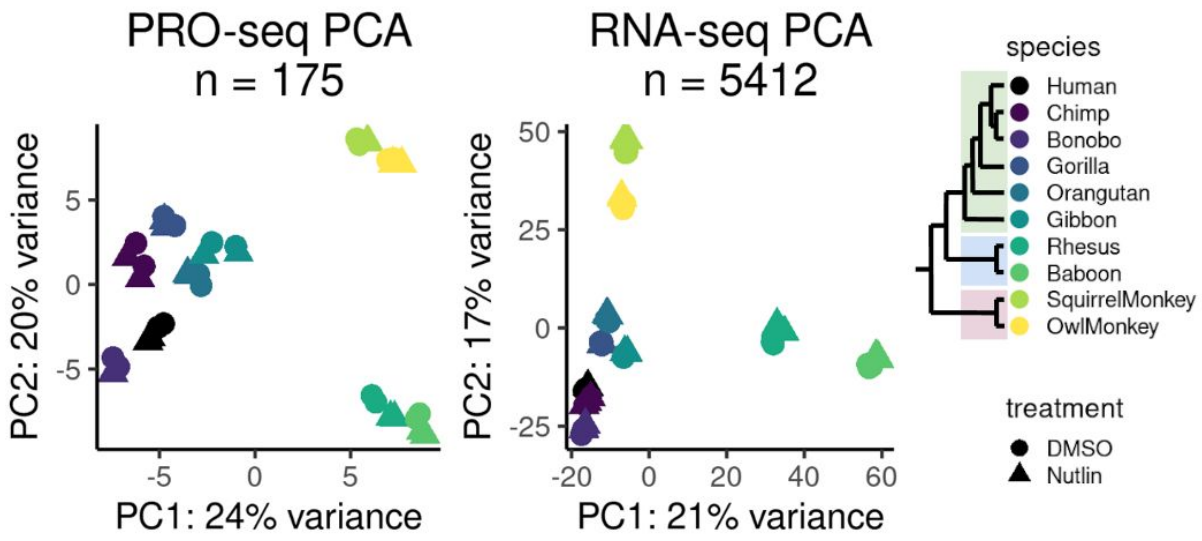


Figure 2.9: Dotplots showing the top gene sets from the Hallmark gene sets from the Molecular Signature Database, with the -log10 adjusted p-value (family-wise error rate) in the horizontal axis as determined by the Gene Set Enrichment Analysis (GSEA). On top are the results from the PRO-seq datasets, on the bottom are the results from the RNA-seq datasets. The color scheme is denoted on the right, with shapes representing the three main clades of the anthropoids sampled: circles being hominoids, triangles being old world monkeys, and squares being new world monkeys.

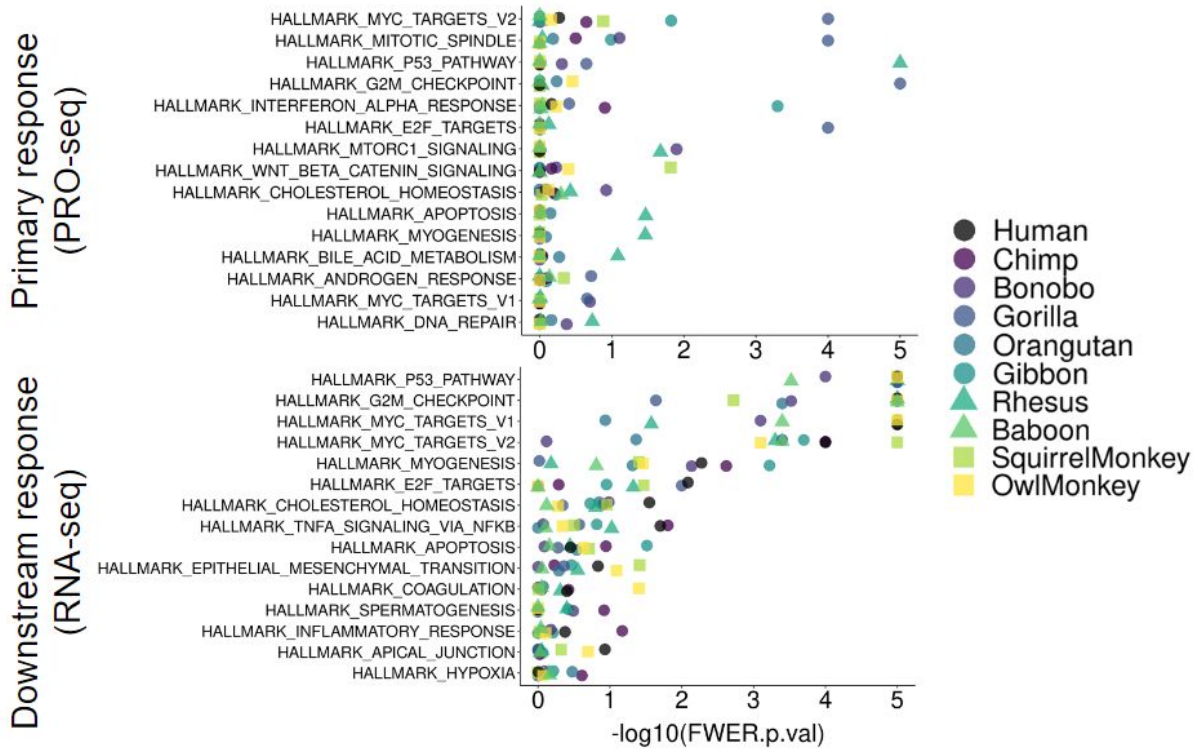


Figure 2.10: MA plots obtained from the Transcription Factor Enrichment Analysis (TFEA) showing on the horizontal axis the number of motif hits (as log10), and in the vertical axis the corrected E-score. Each dot represents a motif from the JASPAR2022 non-redundant vertebrate motif database. Labeled are the motifs from the TP53 and TP63 transcription factors, which have nearly identical motifs. The top two rows are from the PRO-seq datasets, and the bottom top rows are from the RNA-seq datasets.

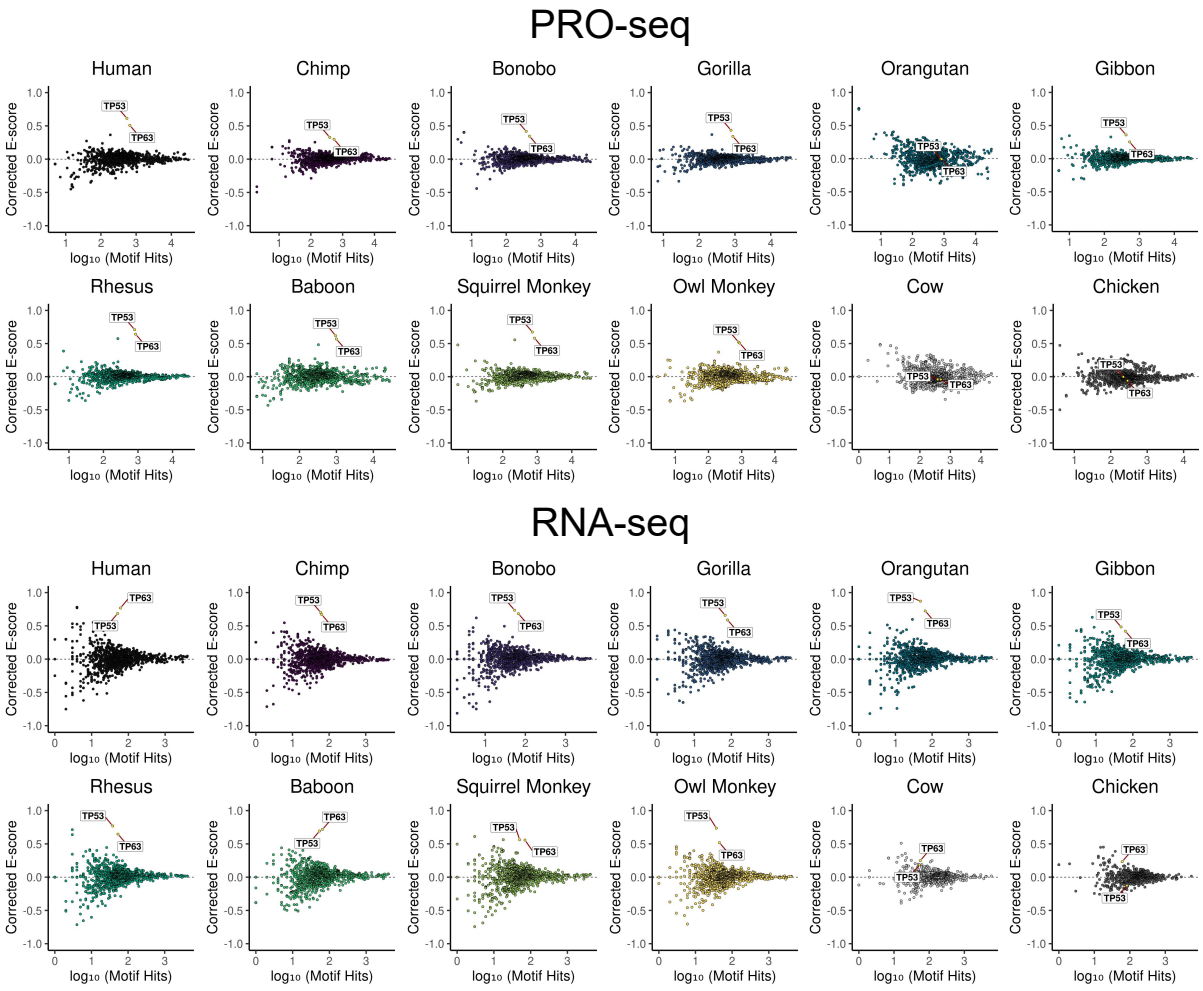
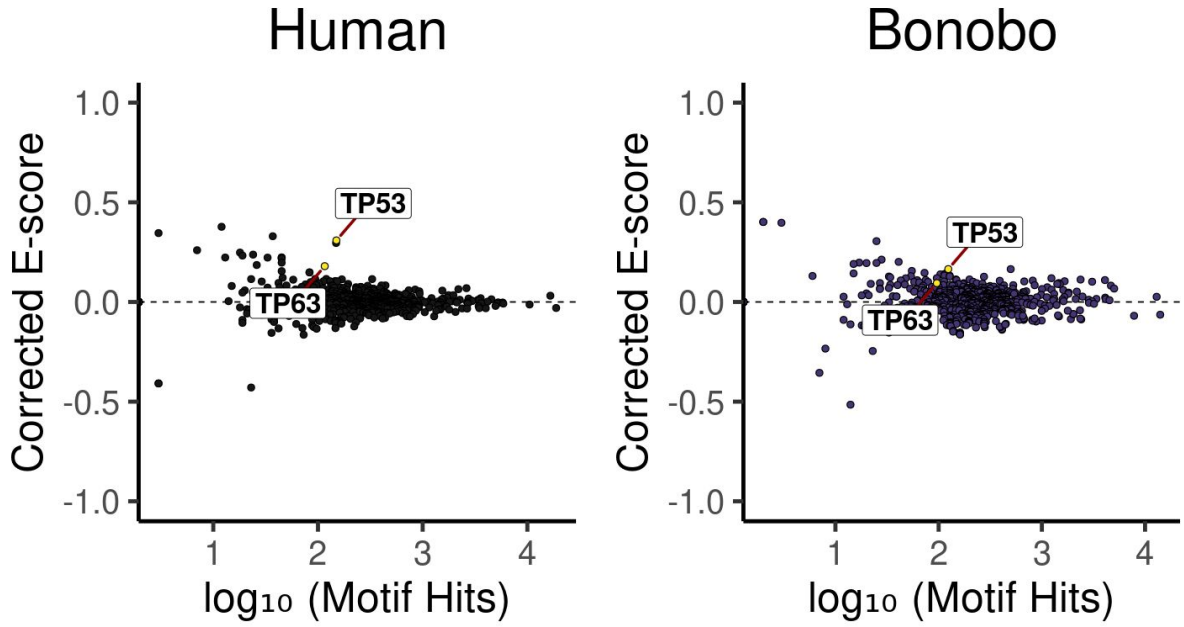


Figure 2.11: MA plots obtained from the Transcription Factor Enrichment Analysis (TFEA) showing on the horizontal axis the number of motif hits (as log10), and in the vertical axis the corrected E-score. Each dot represents a motif from the JASPAR2022 non-redundant vertebrate motif database. Labeled are the motifs from the TP53 and TP63 transcription factors, which have nearly identical motifs. The two plots are from the ATAC-seq datasets.



to assess what TFs are driving the different observed genic responses in each primate using the Transcription Factor Enrichment Analysis (TFEA) [161] (Figure 2.10 and Figure 2.11). TFEA was used with the JASPAR2022 non-redundant vertebrate TF motif database [32]. For the case of the PRO-seq datasets, my loci of interest were the transcribed bidirectional loci detected by Tfit. For the RNA-seq datasets, my loci were all of the annotated gene transcriptional start sites from the primate gene standard annotation. For the ATAC-seq datasets, my loci were the peaks detected by the peak caller MACS2 [212]. The results showed that in both the primary (PRO-seq and ATAC-seq) and downstream (RNA-seq) timepoints, both the TP53 and TP63 motifs appear as the only enriched TF motifs for most species. It should be noted that both p53 and p63, as the paralogs they are, share almost indistinguishable DNA motif sequences.

2.3.2 Gene-centric interrogation of the rewiring of the p53 transcriptional response

After being satisfied that my primates had been sufficiently induced by Nutlin to activate p53 with the above results, I wanted to investigate to what extent the primary and downstream p53-driven transcriptional response was shared across the 10 primates. Because of the difference in the overall magnitude of response across the primates, I decided not to focus on differences in fold-change magnitude across orthologous genes, but rather just consider if the orthologous genes were transcriptionally induced or not in a binary fashion. I compiled the primary and downstream gene sets by using the union of DEGs across the 10 primates, from the PRO-seq and RNA-seq datasets, respectively. To assess if a gene was induced in either gene set, I relied on the adjusted p-values being less than a fixed alpha of 0.05 using the RNA-seq datasets for both the primary and downstream gene sets. This was done because the PRO-seq datasets are not sufficiently sensitive due to their intrinsic low fold-change and the fact that I only obtained two replicates versus the three replicates for RNA-seq.

I show the binary induction of both the primary and downstream p53 gene sets in Figure 2.12, ordered by the number of species in which a given gene is induced. I interrogated how often the genes in both gene sets have a TP53 motif instance in their promoter sequences, finding no

Figure 2.12: Binary heatmaps showing in the horizontal axis genes and in the vertical axis the species LCLs tested. White color means the gene in that species was not induced by Nutlin, black color means the gene in that species was induced by Nutlin. Induction is defined by an adjusted p-value less than alpha = 0.05 by DESeq2. The purple and green track on top of the heatmaps denotes if the gene has a TP53 motif located +/- 1.5 kb from the annotated transcription start site as defined by FIMO using a threshold of 10^{-6}. The track with the coloration going from black to light yellow denotes the number of primates where the gene was found to be induced by Nutlin. The top heatmap is obtained from the PRO-seq datasets, and the bottom heatmap is obtained from the RNA-seq datasets.

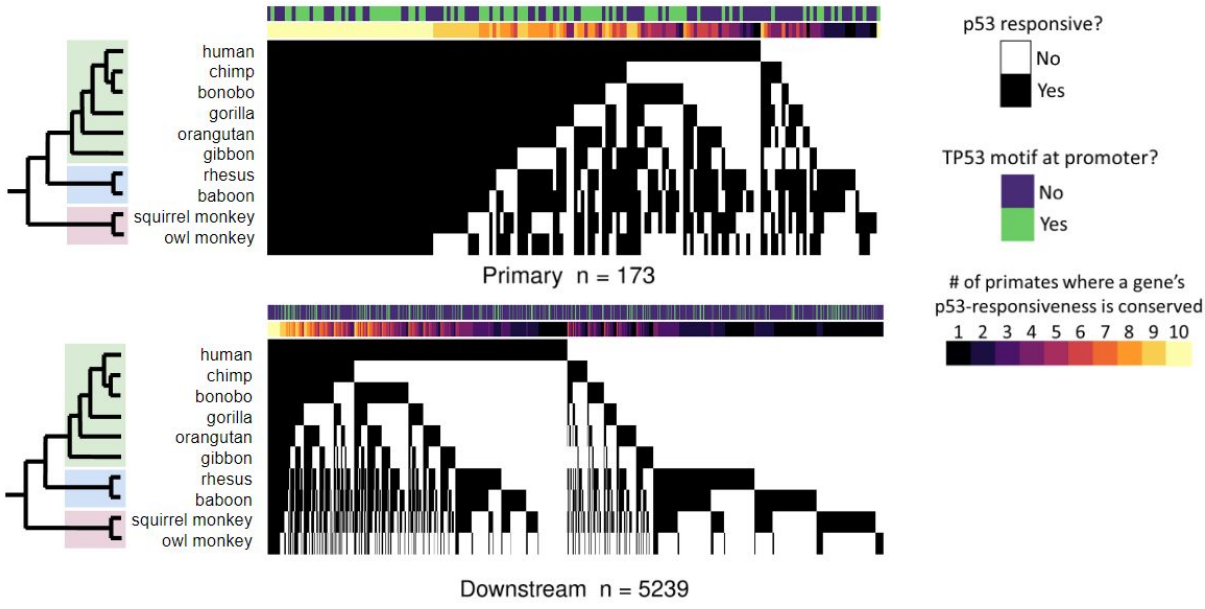
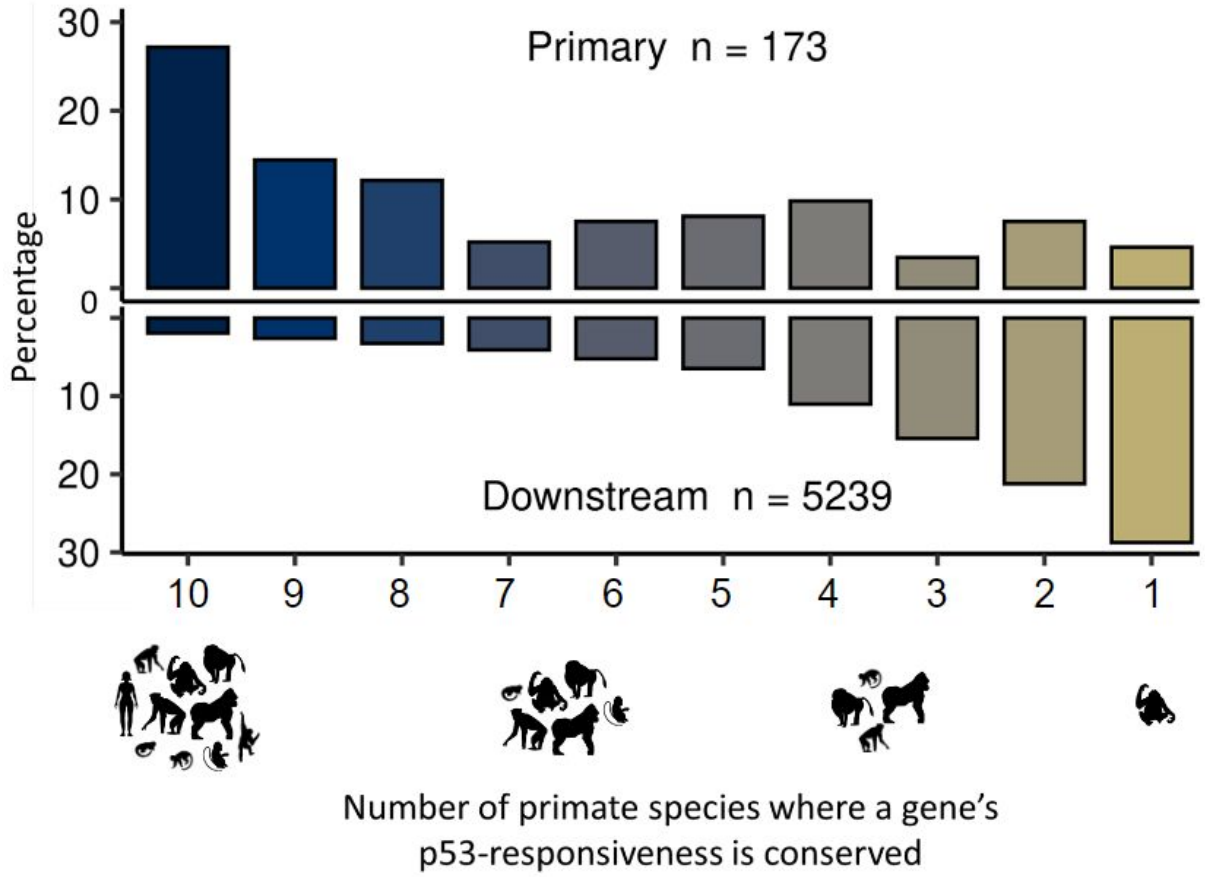


Figure 2.13: Histograms showing in the horizontal axis the number of primates where a given gene is induced by Nutlin and in the vertical axis the percentage from the total of genes in the primary gene set (top) or the downstream gene set (bottom). Note that cartoons indicate the number of species rather than specific species.



obvious enrichment for the primary relative to the downstream gene sets. These results seem to indicate that the primary response is conserved more than the downstream response. Indeed, after quantifying these conservation patterns (Figure 2.13), I clearly observed that the primary response is much more conserved than the downstream response, with a very interesting inverse relationship, with the downstream response having most of its genes induced in only a single or a few primates.

The pair-wise comparison between human and each non-human primate is shown in Figure 2.14, and it further suggests that as the evolutionary distance increases, there are fewer genes whose induction is shared (gray color dots), and more species-specific induced genes (non-gray colored dots).

One possible explanation for the opposite conservation patterns between the primary and downstream gene sets is that differences in the primary response are TF themselves, and that these TFs drive the gene diversity that is expressed shortly afterwards. To test this possibility, I checked how often the genes in each primary and downstream categories are classified as TFs using two different published TF catalogs [191, 106] (Figure 2.15). The results showed that there are not a lot of primary genes classified as TF, and many in the downstream gene sets are TFs themselves. In addition, as a sanity control, I looked at the proportion of the primary and downstream gene sets that I obtained in this study that match the p53 core gene set as proposed by [5], and I saw that the conserved genes from my primary gene set were in strong agreement with with this p53 core gene set.

Next, I wanted to interrogate if there are epigenetic features that correlate with the extent with which a gene's induction by p53 is conserved across primates (Figure 2.16). I divided my set of p53-responsive genes into 10 bins, denoted by the number of primates where the induction is observed in RNA-seq. And I looked at different aggregate coverage signals on human LCLs over each of these 10 gene bins, including DNA accessibility (my ATAC-seq dataset), histone marks associated with regulatory elements (ChIP-seq for H3K4me3 and H3K27ac), DNA methylation (bisulfite sequencing), as well as p53 binding (ChIP-seq for p53 upon activation), and as a positive control I looked at DNA sequence conservation (PhastCons score across mammals [171]). The

Figure 2.14: Scatterplots showing the pairwise log2 fold-change values of genes of human versus all other nine non-human primates. Each dot is a gene, and it is colored black if the gene is significant only in human, colored gray if it is significant in both primates, and another color if the gene is significant only in the other primate. Significance is defined by an adjusted p-value less than an alpha = 0.05 by DESeq2. The top two rows are from the PRO-seq datasets, and the bottom two rows are from the RNA-seq datasets.

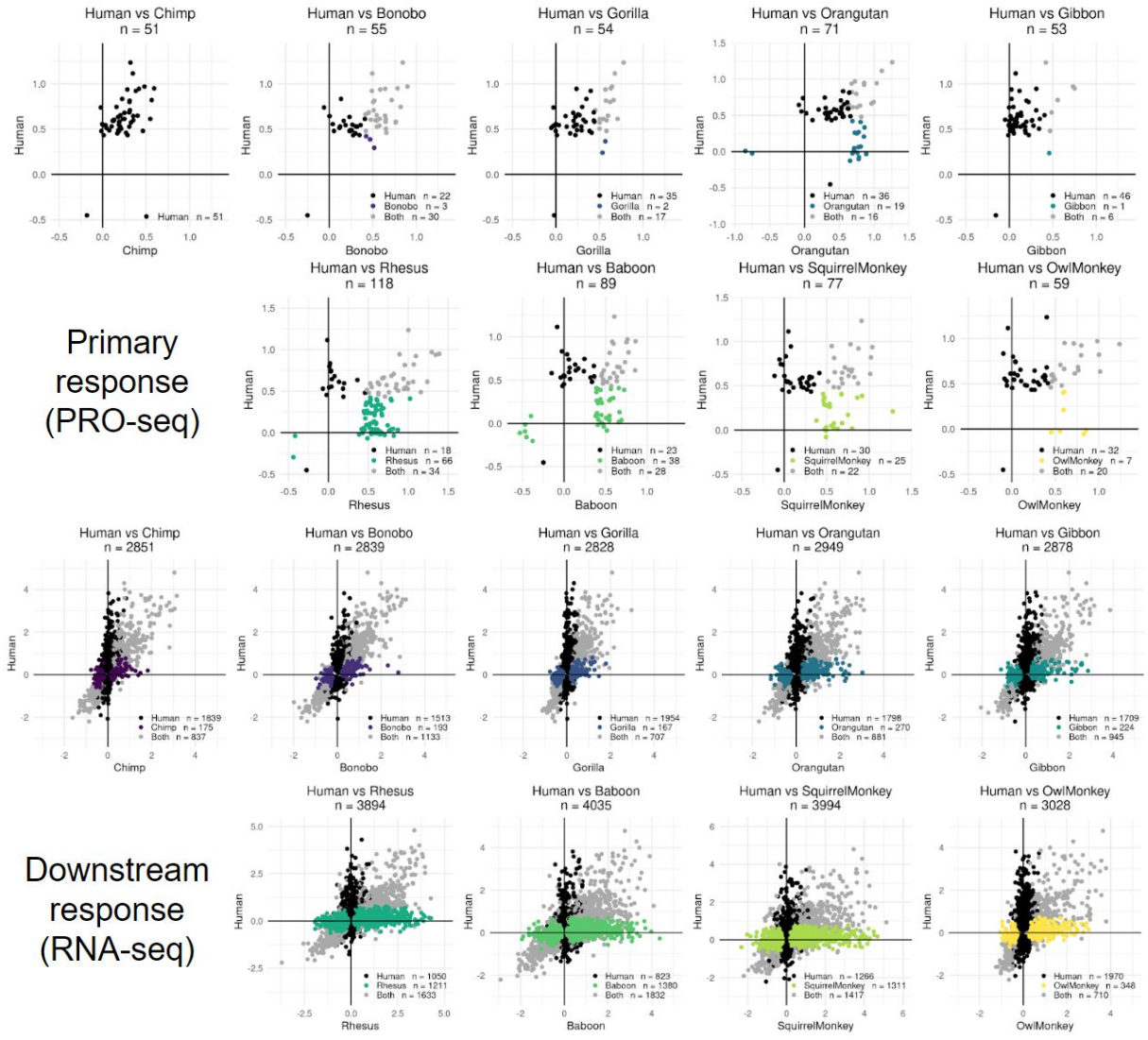


Figure 2.15: Binary heatmaps showing in the horizontal axis genes and in the vertical axis three different binary categories, where white denotes the gene is not part of the category and black denotes the gene is included in the category. The top category is the gene set defined as part of the p53 core set by [Andrysik2017], the middle category is the gene set defined as being transcription factors by [Vaquerizas2009], the bottom gene set is defined as being transcription factors by [Lambert2018]. The track with the coloration going from black to light yellow denotes the number of primates where the gene was found to be induced by Nutlin. The top heatmap is obtained from the PRO-seq datasets, and the bottom heatmap is obtained from the RNA-seq datasets.

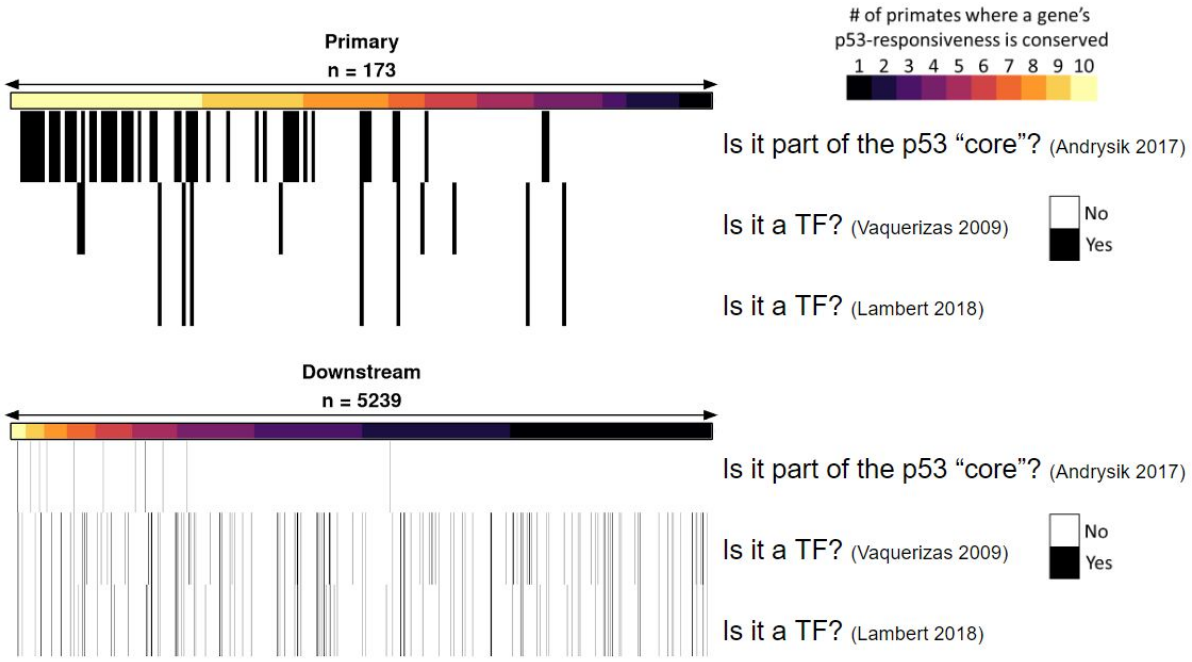
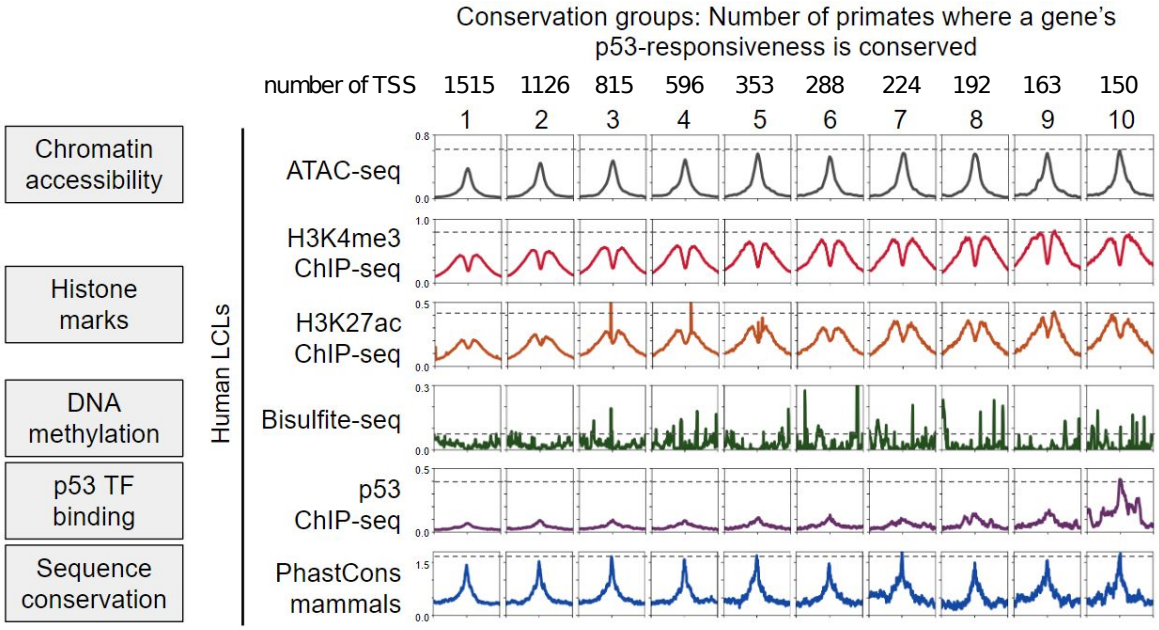


Figure 2.16: Metaplots showing the aggregate signal by different genomic assays in each row and by each of the 10 conservation bins in each column. The ATAC-seq bigwig coverage file was obtained from the human (GM12878) treated with DMSO from this study. The SRA numbers for the ChIP-seq for H3K4me3 from cell line GM12878 are SRR577356 and SRR577357. The SRA numbers for the ChIP-seq for H3K27ac from cell line GM12878 are SRR227633 and SRR227634. The SRA number for the whole-genome bisulfite sequencing from cell line GM12878 is SRR4235788. The SRA numbers for the ChIP-seq for p53 from cell line GM12878 is SRR851807. The human PhyloP scores bigwig coverage file was obtained from UCSC Zoonomia Cactus files (https://cglgenomics.ucsc.edu/data/cactus/).



results showed that there is no discernible difference in DNA methylation or sequence conservation, whereas there is a positive correlation in the aggregate signature for ATAC-seq and ChIP-seq for H3K4me3, H3K27ac, and p53 as the genes' induction gets more conserved.

While my analyses so far suggest that the conservation of the p53 responsive transcriptional network is relatively more conserved at 1 hour compared to 6 hours, I decided to further test these observations by paying a closer look at potential false positive cases of gene induction. In particular, using the principle of parsimony, I decided to categorize each of the genes in the primate standard gene annotation with the fewest numbers of evolutionary events that could explain their current induction patterns (Figure 2.17). An evolutionary event is defined as an instance in the past where a last common ancestor in an internal branch in the primate phylogeny endured a mutation that made a change in a given gene's p53-responsiveness. For example, a gene that is p53-responsive in rhesus and baboon but not in the hominoids or new world monkeys can be thought of having had single evolutionary event in the common ancestor of old world monkeys where that gene was added into the p53-responsive network. From the ∼8,000 genes in the primate standard gene annotation, around one third show no p53-responsiveness in any of the 10 primates tested. The other two thirds show p53-triggered induction in at least one primate (Figure 2.17 top). Using the principle of parsimony, I posit that single evolutionary events are much more likely than multiple evolutionary events to explain the current induction patterns. Further, if a given gene's induction pattern cannot be explained by a single evolutionary event (e.g. a gene is only induced in rhesus and owl monkey), I cannot be certain if the observed pattern is a real unlikely evolutionary scenario or if it appears as such due to their induction classification being a false positive.

Having removed the p53-induced genes that show unlikely evolutionary trajectories (e.g. with more than one evolutionary event to explain its current induction patterns), I tested if the inverse conservation patterns for the primary and downstream gene sets are maintained (Figure 2.18 top). Though this stringent filter removed almost one third of the genes, I still observed a similar pattern, with the primary genes' induction being conserved in many primates while the downstream genes being induced only in a few or a single primate. Moreover, I tested the persistence of these

Figure 2.17: Classification of all 8028 genes from the primate standard annotation by their parsimonious scenarios in which the p53-induction was acquired in the primate evolutionary tree. Top, histogram with the number of genes categorized in each of the 36 likely evolutionary scenarios where there was a single mutational event. The scenarios are labeled with single letter codes, with both upper and lower case letters denoting one possible scenario and its inverted scenario representing a loss or a gain of p53-responsiveness. Scenarios Aa to Gg denote events in the inner nodes of the evolutionary tree, or of common ancestors; scenarios Hh to Qr denote events in the final nodes of the evolutionary tree or occurring in the extant species; scenarios Rr denote scenarios where the gene is either not induced in any primate or induced in all 10 primates. In the top inset the venn diagram shows in light green the number of genes from the standard annotation with no p53-responsiveness in any of the 10 primates tested, in dark green the number of genes that show an induction change explained by only 1 change under parsimony, in red the number of genes that can only be explained by at least two or more evolutionary events and are thus discarded due to them being unlikely and being potentially false-positives.

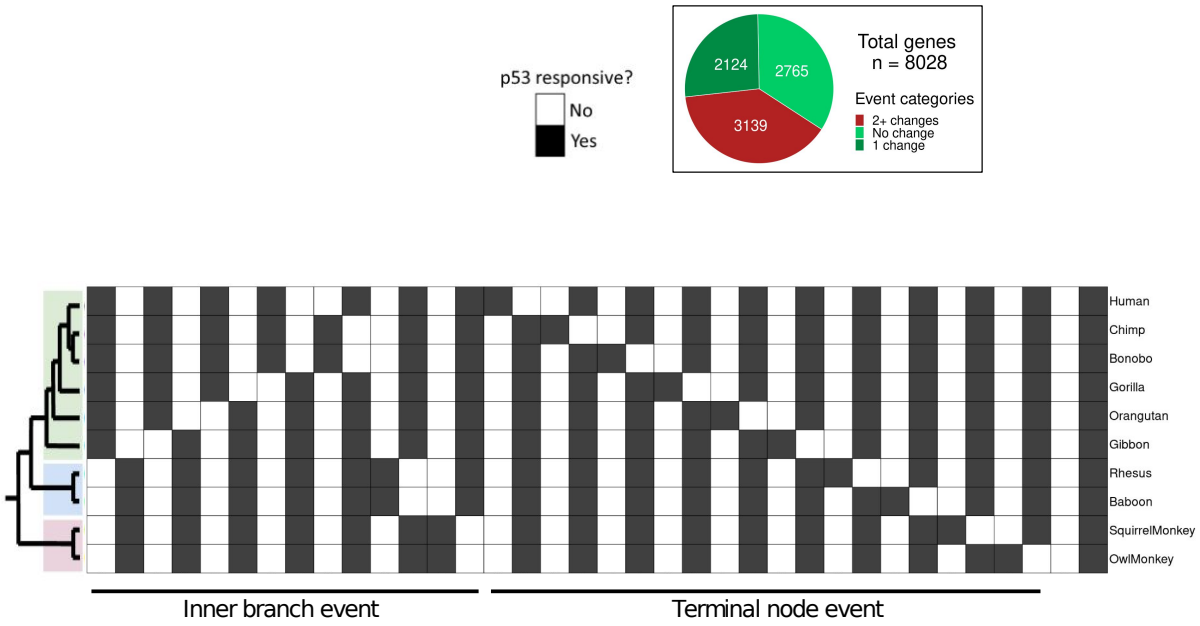
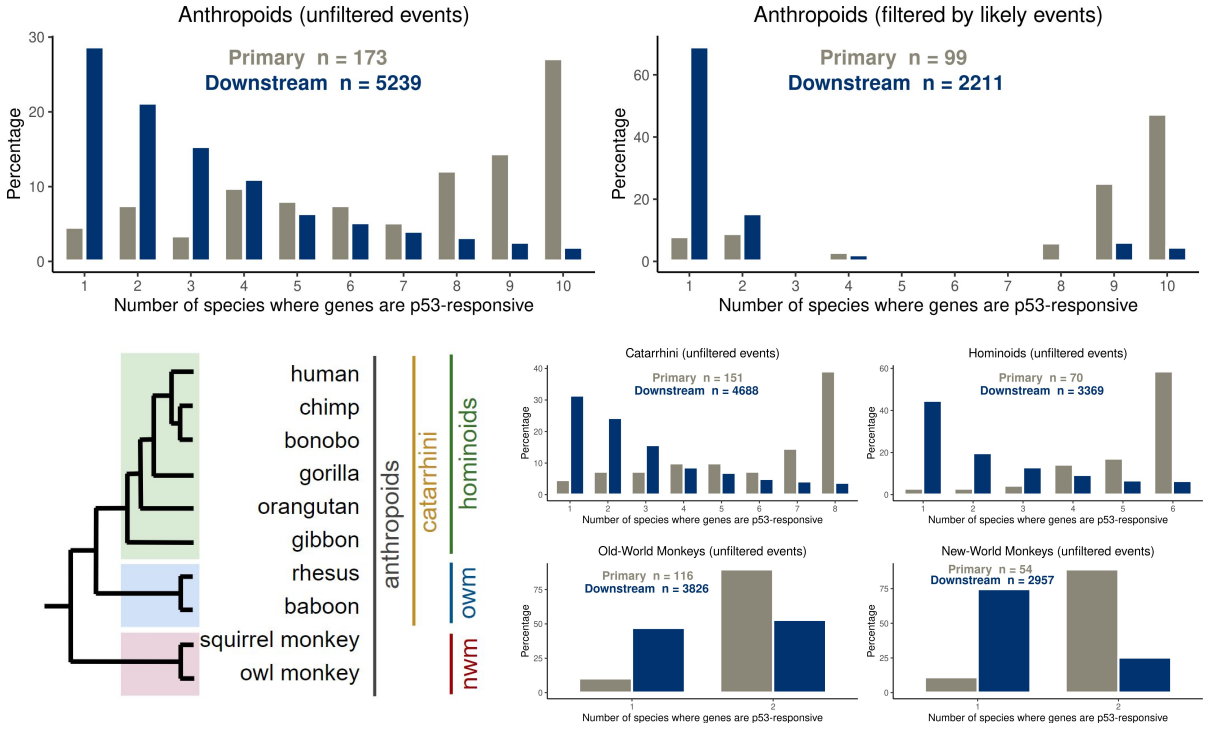


Figure 2.18: Top row; left, histogram showing the percentage of the total number of gene per gene set (primary in gray and downstream in blue) in the vertical axis, and the total number of primates where genes are p53-responsive in the horizontal axis, when considering all genes in the standard annotation considering all 10 anthropoids. Top row; right, similar to left but considering only genes whose induction can be explained by a single evolutionary event. Bottom, left; cladogram of the 10 anthropoids in this study, further categorized in subgroups. Bottom, right; four histograms that contain the p53-induced genes when considering only the four phylogenetic subgroups: catarrhini, hominoids, old world monkeys, and new world monkeys.



conservation patterns by sampling the primates to only keep the inner clades (e.g. hominoids-only, old world monkeys-only, new world monkeys-only), and the pattern persisted. (Figure 2.18 bottom).

To finish the gene-centric interrogation of the rewiring of the p53-responsive transcriptional response, I decided to test the extent with which the ranking of the responsive genes has diverged between the tested primates. Meaning, to see if the most induced genes are also the top genes across the other species. To achieve this, I used GSEA again but instead of using public gene sets, I defined my own reference gene sets. I fixed the gene set as being that of the human LCL DEG from, and I tested the enrichment curves and score for each other non-human primate ranked gene lists, for either the primary (Figure 2.19) or downstream (Figure 2.20) datasets.

2.3.3 Regulatory element-centric interrogation of the rewiring of the p53 transcriptional response

Afterwards, I set out to explore the differential usage of transcribed regulatory elements. I defined the set of human induced transcribed bidirectional loci using TFEA with the human PRO-seq datasets. TFEA uses DESeq2 internally to rank bidirectionals based on their transcription levels. I took the top 1000 out of ∼30,000 as my set of induced bidirectionals, and used liftOver [92] to obtain the orthologous loci in the other primate reference genomes. Owl monkey was not used for this analysis, as there are no existing chain files with which to cross-map loci between it and other reference genomes. I compared the expression levels, normalized by RPKM, between human and each of the other 8 non-human primates (Figure 2.21).

Some TFs are thought to bind preferentially to distal enhancer regions, whereas other TFs bind preferentially to proximal promoter regions, when regulating their target genes [4]. Next, I decided to test how this predilection of what regions TFs bind to in the genome changed throughout the phylogeny of the 12 species in the original study, including cow and chicken. I took the full gene annotation from each animal (not the standard annotation), and obtained a list of all gene transcription start sites (TSS), covering a window of +/- 1.5 kb. I filtered this TSS list so that

Figure 2.19: Left, enrichment curves from the Gene Set Enrichment Analysis (GSEA) from the PRO-seq datasets for the ranked gene lists for each of the 10 primates using the gene set fixed as the DEGs from the human LCL treated with Nutlin. A horizontal red dotted line denotes the enrichment score as a reference line. Right, barplot showing the normalized enrichment score obtained from the leading edges of the enrichment curves on the left. A vertical red dotted line denotes the score for human as a reference line.

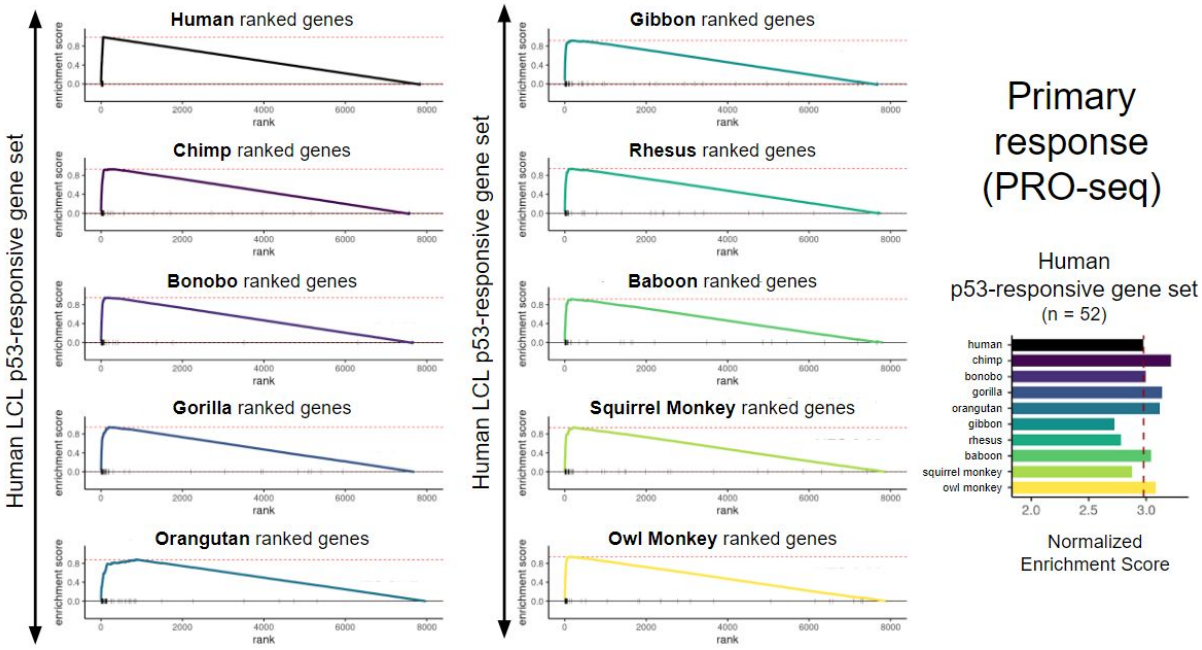


Figure 2.20: Left, enrichment curves from the Gene Set Enrichment Analysis (GSEA) from the RNA-seq datasets for the ranked gene lists for each of the 10 primates using the gene set fixed as the DEGs from the human LCL treated with Nutlin. A horizontal red dotted line denotes the enrichment score as a reference line. Right, barplot showing the normalized enrichment score obtained from the leading edges of the enrichment curves on the left. A vertical red dotted line denotes the score for human as a reference line.

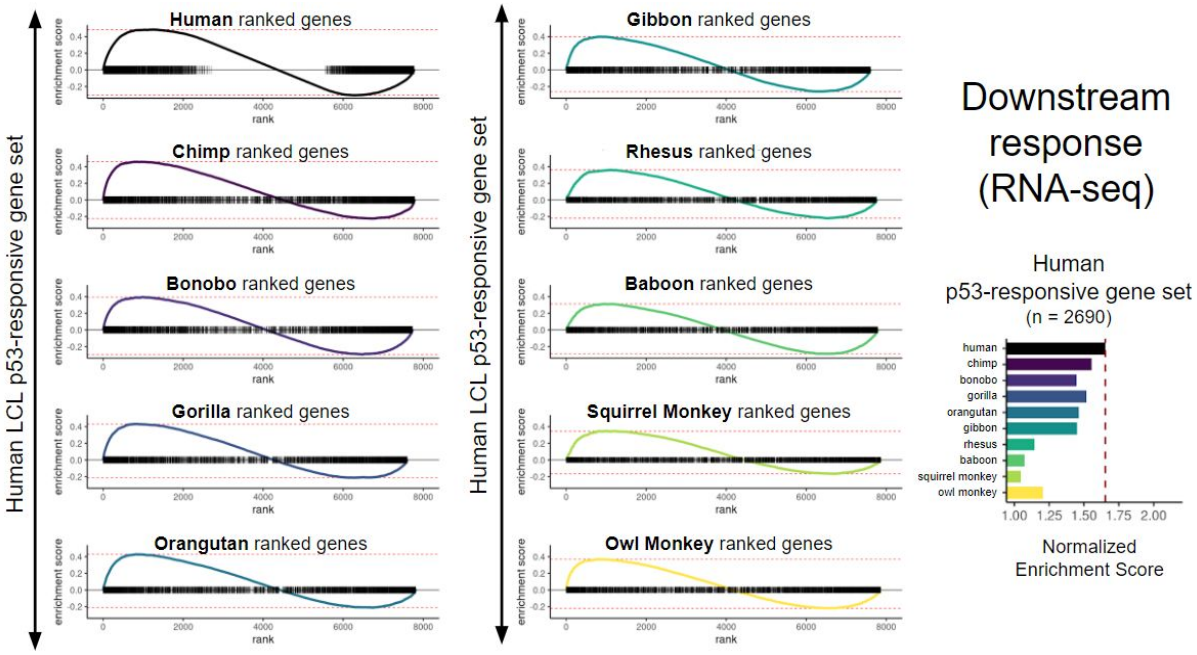
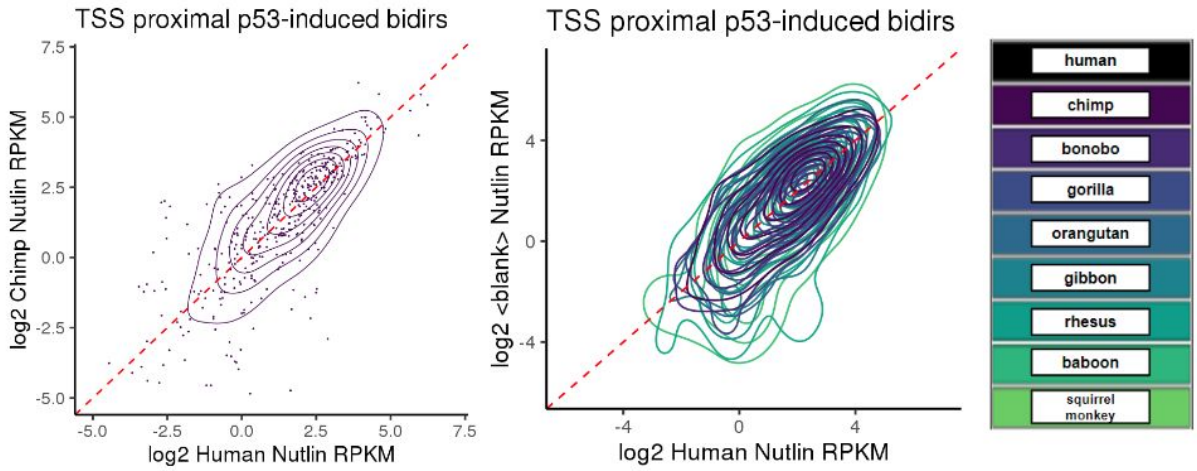
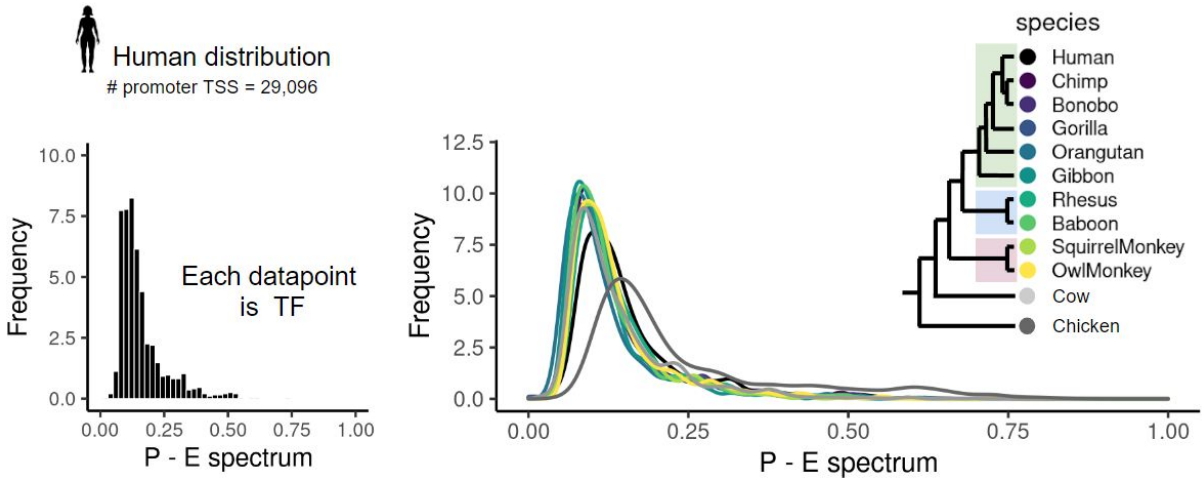


Figure 2.21: Left, scatterplot showing the log2 RPKM values of p53-responsive bidirectionals in the human LCL on the horizontal axis and the log2 RPKM values of the corresponding orthologous bidirectionals found in the chimp LCLs. Each dot is a bidirectional loci as detected by Tfit. Orthologous loci are defined using liftOver. Right, overlay of the density contours of all scatterplots of all 8 non-human primates (except for owl monkey) relative to the human expression values. Density contours are displayed with the color scheme depicted in the far right. Owl monkey is missing as there are no liftOver chains between owl monkey and any of the other primates in the study.



I only kept the promoters that are actually being transcribed by LCLs by overlapping the TSS list with Tfit-detected bidirectional loci. I scanned each species reference genome with each motif from the JASPAR2022 non-redundant vertebrate TF motif database with FIMO using a stringent significance threshold of -10^{-6}. I then simply categorized each motif instance from each TF as either being associated with a promoter region (the transcribed TSS list) or if they are distant from these promoter regions. This gives a fraction which I termed the Promoter-Enhancer score, where a score closer to 0 indicates the TF prefers to bind to distal enhancers, and a score closer to 1 indicates the TF prefers to bind to proximal promoters (Figure 2.22). The results indicated that most TFs have a score of around 0.15, which is very biased towards motifs rarely overlapping transcribed promoter regions.

Figure 2.22: Left, histogram showing in the horizontal axis the numerical value expressed from 0 to 1 on the fraction of times a given transcription factor motif is localized at the promoter of genes (+/- 1.5 kb from the gene transcription start site), and in the vertical axis.the frequency of transcription factor motifs with that fraction value. Right an overlay of the histogram contours for each of the 12 species used in this study, following the color scheme as denoted in the cladogram shown on the far right.



I examined what TFs have the most varied Promoter-Enhancer scores across the 12 species (Figure 2.23). Intriguingly, the top 5 TFs with the highest score variance were all zinc finger TFs, which are known to evolve rapidly because of its activity to silence ever changing transposable

elements [186, 204]. On the other hand, the top 5 TFs with the lowest score variance show that TP53 itself is the least variable TF in terms of its predilection for using promoters or enhancers.

2.3.4 Exploring confounding factors affecting the interpretation of the p53 transcriptional response

The results so far suggest that the transcriptional response upon p53 activation has been rewired in the primate phylogeny. However, as no study system is perfect, there are some worrying confounding factors that affect the confidence in my interpretations.

The primate LCLs used in this study are transformed by Epstein-Barr Virus (EBV), except for the rhesus LCL which was transformed by Papiine Herpesvirus 1. It is safe to assume that the LCLs are not identical to the primary B-cells they were derived from [215]. However, not only are they not naturally occurring cells, but the viral activity may differ across the LCLs. To approximate their activity, I determined the percentage of reads coming from each transformant virus (Figure 2.24). The results indicate that, indeed, the viral load is not the same across the 10 primate LCLs. For example, the human, squirrel monkey, and owl monkey LCLs have similarly high percentages of EBV-derived reads; whereas the chimp, gorilla, and baboon LCLs have similarly low percentages of EBV-derived reads.

EBV can exist in at least three known latency programs which differ in the degree of their intracellular activity [91]. I then determined the expression values of key EBV genes that are known to control the viral latency stage (Figure 2.25). Across the 9 primate LCLs transformed with EBV, I observed that they all display a similar latency state as approximated by the expression of the EBV TFs EBNA1, EBNA2, and EBNALP.

Though EBV may be active in the LCLs with different magnitudes, what is of special concern is the degree with which the virus is directly implicated in the disruption of the p53 transcriptional response. To test this, I checked the colocalization of EBV TFs with p53 regulatory elements, by relying on publicly available ChIP-seq for the EBV TF EBNA2, as well as positive controls such as ChIP-seq for the enhancer-associated mark H3K27ac and for the p53 protein itself (Figure

Figure 2.23: Left, cladogram of the 12 species used in this study with their respective assigned colors. Middle, barcode plots of the five transcription factor motifs with the most variable (i.e. highest standard deviation across the 12 values) fraction of motifs localized at the promoters of genes (+/- 1.5 kb from the gene transcription start site). Right, barcode plots of the five transcription factor motifs with the least variable fractions.

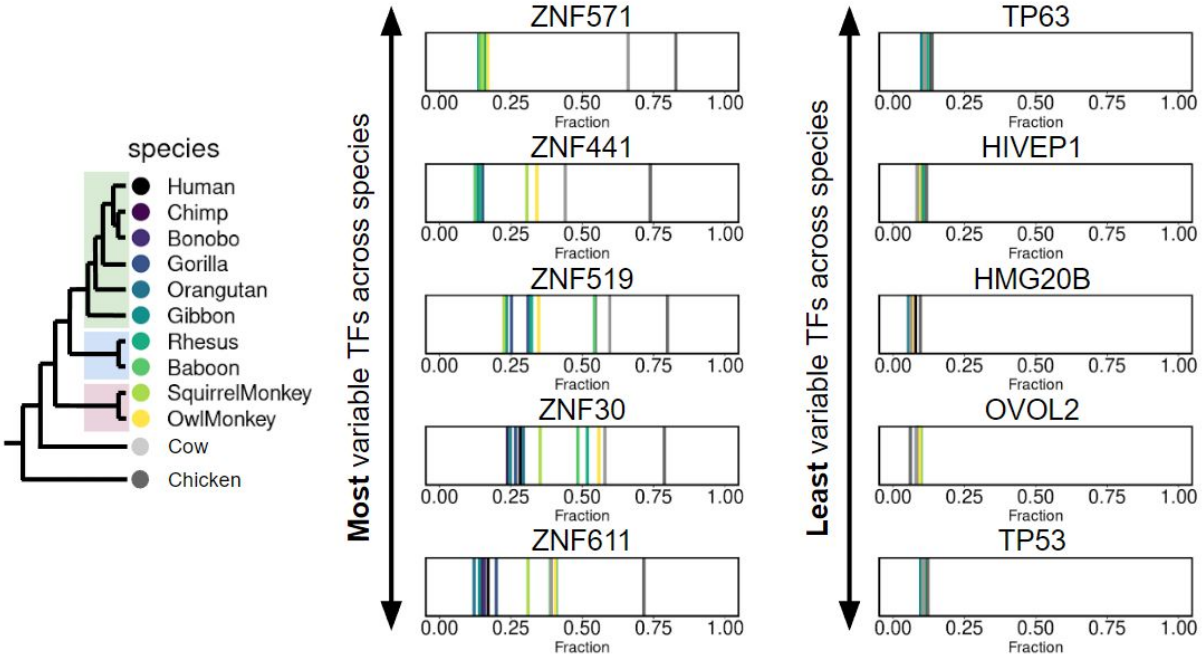


Figure 2.24: Dotplots of the fraction of reads per library that mapped to the Epstein-Barr Virus genome, with the percentage expressed as log2 in the horizontal axis and the primate species in the vertical axis. The DMSO-treated datasets are shown in purple, and the Nutlin-treated datasets are shown in green. The RNA-seq datasets are shown to the left, and the PRO-seq datasets are shown to the right. For rhesus, the fraction shown represents reads mapped to Papiine Herpesvirus 1.

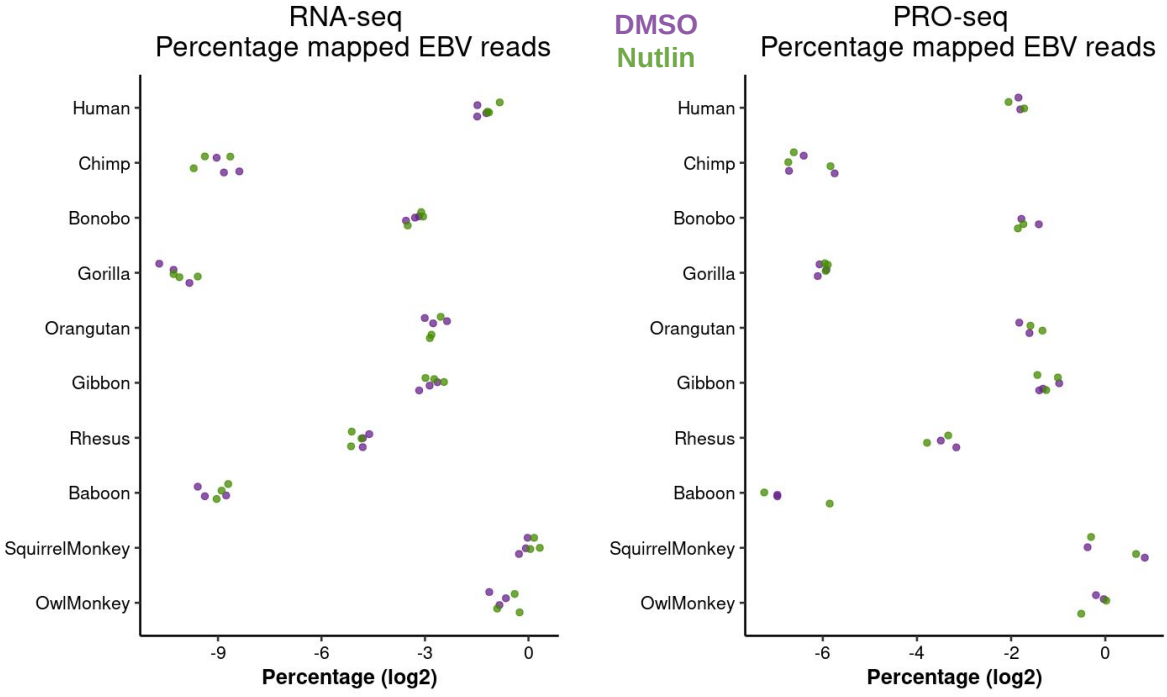


Figure 2.25: Dotplots showing the mRNA expression levels in transcripts per million (TPM) of 11 Epstein-Barr Virus latency genes across the RNA-seq datasets for 9 of the primates LCLs used in this study. The DMSO-treated datasets are shown in purple, and the Nutlin-treated datasets are shown in green. For rhesus, the values are not shown as that LCL was transformed with Papiine Herpesvirus 1.

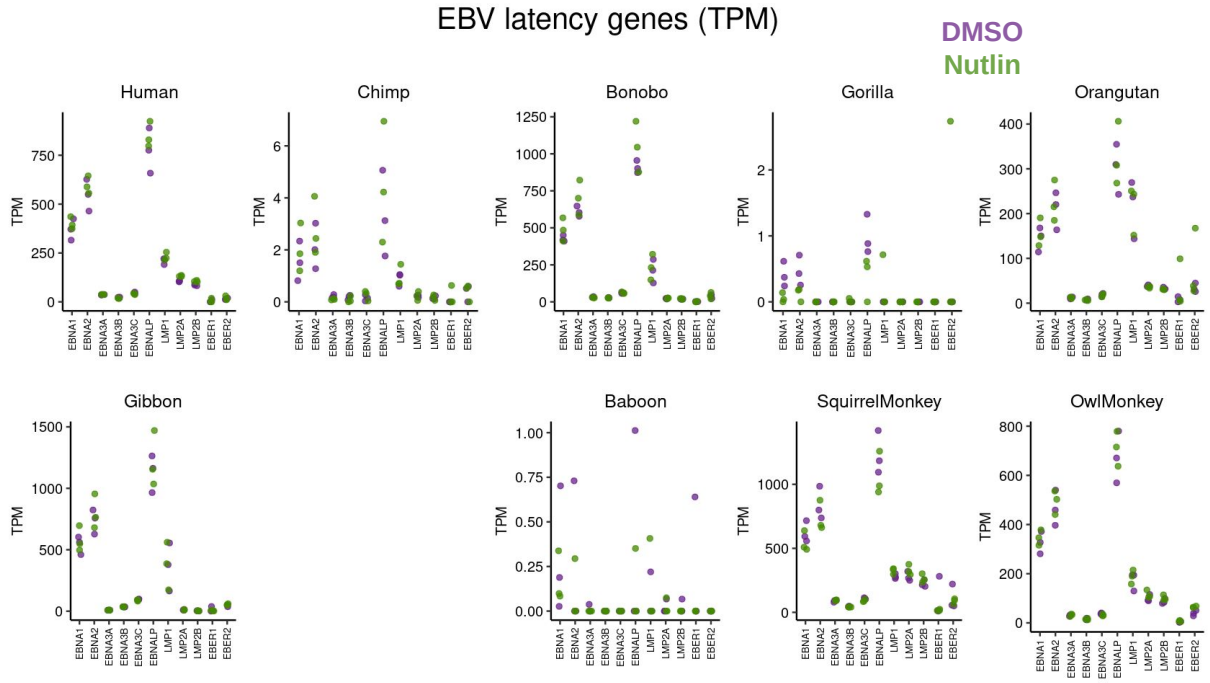
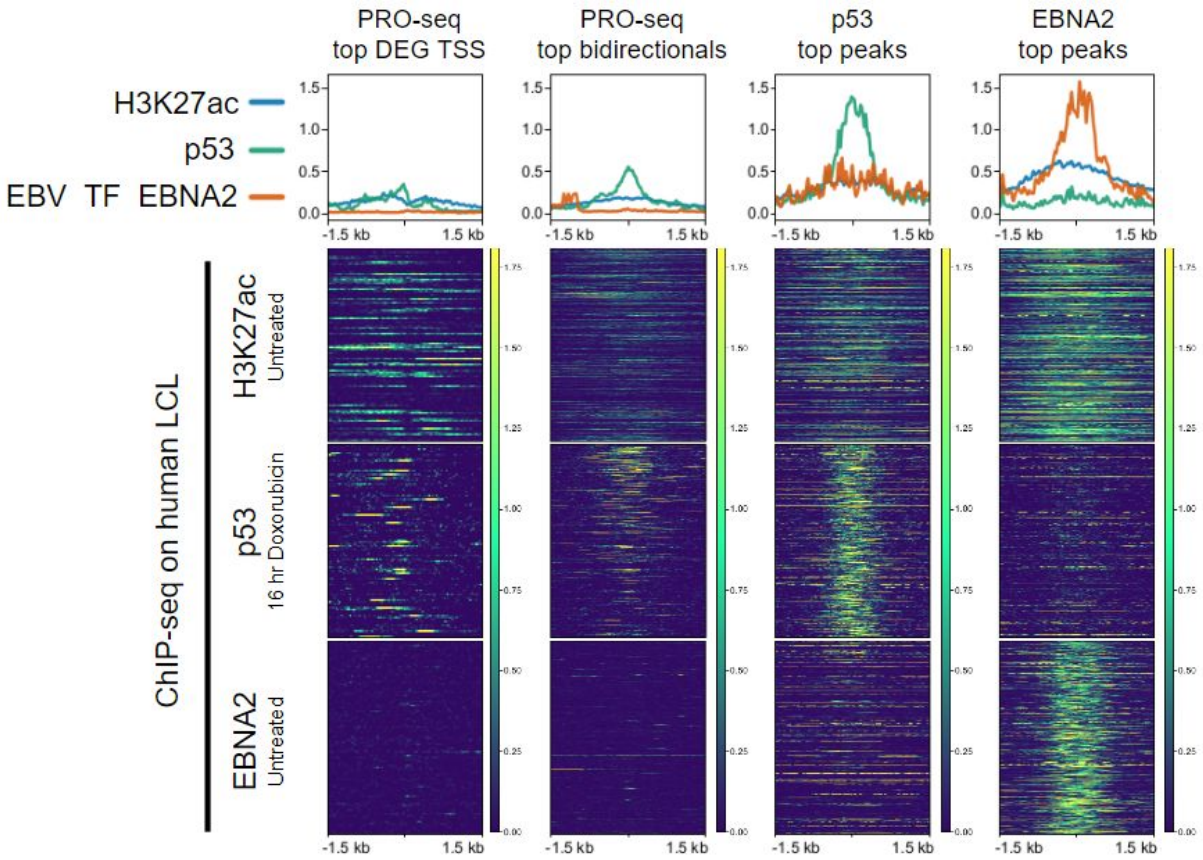


Figure 2.26: Aggregate signal of public ChIP-seq datasets with three types of features (rows) relative to four sets of loci (columns). The three ChIP-seq datasets are: ChIP-seq for H3K27ac from cell line GM12878 obtained using the SRA numbers SRR227633 and SRR227634. ChIP-seq for p53 from cell line GM12878 obtained using the SRA number SRR851807. ChIP-seq for EBNA2 from cell line GM12878 obtained using the SRA numbers SRR332245 and SRR332246. The four sets of loci are from left to right: the transcription start sites (+/- 1.5 kb) of the top 100 differentially expressed genes (DEGs) from the Nutlin-treated human PRO-seq dataset, the top 1000 bidirectional loci ranked by TFEA from the Nutlin-treated human PRO-seq dataset, the top 1000 peaks ranked in decreasing order by ChIP-seq p53 signal, and the top 1000 peaks ranked in decreasing order by ChIP-seq EBNA2 signal. Top, metaplots showing the aggregate ChIP-seq signals, with H3K27ac in blue, p53 in green, and EBNA2 in orange. Bottom, heatmaps of individual loci as each row showing their respective ChIP-seq signal per column.



2.26). I observed that there does not appear to be significant occupancy of the EBNA2 on neither p53-induced bidirectional loci as defined by TFEA, nor on p53-induced gene promoters.

I decided to bypass the need to identify all viral TF binding sites and interrogate the ultimate goal that EBV has, which is to force their host cells to continuously replicate. If the EBV is acting differently across the LCLs, then this activity may be observed by differences in their host cells' replication rate. Given that I lack experimental data for directly testing replication rates, I decided to approximate this measurement by focusing on the transcription levels of key genes controlling cell proliferation. I obtained this gene list from [207] and compared the expression distribution of these genes across the 10 primate LCLs, as well as against a set of controls consisting of cell populations known to be proliferating rapidly and slowly through the action of replication inhibitors [127] (Figure 2.27). The results indicate that relative to the controls, all 10 primate LCLs have very similar gene expression distributions.

Finally, as I observed that the primate LCLs responded to Nutlin with different magnitudes (Figure 2.6 and Figure 2.7), I wanted to see if the intrinsic expression of TP53 itself was different across the LCLs at the moment of the Nutlin treatment (Figure 2.28). Perhaps the reason the response was different is because there were significantly different p53 molecules in the cytoplasm ready to be used as transcriptional activators. The results show that indeed, based on TPM normalized mRNA expression values from the RNA-seq datasets, p53 was not being expressed similarly across the LCLs. I can quantify the response magnitude across the PRO-seq and RNA-seq by simply averaging the absolute values of all fold-changes, and there seems to be a correlation with the p53 steady-state mRNA levels and the overall Nutlin-responsiveness.

All things considered, here I report an exploration of the evolution of the p53-triggered transcriptional response across the anthropoid lineage, where I support the field by the release of PRO-seq and RNA-seq datasets upon the treatment with Nutlin-3a across LCLs derived from 12 animal species.

Figure 2.27: Violin distribution plots of 51 genes involved in cell cycle progression regulation across RNA-seq datasets of the 10 primate LCLs used in this study, and 7 MCF10A datasets from [Min2019]. Samples Control refers to MCF10A with sorting, p21High refers to MCF10A cells that were sorted for cells displaying elevated p21 protein levels, p21Low refers to cells that were sorted for cells displaying low p21 protein levels, ContactInhibition refers to cells that were let to grow to such an extent that they became over confluent, SerumStarvation refers to cells whose media was removed of growth serum, CDK46i refers to cells that were treated with an inhibitor for CDK4/6, Meki refers to cells that were treated with a Mek inhibitor. The mRNA transcript levels are expressed as log2 of transcripts per million (TPM) normalized values. Overlaid in red, blue, and green, are line plots of the key cell cycle regulators CDK1, CDK2, and p21, respectively.

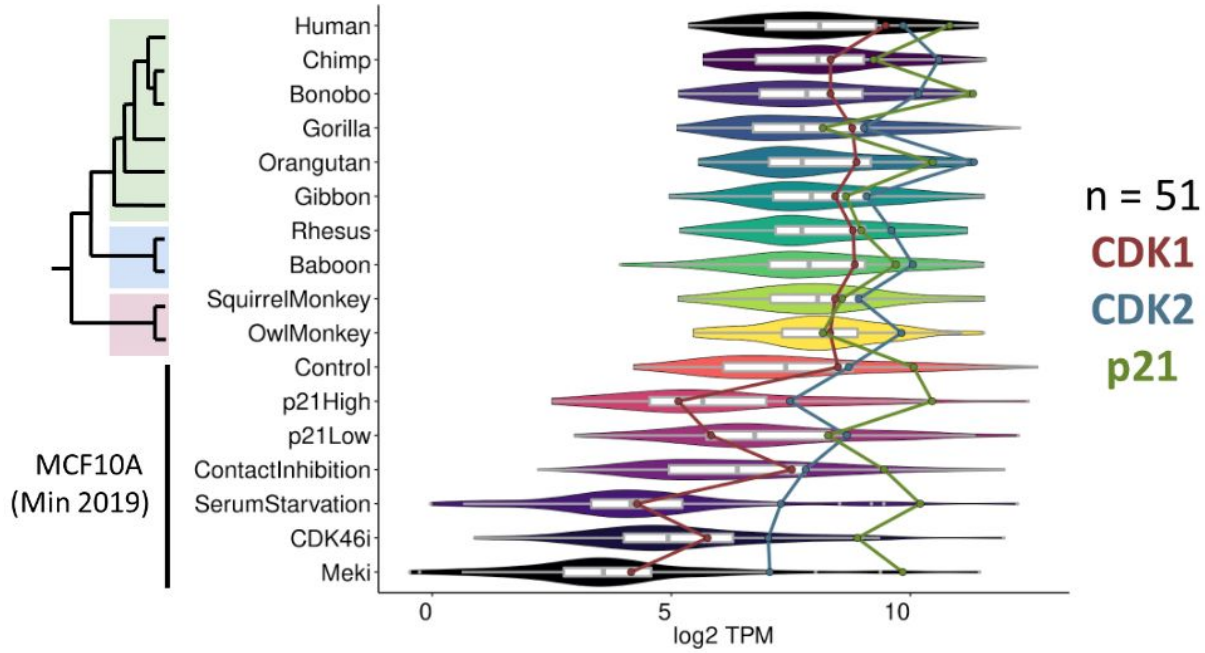
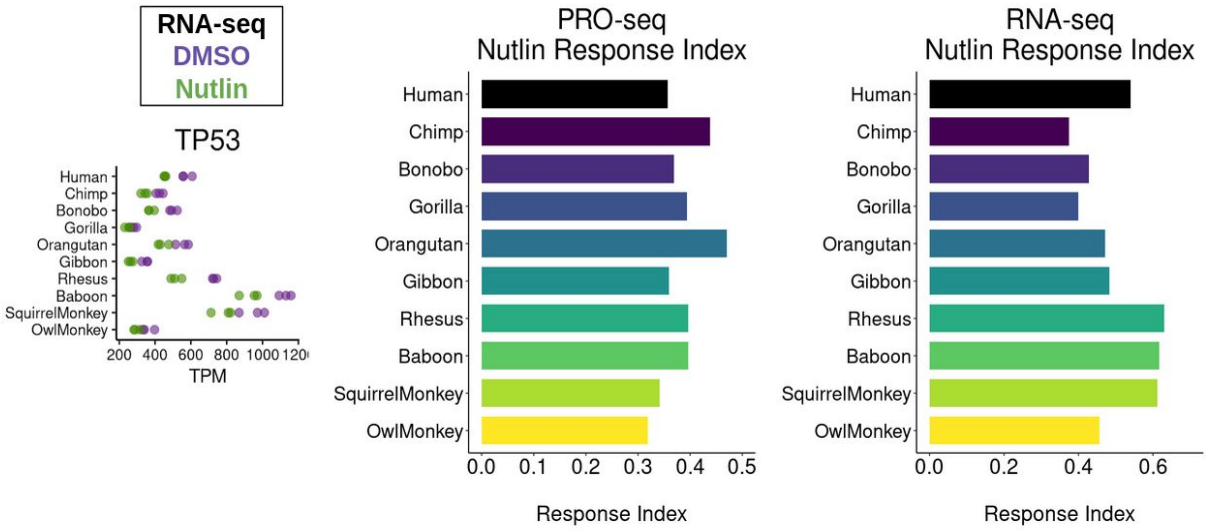


Figure 2.28: Left, transcript per million (TPM) normalized expression values of the TP53 gene across the RNA-seq datasets from the 10 primate LCLs used in this study. The DMSO-treated datasets are shown in purple, and the Nutlin-treated datasets are shown in green. Right and middle, barplots showing the response index (i.e. the average of the absolute log2 fold-change for all genes) for the PRO-seq (middle) and RNA-seq (right) Nutlin-treated primate LCLs datasets.



2.4 Discussion

In regards to the results shown on Figure 2.4, the chicken datasets showed a lower number of bidirectional calls. This could simply be explained by the fact that the chicken genome (galGal6) is 1.05 billion base pairs (bp) long, whereas the average length of the other 11 species is 2.78 billion bp. Also, I noticed that the first human LCL PRO-seq replicates did not have a huge difference in their number of bidirectionals to the other samples despite its higher sequencing depth, as seen in Figure 2.3.

On Figure 2.6, it is clear that the degree with which the species reacted to Nutlin was not the same, with LCLs from human and rhesus responding with a greater magnitude than that of chimp and gorilla, for example. Even though all LCLs were treated with the same Nutlin concentration, perhaps some LCLs were not able to respond with the same magnitude as they had different numbers of p53 molecules to control the strength of the ensuing transcriptional response. Interestingly, and in agreement with previous observations [3, 5, 181], I saw that the primary response (using PRO-seq at 1 hour as proxy) was composed of much fewer genes than the downstream response (using RNA-seq at 6 hours as proxy), with the latter being in the order of a few thousand genes.

The observations on Figure 2.7 suggest that some of the key players in the canonical p53-driven response are quite conserved across anthropoids, not surprisingly. I therefore expect that if there is rewiring in the p53 response gene network, that they may have occurred in less crucial aspects of the response. After all, all primates probably have had the same evolutionary pressures in the context of dealing with DNA damage and avoiding uncontrolled cell growth that negatively impacts passing on their genes to the next generation; but they may have experienced different evolutionary pressures in less crucial organismal roles that p53 is known to be involved in.

The PCA results on Figure 2.8 suggest that there are differences in the p53 transcriptional response, and that the differences follow the expected degree of evolutionary divergence time that the primates have experienced.

Figure 2.9 GSEA enrichment is much more pronounced for the RNA-seq/downstream datasets,

which is not surprising, as these gene sets are mostly derived from curated RNA-seq experiments. Though there are gene sets that appear to be enriched in specific species (e.g. coagulation in owl monkey), these results should be taken with caution, as the gene sets have been defined in human experiments, and are therefore not necessarily suitable to non-human species.

On the other hand, the TFEA enrichment displayed in Figure 2.10 and 2.11 showed that the orangutan PRO-seq plot looks very weird, and no obvious reason that I can think of explains its shape, even though both its PRO-seq and RNA-seq datasets show upregulation of canonical p53-responsive genes. The plots also show that, in agreement with the DESeq2 number of DEGs, the cow and chicken samples show no activation of p53, which supports my decision to drop them from the rest of the analysis.

The fraction of responsive genes by the number of species they were induced on, as shown in Figure 2.13, suggest that evolution has maintained a very streamlined primary response, with only a few differences in it that later in the timelapse of the p53 response become amplified and diversify the response that each primate deploys.

The classification of the primary and downstream genes as being TFs in Figure 2.15 may even suggest that having picked a later time point (e.g. 12 hours after p53 activation) may have yielded an even greater induced gene diversity across primates.

Though no definitive chromatin feature was ascertained to be the cause for the different gene induction conservation patterns, as seen in Figure 2.16; the results could be explained by the fact that genes whose induction by p53 is more conserved are simply much more readily prone for activation as the cells across primates keep them ready "to go" at the moment's notice.

The analysis on the GSEA when fixing the human DEGs as the gene set compared across the primate species (Figure 2.19 and 2.20) suggest that the ranking of the p53-responsive genes in humans is quite similar across hominoids, but this similarity decreases when testing the old and new world monkeys, which is in agreement with their evolutionary divergence. This pattern is more pronounced in the RNA-seq set, probably due to its greater gene set size.

By focusing on the promoter's activity across species, shown in Figure 2.21, suggest that as the

evolutionary distance increases between human and its ortholog pair, there is an overall difference in the transcription of orthologous loci, which in turns suggest that the regulatory elements themselves are more differentially used as a function of the evolutionary distance.

A big caveat for the promoter bias analysis in Figure 2.22 is that not all predicted motif instances are used by a given TF. There is need to have binding evidence for each TF such as ChIP-seq data, which ENCODE happens to have for many TFs on the human LCL used in this study. That said, ENCODE reached a similar conclusion using only ChIP-seq data, namely that most TFs bind predominantly to enhancers rather than promoters [142, 193]. And more specifically, when looking at specific TF cases (Figure 2.23), the results suggest that the p53 tetramer is not permitted to change its bias between promoters and enhancers because of adverse fitness effects on its host.

The examination of the differential viral activity on the transformed LCLs revealed that EBV may not be acting in an identical fashion among all the animal LCLs used in the study. In Figure 2.24, if the amount of mRNA transcripts that is collected from a random sampling of reads from a cell culture is a reliable approximation of the viral activity, then it is worrisome that the viral activity is not at least the same across the cells composing the study system. Though Figure 2.25 suggests that at least all primate LCLs may be equally compromised in that all EBV show similar latency programs. Furthermore, the EBV TF EBNA2 seems to not bind nearby the host p53 binding sites (Figure 2.26), which is a hopeful result, but it relies on the correct identification of all p53 regulatory binding sites. In addition, Figure 2.27 suggests that the primate LCLs are proliferating at similar rates. Finally, the results in Figure 2.28 show that p53 is transcribed at different levels across the primate LCLs. However, to truly measure p53 protein abundance, I should measure p53 levels directly, as it is known that mRNA levels do not correlate with protein levels [116].

2.5 Limitations

Here, I investigate the primary and downstream transcriptional response of LCL derived from 10 anthropoid species by analyzing PRO-seq and RNA-seq datasets treated with Nutlin for 1 hour and 6 hours, respectively. Though my results suggest unexpected evolutionary dynamics in the form of distinct conservation patterns between the primary and downstream gene responses, there are some potential variables that may have introduced confounding factors in my analysis.

The PRO-seq and RNA-seq datasets show unequal magnitudes in their transcriptional responses upon the Nutlin treatments (e.g. the human LCL shows a greater overall transcriptional response to Nutlin than the chimp LCL). These results make it hard to make interspecies comparisons of how p53-responsive genes are differentially regulated when the observed differences may be experimental artifacts instead.

The current datasets interrogated a single individual among the species tested, so any claims in species-specificity should be taken with caution. Had another individual been tested instead, the potential species-specific differences may not have been observed. A bigger sampling of individuals of both sexes and ages is needed to properly ascertain when differences have been fixed in a population and are therefore species-specific.

The usage of LCLs derived from EBV-infected quiescent B-cells is another source of caution for my evolutionary claims. EBV is known to persist in LCLs in a latent dormant state as circular DNA epiblasts or even integrated into the host genome [124, 87]. Upon infection, EBV hijacks the B-cells' proliferation machinery and kicks them into overdrive, forcing the otherwise quiescent cells to progress through their cell cycle and in turn making EBV proliferate, which is the very phenotype that researchers have exploited to use it to easily obtain LCLs from primates. Unfortunately, cell cycle progression is one of the main phenotypes that p53 controls upon its activation, and so it is not unreasonable to think that EBV compromises the natural effects of p53 activation. In addition, the rhesus macaque, cow, and chicken LCLs were not transformed with EBV; but with Papiine Herpesvirus 1, Bovine Leukemia Virus and Avian Leukosis Virus, respectively. This may

further complicate the interspecies comparisons as interspecies LCLs do not even harbor the same transformant viruses.

The above caveats may be the reason I observe a difference in the magnitude of the response even when samples were treated with the same Nutlin concentration. It should be carefully considered how to assess when a given feature, a gene or a regulatory element, is differentially transcribed across the samples in consideration. It may be that a given gene, for instance, is statistically non-induced in a given sample, but that this is due not to its induction absence in the p53-responsive network, but because the sample's response was not sufficiently big for detection. Therefore, I believe that there is a need to validate the results with the use of primary B-cells to bypass most of the above limitations.

The differentiation of primary versus downstream genes in the p53-transcriptional network is based mostly on the temporal induction of a gene. If a gene is not observed in the 1 hour PRO-seq datasets but it is induced in the 6 hour RNA-seq datasets, then it is assumed to be a downstream p53-responsive gene. A proper classification of primary and non-primary targets of gene regulation by a TF is the evidence that that TF was bound to regulatory elements of such a gene. p53 is known to preferentially bind to distal enhancers to the genes it regulates [3, 5, 181] instead to the proximal enhancers. Because I do not have suitable evidence to link bidirectional transcription loci (putative enhancers) to their target genes, I cannot unambiguously tell what if a gene is directly regulated by p53 or by a downstream TF. To overcome this, I propose the obtention of ChIP-seq datasets querying the binding of p53 to each species LCLs upon p53 activation, as well as obtaining a genome-wide chromatin conformation capture assay (e.g. HiChIP [134]) to link the putative enhancer regions to their target genes.

Finally, the decision to interrogate the primary and downstream transcriptional response with two different transcriptomic assay, PRO-seq and RNA-seq respectively, brings up the possibility that the differences in conservation patterns between the two time points may be confounded by intrinsic differences in the detection sensibilities of the techniques themselves. To overcome this potential scenario, I propose validating my results by obtaining either a 6 hour PRO-seq dataset

to match the current 6 hour RNA-seq dataset, or a 1 hour RNA-seq to match the current 1 hour PRO-seq datasets.

2.6 Methods

2.6.1 Cell lines information for the Nutlin interspecies datasets

Table 2.1 describes the information of the LCLs used to generate the Nutlin interspecies dataset; including the date they were received, the species, the ID, the biological sex, the age of the animal at the time of the cell transformation,and the source.

Table 2.1: Cell lines information for interspecies dataset used in chapter 2

Received	Species	ID	Sex	Age	Source
2019/01/29	Human	GM12878	F	NA	Coriell NIGMS
2019/10/04	Chimp	AG18358	F	22 years	Yoav Gilad Lab
2019/01/29	Bonobo	PR00748	F	11 years	Sara Sawyer Lab
2019/19/12	Gorilla	GG05	F	NA	Evan Eichler Lab
2019/01/29	Orangutan	PR00650	M	14 years	Sara Sawyer Lab
2020/01/15	Gibbon	Ricky Vok	F M	NA	Lucia Carbone Lab
2019/10/04	Rhesus	Mm 290-96	M	NA	Yoav Gilad Lab
2020/02/11	Baboon	AG17874	M	6 years	Coriell NIA
2019/12/12	Squirrel Monkey	SML clone 4D8	M	Adult	ATCC Ref. CRL-2311
2019/12/12	Owl Monkey	OML clone 13C	F	Adult	ATCC Ref. CRL-2312
2019/05/16	Cow	BL3.1	M	3 months	ATCC Ref. CRL-2306
2019/12/27	Chicken	DT40	F	1 day	ATCC Ref. CRL-2111

2.6.2 PRO-seq, ATAC-seq, and RNA-seq growth conditions

The human, chimp, bonobo, gorilla, orangutan, gibbon, rhesus, baboon, squirrel monkey, owl monkey, and cow LCLs were cultured in RPMI-1640 media (Gibco Ref. 72400-047), 15% FBS (R&D Systems Ref. S11150), and 100 U/mL Penicillin-Streptomycin (Gibco Ref. 15140-122). The chicken LCLs were cultured in RPMI-1640 media (Gibco Ref. 72400-047), 10% FBS (R&D Systems Ref. S11150), 5% Chicken serum (Sigma Ref. C5405-100ML), 10% Tryptose phosphate broth

(Sigma Ref. T8154), 0.05 mM β-mercaptoethanol (MP Ref. 194834), and 100 U/mL Penicillin-Streptomycin (Gibco Ref. 15140-122). All LCLs were cultured using vent-cap T-25 flasks (Corning 430639), and kept at a confluency between 400,000 cells/mL and 800,000 cells/mL during cell culture at 37°C with 5% CO_2, except for the chicken LCLs which were kept between 1,000,000 cells/mL and 2,500,000 cells/mL.

2.6.3 PRO-seq treatment conditions

Each of the 12 animal LCLs were treated for 1 hour with either 20 μM Nutlin-3a (Sigma Ref. SML0580), or with 0.001% DMSO as a negative control. 2 T-25 cultures per LCLs were used per treatment.

2.6.4 PRO-seq nuclei extraction

Nuclei isolation was done as described in [Core2008] with some modifications. Briefly, LCL cultures ranging from 5 to 36 million cells were used for each condition. After each culture was treated for 1 hour, the cultures were washed twice with ice-cold PBS. Then, the cell pellets were carefully resuspended in 6 mL of lysis buffer (0.1% DEPC-DI water with 10 mM Tris-HCl pH 7.4, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 1 mM DTT, Invitrogen Ref. AM2696 SUPERase-IN RNAse inhibitor, and with Roche Ref. 11836170001 protease inhibitor cocktail) and centrifuged for 15 minutes at 4°C at 1000 x g. The pellets were carefully resuspended in 1 mL lysis buffer using Finntip wide orifice pipette tips (Thermo Scientific Ref. 9405163), were mixed with 4 mL more of lysis buffer, and centrifuged a second time for 15 minutes at 4°C at 1000 x g. The pellets were carefully resuspended a second time in 1 mL lysis buffer using Finntip wide orifice pipette tips, transferred to low binding 1.7 mL eppendorf tubes (Costar Ref. 3207), and centrifuged for 5 minutes at 4°C at 1000 x g. The pellets were carefully resuspended using Finntip wide orifice pipette tips in 500 μL freezing buffer (0.1% DEPC-DI water with 50 mM Tris-HCl pH 8.0, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, and SUPERase-IN RNAse inhibitor), and centrifuged for 2 minutes at 4°C at 2000 x g. The resulting nuclei pellets were resuspended

a final time in 110 μL of freezing buffer using Finntip wide orifice pipette tips. I mixed 10 μL of the resuspended nuclei with 990 μL of PBS for counting the nuclei yield. The remaining 100 μL resuspended nuclei were snap-frozen in liquid nitrogen and stored at -70°C before being used for the PRO-seq nuclear-run on reactions.

Table 2.2 describes the dates when the nuclei were extracted for each of the PRO-seq Nutlin interspecies datasets.

Table 2.2: Nuclei isolation dates of the PRO-seq Nutlin interspecies dataset used in chapter 2

Nuclei extraction	Samples
2019/10/29	Human-1, Chimp-1, Bonobo-1, Rhesus-1
2019/11/26	Cow-1
2022/01/02	Gorilla-1, SquirrelMonkey-1
2020/01/30	Gibbon-1 (Ricky), OwlMonkey-1, Chicken-1
2020/02/17	Orangutan-1
2020/06/14	Human-2, Chimp-2, Bonobo-2, Gorilla-2, Orangutan-2, Gibbon-2 (Ricky)
2020/06/14	Rhesus-2, Baboon-2, SquirrelMonkey-2, OwlMonkey-2, Cow-2, Chicken-2
2020/09/03	Chimp-3, Gorilla-3, Gibbon-3 (Vok), Baboon-1

2.6.5 PRO-seq library preparation

PRO-seq datasets were prepared as described in [60], which in turn is a modified protocol from [122]. Briefly, between 3 to 15 million nuclei per dataset were used for the PRO-seq transcription run-on using a mixture of rNTP and Biotin-11-CTP (Biotin-11-CTP at 0.025 mM from PerkinElmer Ref. NEL542001EA; rCTP at 0.025 mM from Promega Ref. E604B, rATP at 0.125 mM Ref. E601B, rGTP at 0.125 mM Ref. E603B, and rUTP at 0.125 mM Ref. E6021). 1% of S2 Drosophila melanogaster nuclei relative to the number of the sample nuclei were added during the run-on reaction as a normalization spike-in. Total RNA was extracted using a phenol/chloroform precipitation. Isolated RNA was fragmented using base hydrolysis with NaOH. Biotinylated fragmented nascent transcripts were isolated using a first streptavidin Dynabeads M-280 (Invitrogen Ref. 11206D) pull down, and the VRA3 RNA adaptor was ligated at their 3' end. A second

streptavidin bead pull down was performed, followed by the enzymatic modifications of the RNA fragment 5' ends with a pyrophosphohydrolase and a polynucleotide kinase, and the VRA5 RNA adaptor was ligated at their fixed 5' ends. A third streptavidin bead pull down was performed, followed by the reverse transcription of the resulting adaptor-ligated libraries. The libraries were cleaned up with AMPure XP beads (Beckman Coulter Ref. A63881). Then, the libraries were amplified using 13 PCR cycles, and cleaned up again with another round of AMPure XP beads. The resulting library concentrations were measured with the Qubit dsDNA high sensitivity assay (Invitrogen Ref. Q32851), and their size distributions assessed using the Agilent High Sensitivity D1000 ScreenTape.

I prepared a third Nutlin and DMSO treated PRO-seq datasets for Chimp, Gorilla, and Gibbon, as one of the two first replicates did not get adequate Nutlin induction and were therefore not useful replicates.

Table 2.3 describes the dates when the PRO-seq Nutlin interspecies datasets were prepared.

Table 2.3: Library preparation dates of the PRO-seq Nutlin interspecies dataset used in chapter 2

Library prep date	Samples
2019/11/06	Human-1, Chimp-1, Bonobo-1, Rhesus-1
2019/11/28	Cow-1
2020/01/06	Gorilla-1, SquirrelMonkey-1
2020/02/21	Orangutan-1, Gibbon-1, OwlMonkey-1, Chicken-1
2020/06/15	Human-2, Chimp-2, Bonobo-2, Gorilla-2, Orangutan-2, Gibbon-2, Rhesus-2
2020/06/15	Baboon-2, SquirrelMonkey-2, OwlMonkey-2, Cow-2, Chicken-2
2020/09/04	Chimp-3, Gorilla-3, Gibbon-3, Baboon-1

2.6.6 PRO-seq sequencing information

Table 2.4 describes the dates when the PRO-seq Nutlin interspecies datasets were sequenced. Samples that appear twice in the sample indicate that the sample was further re-sequenced, and also concatenated. Samples sequenced on the consecutive days 2020/07/16 and 2020/07/22 were

concatenated into merged FASTQ files, were pooled once and ran in two sequencing lanes to obtain the desired number of reads without introducing potential batch effects. Boulder refers to the University of Colorado Boulder, Next Generation Sequencing Facilities. Reno refers to the University of Nevada Reno, Nevada Genomics Center.

Table 2.4: Library sequencing dates of the PRO-seq Nutlin interspecies dataset used in chapter 2

Sequencing date	Place	Samples
2019/11/21	Boulder	Human-1 (D/N), Chimp-1 (D/N), Bonobo-1 (D/N)
2019/11/21	Boulder	Rhesus-1 (D/N)
2019/12/17	Boulder	Cow-1 (D/N)
2020/03/09	Boulder	Gorilla-1 (D/N), Orangutan-1 (D/N), Gibbon-1 (D/N)
2020/03/09	Boulder	SquirrelMonkey-1 (D/N), OwlMonkey-1 (D/N), Chicken-1 (D/N)
2020/07/16 & 22	Boulder	Human-2 (D/N), Chimp-2 (D/N), Bonobo-2 (D/N)
2020/07/16 & 22	Boulder	Gorilla-2 (D/N), Orangutan-2 (D/N), Gibbon-2 (D/N)
2020/07/16 & 22	Boulder	Rhesus-2 (D/N), Baboon-2 (D/N), SquirrelMonkey-2 (D/N)
2020/07/16 & 22	Boulder	OwlMonkey-2 (D/N), Cow-2 (D/N), Chicken-2 (D/N)
2020/09/14	Boulder	Human-1 (D/N), Chimp-3 (D/N), Gorilla-3 (D/N)
2020/09/14	Boulder	Gibbon-3 (D/N), Baboon-1 (D/N)
2021/12/07	Reno	Human-2 (N), Bonobo-2 (D/N), Gorilla-2 (D/N), Gorilla-3 (N)
2021/12/07	Reno	Gibbon-2 (D/N), Rhesus-2 (N), Baboon-2 (N)
2021/12/07	Reno	SquirrelMonkey-2 (D/N), OwlMonkey-1 (N), OwlMonkey-2 (D)
2021/12/07	Reno	Cow-2 (D), Chicken-2 (N)

Datasets sequenced on a NextSeq 500 (Boulder) or on a NextSeq 2000 (Reno) as single-end 76 bp reads. Base calls and demultiplexing was done using Bcl2Fastq2 (v2.2.0). The letters "D" and "N" in parentheses denote "DMSO sample" and "Nutlin sample", respectively.

Table 2.5 describes the number of reads per PRO-seq library in the Nutlin interspecies dataset.

2.6.7 PRO-seq datasets processing

- PRO-seq datasets were processed using the Nextflow pipeline found in `https://github.com/Dowell-Lab/Nascent-Flow`.
- Read quality was assessed using FastQC (v0.11.5)

- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r, qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.

- Mapping was done using HISAT2 (v2.1.0) with options –very-sensitive –no-spliced-alignment on each species' respective reference genomes. The human reference genome hg38 was obtained from `https://hgdownload.soe.ucsc.edu/goldenPath` (referred to as GP in the rest of this document) `GP/hg38/bigZips/hg38.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY. The chimp reference genome panTro6 was obtained from `GP/panTro6/bigZips/panTro6.fa.gz`, and was modified so that only contained the main chromosome contigs (chr1, chr3, chr4, chr6, chr5, chr7, chrX, chr8, chr12, chr11, chr10, chr2B, chr9, chr2A, chr13, chr14, chr15, chr17, chr16, chr18, chr20, chr19, chr22, chr21, chrY, chrM). The bonobo reference genome panPan3 was obtained from `GP/panPan3/bigZips/panPan3.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr3, chr4, chr5, chr6, chr7, chrX, chr8, chr12, chr11, chr2B, chr10, chr9, chr2A, chr13, chr14, chr15, chr17, chr18, chr16, chr20, chr19, chr21, chr22, chrM). The gorilla reference genome gorGor4 was obtained from `GP/gorGor4/bigZips/gorGor4.fa.gz`, and was modified so that it only contained the main chromosome contigs (chr1, chr4, chr3, chr6, chr5, chr7, chrX, chr10, chr8, chr2B, chr11, chr12, chr9, chr2A, chr13, chr17, chr14, chr16, chr15, chr18, chr20, chr19, chr21, chr22, chrM). The orangutan reference genome ponAbe3 was obtained from `GP/ponAbe3/bigZips/ponAbe3.fa.gz`, and was modified so that it only contained the main chromosome contigs (chr1, chr3, chr4, chr5, chr6, chrX, chr7,

chr8, chr12, chr10, chr2B, chr11, chr9, chr2A, chr13, chr14, chr15, chr18, chr17, chr16, chr20, chr19, chr22, chr21, chrM). The gibbon reference genome nomLeu3 was obtained from GP/nomLeu3/bigZips/nomLeu3.fa.gz and was modified so that it only contained the main chromosome contigs chr1a, chr2, chr3, chr4, chr5, chr6, chr7b, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22a, chr23, chr24, chr25, chrX. The rhesus reference genome rheMac10 was obtained from GP/rheMac10/bigZips/rheMac10.fa.gz and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chrM, chrX, chrY. The baboon reference genome papAnu4 was obtained from GP/papAnu4/bigZips/papAnu4.fa.gz, and was modified so that it only contained the main chromosome contigs (chr1, chr2, chr5, chr3, chr6, chr4, chr7, chrX, chr8, chr11, chr12, chr9, chr14, chr15, chr13, chr17, chr10, chr16, chr18, chr20, chr19, chrM). The squirrel monkey reference genome saiBol1 was obtained from GP/saiBol1/bigZips/saiBol1.fa.gz and was modified so that it only contained the contigs numbered from JH378105 to JH378420, which were renamed as chr1 to chr316, respectively. The owl monkey reference genome Anan_2.0 was obtained from https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Aotus_nancymaae/all_assembly_versions/GCF_000952055.2_Anan_2.0 (referred to as AM in the rest of this document) AM/GCF_000952055.2_Anan_2.0_genomic.fna.gz. A file listing all the contigs was downloaded from AM/GCF_000952055.2_Anan_2.0_assembly_report.txt, was sorted by the contig size in descending order, and only kept the first 871 contigs, a threshold that was defined as the last contig with an annotated gene, and the contigs were renamed as chr1 to chr871. The annotated GTF file was obtained from AM/GCF_000952055.2_Anan_2.0_genomic.gtf. The mitochondrial genome was concatenated to the resulting genome file, and was obtained from AM/GCF_000952055.2_Anan_2.0_assembly_structure/non-nuclear/assembled_chromosomes/FASTA/chrMT.fna.gz. The cow reference genome bosTau9 was obtained from GP/bosTau9/bigZips/bosTau9.fa.gz and was modi-

fied so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr29, chrM, chrX. The chicken reference genome galGal6 was obtained from `GP/galGal6/bigZips/galGal6.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr30, chr31, chr32, chr33, chrM, chrW, chrZ.

- SAM to BAM conversion was done using Samtools (v1.8).

- Bigwigs were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1 –normalizeUsing RPKM –filterRNAstrand reverse (for pos file) or forward (for neg file) –scaleFactor 1 (for pos file) or -1 (for neg file).

- Bidirectional loci were determined using Tfit and dREG as described in the Nextflow pipeline https://github.com/Dowell-Lab/Bidirectional-Flow. It removes multimapped reads using Samtools (v1.8) view -h -q 1 'bam file' | grep -P '(NH:i:1| ∧@)' | samtools view -h -b. Tfit calls were obtained by first using the Tfit bidir module to call prelim regions. The annotation was used to add 3 kb-wide TSS regions to the prelim file and removed any part of the prelim regions that overlap with the TSS regions. Prelim regions $>$ 10 kb were then fragmented down to equal size regions ($<$ 10kb) with 50% overlap and then coverage filtered to keep prelim regions having $>$ 9 mapped reads. Finally, the adjusted prelim regions were used as regions of interest to the Tfit model module to obtain Tfit calls. dREG calls were filtered as having FDR $<$ 0.05, merged if within 20bp of each other, and having $>$ 9 mapped reads. Bidirectional transcription calls were combined using muMerge (v1.1.0) across experimental conditions.

- Read counts over genes were obtained using R (v3.6.0) Rsubread featureCounts (v1.32.4)

with the options isGTFAnnotationFile=FALSE, useMetaFeatures=TRUE, allowMultiOverlap=TRUE, largestOverlap=TRUE, isPairedEnd=FALSE, strandSpecific=1; using the multimapped reads filtered BAM files; and using a custom SAF file that contains the longest annotated entry per gene from the RefSeq annotation, without the initial 25% genic region starting from the 5' end to remove the RNA polymerase pausing region. The specific steps to produce this SAF file are as follows, with the human hg38 annotation as example: wget GP/hg38/bigZips/genes/hg38.ncbiRefSeq.gtf.gz convert2bed –input=gtf –output=bed –do-not-sort < hg38.ncbiRefSeq.gtf > hg38.ncbiRefSeq.bed grep -w transcript hg38.ncbiRefSeq.bed | grep -v chr[0-9]*_ | cut -f1,2,3,4,5,6 | awk -v OFS='\t' 'print $1, $2, $3, $4, $5, $6, $3-$2' > hg38.ncbiRefSeq.bed.tmp sort -nk7r hg38.ncbiRefSeq.bed.tmp | sort -u -k4,4 | awk -v OFS='\t' 'print $1, $2, $3, $4, $5, $6' | sort -k 1,1 -k2,2n > hg38.ncbiRefSeq.oneEntry.bed awk -v OFS='\t' 'print $4, $1, $2, $3, $6' hg38.ncbiRefSeq.oneEntry.bed > hg38.ncbiRefSeq.oneEntry.saf awk -v OFS='\t' ' if ($5 == "+") printf "%s\t%s\t%.0f\t%.0f\t%s\n", $1, $2, $3+(($4-$3)*0.25), $4, $5; else print $0 ' hg38.ncbiRefSeq.oneEntry.saf > hg38.ncbiRefSeq.without5prime25.oneEntry.saf.tmp awk -v OFS='\t' ' if ($5 == "-") printf "%s\t%s\t%.0f\t%.0f\t%s\n", $1, $2, $3, $4-(($4-$3)*0.25), $5; else print $0 ' hg38.ncbiRefSeq.without5prime25.oneEntry.saf.tmp > hg38.ncbiRefSeq.without5prime25.oneEntry.saf

- Differential gene expression was done using DESeq2 (v1.26.0). Gene set enrichment was done using the GSEA GUI (v4.3.2) with the Human MSigDB Collections (v7.5.1).

- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the multimapped reads filtered BAM files and the muMerged Tfit or dREG bidirectionals.

2.6.8 ATAC-seq treatment conditions

Only the human and bonobo LCLs were used for obtaining ATAC-seq libraries. They were treated for 1 hour with either 20 μM Nutlin-3a (Sigma Ref. SML0580), or with 0.001% DMSO as a negative control. The treatments were done in 12-well plate wells so that each well had 100,000 cells in 2 mL volume. Each condition was prepared in duplicates, and all samples were processed in parallel.

2.6.9 ATAC-seq library preparation

The ATAC-seq libraries were made following the [42] protocol. Briefly, after the 1 hour treatments, the 100,000 cells were transferred from their 12-well plate wells to 1.8 mL eppendorf tubes, and centrifuged at 500 x g for 7 minutes at 4°C. The supernatant was carefully removed and replaced with 50 μL of ice-cold lysis buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL, 0.1% Tween-20, 0.01% Digitonin), the cells resuspended 4 times pipetting up and down, and incubated on ice for 5 minutes. Then, added 1 mL of wash buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) and the tubes inverted 4 times to mix. The tubes were centrifuged at 500 x g for 10 minutes at 4°C and the supernatant was carefully removed without disturbing the small cell pellet. The pellets were then carefully resuspended by pipetting 6 times with 50 μL of the transposition mix (25 μL Tagment DNA Buffer Illumina Ref. 15027866, 2.5 μL Tagment DNA Enzyme 1 Illumina Ref. 15027865, 0.5 μL Digitonin diluted 1:1 with water, 0.5 μL 10% Tween-20, 5 μL water, 16.5 μL PBS), and were incubated for 30 minutes in a heat block at 37°C, flicking the tube often. Afterwards, the samples were cleaned using the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014) following the manufacturer's instructions, and eluted in 21 μL elution buffer. Then, a PCR pre-amplification was done using NEBNext Ultra II Q5 Master Mix (NEB Ref. M0544S) using 5 cycles. Then, a qPCR was done using NEBNext Ultra II Q5 Master Mix, SYBR Gold (Life Tech Ref. S11494), and 5 μL of the pre-amplified sample, and the results used to determine the additional number of extra PCR cycles

using Nextera DNA CD Indices (Illumina Ref. 20015882), which was just 1 additional cycle. The post-amplified ATAC-libraries were cleaned-up with the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014). The libraries were size-selected to remove DNA fragments greater than 1000 bp with a Sage Science BluePippin. The ATAC-seq libraries were quantified with Qubit HS DNA assay and their fragment size-distributions determined with Agilent HS D5000 ScreenTape. All samples were processed in parallel. After the samples were pooled and size-selected, they were cleaned-up using AMPure XP beads (Beckman Coulter Ref. A63881) at 1.5x volume and eluted into 20 μL of EB buffer (Qiagen Ref. 19086).

2.6.10 ATAC-seq sequencing information

The pooled 1st and 2nd replicates were sequenced on 2019/03/15 on a NextSeq 500 as paired-end 150 bp reads.

Table 2.6 describes the number of reads per ATAC-seq library in the Nutlin interspecies dataset.

2.6.11 ATAC-seq datasets processing

- Read quality was assessed using FastQC (v0.11.5).
- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tpe, tbo, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.
- Mapping was done using HISAT2 (v2.1.0) with options –new-summary –very-sensitive –no-spliced-alignment. The human reference genome hg38 was obtained from `GP/hg38/bigZips/hg38.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16,

chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY). The bonobo reference genome panPan3 was obtained from `GP/panPan3/bigZips/panPan3.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr3, chr4, chr5, chr6, chr7, chrX, chr8, chr12, chr11, chr2B, chr10, chr9, chr2A, chr13, chr14, chr15, chr17, chr18, chr16, chr20, chr19, chr21, chr22, chrM).

- Converted mapped SAM to BAM files using Samtools (v1.8) view -F 4 to remove unmapped reads.
- Read duplicates were removed using Sambamba (v0.6.6) markdup with options –remove-duplicates, –overflow-list-size=300000.
- Bedgraph files were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1, –normalizeUsing CPM.
- Peaks were determined using MACS2 (v2.1.1.20160309) callpeak with options –nolambda, –nomodel, –keep-dup all, –call-summits, and filtered out narrowPeaks with a score < 100.
- Peaks were merged across the species datasets using muMerge (v1.1.0) using options –save_sampids, –verbose.
- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the deduplicated BAM files and the muMerged MACS2 peaks.

2.6.12 RNA-seq treatment conditions

On the day of the treatments, each of the 12 animal LCLs were transferred to 24-well plate wells with 250,000 cells each in a total volume of 250 μL. Each LCLs was treated for 6 hour with either 20 μM Nutlin-3a (Sigma Ref. SML0580), or with 0.001% DMSO as a negative control. After the 6 hours, 2 mL of RNA lysis buffer was added to the 24-well plate wells for a total volume of 2,250 μL, and the plates were stored at -70°C until all 3 replicates were ready to be processed together.

The 1st replicates were treated on 2020/06/11, the 2nd replicates were treated on 2020/06/12, and the 3rd replicates were treated on 2020/06/13. All LCLs were processed in parallel.

2.6.13 RNA-seq library preparation

Total RNA was extracted using the Quick-RNA MiniPrep Plus (Zymo Research Ref. R1058) following the manufacturer's instructions, and the RNA concentrations were measured using a Qubit HS RNA kit, yielding concentrations ranging from 2 ng/μL to 18 ng/μL. The RNA-seq libraries were prepared using the KAPA mRNA HyperPrep Kit (Roche Ref. KK8581), KAPA mRNA Capture Kit (Roche Ref. KK8441), and KAPA Pure Beads (Roche Ref. KK8545); following the manufacturer's instructions (KR1352 – v7.21) using 250 ng of total RNA from most samples (though a few with low concentration had only 150-100 ng) as input with an RNA fragmentation step of 6 minutes at 94°C, and using 11 cycles in the amplification step for the samples that had 250 ng of input RNA or 12-14 cycles for those samples with less input RNA. The finalized libraries concentrations were obtained using the Qubit dsDNA high sensitivity assay kit (Invitrogen Ref. Q32851), with final concentrations ranging from 7 ng/μL to 56 ng/μL. The 1st replicates were processed in parallel on 2020/11/21, the 2nd replicates were processed in parallel on 2020/11/22, and the 3rd replicates were processed in parallel on 2022/11/23.

2.6.14 RNA-seq sequencing information

The 1st, 2nd, and 3rd replicates were pooled together and were sequenced on 2021/02/17 on a NovaSeq 6000 as paired-end 150 bp reads.

Table 2.7 describes the number of reads per RNA-seq library in the Nutlin interspecies dataset.

2.6.15 RNA-seq datasets processing

- Read quality was assessed using FastQC (v0.11.5).

- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r, qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tbo, tpe, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.

- Mapping was done using HISAT2 (v2.1.0) with options –very-sensitive on each species' respective reference genomes. The human reference genome hg38 was obtained from `GP/hg38/bigZips/hg38.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY. The chimp reference genome panTro6 was obtained from `GP/panTro6/bigZips/panTro6.fa.gz`, and was modified so that only contained the main chromosome contigs (chr1, chr3, chr4, chr6, chr5, chr7, chrX, chr8, chr12, chr11, chr10, chr2B, chr9, chr2A, chr13, chr14, chr15, chr17, chr16, chr18, chr20, chr19, chr22, chr21, chrY, chrM). The bonobo reference genome panPan3 was obtained from `GP/panPan3/bigZips/panPan3.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr3, chr4, chr5, chr6, chr7, chrX, chr8, chr12, chr11, chr2B, chr10, chr9, chr2A, chr13, chr14, chr15, chr17, chr18, chr16, chr20, chr19, chr21, chr22, chrM). The gorilla reference genome gorGor4 was obtained from `GP/gorGor4/bigZips/gorGor4.fa.gz`, and was modified so that it only contained the main chromosome contigs (chr1, chr4, chr3, chr6, chr5, chr7, chrX, chr10, chr8, chr2B, chr11, chr12, chr9, chr2A, chr13, chr17, chr14, chr16, chr15, chr18, chr20, chr19, chr21, chr22, chrM). The orangutan reference genome ponAbe3 was obtained from `GP/ponAbe3/bigZips/ponAbe3.fa.gz`, and was modified so that it only contained the main chromosome contigs (chr1, chr3, chr4, chr5, chr6, chrX, chr7, chr8, chr12, chr10, chr2B, chr11, chr9, chr2A, chr13, chr14, chr15, chr18, chr17, chr16, chr20, chr19, chr22, chr21, chrM). The gibbon reference genome nomLeu3 was obtained from `GP/nomLeu3/bigZips/nomLeu3.fa.gz` and

was modified so that it only contained the main chromosome contigs chr1a, chr2, chr3, chr4, chr5, chr6, chr7b, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22a, chr23, chr24, chr25, chrX. The rhesus reference genome rheMac10 was obtained from `GP/rheMac10/bigZips/rheMac10.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chrM, chrX, chrY. The baboon reference genome papAnu4 was obtained from `GP/papAnu4/bigZips/papAnu4.fa.gz`, and was modified so that it only contained the main chromosome contigs (chr1, chr2, chr5, chr3, chr6, chr4, chr7, chrX, chr8, chr11, chr12, chr9, chr14, chr15, chr13, chr17, chr10, chr16, chr18, chr20, chr19, chrM). The squirrel monkey reference genome saiBol1 was obtained from `GP/saiBol1/bigZips/saiBol1.fa.gz` and was modified so that it only contained the contigs numbered from JH378105 to JH378420, which were renamed as chr1 to chr316, respectively. The owl monkey reference genome Anan_2.0 was obtained from `AM/GCF_000952055.2_Anan_2.0_genomic.fna.gz`. A file listing all the contigs was downloaded from `AM/GCF_000952055.2_Anan_2.0_assembly_report.txt`, was sorted by the contig size in descending order, and only kept the first 871 contigs, a threshold that was defined as the last contig with an annotated gene, and the contigs were renamed as chr1 to chr871. The annotated GTF file was obtained from `AM/GCF_000952055.2_Anan_2.0_genomic.gtf`. The mitochondrial genome was concatenated to the resulting genome file, and was obtained from `AM/GCF_000952055.2_Anan_2.0_assembly_structure/non-nuclear/assembled_chromosomes/FASTA/chrMT.fna.gz`. The cow reference genome bosTau9 was obtained from `GP/bosTau9/bigZips/bosTau9.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr29, chrM, chrX. The chicken reference genome galGal6 was obtained from `GP/galGal6/bigZips/galGal6.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4,

chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr30, chr31, chr32, chr33, chrM, chrW, chrZ.

- SAM to BAM conversion was done using Samtools (v1.8).

- Bigwigs were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1, –normalizeUsing RPKM, –filterRNAstrand forward (for the positive strand file) or reverse (for the negative strand file), –scaleFactor 1 (for the positive strand file) or -1 (for the negative strand file).

- Read counts over genes were obtained using R (v3.6.0) Rsubread featureCounts (v1.32.4) with the options isGTFAnnotationFile=TRUE, useMetaFeatures=TRUE, GTF.featureType="exon", GTF.attrType="gene_id", allowMultiOverlap=TRUE, largestOverlap=TRUE, isPairedEnd=TRUE, strandSpecific=2; using each species GTF annotation file. The human hg38 GTF annotation file was obtained from `GP/hg38/bigZips/genes/hg38.ncbiRefSeq.gtf.gz`, the chimp panTro6 GTF annotation file was obtained from `GP/panTro6/bigZips/genes/panTro6.ncbiRefSeq.gtf.gz`, the bonobo panPan3 GTF annotation file was obtained from `GP/panPan3/bigZips/genes/ncbiRefSeq.gtf.gz`, the gorilla gorGor4 GTF annotation file was obtained from `GP/gorGor4/bigZips/genes/gorGor4.ensGene.gtf.gz`, the orangutan ponAbe3 GTF annotation file was obtained from `GP/ponAbe3/bigZips/genes/ponAbe3.ncbiRefSeq.gtf.gz`, the gibbon nomLeu3 GTF annotation file was obtained from `GP/nomLeu3/bigZips/genes/nomLeu3.ensGene.gtf.gz`, the rhesus rheMac10 GTF annotation file was obtained from `GP/rheMac10/bigZips/genes/rheMac10.ncbiRefSeq.gtf.gz`, the baboon papAnu4 GTF annotation file was obtained from `GP/papAnu4/bigZips/genes/papAnu4.ncbiRefSeq.gtf.gz`, the squirrel monkey saiBol1 GTF annotation file was obtained from `GP/saiBol1/bigZips/genes/saiBol1.ensGene.gtf.gz` and was modified so that the contig names reflect the chrN names just as they were assigned in the genome FASTA file, the owl monkey Anan_2.0 GTF annotation file was obtained from AM/GCF_000952055.2_Anan_2.0_genomic.gtf.gz

and was modified so that the contig names reflect the chrN names just as they were assigned in the genome FASTA file, the cow bosTau9 GTF annotation file was obtained from `GP/bosTau9/bigZips/genes/bosTau9.ncbiRefSeq.gtf.gz`, and the chicken galGal6 GTF annotation file was obtained from `GP/galGal6/bigZips/genes/galGal6.ncbiRefSeq.gtf.gz`.

- Differential gene expression was done using DESeq2 (v1.26.0). Gene set enrichment was done using the GSEA GUI (v4.3.2) with the Human MSigDB Collections (v7.5.1).

- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the BAM files, and a BED file containing the annotated gene TSSs that was obtained by further processing the above GTF files as follows, using the human hg38 annotation as an example: convert2bed –input=gtf –output=bed –do-not-sort < hg38.ncbiRefSeq.gtf > hg38.ncbiRefSeq.bed grep -w transcript hg38.ncbiRefSeq.bed | grep -v chr[0-9]*_ | cut -f1,2,3,4,5,6 | awk -v OFS='\t' 1print $1, $2, $3, $4, $5, $6, $3-$2' > hg38.ncbiRefSeq.bed.tmp sort -nk7r hg38.ncbiRefSeq.bed.tmp | sort -u -k4,4 | awk -v OFS='\t' 'print $1, $2, $3, $4, $5, $6' | sort -k 1,1 -k2,2n > hg38.ncbiRefSeq.oneEntry.bed awk -v OFS='\t' '{if ($6 == "+") print $1,$2-1500,$2+1500,$4; if ($6 == "-") print $1,$3-1500,$3+1500,$4}' hg38.ncbiRefSeq.oneEntry.bed >hg38.ncbiRefSeq.TSS.oneEntry.bed.tmp awk -v OFS='\t' '{if ($2 < 0) print $1,"0",$3,$4; else if ($2 > 0) print $0}' hg38.ncbiRefSeq.TSS.oneEntry.bed.tmp > hg38.ncbiRefSeq.TSS.oneEntry.bed

2.6.16 Defining a standard gene annotation for all 10 primates

Because not all of the public primate gene annotations are equally complete (i.e. the human hg38 reference genome has more details on where genes are located relative to less studied primates such as squirrel monkey), a standard gene annotation was made for all 10 primates such that they each have roughly the same number of genes. In order to make this standard annotation, a series of filters were done on the current available public annotations. Pairwise gene orthology tables between the human gene annotation (GRCh38.p13) and each of the other 9 primates gene annotations (Pan_tro 3.0/panTro5 for chimp, MPI-EVA panpan1.1/panPan2 for bonobo, gorGor4.1/gorGor4 for gorilla, PPYG2 for orangutan, Nleu_3.0/nomLeu3 for gibbon, Mmul_10/rheMac10 for rhesus,

Panu_3.0/papAnu4 for baboon, SaiBol1.0 for squirrel monkey, and Anan_2.0 for owl monkey) were obtained from Ensembl BioMart (http://uswest.ensembl.org/biomart/martview/), adding the "homology type" information corresponding to one-to-one, one-to-many, or many-to-many orthology classification. Using the human gene annotation as the baseline, non-human primates genes were removed from the standard annotation if they did not have a defined ortholog with any human genes. Only the longest gene entry per gene (e.g. from all transcript isoforms per gene) were kept, and the whole gene region was considered without considering intron/exon boundaries. Genes were also removed from the standard annotation if they had a defined orthology of one-to-many and many-to-many to remove potential missasignments of orthology. For example, gene X1 in human is part of a gene family with multiple paralogs (e.g. X1, X2, X3); and their orthology ascertainment is not clear with respect to the orthologous gene family in another primate, such that it is hard to assess if X1 corresponds to X1' or X2' or X3' in the other species. These stringent filters potentially remove genes responding to Nutlin-3a that may be of interest to study the p53-responsive gene network (e.g. the TP53 gene itself is removed due to it being paralog to TP63 and TP73), so some genes were manually added back into each of the standard gene annotations. To define which genes, if any, were needed to be added back, DESeq2 (v1.26.0) was used on each primate RNA-seq Nutlin-treated datasets using the same GTF files defined in the previous section "RNA-seq datasets processing" (hg38.ncbiRefSeq.gtf for human, panTro6.ncbiRefSeq.gtf for chimp, panPan3.ncbiRefSeq.gtf for bonobo, gorGor4.ensGene.gtf for gorilla, ponAbe3.ncbiRefSeq.gtf for orangutan, nomLeu3.ensGene.gtf for gibbon, rheMac10.ncbiRefSeq.gtf for rhesus, papAnu4.ncbiRefSeq.gtf for baboon, saiBol1.ensGene.gtf for squirrel monkey, and GCF_000952055.2_Anan_2.0_genomic.gtf for owl monkey). A union of the differentially expressed genes from all 10 primates was made to account for genes that may be regulated by p53 in a single species but not in any other species, and those genes were added back into each species standard annotation. If some genes could not be added back because their annotation did not exist (e.g. the MDM2 annotation is missing from the public owl monkey GTF annotation), then the genomic coverage was inspected visually to see gene itself existed in the expected genomic neighborhood

and the gene annotation manually added in. The resulting standard annotations per primate had a slightly different number of genes, shown in the table below.

Table 2.8 describes the number genes in both the full public gene annotation and in the primate standard gene annotation.

2.7 Data availability

The sequencing datasets described here were deposited to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database.

The PRO-seq, ATAC-seq, and RNA-seq datasets for the interspecies (human, chimp, bonobo, gorilla, orangutan, gibbon, rhesus, baboon, squirrel monkey, owl monkey, cow, and chicken) LCLs treated with Nutlin-3a were deposited under the GEO accession number GSE217051, with the PRO-seq datasets having the series GSE217034, the ATAC-seq datasets having the series GSE217032, and the RNA-seq datasets having the series GSE217047.

Table 2.5: Sequencing depth of the Nutlin PRO-seq interspecies datasets used in chapter 2

Dataset	Read number	Dataset	Read number
PRO-DMSO-Human-1	128,551,825	PRO-Nutlin-Human-1	136,598,413
PRO-DMSO-Human-2	53,221,267	PRO-Nutlin-Human-2	33,652,094
PRO-DMSO-Chimp-1	36,328,857	PRO-Nutlin-Chimp-1	40,076,454
PRO-DMSO-Chimp-2	30,812,332	PRO-Nutlin-Chimp-2	32,056,149
PRO-DMSO-Chimp-3	40,473,550	PRO-Nutlin-Chimp-3	43,125,129
PRO-DMSO-Bonobo-1	39,722,684	PRO-Nutlin-Bonobo-1	41,785,354
PRO-DMSO-Bonobo-2	39,977,560	PRO-Nutlin-Bonobo-2	41,063,040
PRO-DMSO-Gorilla-1	42,791,897	PRO-Nutlin-Gorilla-1	39,263,818
PRO-DMSO-Gorilla-2	48,613,736	PRO-Nutlin-Gorilla-2	53,464,158
PRO-DMSO-Gorilla-3	39,341,676	PRO-Nutlin-Gorilla-3	102,878,564
PRO-DMSO-Orangutan-1	39,953,141	PRO-Nutlin-Orangutan-1	41,573,013
PRO-DMSO-Orangutan-2	49,145,494	PRO-Nutlin-Orangutan-2	39,759,861
PRO-DMSO-Gibbon-1	42,440,285	PRO-Nutlin-Gibbon-1	38,533,349
PRO-DMSO-Gibbon-2	43,843,392	PRO-Nutlin-Gibbon-2	82,972,298
PRO-DMSO-Gibbon-3	36,929,840	PRO-Nutlin-Gibbon-3	41,247,219
PRO-DMSO-Rhesus-1	41,659,654	PRO-Nutlin-Rhesus-1	39,699,317
PRO-DMSO-Rhesus-2	39,662,853	PRO-Nutlin-Rhesus-2	55,227,424
PRO-DMSO-Baboon-1	40,987,089	PRO-Nutlin-Baboon-1	40,383,770
PRO-DMSO-Baboon-2	40,362,782	PRO-Nutlin-Baboon-2	41,263,485
PRO-DMSO-SquirrelMonkey-1	40,622,235	PRO-Nutlin-SquirrelMonkey-1	44,792,266
PRO-DMSO-SquirrelMonkey-2	53,565,337	PRO-Nutlin-SquirrelMonkey-2	76,728,427
PRO-DMSO-OwlMonkey-1	36,173,037	PRO-Nutlin-OwlMonkey-1	72,901,433
PRO-DMSO-OwlMonkey-2	79,019,096	PRO-Nutlin-OwlMonkey-2	45,043,011
PRO-DMSO-Cow-1	51,448,476	PRO-Nutlin-Cow-1	48,551,650
PRO-DMSO-Cow-2	68,896,589	PRO-Nutlin-Cow-2	43,544,365
PRO-DMSO-Chicken-1	39,895,173	PRO-Nutlin-Chicken-1	40,736,855
PRO-DMSO-Chicken-2	43,690,552	PRO-Nutlin-Chicken-2	91,729,704

Table 2.6: Sequencing depth of the Nutlin ATAC-seq interspecies datasets used in chapter 2

Dataset	Read number	Dataset	Read number
ATAC-DMSO-Human-1	17,181,112	ATAC-Nutlin-Human-1	23,374,036
ATAC-DMSO-Human-2	15,614,866	ATAC-Nutlin-Human-2	20,091,690
ATAC-DMSO-Bonobo-1	11,500,310	ATAC-Nutlin-Bonobo-1	11,834,364
ATAC-DMSO-Bonobo-2	15,476,046	ATAC-Nutlin-Bonobo-2	22,325,385

Table 2.7: Sequencing depth of the Nutlin RNA-seq interspecies datasets used in chapter 2

Dataset	Read number	Dataset	Read number
RNA-DMSO-Human-1	27,172,763	RNA-Nutlin-Human-1	31,879,772
RNA-DMSO-Human-2	30,865,729	RNA-Nutlin-Human-2	26,486,304
RNA-DMSO-Human-3	23,451,032	RNA-Nutlin-Human-3	26,825,322
RNA-DMSO-Chimp-1	35,779,563	RNA-Nutlin-Chimp-1	36,153,728
RNA-DMSO-Chimp-2	42,870,297	RNA-Nutlin-Chimp-2	30,171,049
RNA-DMSO-Chimp-3	31,406,200	RNA-Nutlin-Chimp-3	34,246,769
RNA-DMSO-Bonobo-1	33,334,490	RNA-Nutlin-Bonobo-1	35,863,758
RNA-DMSO-Bonobo-2	29,507,586	RNA-Nutlin-Bonobo-2	29,555,100
RNA-DMSO-Bonobo-3	28,983,577	RNA-Nutlin-Bonobo-3	29,478,273
RNA-DMSO-Gorilla-1	24,696,062	RNA-Nutlin-Gorilla-1	28,239,251
RNA-DMSO-Gorilla-2	32,811,403	RNA-Nutlin-Gorilla-2	32,740,210
RNA-DMSO-Gorilla-3	27,158,118	RNA-Nutlin-Gorilla-3	28,979,991
RNA-DMSO-Orangutan-1	28,355,108	RNA-Nutlin-Orangutan-1	33,762,440
RNA-DMSO-Orangutan-2	30,970,138	RNA-Nutlin-Orangutan-2	28,737,830
RNA-DMSO-Orangutan-3	28,431,272	RNA-Nutlin-Orangutan-3	32,112,614
RNA-DMSO-Gibbon-1	29,734,744	RNA-Nutlin-Gibbon-1	29,222,784
RNA-DMSO-Gibbon-2	35,767,335	RNA-Nutlin-Gibbon-2	32,404,840
RNA-DMSO-Gibbon-3	26,942,341	RNA-Nutlin-Gibbon-3	30,538,286
RNA-DMSO-Rhesus-1	35,271,847	RNA-Nutlin-Rhesus-1	34,121,235
RNA-DMSO-Rhesus-2	30,252,600	RNA-Nutlin-Rhesus-2	47,115,388
RNA-DMSO-Rhesus-3	26,649,442	RNA-Nutlin-Rhesus-3	39,834,628
RNA-DMSO-Baboon-1	35,473,612	RNA-Nutlin-Baboon-1	43,587,046
RNA-DMSO-Baboon-2	31,851,633	RNA-Nutlin-Baboon-2	31,679,505
RNA-DMSO-Baboon-3	31,213,233	RNA-Nutlin-Baboon-3	26,105,418
RNA-DMSO-SquirrelMonkey-1	27,555,746	RNA-Nutlin-SquirrelMonkey-1	28,834,275
RNA-DMSO-SquirrelMonkey-2	30,275,157	RNA-Nutlin-SquirrelMonkey-2	29,849,683
RNA-DMSO-SquirrelMonkey-3	27,993,242	RNA-Nutlin-SquirrelMonkey-3	28,740,278
RNA-DMSO-OwlMonkey-1	28,561,055	RNA-Nutlin-OwlMonkey-1	33,724,364
RNA-DMSO-OwlMonkey-2	31,680,120	RNA-Nutlin-OwlMonkey-2	27,982,205
RNA-DMSO-OwlMonkey-3	26,491,154	RNA-Nutlin-OwlMonkey-3	31,696,573
RNA-DMSO-Cow-1	30,062,880	RNA-Nutlin-Cow-1	29,857,796
RNA-DMSO-Cow-2	30,502,323	RNA-Nutlin-Cow-2	32,375,715
RNA-DMSO-Cow-3	31,237,097	RNA-Nutlin-Cow-3	36,811,566
RNA-DMSO-Chicken-1	27,890,812	RNA-Nutlin-Chicken-1	21,928,188
RNA-DMSO-Chicken-2	29,351,988	RNA-Nutlin-Chicken-2	28,046,475
RNA-DMSO-Chicken-3	27,980,639	RNA-Nutlin-Chicken-3	32,824,773

Table 2.8: Number of genes in the full public and in the standard annotation

Species	Assembly	Genes in full annotation	Genes in standard annotation
Human	hg38	38,258	8,079
Chimp	panTro6	33,927	8,081
Bonobo	panPan3	31,166	8,078
Gorilla	gorGor4	30,025	8,080
Orangutan	ponAbe3	27,037	8,081
Gibbon	nomLeu3	27,386	8,074
Rhesus	rheMac10	33,673	8,080
Baboon	papAnu4	30,764	8,080
Squirrel monkey	saiBol1	27,390	8,077
Owl monkey	Anan_2.0	31,324	8,048

Chapter 3

The evolution of the type I interferon transcriptional response

A portion of this chapter was published as:

Ramirez, D., Chuong, E.B., Dowel, R.D. Nascent transcription upon interferon-α2 stimulation on human and rhesus macaque lymphoblastoid cell lines. In revision at *BMC Research Notes*.

3.1 Introduction

Probably ever since the primeval life forms emerged on Earth, hosts and their pathogens have been engaged in a perpetual battle where one needs to defeat the other for survival. These pathogens vary in size, from relatively big multicellular parasites, protozoa, fungi, bacteria, archaea, viruses, down to even the tiniest known infectious entities known as viroids, which are as simple as single-stranded circular RNA molecules [55, 1]. Hosts have evolved a complex system comprising the range from entire organs and cell types in the case of eukaryotes, to relatively simple protein-based defenses in the case of unicellular life forms, whose job is to defend their host against the myriad of pathogenic threats they face on a daily basis. The collection of these systems is referred to as the immune system [20, 135].

The immune system is roughly divided into two broad branches. The innate immune system is the first line of defense, and it is constant throughout the lifetime of its host. It comprises physical barriers such as the skin, mucosa surrounding the inner pipings that are in contact with

elements coming from the outside world; all the way to specialized cells whose job is to gobble up and destroy intracellular intruders, and intricate molecular devices that sense pathogenic signals outside and inside of cells and in turn deploy soluble proteins that can tear apart the trespassers [155]. In contrast, the adaptive immune system is a second layer of defense that learns from its enemies and tailors a specific response to clear up the pathogens that the innate immune system failed to take care of. It takes time to adapt, and it changes throughout the life of its host. This adaptive system comprises a convoluted network of cell types that circulate through blood and lymph vessels, sample their surroundings for signatures of pathogens, and amplify defenses in the form of immunoglobulins with precise specificity that either directly degrade pathogens or tag them for elimination by other immune cells [63].

One crucial element that regulates the immune response is the family of soluble cytokines called interferons (IFN). They are divided into three types (Type I, Type II, and Type III) depending on the transmembrane receptors that they bind to. Here I focus on Type I IFN, which are deployed when cells detect signs of viral pathogens through surveillance proteins such as Toll-like receptors or cytosolic nucleic-acid sensors. Upon their synthesis, these IFN proteins are released from infected cells and are recognized by neighboring cells by membrane receptors. The receptors then trigger a signaling response that culminates with the assembly of specialized protein complexes formed by members of the STAT and IRF families of transcription factors (TFs), namely the complex ISGF3, that in turn orchestrate the gene transcription of a myriad of interferon-stimulated genes (ISGs) to fight the incoming pathogens [126]. The expression of ISGs comes in waves, with a set of primary ISGs appearing as soon as a fraction of an hour after IFN stimulation, to further downstream ISGs coming into play many hours afterwards [131].

The type I IFN gene family has undergone extensive gene duplication in many parts of the metazoan branch of life. In humans, for instance, the family is composed of 17 genes: 13 paralogs of IFN-α, and single copies of IFN-ϵ, IFN-κ, IFN-ω, and IFN-β; the latter known to induce a very robust immune response [75]. It has remained a debated topic in the field the physiological importance of having multiple type I IFN proteins all binding to the same IFNAR1 and IFNAR2

heteromeric membrane receptor. It has been proposed that different IFN cytokines bind to the receptors with different binding affinities, which in turn may lead to somewhat different cellular responses in the form of distinct ISGs being expressed [147, 167].

In addition to the diversity observed in the members of the IFN gene families, the set of ISGs triggered upon sensing pathogens has also been subject to continuous change throughout evolutionary time [168]. This comes at no surprise, as the immune response is subjected to great evolutionary pressure to diversify in face of the biological threats that species are subjected to in their own ecological niches.

These changes in gene expression occur, in part, due to changes in the transcription regulatory elements, namely promoters and enhancers, that dictate when and what genes are transcribed. Indeed, changes in the signatures of regulatory elements across species, even closely related ones, has been a common finding ever since high-throughput genomic assays have been widely used to do comparative genomic analysis [194]. The mechanism underlying this rewiring of the transcription regulatory elements has not been fully understood, though instances have been described of selfish genetic elements introducing new DNA regulatory sequences to new loci when they replicate, or on the intrinsic sequence malleability of enhancers that can erode or arise by random mutations [28, 114, 73].

Indeed, there is extensive genetic variation within a single species, even in animals that have undergone significant ecological bottlenecks as is the case of humans after their migration out of Africa to all corners of Earth [81, 40]. Compelling evidence has been recently brought up that describes how the immune system across ethnic human populations have been uniquely shaped by the historical pathogens that our ancestors have encountered [71, 115, 101].

3.2 Experimental system

To study the differences in the IFN-responsive transcriptional regulation, I made four distinct types of experimental set-ups.

1) I obtained ATAC-seq datasets, in duplicates, of human and bonobo LCLs treated for 1

hour with human IFN-α2, or without IFN. These were made first as I wanted to test the degree of chromatin accessibility differences between two closely related primates. I also only had those two primate LCLs at the time.

2) I obtained PRO-seq datasets of human and rhesus macaque LCLs (one female and one male individuals per species) on three experimental conditions: treating both primates LCLs with human IFN-α2, treating both primates LCLs with rhesus IFN-α2, or treating both primates LCLs with the carrier BSA as a negative control. These datasets are meant to test if there is a difference between using each species with their cognate IFN-α2 or if the observed transcriptional response is independent of the species-source of IFN-α2.

3) I set out to investigate the degree of diversity in a set of 6 metazoan species in the transcriptional response triggered type I IFN molecules. My chosen animal species represent a significant evolutionary timescale: Chicken, cow, squirrel monkey, rhesus macaque, gibbon, and human; with the most recent common ancestor estimated to have lived around 300 million years ago [137].

4) Finally, I also set out to investigate the degree of diversity in a set of human ethnic populations represented by 8 individuals. My sampling is intentionally broad and includes representatives of the Yoruba people from Nigeria, the Mende people from Sierra Leone, the Luhya people from Kenya, Tamil from Sri Lanka, Han from China, indigenous people from Peru, Caucasian from the United States, and a mixed individual from Mexico. Lymphoblastoid cell lines (LCLs) were obtained from each of the 6 species and 8 humans to generate the transcriptomic datasets.

For the 3rd and 4th experimental set-ups, IFN-β was chosen among the other members of the type I IFN family because it elicits a strong response relative to the other proteins, as well as because the chosen species each have a single copy of IFN-β which renders comparisons unambiguous in that the proteins tested are truly orthologous. I stimulated each of the 6 animal LCLs with their cognate IFN-β proteins as an attempt to overcome the potential confounding factor that may arise from having the IFNAR1 and IFNAR2 heteromeric membrane receptor from one species binding to a soluble extracellular IFN-β from another species with potential mismatches in the optimal

binding affinity that each species has evolved to retain.

To obtain a clearer picture of the transcriptional response to IFN-β, for the 3rd and 4th experimental set-ups, I obtained datasets at 1 hour and at 3 hours after the stimulations to approximate the immediate primary and downstream responses. For the 1 hour time point I obtained PRO-seq datasets in duplicates, such that I could study the transcribed regulatory elements controlling ISGs transcription. For the 3 hour time point I obtained RNA-seq datasets in triplicates, so that I could study matured and processed ISGs.

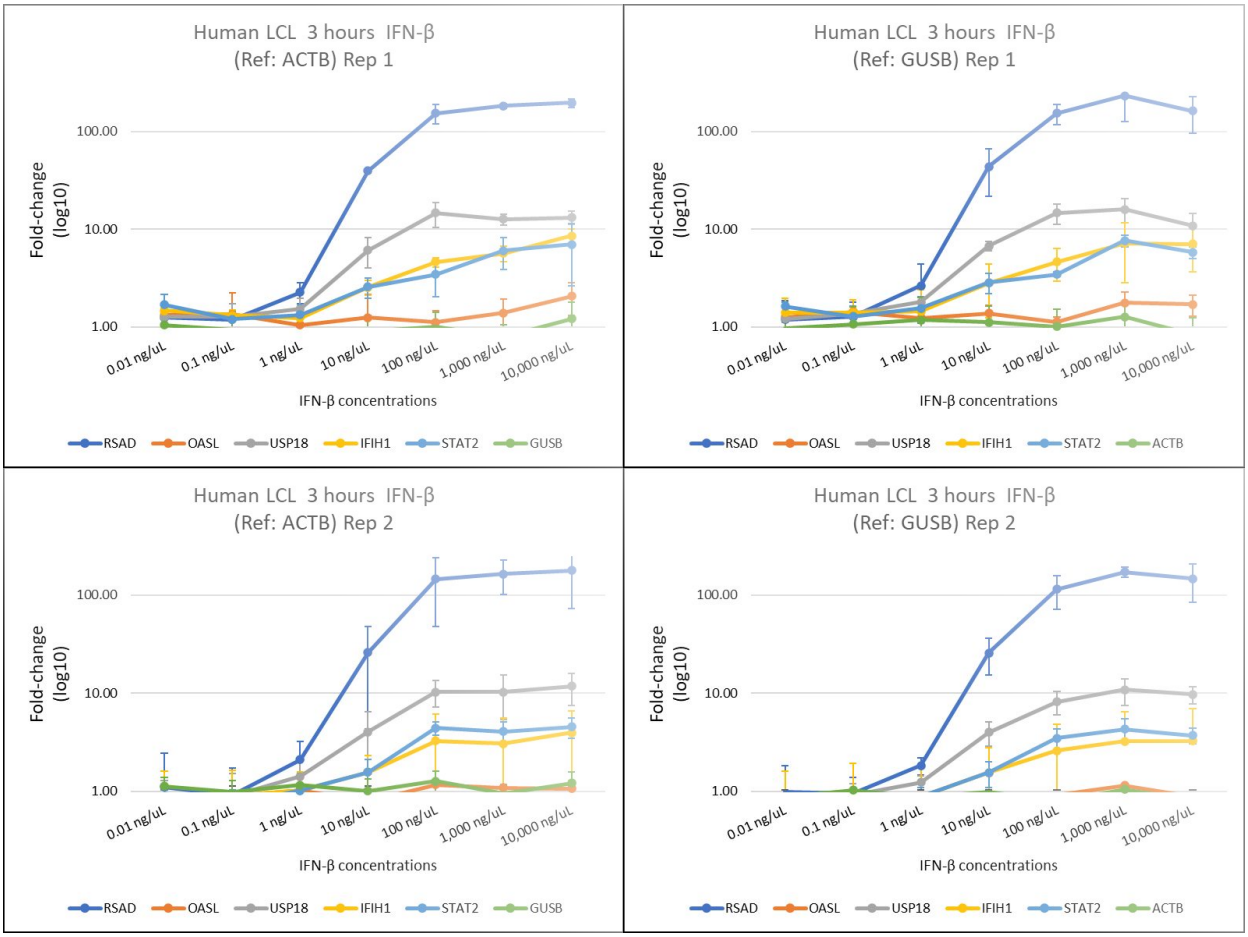
3.3 Results

3.3.1 Cross-species transcriptional response to interferon

To obtain the interspecies comparisons on the IFN-β transcriptional responses between human, gibbon, rhesus monkey, squirrel monkey, cow, and chicken; I set out to estimate the appropriate concentrations of the IFN-β protein purifications purchased from Kingfisher Biotech Inc. that would yield comparable induction magnitudes of ISGs. To this end, I decided to test the induction of a few ISGs in all 6 species with different IFN-β concentrations using quantitative reverse transcription polymerase chain reaction (RT-qPCR).

I designed primers for five well known ISGs in human that I posited their induction with IFN-β would be conserved: RSAD, OASL, USP18, IFIH1, and STAT2. In addition, I designed primers to two housekeeping genes, ACTB and GUSB, to test if the results were consistent independent of the chosen reference point. Before proceeding on doing the large-scale experiment with all six species LCLs, I did a pilot with only the human LCL using IFN-β concentrations spanning 0.01 ng/µL to 10,000 ng/µL in ten-fold increments (Figure 3.1). I observed that the fold-change of ISGs did not change in a significant way by using either of the two housekeeping genes. I also observed that the concentrations used covered the whole dynamic range of the ISGs fold changes, with the lowest concentration of 0.01 ng/µL not showing any ISG induction, whereas the highest concentration of 10,000 ng/µL showing a fold change that had already plateaued at the maximum

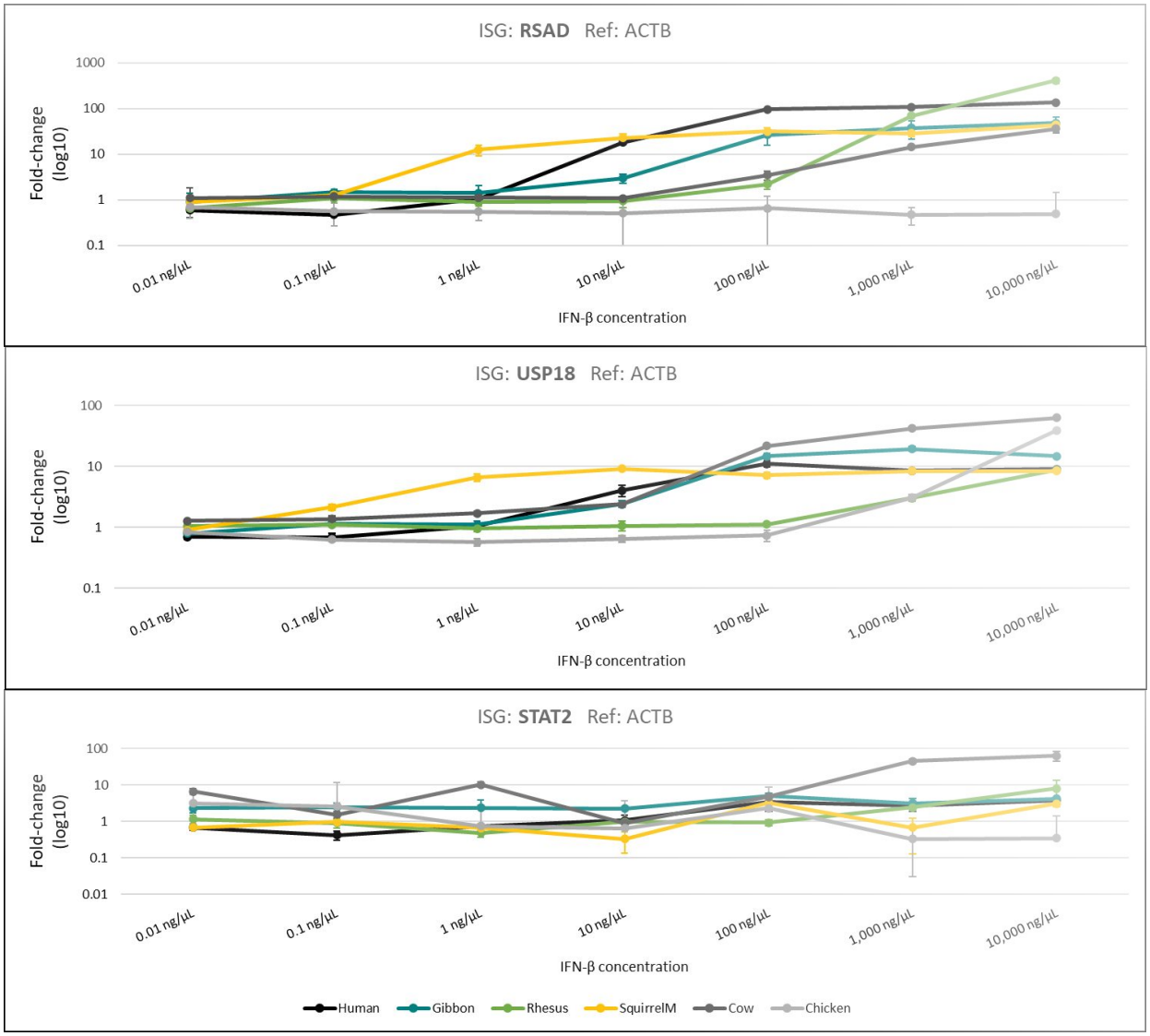
Figure 3.1: Reverse transcription quantitative polymerase chain reaction (RT-qPCR) results showing the fold change of five tested ISGs in the human LCLs using ten-fold increase in IFN-β concentrations. Two different housekeeping genes were used as reference, ACTB (left) and GUSB (right). A first replicate is shown on top, and a second replicate shown on the bottom.



of the sensitivity of the instrument or of the cellular response. The results further showed that not all 6 ISGs showed IFN-β induction, such as OASL which showed no induction even at the highest IFN-β concentration. The induction curves were very well replicated in an independent replicate. To minimize subsequent work, these results helped us narrow down to only using 3 ISGs when testing the 6 species: RSAD, USP18, and STAT2; as they displayed the greatest fold-changes. Moving forward, I used ACTB as the reference housekeeping gene to calculate the fold-changes of the ISGs.

The next RT-qPCR experiment then was done on the 6 species LCLs using those 3 ISGs that showed the greatest fold-change in the human LCL, using each species cognate IFN-β protein (Figure 3.2). The fold-change curves showed that STAT2 showed aberrant fold-changes across most samples, so STAT2 was dropped from the analysis. RSAD and USP18, however, showed informative fold-change curves. Interestingly, RSAD showed no upregulation with any chicken IFN-β concentration in the chicken LCL (light gray color), which suggests that either the primers did not work for that sample or that RSAD is not inducible in the chicken LCL. The squirrel monkey LCL (yellow) showed upregulation even a low concentration when the other LCLs did not yet show induction, which suggested that the squirrel monkey IFN-β purification bioactivity is much greater than the other purifications, or that the squirrel monkey LCL are much more readily induced by IFN-β. On the contrary, the rhesus LCL showed a delayed induction for both ISGs, not showing a plateau even at the biggest IFN-β concentrations. Based only on USP18, the chicken LCL also showed that it needs a greater IFN-β concentration to achieve a reasonable ISG induction. The gibbon and cow LCLs show relatively similar ISGs induction dynamics to the human LCL. All in all, based on these results, I decided to use the following IFN-β concentrations for each species: The human LCLs would be treated with human IFN-β at 100 ng/mL. The gibbon LCLs would be treated with gibbon IFN-β at 100 ng/mL. The rhesus LCLs would be treated with rhesus IFN-β at 500 ng/mL. The squirrel monkey LCLs would be treated with squirrel monkey IFN-β at 5 ng/mL The cow LCLs would be treated with cow IFN-β at 200 ng/m. And the chicken LCLs would be treated with chicken IFN-β at 500 ng/mL. Both the rhesus and chicken concentrations should be

Figure 3.2: Reverse transcription quantitative polymerase chain reaction (RT-qPCR) results showing the fold-change of 3 tested ISGs in the human, gibbon, rhesus, squirrel monkey, cow, and chicken LCLs using ten-fold increase in IFN-β concentrations. ACTB was used as reference housekeeping gene.



higher given the RT-qPCR results, but it would become unrealistic to get the amount of IFN-β needed, possibly in the dozens of mg per treatment.

I proceeded to obtain the IFN-β-treated 1 hour PRO-seq and 3-hour RNA-seq interspecies datasets, with two and three replicates, respectively. The PRO-seq datasets have an average of 38 million single-end reads, whereas the RNA-seq datasets have an average of 29 million paired-end reads (Figure 3.3). The first PRO-seq replicates were purposely sequenced at a lower depth than the 2nd replicates, because of budget constraints. All PRO-seq and RNA-seq had good sequencing quality control metrics.

Doing differential expression analysis on the PRO-seq and RNA-seq datasets with DESeq2, I observed that the most differentially expressed genes (DEGs) correspond to known ISGs such as RSAD2, MX1, and IFIT1; which show upregulation even just after 1 hour of IFN-β expression as shown in the PRO-seq datasets (Figure 3.4). Most of the DEGs are positively regulated, with a few genes that show downregulation upon IFN-β. Importantly, I observed that not all six species had a response with the same magnitude. The human LCL showed the greatest response, in terms of both the number of DEGs and in their fold-change (Figure 3.5). The cow and the chicken LCLs showed less response, with chicken showing an almost imperceptible induction. The gibbon, rhesus, and squirrel monkey LCLs displayed an intermediate responsiveness to IFN-β. These observations suggest that the RT-qPCR calibration was not performed successfully.

With the additional complication of having datasets that seem to be induced with IFN-β with different magnitude responses across the species LCLs, it becomes difficult to assess when observed changes in DEGs are due to rewiring in the IFN-β responsive network due to evolutionary forces or due to experimental error.

To overcome the differences in the response magnitude across the species LCL datasets, I set out to compare the ranking of the ISGs per species. I examined if the same group of genes are upregulated in the same order in all tested species regardless of their fold change levels. In other words, if the top 10 DEGs in one species are also the top 10 DEGs in the other species. To do this, I relied on the ranking metric of the Gene Set Enrichment Analysis (GSEA), which tests a given

Figure 3.3: Total number of short sequencing reads for the interspecies PRO-seq datasets (top) and the RNA-seq datasets (bottom). In light gray are the untreated samples, and in dark gray are the IFN-β treated samples.

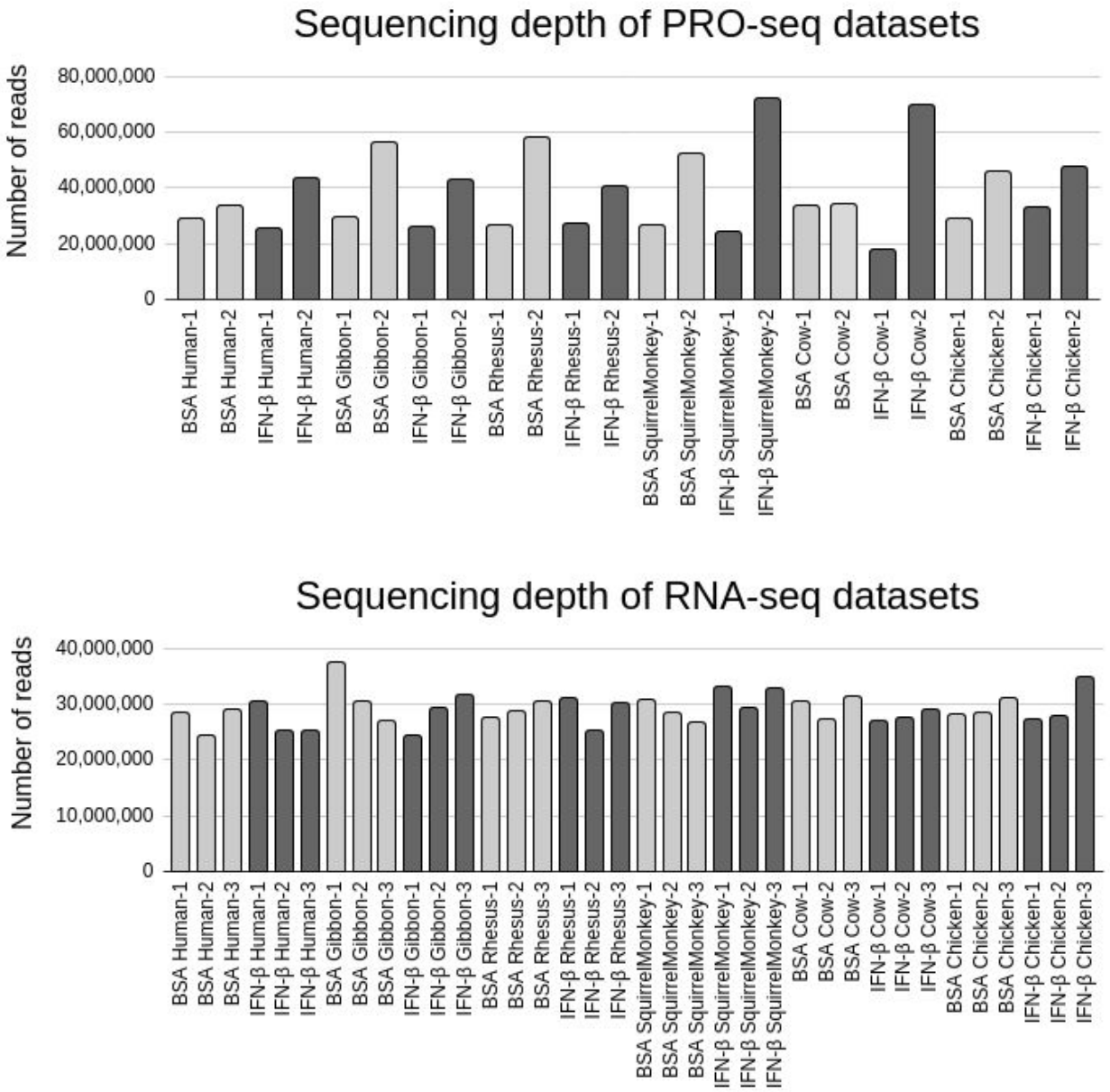


Figure 3.4: Volcano plots showing the log2 fold change in the horizontal axis and the -log10 adjusted p-value for all genes for cells treated with IFN-β. The top two rows show the PRO-seq datasets from the 6 species, and the bottom two rows show the RNA-seq datasets. A few of the top differentially expressed genes in humans are labeled in all samples.

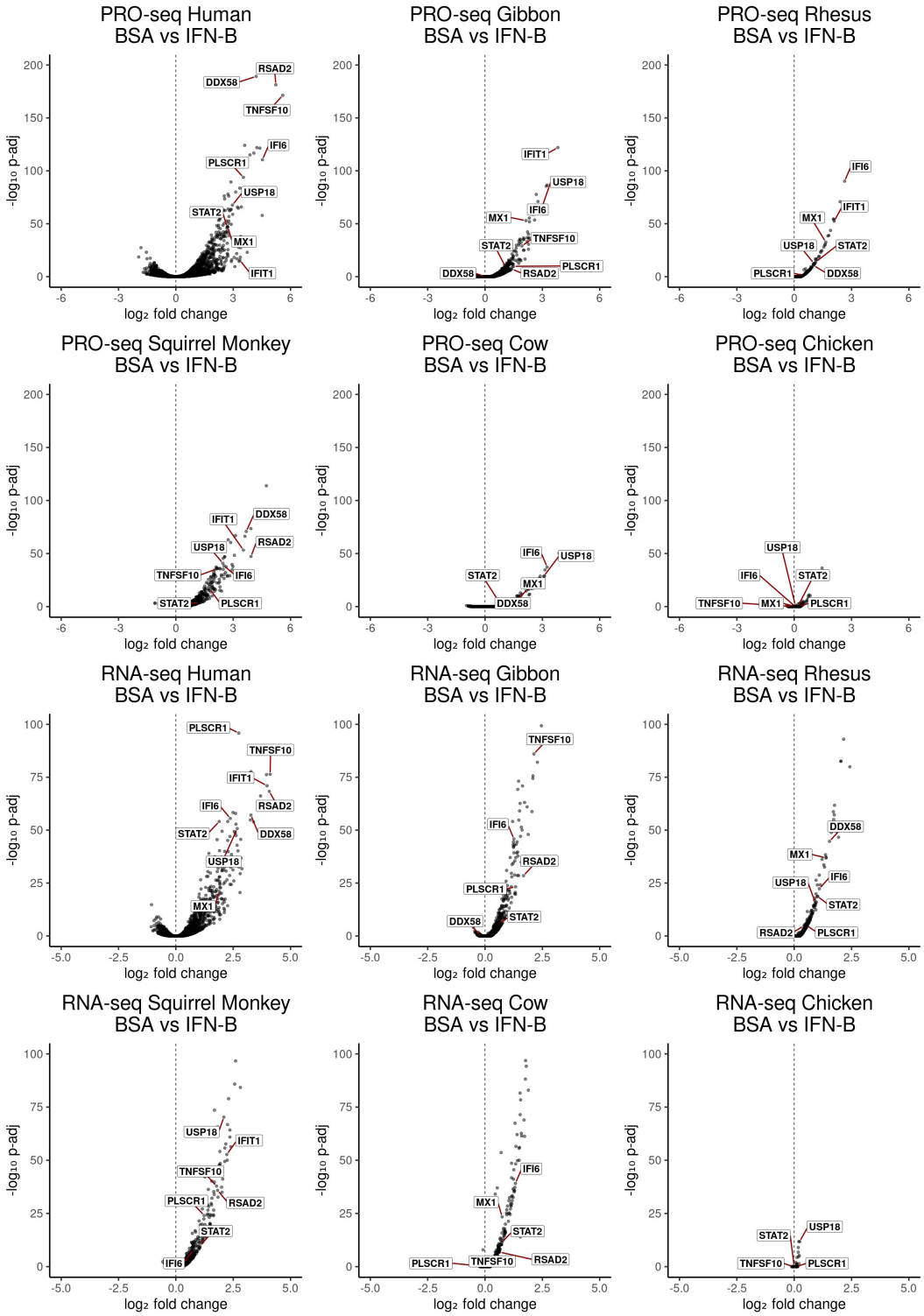
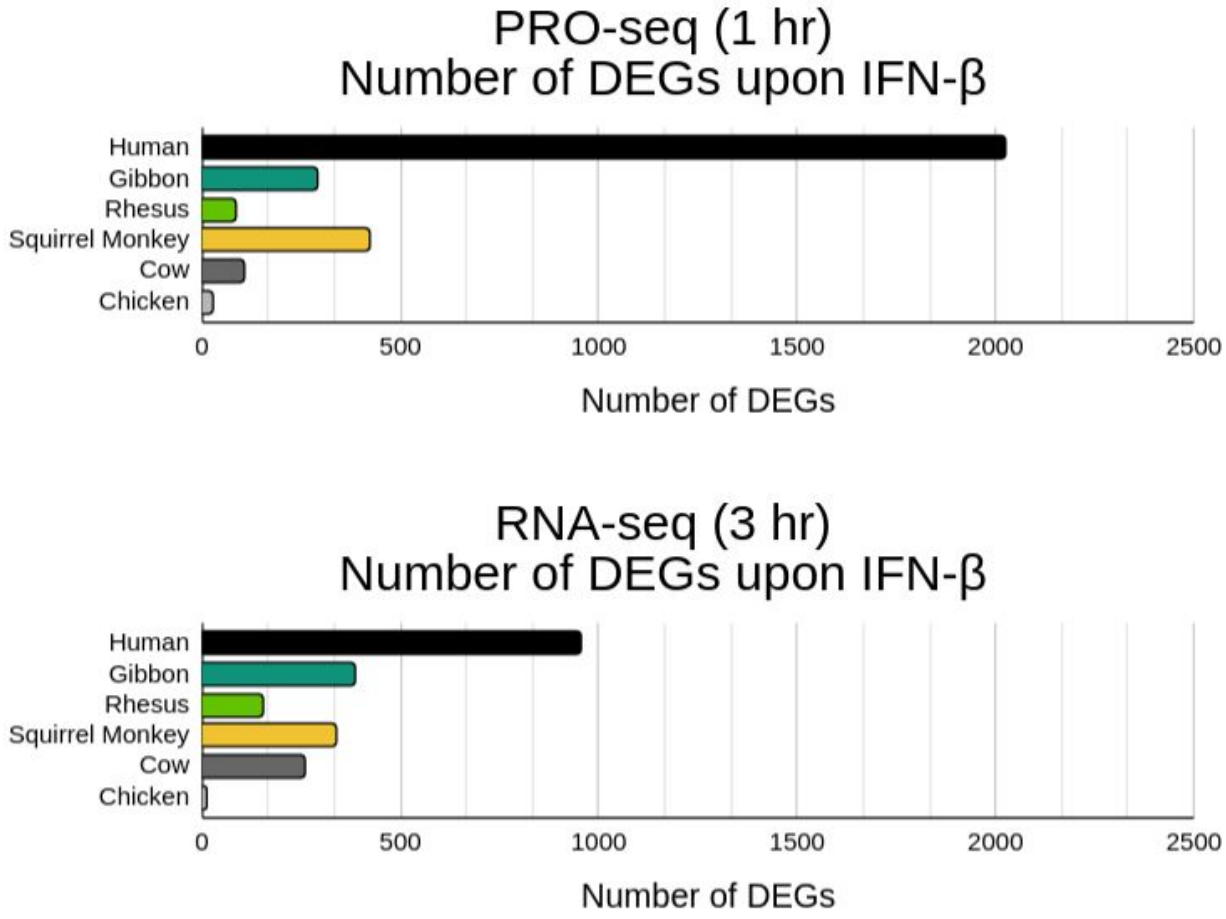


Figure 3.5: Total number of differentially expressed genes (DEGs) in all six species LCLs treated with IFN-β in the PRO-seq (top) and in the RNA-seq (bottom) datasets. DEGs were defined using DESeq2 with an alpha value of 0.05.



set of ranked genes and how they are enriched for pre-existing gene sets, such as those regularly used from the curated Molecular Signature Database.

The GSEA results (Figure 3.6) highlight that among the top significantly enriched gene sets for both the PRO-seq and RNA-seq datasets across the 6 species LCLs treated with IFN-β are the IFN response itself, together with other immune-related sets such as JAK STAT Signaling. However, many other seemingly unrelated gene sets appeared highly enriched as well, such as Angiogenesis, Estrogen Response, and Hedgehog Signaling. These results should be interpreted with caution, as the reference gene sets themselves have been curated from both the literature and genetic screens that are highly biased towards the human response and not necessarily suitable to be tested on non-human animals. Nonetheless, these results indicate that the overall DEG signature is reminiscent of that of the IFN response with potentially interesting differences that warrant further examination.

Finally, besides testing the overall genic-centered response across the 6 species LCLs treated with IFN-β, I performed a preliminary assessment of the potential differential activity of TFs. To this end, I implemented the Transcription Factor Enrichment Analysis (TFEA) on the PRO-seq and RNA-seq datasets. TFEA performs an enrichment statistic on the co-localization of a set of regions of interest with a set of known TF motifs. Once the regions of interests are ranked, such as by the level of differential transcription signal between the untreated and IFN-β-treated conditions, one expects that those motifs belonging to the TFs that are causing the changes in transcription will be spatially co-localized with the top ranked regions of interest. In other words, if I are focusing on bidirectionals induced by IFN-β, then the topmost induced bidirectionals should be enriched with STAT and IRF motifs in their proximity, as ISGF3 (the heteromeric TF complex activated by IFN-β and composed by STAT1, STAT2, and IRF9) will have caused those bidirectionals to be induced in the first place. For the PRO-seq datasets, my regions of interest were the bidirectional transcription loci as detected by Tfit v1.2 (Figure 3.7), with similar number of bidirectionals in five out of the six species at around 30,000 per dataset, except for chicken which had around 20,000. For the RNA-seq the regions were the transcription start sites of all genes.

Figure 3.6: Gene Set Enrichment Analysis results using the Hallmark gene sets from the Molecular Signature Database on gene lists ranked by DESeq2 of IFN-β treated cells. The 10 gene sets with the lowest adjusted p-values are shown transformed as their negative log10 values. The first two rows show the results from the PRO-seq datasets, and the bottom two rows show the results from the RNA-seq datasets

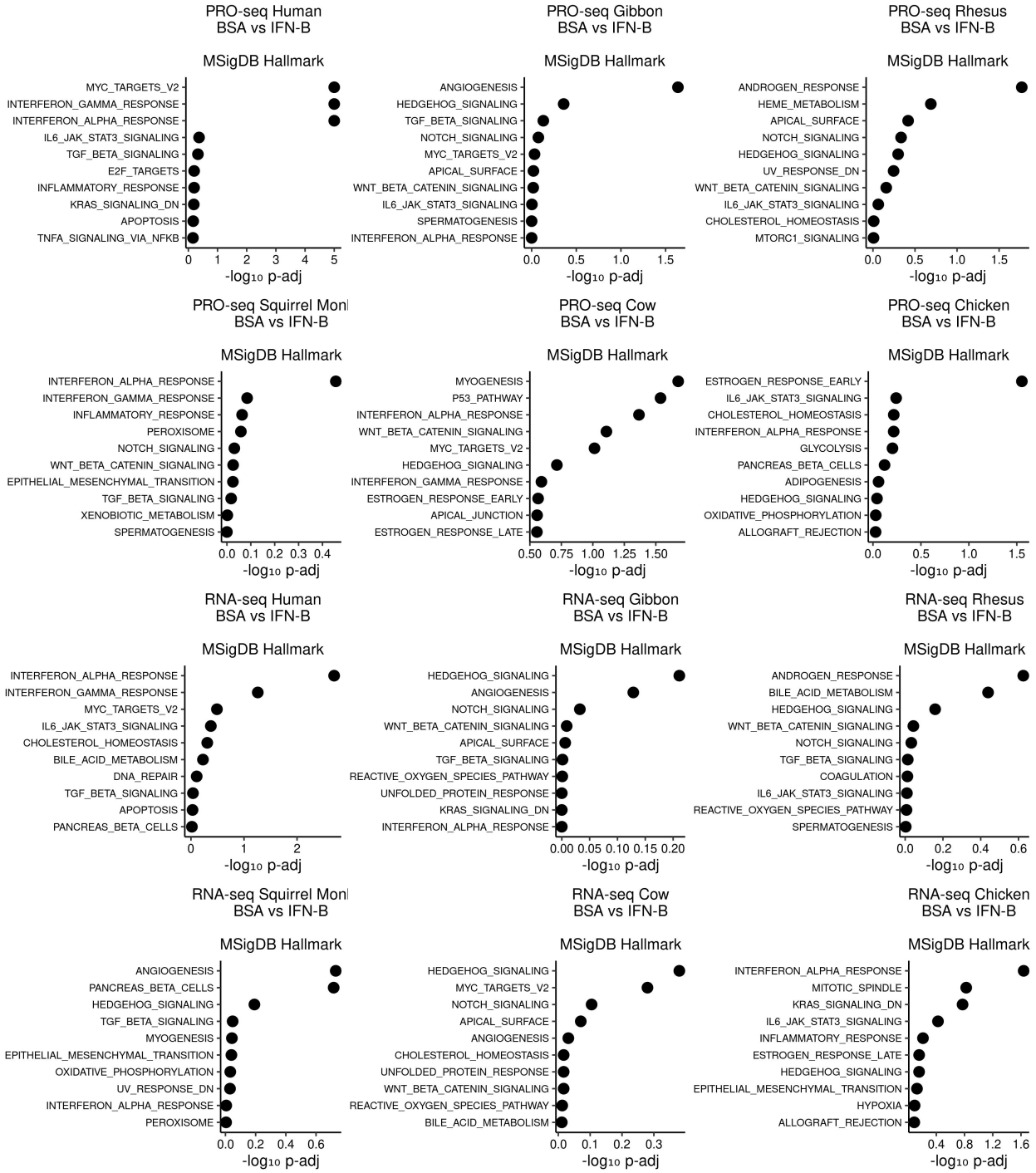


Figure 3.7: Total number of loci identified by Tfit with bidirectional transcription in the IFN-β-treated interspecies PRO-seq datasets.

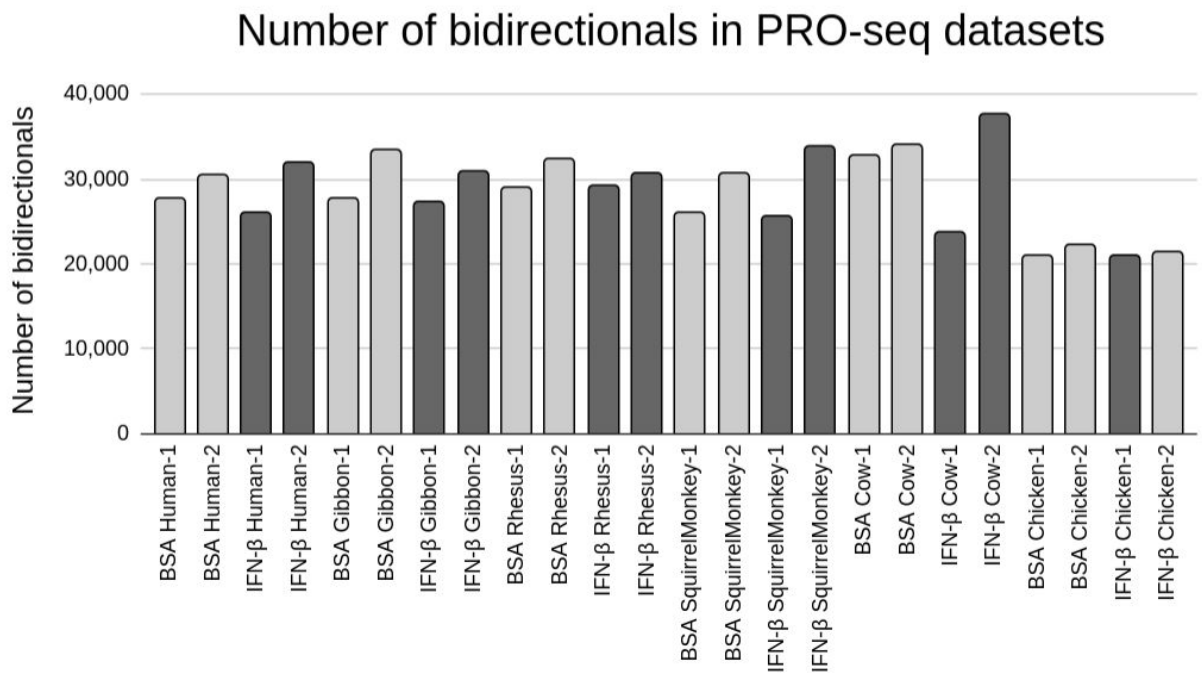
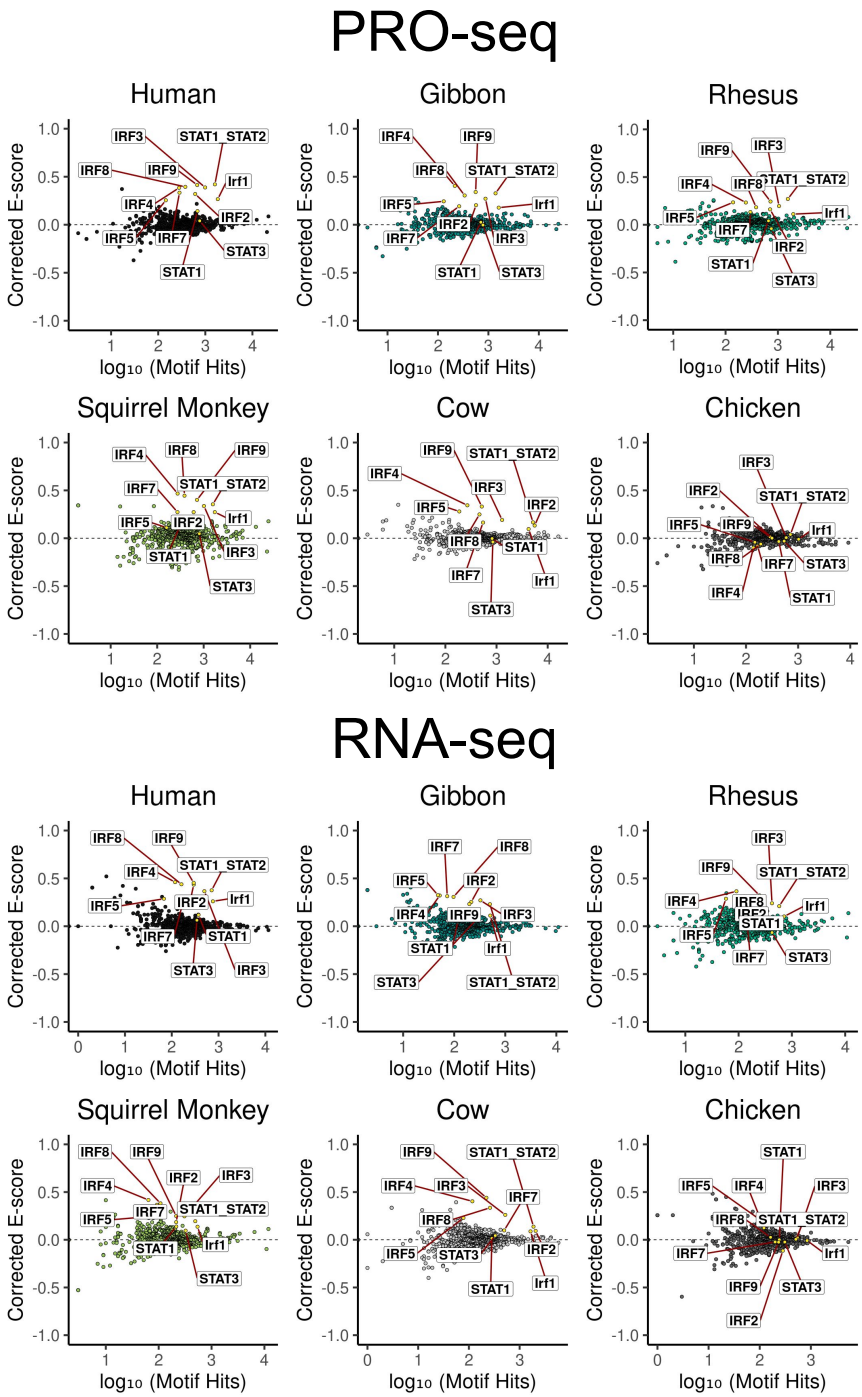


Figure 3.8: Transcription Factor Enrichment Analysis MA plots showing the corrected E-score in the vertical axis and the log10 of the number of motif hits in the horizontal axis. Each green dot represents a TF motif from the JASPAR2022 non-redundant vertebrate motif database. Labeled and colored in yellow are motifs from the IRF and STAT gene families. On the top two rows are the PRO-seq datasets results, and on the bottom two rows are the RNA-seq datasets results.



The TFEA results for the PRO-seq and RNA-seq datasets (Figure 3.8), show that five out of the six species LCLs treated with IFN-β display an enrichment for the STAT and IRF motifs, which suggest that ISGF3 is the main transcription factor driving the transcriptional response at both the 1 hour and 3 hour timepoints, in PRO-seq and RNA-seq, respectively. The chicken LCL, however, did not show any motif enrichment, which agrees with their lack of IFN-β induction.

Altogether, these preliminary results suggest that the interspecies LCL PRO-seq and RNA-seq datasets treated with IFN-β are a good model to study the differences in the gene transcriptional response of the type I IFN-β, alas removing the chicken datasets which did not show a perceivable IFN-β stimulation.

3.3.2 Cross-species (human and bonobo) chromatin accessibility changes upon interferon stimulation

With regards to the ATAC-seq datasets obtained only from the human and bonobo LCLs, I observed that the samples did not have many regions with differences in chromatin accessibility, but differences were observed at highly induced ISGs such as MX1 and STAT1. I used TFEA using this ranking of differentially accessible regions (Figure 3.9), and I obtained an enrichment of IRF and STAT motifs similar to the PRO-seq datasets.

3.3.3 Transcriptional response of human and rhesus to their cis- or trans- interferon

I obtained PRO-seq datasets from LCLs derived from both a male and a female individual from two primates, human and rhesus macaque. The human IFN-α2 and rhesus IFN-α2-treated LCLs display a typical type I interferon stimulation transcriptional response compared to the BSA control datasets. However, the human IFN-α2-treated datasets show a greater interferon stimulation magnitude than the rhesus IFN-α2-treated datasets, regardless of the primate LCL used, evident by looking at the transcription levels of genes (Figure 3.10).

All datasets still displayed a similar set of enriched TF motifs by TFEA, namely the IRF and STAT motifs (Figure 3.11). These results suggest that, though the response magnitude is

Figure 3.9: Transcription Factor Enrichment Analysis MA plots from the ATAC-seq datasets showing the corrected E-score in the vertical axis and the log10 of the number of motif hits in the horizontal axis. Each green dot represents a TF motif from the JASPAR2022 non-redundant vertebrate motif database. Labeled and colored in yellow are the motifs from the IRF and STAT gene families.

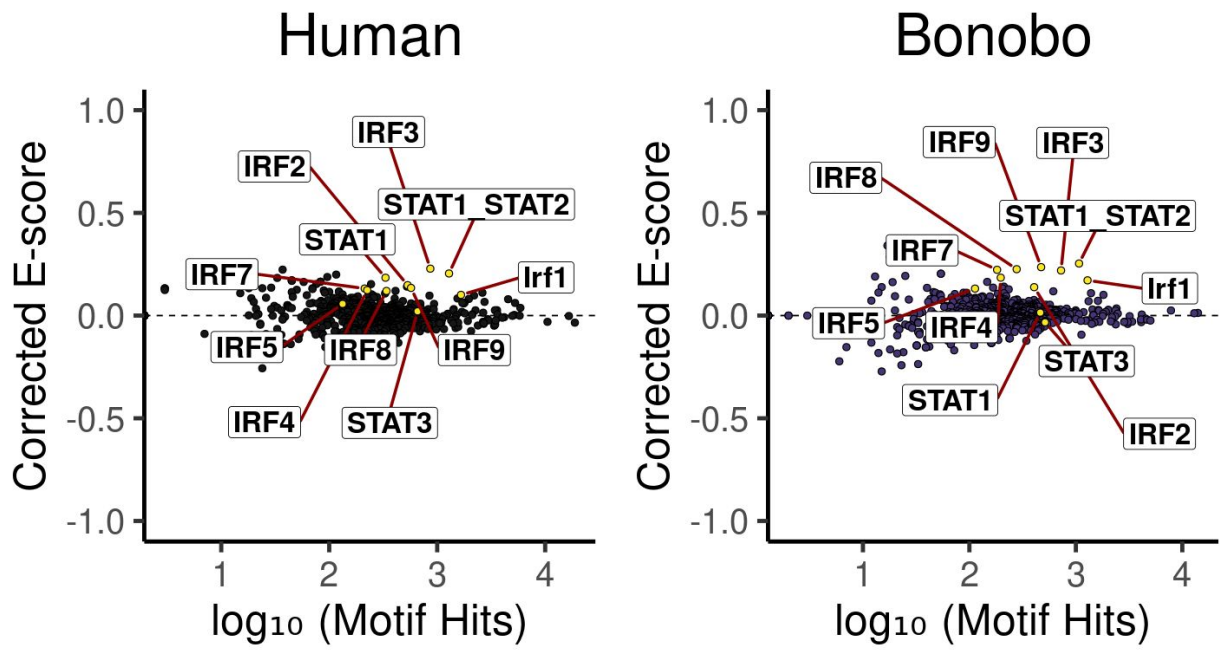
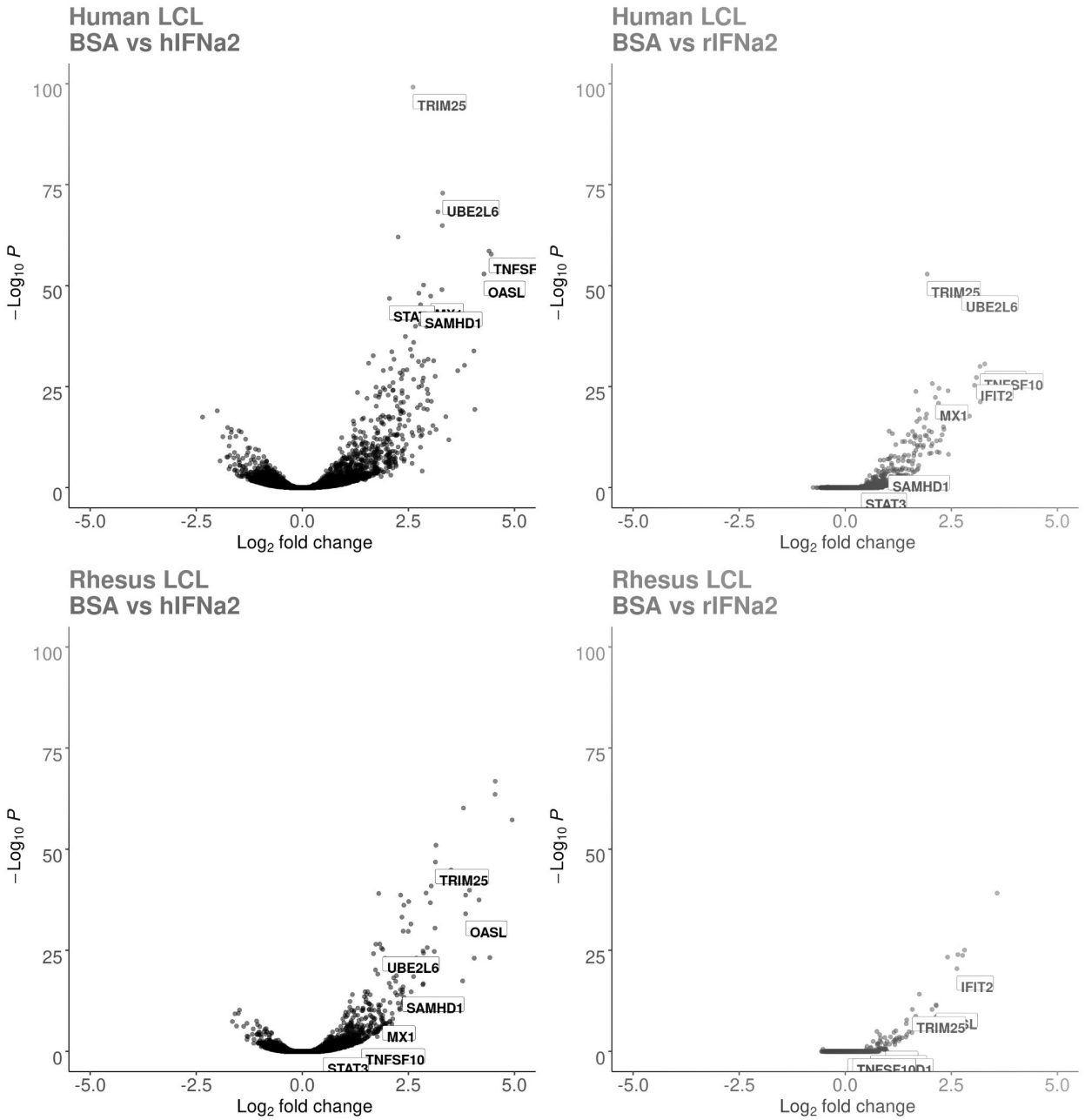


Figure 3.10: DESeq2 volcano plots show upregulation of ISGs for the human (top row) and rhesus LCLs (bottom row) treated with either human IFN-α2 (hIFNa2; left) or rhesus IFN-α2 (rIFNa2; right). Classical ISGs are labeled. Female and male datasets per species were used as replicates.



not the same when using the human or rhesus IFN-α2, subsequent analysis can still compare the differences in ISGs induced between the human and rhesus LCLs by focusing on either the human IFN-α2-treated dataset pairs, or the rhesus IFN-α2-treated dataset pairs.

3.3.4 Intrahuman transcriptional response to interferon

Similar to the interspecies datasets, I proceeded to obtain the IFN-β-treated 1 hour PRO-seq and 3-hour RNA-seq intrahuman datasets with the 8 human individuals, with two and three replicates, respectively. The intrahuman datasets were renamed from their official IDs to mock names whose initial letter starts with their country of origin. The LCL derived from a Mormon female human from the United States was labeled as Ursula, the LCL derived from a male human from the Mende people from Sierra Leone was labeled as Sengbe, the LCL derived from a female human from the Luhya people from Kenya was labeled as Khaondo, the LCL derived from a female human from the Yoruba people from Nigeria was labeled Niyilolawa, the LCL derived from a male human indigenous from Pero was labeled Pedro, the LCL derived from a Tamil female human from Sri Lanka was labeled as Srivathani, the LCL derived from a Han male human from China was labeled ChenChao. The LCL derived from a male from Mexico did not follow this renaming scheme and was simply labeled DR. The DR LCL was transformed by infecting primary B-cells extracted from the thesis author with Epstein-Barr Virus with the intention to have a readily available source of biomaterial to further compare the LCL IFN-β-treated transcriptome with to a genetic background matched primary B-cell IFN-β-treated transcriptome, although the primary B-cell treatments were not obtained. The other human LCLs were purchased from the Coriell cell repository belonging to the NIGMS Human Genetic Cell Repository and the NHGRI Sample Repository for Human Genetic Research. The PRO-seq datasets have an average of 43 million single-end reads, whereas the RNA-seq datasets have an average of 34 million paired-end reads (Figure 3.12).

The second intrahuman PRO-seq replicates had a significant decrease in their quality, likely due to overamplification during the preparation of the sequencing libraries. This low quality was

Figure 3.11: TFEA MA plots with the Tfit-muMerge bidirectional calls show TFs motifs enrichment for the human (top row) and rhesus LCLs (bottom row) treated with either human IFN-α2 (hIFNa2; left) or rhesus IFN-α2 (rIFNa2; right). STAT and IRF motif families are labeled. Female and male datasets per species were used as replicates.

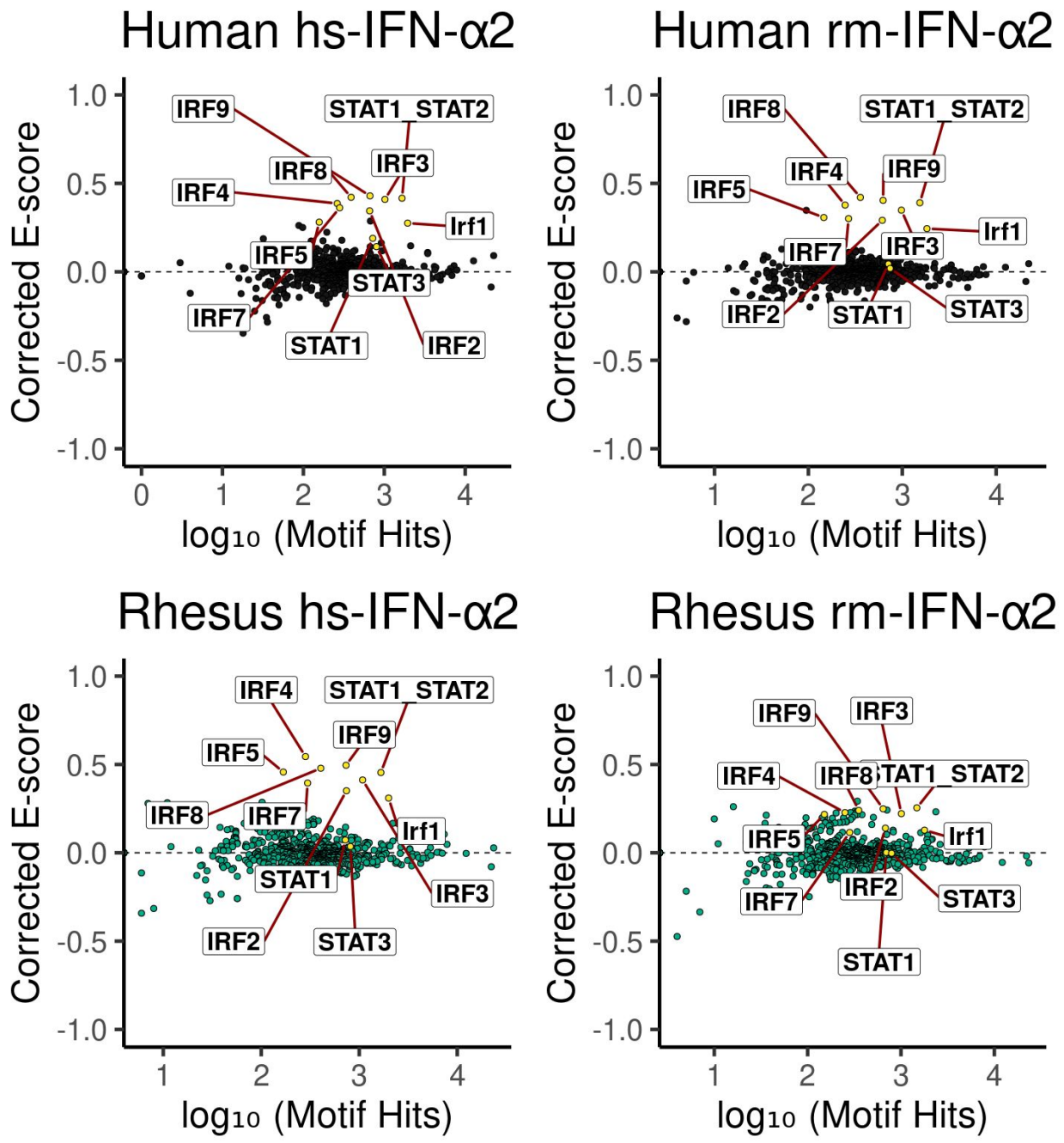
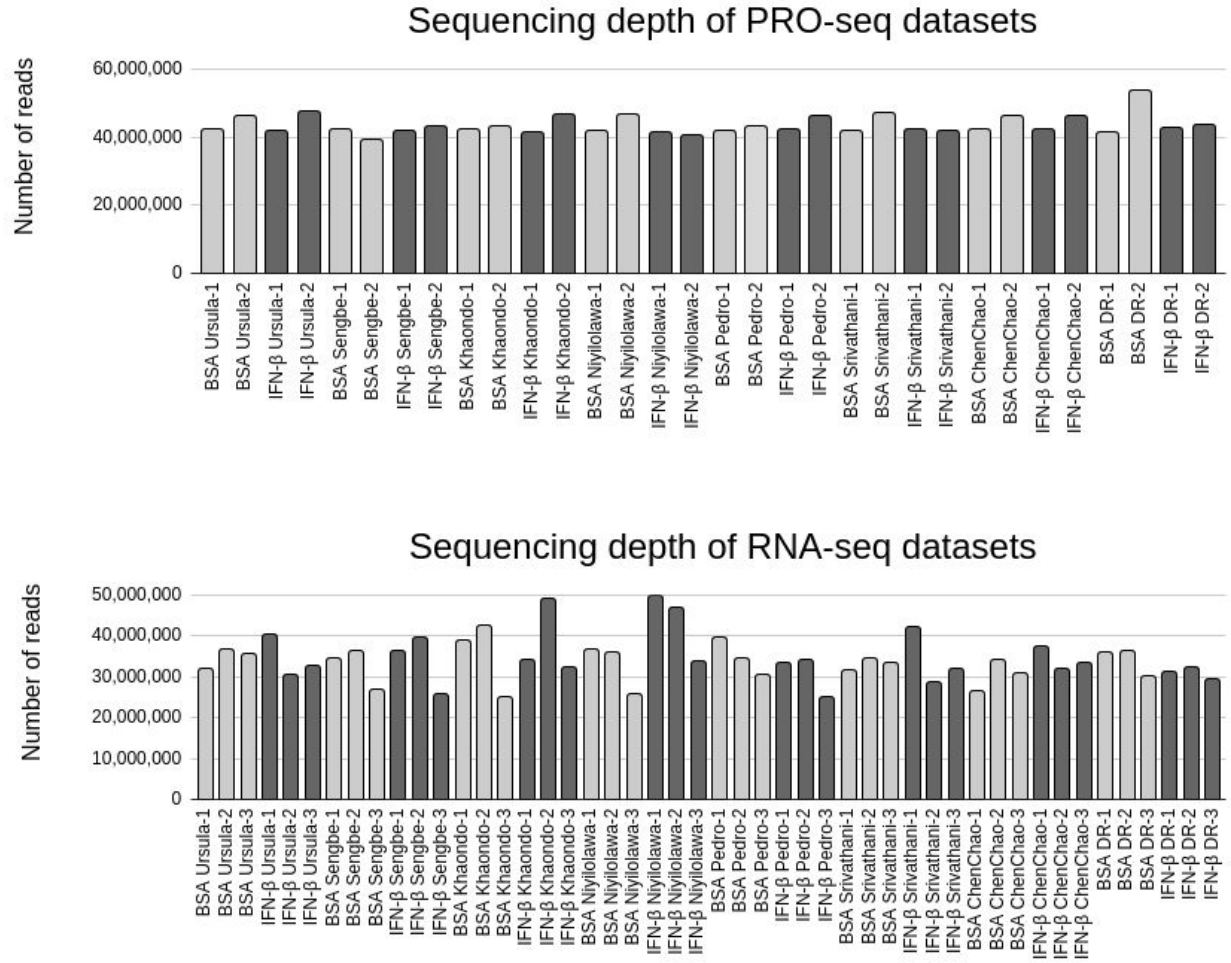


Figure 3.12: Total number of short sequencing reads for the intrahuman PRO-seq datasets (top) and the RNA-seq datasets (bottom). In light gray are the untreated samples, and in dark gray are the IFN-β treated samples.



especially evident in the 2nd Ursula and DR PRO-seq replicates, which directly affected the ability to detect DEGs in these two samples. The rest of the PRO-seq datasets and all the RNA-seq datasets showed sufficiently good quality to detect a similar number of DEGs per individual (Figure 3.13).

Similar to the interspecies datasets, the human PRO-seq and RNA-seq datasets stimulated with IFN-β show mostly upregulated genes with only a few genes whose transcription decrease at the two time points tested. All eight human LCLs highly induced typical ISGs, such as RSAD2, USP18, MX1, TNFSF10, among others (Figure 3.14).

An equal GSEA analysis was done on the human LCLs datasets. The top enriched gene sets also encompass the expected immune and stress-related sets such as the Interferon Response, Inflammatory Response, JAK STAT Signaling, Unfolded Protein Response, Oxidative Phosphorylation, and DNA Repair, among others (Figure 3.15).

Finally, just as with the interspecies datasets, I also used TFEA to check what TFs are driving the induction of the identified DEGs. I found that only the STAT and IRF motifs show an enrichment score in both the human PRO-seq and RNA-seq datasets, which agrees with ISGF3 driving the induction of ISGs (Figure 3.16).

In conclusion, here I present PRO-seq and RNA-seq datasets of LCLs treated with IFN-β for 1 hour and 3 hours, respectively. The datasets span both 6 distinct vertebrate species, and 8 human ethnicities. They will serve to study how the type I IFN transcriptional response has been rewired through evolutionary time in response to the pathogens that the hosts have encountered in their unique ecological niches.

3.4 Limitations

Here, I present PRO-seq and RNA-seq datasets treated with IFN-β for 1 hour and 3 hours, respectively, on LCLs derived from 6 different animal species and 8 different human ethnicities. Though the preliminary results suggest that there are differences in the usage of putative regulatory elements (i.e. bidirectional transcription loci), and in their accompanying regulated ISGs, there are

Figure 3.13: Total number of differentially expressed genes (DEGs) in all 8 human LCLs treated with IFN-β in the PRO-seq (top) and in the RNA-seq (bottom) datasets. DEGs were defined using DESeq2 with an alpha value of 0.05.

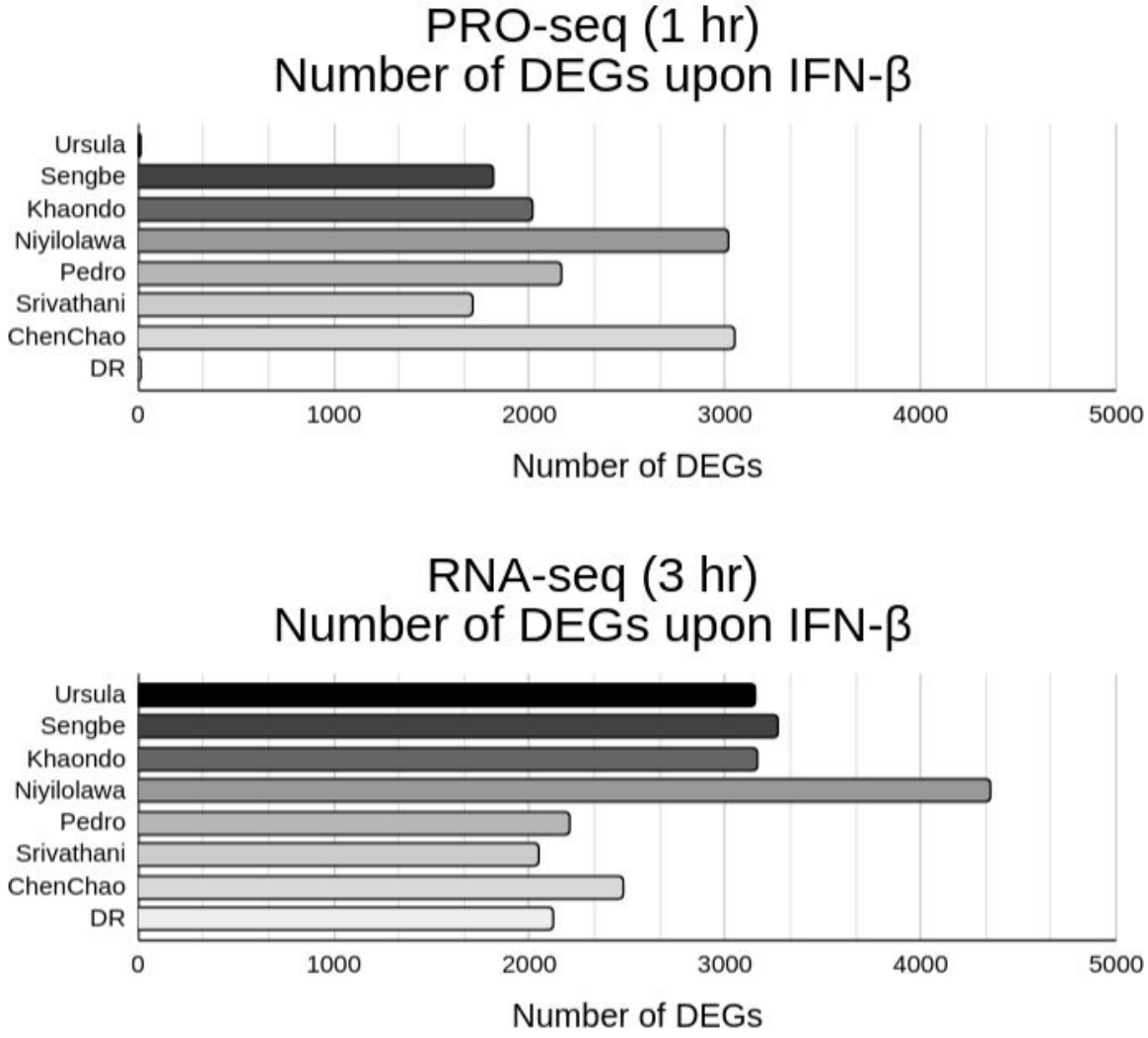


Figure 3.14: Volcano plots showing the log2 fold change in the horizontal axis and the -log10 adjusted p-value for all genes for the human LCLs treated with IFN-β. The top two rows show the PRO-seq datasets, and the bottom two rows show the RNA-seq datasets. A few of the top differentially expressed genes are labeled in all samples.

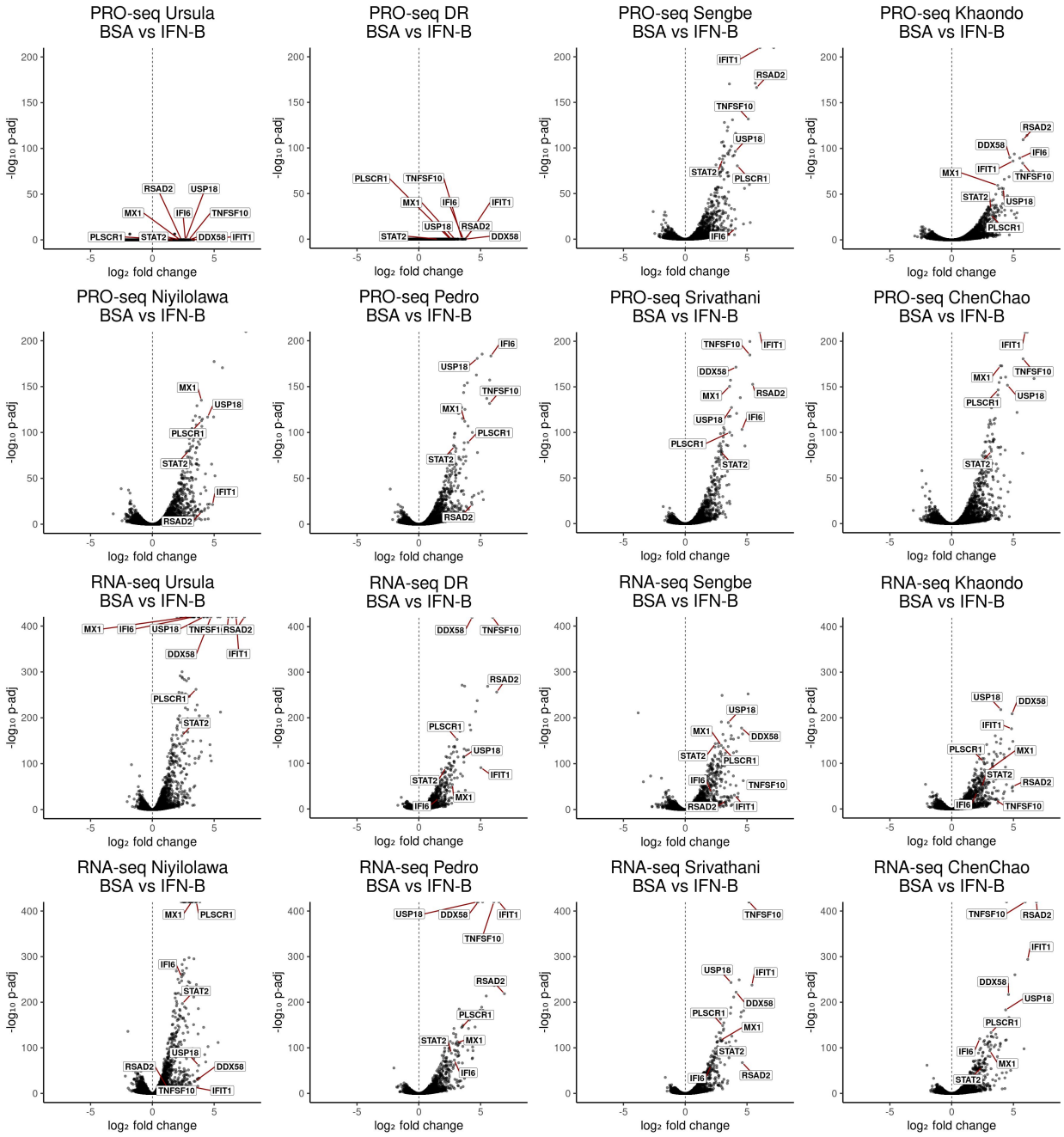


Figure 3.15: Gene Set Enrichment Analysis results using the Hallmark gene sets from the Molecular Signature Database on gene lists ranked by DESeq2 of IFN-β treated human cells. The 10 gene sets with the lowest adjusted p-values are shown transformed as their negative log10 values. The first two rows show the results from the PRO-seq datasets, and the bottom two rows show the results from the RNA-seq datasets.

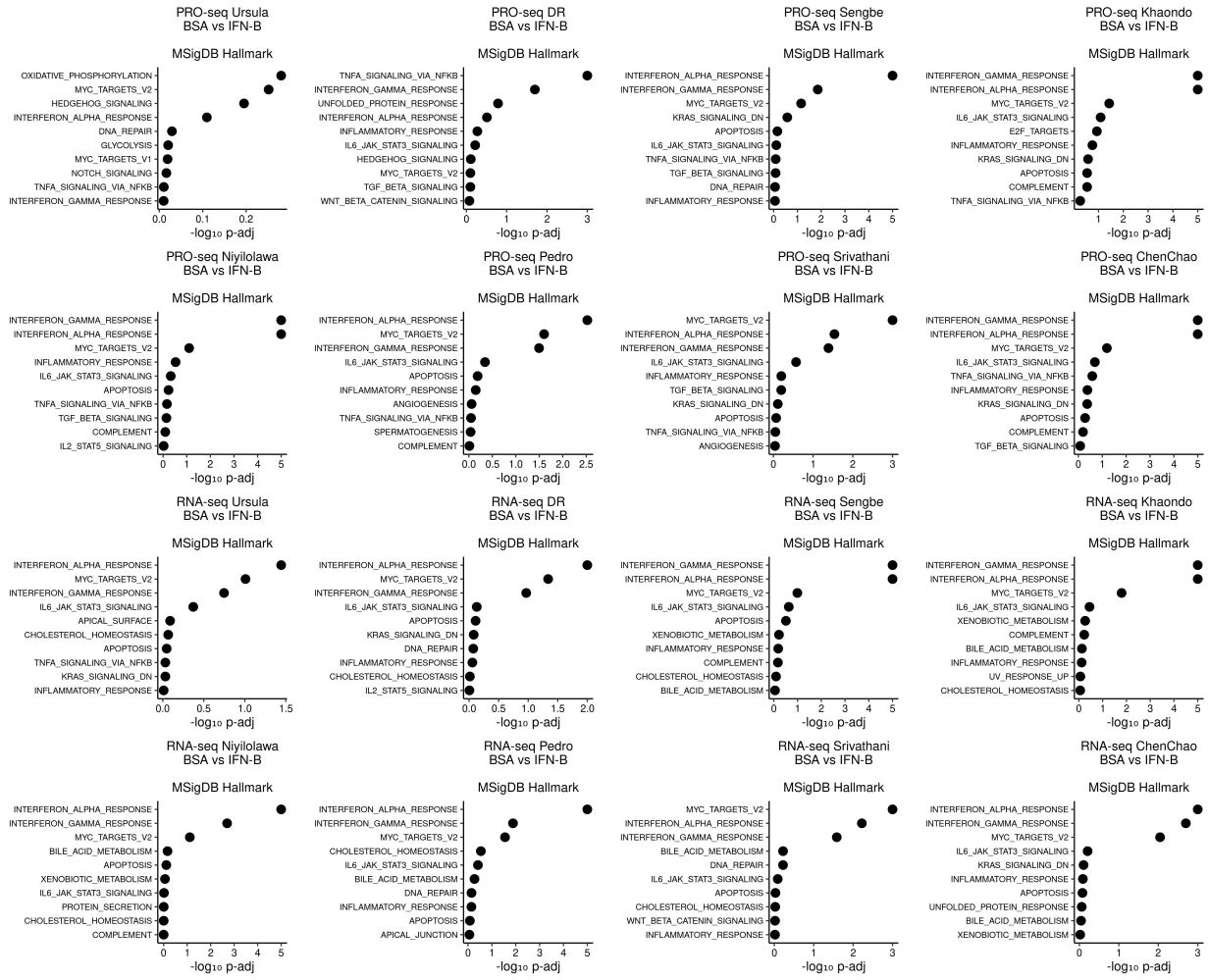
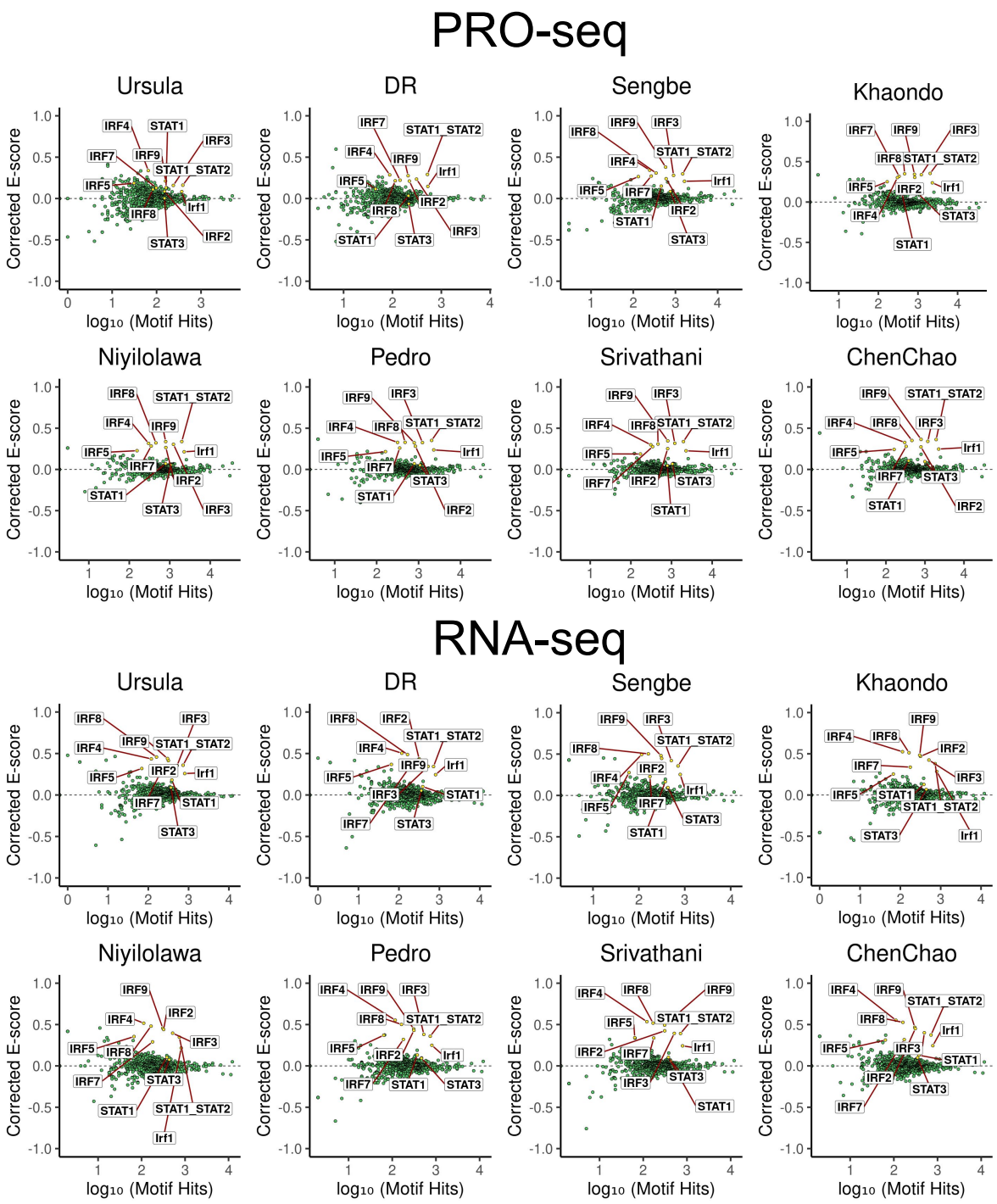


Figure 3.16: Transcription Factor Enrichment Analysis MA plots showing the corrected E-score in the vertical axis and the log10 of the number of motif hits in the horizontal axis. Each green dot represents a TF motif from the JASPAR2022 non-redundant vertebrate motif database. Labeled and colored in yellow are motifs from the IRF and STAT gene families. On the top two rows are the PRO-seq datasets results, and on the bottom two rows are the RNA-seq datasets results.



some potential variables that may have introduced confounding factors in subsequent analysis.

The experimental approach to test each species LCLs with their species-matched IFN-β was done to bypass any potential confounding effect of disrupting the fine-tuned binding affinity of IFN-β with their membrane receptors. And the IFN-β dosage per species was calibrated with RT-qPCR using a few ISGs. However, it is obvious that the PRO-seq and RNA-seq datasets show unequal magnitudes in their transcriptional responses, which suggest that either the calibration failed or the chosen ISGs were inappropriately chosen. These results make it hard to make inter-species comparisons of how ISGs are differentially regulated when the observed differences may be experimental artifacts instead.

In the human – rhesus cis/trans study, the human IFN-α2 protein was obtained from Proteintech Cat. no. HZ-1066, whereas the rhesus macaque IFN-α2 was obtained from PBL Assay Science Ca. no. 16105-1. Each manufacturer tested their purified protein activities using different assays, with the human IFN-α2 protein purification having been tested with a "dose-dependent cytotoxicity of the human TF-1 cell line (human erythroleukemic indicator cell line)" [78], and the rhesus IFN-α2 protein purification with a "cytopathic inhibition assay on Bovine (MDBK) kidney cells with vesicular stomatitis [virus] (VSV)" [77]. Discrepancies in the bioactivity assay details may have resulted in unequal magnitude of IFN-dependent transcriptional responses even when using 100 units/mL for both the human and rhesus IFN-α2 protein treatments. To this end, I observe that both cell lines responded more strongly to the human IFN-α2, as observed by the number of differentially transcribed genes and by the magnitude of the ISGs fold-change. Likeiwise, while I assayed two distinct cell lines per species, one female and one male, each biological sex was only assayed once.

As it is clearly shown in the intrahuman LCL IFN-β-treated datasets, there is significant interindividual variability in the IFN-β transcriptional response among individuals of a given species. It stands to reason that a similar or greater variability is expected in the natural gibbon, rhesus macaque, squirrel monkey, cow, and chicken populations. The current datasets interrogated a single individual among these species, so any claims in species-specificity should be taken with

caution. Had another individual been tested instead, the potential differences may not have been observed. A bigger sampling of individuals of both sexes and ages is needed to properly ascertain when differences have been fixed in a population and are therefore species-specific.

The usage of LCLs derived from Epstein-Barr Virus (EBV)-infected quiescent B-cells is another source of caution for evolutionary claims. EBV is known to persist in LCLs in a latent dormant state as circular DNA epiblasts or even integrated into the host genome [124, 87]. And their continuous presence in LCLs has a potential to compromise to an extent the IFN-controlled transcriptional response, as EBV has evolved to bypass the host defenses to subsist in such latent state. EBV also controls the cell proliferation, which is the very phenotype that researchers have exploited to use it to easily obtain LCLs from primates, and there is a possibility that the different LCLs are progressing through their cell-cycle at sufficiently different rates to affect their ability to respond to a stimulus such as IFN-β. In addition, the rhesus macaque, cow, and chicken LCLs were not transformed with EBV; but with Papiine Herpesvirus 1, Bovine Leukemia Virus and Avian Leukosis Virus, respectively. This may further complicate the interspecies comparisons as interspecies LCLs do not even harbor the same transformant viruses.

The above caveats may be the reason I observe a difference in the magnitude of the response even when samples were treated with the same IFN-β proteins, such as in the intrahuman datasets. It should be carefully considered how to assess when a given feature, a gene or a regulatory element, is differentially transcribed across the samples in consideration. It may be that a given gene, for instance, is statistically non-induced in a given sample, but that this is due not to its induction absence in the IFN-β responsive network, but because the sample's response was not sufficiently big for detection. Therefore, I believe that there is a need to validate the results with the use of primary B-cells to bypass most of the above limitations.

3.5 Methods

3.5.1 Cell lines information for IFN interspecies dataset

Table 3.1 describes the information of the LCLs used to generate the IFN interspecies dataset; including the date they were received, the species, the ID, and the source.

Table 3.1: Cell lines information for interspecies dataset used in chapter 3

Received	Species	ID	Source
2019/01/29	Human	GM12878	Coriell/NIGMS
2019/01/29	Bonobo	PR00748	Sara Sawyer Lab
2020/01/15	Gibbon	Ricky	Lucia Carbone Lab
2019/10/04	Rhesus	Mm 290-96	Yoav Gilad Lab
2019/12/12	Squirrel Monkey	SML clone 4D8	ATCC Ref. CRL-2311
2019/05/16	Cow	BL3.1	ATCC Ref. CRL-2306
2019/12/27	Chicken	DT40	ATCC Ref. CRL-2111

3.5.2 PRO-seq, ATAC-seq, and RNA-seq growth conditions for IFN interspecies dataset

The human, bonobo, gibbon, rhesus, squirrel monkey, and cow LCLs were cultured in RPMI-1640 media (Gibco Ref. 72400-047), 15% FBS (R&D Systems Ref. S11150), and 100 U/mL Penicillin-Streptomycin (Gibco Ref. 15140-122). The chicken LCLs were cultured in RPMI-1640 media (Gibco Ref. 72400-047), 10% FBS (R&D Systems Ref. S11150), 5% Chicken serum (Sigma Ref. C5405-100ML), 10% Tryptose phosphate broth (Sigma Ref. T8154), 0.05 mM β-mercaptoethanol (MP Ref. 194834), and 100 U/mL Penicillin-Streptomycin (Gibco Ref. 15140-122). All LCLs were cultured using vent-cap T-25 flasks (Corning 430639), and kept at a confluency between 400,000 cells/mL and 800,000 cells/mL during cell culture at 37°C with 5% CO_2, except for the chicken LCLs which were kept between 1,000,000 cells/mL and 2,500,000 cells/mL.

3.5.3 RT-qPCR to define IFN concentrations per species for IFN interspecies dataset

Reverse transcription quantitative polymerase chain reaction (RT-qPCR) was used to determine the concentrations at which different orthologous IFN-β elicit equivalent mRNA expression levels of a few tested Interferon-Stimulated Genes (ISGs) on each of the LCLs from the six different species. The ISGs were picked on the basis that all the species (human, gibbon, rhesus, squirrel monkey, cow, and chicken) should have annotated orthologs, as well as having evidence of induction upon interferon in human LCLs. The five picked ISGs were RSAD2, OASL, USP18, IFIH1, and STAT2; as well as two housekeeping genes ACTB and GUSB. To design the primers, multiple sequence alignments were done using MEGA (v10.0.5) with the orthologous mRNA sequences of the seven chosen genes, and regions with the most sequence conservation were chosen. However, no single primer sequence was fully conserved in the six species, which yielded degenerate primer sequences which are described next. All primers had an average GC content of 54% (with 5.6% of st.dev.) and an average melting temperature of 60.8°C (with 1.2°C of st.dev.). The RSAD2 forward primer had the sequence 5'-TGG YCA AGG AAR GAA GAA CC-3' spanning the sequence coordinates relative to the human gene (NCBI reference sequence NM_080657.5) from nucleotide 583 to 602, and a reverse primer with the sequence 5'-CAC TGG AAS ACY TTC CAG CG-3' corresponding to nucleotides 730 to 749, yielding an amplicon of length 166 bp. The OASL forward primer had the sequence 5'-CTT CAS CGA RCT GCA G-3' spanning the coordinates relative to the human gene (NCBI reference sequence NM_003733.4) from nucleotide 570 to 585, and a reverse primer with the sequence 5'-CCC AGG CRT AGA TGG TYA G-3' corresponding to nucleotides 715 to 733, yielding an amplicon of length 163 bp. The USP18 forward primer had the sequence 5'-ACA TTG GAC AGA CMT GCT G-3' spanning the coordinates relative to the human gene (NCBI reference sequence NM_017414.4) from nucleotide 176 to 194, and a reverse primer with the sequence 5'-CTG CAT CTT CTC CAR CAG C-3' corresponding to nucleotides 315 to 333, yielding an amplicon of length 157 bp. The IFIH1 forward primer had the sequence 5'-AAC CAG AGT

GGC YGT TTA C-3' spanning the coordinates relative to the human gene (NCBI reference sequence NM_022168.4) from nucleotide 1005 to 1023, and a reverse primer with the sequence 5'-GCT GTT CMA CTA RCR GTA CC-3' corresponding to nucleotides 1095 to 1114, yielding an amplicon of length 109 bp. The STAT2 forward primer had the sequence 5'-CWC CTG GGT GGA RCA C-3' spanning the coordinates relative to the human gene (NCBI reference sequence NM_005419.4) from nucleotide 1836 to 1851, and a reverse primer with the sequence 5'-TAG AGR AAG MGC ART GG-3' corresponding to nucleotides 1978 to 1994, yielding an amplicon of length 158 bp. The ACTB forward primer had the sequence 5'-GAG AAG ATG ACM CAG ATC ATG-3' spanning the coordinates relative to the human gene (NCBI reference sequence NM_001101.5) from nucleotide 349 to 369, and a reverse primer with the sequence 5'-CCA GRT CCA GAC GSA GGA TG-3' corresponding to nucleotides 522 to 541, yielding an amplicon of length 192 bp. The GUSB forward primer had the sequence 5'-GCC DTA YCT GAT GCA CG-3' spanning the coordinates relative to the human gene (NCBI reference sequence NM_000181.4) from nucleotide 867 to 883, and a reverse primer with the sequence 5'-GCR TCC TCR TGC TTG TTG AC-3' corresponding to nucleotides 1042 to 1061, yielding an amplicon of length 194 bp. Where the non-standard nucleotide letters follow the capital IUB (International Union of Biochemistry) code with the standard equivalent ATGC nucleotides in parenthesis: Y (C/T), R (A/G), S (G/C), M (A/C), W (A/T), K (G/T), and D (A/G/T). A primer with such a letter was then synthesized with an equal proportion of either of the standard nucleotides in the resulting primer tube, such that if a primer sequence had a single "W" then 50% of the synthesized primer sequences contained an "A" and the other 50% of the primer sequences contained a "T". LCLs from human, gibbon, rhesus, squirrel monkey, cow, and chicken, were seeded into 48-well plate wells, each well with 150,000 cells in 225 uL of volume, eight wells per species. IFN-β stock aliquots were prepared at 250,000 ng/mL for all six IFN-β proteins. Each species LCL was treated for 3 hours with different concentrations of their species-matching IFN-β proteins. The human LCLs were treated with human IFN-β (Kingfisher Biotech Ref. RP1788H-100 Lot. KU4428KU), the gibbon LCLs were treated with gibbon IFN-β (Kingfisher Biotech Ref. RP1791GB-025 Lot. LU4443KU), the rhesus LCLs were treated with rhesus

IFN-β (Kingfisher Biotech Ref. RP1709Y-025 Lot. CU4126BU), the squirrel monkey LCLs were treated with squirrel monkey IFN-β (Kingfisher Biotech Ref. RP1829SM-025 Lot. BV4549LU), the cow LCLs were treated with cow IFN-β (Kingfisher Biotech Ref. RP0298B-025 Lot. FO1566FL), and the chicken LCLs were treated with chicken IFN-β (Kingfisher Biotech Ref. RP1786C-025 Lot. KU4407KU). The highest IFN-B concentration tested was 10,000 ng/mL, and serial 10-fold dilutions were made from this one yielding seven different concentrations: 10,000 ng/mL, 1,000 ng/mL, 100 ng/mL, 1 ng/mL, 10 ng/mL, 0.1 ng/mL, 0.01 ng/mL, as well as an 8th dilution with no IFN-B as a baseline reference to calculate fold-change values. After the 3 hour incubation, 850 µL of RNA lysis buffer was mixed into all wells, and total RNA was extracted using the Quick-RNA MiniPrep Plus (Zymo Research Ref. R1058) following the manufacturer's instructions. The RNA purity was determined using a Nanodrop with a 260 nm and 280 nm absorbance ratio ranging from 1.87 to 2.29 for all samples. All RNA samples were diluted to obtain 5 ng/µL, and 18.5 ng of RNA were used per reaction. The RT-qPCR reactions were set up using the Luna Universal One-Step RT-qPCR Kit (NEB Ref. E3005L) following the manufacturer's instructions, with a total volume of 10 µL per reaction, 50 cycles, and 2 or 3 replicates per condition using a Bio-Rad CFX384 Touch Real-Time PCR System.

3.5.4 PRO-seq treatment conditions for IFN interspecies dataset

Each of the 6 animal LCLs were treated with their species-specific IFN-β, or with BSA, for 1 hour prior to the nuclei isolation. The human LCLs were treated with human IFN-β at 100 ng/mL (Kingfisher Biotech Ref. RP1788H-100 Lot. KU4428KU, resuspended in 400 µL of 0.1% BSA PBS on 2020/12/23 at 250,000 ng/mL). The gibbon LCLs were treated with gibbon IFN-β at 100 ng/mL (Kingfisher Biotech Ref. RP1791GB-025 Lot. LU4443KU, resuspended in 100 µL of 0.1% BSA PBS on 2020/12/23 at 250,000 ng/mL). The rhesus LCLs were treated with rhesus IFN-β at 500 ng/mL (Kingfisher Biotech Ref. RP1709Y-025 Lot. CU4126BU, resuspended in 100 µL of 0.1% BSA PBS on 2021/05/09 at 250,000 ng/mL). The squirrel monkey LCLs were treated with squirrel monkey IFN-β at 5 ng/mL (Kingfisher Biotech Ref. RP1829SM-025 Lot. BV4549LU, resuspended

in 100 µL of 0.1% BSA PBS on 2021/05/09 at 250,000 ng/mL). The cow LCLs were treated with cow IFN-β at 200 ng/mL (Kingfisher Biotech Ref. RP0298B-025 Lot. FO1566FL, resuspended in 100 µL of 0.1% BSA PBS on 2021/05/09 at 250,000 ng/mL). The chicken LCLs were treated with chicken IFN-β at 500 ng/mL (Kingfisher Biotech Ref. RP1786C-025 Lot. KU4407KU, resuspended in 100 µL of 0.1% BSA PBS on 2020/12/23 at 250,000 ng/mL). For the untreated BSA negative controls, each species LCLs were treated with an equal volume of BSA PBS (final 0.00004% similar to the IFN-β treatments). 3 T-25 cultures per LCLs were used per treatment, except the chicken LCLs, from which 2 T-25 cultures were used. The 1st replicates were processed on 2021/05/28, and the 2nd replicates were processed on 2021/06/01. All cultures and treatments were processed in parallel.

3.5.5 PRO-seq nuclei extraction for IFN interspecies dataset

Nuclei isolation was done as described in [44] with some modifications. Briefly, LCL cultures ranging from 10 to 30 million cells were used for each condition. After each culture was treated for 1 hour, the cultures were washed twice with ice-cold PBS. Then, the cell pellets were carefully resuspended in 6 mL of lysis buffer (0.1% DEPC-DI water with 10 mM Tris-HCl pH 7.4, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 1 mM DTT, Invitrogen Ref. AM2696 SUPERase-IN RNAse inhibitor, and with Roche Ref. 11836170001 protease inhibitor cocktail) and centrifuged for 15 minutes at 4°C at 1000 x g. The pellets were carefully resuspended in 1 mL lysis buffer using Finntip wide orifice pipette tips (Thermo Scientific Ref. 9405163), were mixed with 4 mL more of lysis buffer, and centrifuged a second time for 15 minutes at 4°C at 1000 x g. The pellets were carefully resuspended a second time in 1 mL lysis buffer using Finntip wide orifice pipette tips, transferred to low binding 1.7 mL eppendorf tubes (Costar Ref. 3207), and centrifuged for 5 minutes at 4°C at 1000 x g. The pellets were carefully resuspended using Finntip wide orifice pipette tips in 500 µL freezing buffer (0.1% DEPC-DI water with 50 mM Tris-HCl pH 8.0, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, and SUPERase-IN RNAse inhibitor), and centrifuged for 2 minutes at 4°C at 2000 x g. The resulting nuclei pellets were resuspended

a final time in 110 µL of freezing buffer using Finntip wide orifice pipette tips. I mixed 10 µL of the resuspended nuclei with 990 µL of PBS for counting the nuclei yield. The remaining 100 µL resuspended nuclei were snap-frozen in liquid nitrogen and stored at -70°C before being used for the PRO-seq nuclear-run on reactions.

3.5.6 PRO-seq library preparation for IFN interspecies dataset

PRO-seq datasets were prepared as described in [60], which in turn is a modified protocol from [122]. Briefly, between 3 to 16 million nuclei per dataset were used for the PRO-seq transcription run-on using a mixture of rNTP and Biotin-11-CTP (Biotin-11-CTP at 0.025 mM from PerkinElmer Ref. NEL542001EA; rCTP at 0.025 mM from Promega Ref. E604B, rATP at 0.125 mM Ref. E601B, rGTP at 0.125 mM Ref. E603B, and rUTP at 0.125 mM Ref. E6021). 1% of S2 Drosophila melanogaster nuclei relative to the number of the sample nuclei were added during the run-on reaction as a normalization spike-in. Total RNA was extracted using a phenol/chloroform precipitation. Isolated RNA was fragmented using base hydrolysis with NaOH. Biotinylated fragmented nascent transcripts were isolated using a first streptavidin Dynabeads M-280 (Invitrogen Ref. 11206D) pull down, and the VRA3 RNA adaptor was ligated at their 3′ end. A second streptavidin bead pull down was performed, followed by the enzymatic modifications of the RNA fragment 5′ ends with a pyrophosphohydrolase and a polynucleotide kinase, and the VRA5 RNA adaptor was ligated at their fixed 5′ ends. A third streptavidin bead pull down was performed, followed by the reverse transcription of the resulting adaptor-ligated libraries. The libraries were cleaned up with AMPure XP beads (Beckman Coulter Ref. A63881). Then, the libraries were amplified using 13 PCR cycles, and cleaned up again with another round of AMPure XP beads. The resulting library concentrations were measured with the Qubit dsDNA high sensitivity assay (Invitrogen Ref. Q32851), and their size distributions assessed using the Agilent High Sensitivity D1000 ScreenTape. The 1st and 2nd replicates were processed together on 2021/05/29.

3.5.7 PRO-seq sequencing information for IFN interspecies dataset

The 1st replicates were sequenced on 2022/10/14, and the 2nd replicates were sequenced on 2022/07/22, both on a NextSeq 2000 as single-end 76 bp reads.

Table 3.2 describes the number of reads per PRO-seq library in the IFN interspecies dataset.

Table 3.2: Sequencing depth of the IFN PRO-seq interspecies datasets used in chapter 3

Dataset	Read number	Dataset	Read number
PRO-BSA-Human-1	28,963,929	PRO-IFN-Human-1	25,226,027
PRO-BSA-Human-2	33,553,035	PRO-IFN-Human-2	43,479,821
PRO-BSA-Gibbon-1	29,720,802	PRO-IFN-Gibbon-1	25,800,568
PRO-BSA-Gibbon-2	56,222,892	PRO-IFN-Gibbon-2	42,612,968
PRO-BSA-Rhesus-1	26,671,442	PRO-IFN-Rhesus-1	27,101,404
PRO-BSA-Rhesus-2	58,274,611	PRO-IFN-Rhesus-2	40,796,870
PRO-BSA-SquirrelMonkey-1	26,521,453	PRO-IFN-SquirrelMonkey-1	24,165,862
PRO-BSA-SquirrelMonkey-2	52,438,859	PRO-IFN-SquirrelMonkey-2	72,146,961
PRO-BSA-Cow-1	33,588,829	PRO-IFN-Cow-1	17,583,572
PRO-BSA-Cow-2	33,831,547	PRO-IFN-Cow-2	69,792,582
PRO-BSA-Chicken-1	29,027,720	PRO-IFN-Chicken-1	33,035,940
PRO-BSA-Chicken-2	45,561,097	PRO-IFN-Chicken-2	47,643,589

3.5.8 PRO-seq datasets processing for interspecies dataset

- PRO-seq datasets were processed using the Nextflow pipeline found in `https://github.com/Dowell-Lab/Nascent-Flow`.
- Read quality was assessed using FastQC (v0.11.5)
- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tpe, tbo, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.

- Mapping was done using HISAT2 (v2.1.0) with options –very-sensitive –no-spliced-alignment on each species' respective reference genomes. The human reference genome hg38 was obtained from `GP/hg38/bigZips/hg38.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY. The gibbon reference genome nomLeu3 was obtained from `GP/nomLeu3/bigZips/nomLeu3.fa.gz` and was modified so that it only contained the main chromosome contigs chr1a, chr2, chr3, chr4, chr5, chr6, chr7b, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22a, chr23, chr24, chr25, chrX. The rhesus reference genome rheMac10 was obtained from `GP/rheMac10/bigZips/rheMac10.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chrM, chrX, chrY. The squirrel monkey reference genome saiBol1 was obtained from `GP/saiBol1/bigZips/saiBol1.fa.gz` and was modified so that it only contained the contigs numbered from JH378105 to JH378420, which were renamed as chr1 to chr316, respectively. The cow reference genome bosTau9 was obtained from `GP/bosTau9/bigZips/bosTau9.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr29, chrM, chrX. The chicken reference genome galGal6 was obtained from `GP/galGal6/bigZips/galGal6.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr30, chr31, chr32, chr33, chrM, chrW, chrZ.

- SAM to BAM conversion was done using Samtools (v1.8).

- Bigwigs were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1 –

normalizeUsing RPKM –filterRNAstrand reverse (for pos file) or forward (for neg file) –scaleFactor 1 (for pos file) or -1 (for neg file).

- Bidirectional loci were determined using Tfit and dREG as described in the Nextflow pipeline `https://github.com/Dowell-Lab/Bidirectional-Flow`. It removes multimapped reads using Samtools (v1.8) view -h -q 1 'bam file' | grep -P '(NH:i:1| ∧@)' | samtools view -h -b. Tfit calls were obtained by first using the Tfit bidir module to call prelim regions. The annotation was used to add 3 kb-wide TSS regions to the prelim file and removed any part of the prelim regions that overlap with the TSS regions. Prelim regions > 10 kb were then fragmented down to equal size regions (< 10kb) with 50% overlap and then coverage filtered to keep prelim regions having > 9 mapped reads. Finally, the adjusted prelim regions were used as regions of interest to the Tfit model module to obtain Tfit calls. dREG calls were filtered as having FDR < 0.05, merged if within 20bp of each other, and having > 9 mapped reads. Bidirectional transcription calls were combined using muMerge (v1.1.0) across experimental conditions.

- Read counts over genes were obtained using R (v3.6.0) Rsubread featureCounts (v1.32.4) with the options isGTFAnnotationFile=FALSE, useMetaFeatures=TRUE, allowMultiOverlap=TRUE, largestOverlap=TRUE, isPairedEnd=FALSE, strandSpecific=1; using the multimapped reads filtered BAM files; and using a custom SAF file that contains the longest annotated entry per gene from the RefSeq annotation, without the initial 25% genic region starting from the 5′ end to remove the RNA polymerase pausing region. The specific steps to produce this SAF file are as follows, with the human hg38 annotation as example: wget `https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/hg38.ncbiRefSeq.gtf.gz` /Users/dara6367/miniconda2/bin/convert2bed input=gtf output=bed do-not-sort < hg38.ncbiRefSeq.gtf > hg38.ncbiRefSeq.bed grep -w transcript hg38.ncbiRefSeq.bed | grep -v chr[0-9]*_ | cut -f1,2,3,4,5,6 | awk -v OFS='\t' '{print $1, $2, $3, $4, $5, $6, $3-$2}' > hg38.ncbiRefSeq.bed.tmp sort -nk7r

hg38.ncbiRefSeq.bed.tmp | sort -u -k4,4 | awk -v OFS='\t' '{print $1, $2, $3, $4, $5, $6}' | sort -k 1,1 -k2,2n > hg38.ncbiRefSeq.oneEntry.bed awk -v OFS='\t' '{print $4, $1, $2, $3, $6}' hg38.ncbiRefSeq.oneEntry.bed > hg38.ncbiRefSeq.oneEntry.saf awk -v OFS='\t' '{ if ($5 == "+") printf "%s\t%s\t%.0f\t%.0f\t%s\n", $1, $2, $3+(($4-$3)*0.25), $4, $5; else print $0 }' hg38.ncbiRefSeq.oneEntry.saf > hg38.ncbiRefSeq.without5prime25.oneEntry.saf awk -v OFS='\t' '{ if ($5 == "-") printf "%s\t%s\t%.0f\t%.0f\t%s\n", $1, $2, $3, $4-(($4-$3)*0.25), $5; else print $0 }' hg38.ncbiRefSeq.without5prime25.oneEntry.saf.tmp > hg38.ncbiRefSeq.without5prime25.oneEntry.saf

- Differential gene expression was done using DESeq2 (v1.26.0). Gene set enrichment was done using the GSEA GUI (v4.3.2) with the Human MSigDB Collections (v7.5.1).

- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the multimapped reads filtered BAM files and the muMerged Tfit or dREG bidirectionals.

3.5.9 ATAC-seq treatment conditions for interspecies dataset

Only the human and bonobo LCLs were used for obtaining ATAC-seq libraries. They were treated for 1 hour with either 100 U/mL of human IFN-α2 (Proteintech Ref. HZ-1066), or with 0.001% DMSO as a negative control. The treatments were done in 12-well plate wells so that each well had 100,000 cells in 2 mL volume. Each condition was prepared in duplicates, and all samples were processed in parallel.

3.5.10 ATAC-seq library preparation for interspecies dataset

The ATAC-seq libraries were made following the [42] protocol. Briefly, after the 1 hour treatments, the 100,000 cells were transferred from their 12-well plate wells to 1.8 mL eppendorf tubes, and centrifuged at 500 x g for 7 minutes at 4°C. The supernatant was carefully removed and replaced with 50 µL of ice-cold lysis buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM

$MgCl_2$, 0.1% IGEPAL, 0.1% Tween-20, 0.01% Digitonin), the cells resuspended 4 times pipetting up and down, and incubated on ice for 5 minutes. Then, added 1 mL of wash buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) and the tubes inverted 4 times to mix. The tubes were centrifuged at 500 x g for 10 minutes at 4°C and the supernatant was carefully removed without disturbing the small cell pellet. The pellets were then carefully resuspended by pipetting 6 times with 50 µL of the transposition mix (25 µL Tagment DNA Buffer Illumina Ref. 15027866, 2.5 µL Tagment DNA Enzyme 1 Illumina Ref. 15027865, 0.5 µL Digitonin diluted 1:1 with water, 0.5 µL 10% Tween-20, 5 µL water, 16.5 µL PBS), and were incubated for 30 minutes in a heat block at 37°C, flicking the tube often. Afterwards, the samples were cleaned using the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014) following the manufacturer's instructions, and eluted in 21 µL elution buffer. Then, a PCR pre-amplification was done using NEBNext Ultra II Q5 Master Mix (NEB Ref. M0544S) using 5 cycles. Then, a qPCR was done using NEBNext Ultra II Q5 Master Mix, SYBR Gold (Life Tech Ref. S11494), and 5 µL of the pre-amplified sample, and the results used to determine the additional number of extra PCR cycles using Nextera DNA CD Indices (Illumina Ref. 20015882), which was just 1 additional cycle. The post-amplified ATAC-libraries were cleaned-up with the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014). The libraries were size-selected to remove DNA fragments greater than 1000 bp with a Sage Science BluePippin. The ATAC-seq libraries were quantified with Qubit HS DNA assay and their fragment size-distributions determined with Agilent HS D5000 ScreenTape. All samples were processed in parallel. After the samples were pooled and size-selected, they were cleaned-up using AMPure XP beads (Beckman Coulter Ref. A63881) at 1.5x volume and eluted into 20 µL of EB buffer (Qiagen Ref. 19086).

3.5.11 ATAC-seq sequencing information for interspecies dataset

The pooled 1st and 2nd replicates were sequenced on 2019/03/15 on a NextSeq 500 as paired-end 150 bp reads.

Table 3.3 describes the number of reads per ATAC-seq library in the IFN interspecies dataset.

Table 3.3: Sequencing depth of the IFN ATAC-seq interspecies datasets used in chapter 3

Dataset	Read number	Dataset	Read number
ATAC-DMSO-Human-1	17,181,112	ATAC-Nutlin-Human-1	23,374,036
ATAC-DMSO-Human-2	15,614,866	ATAC-Nutlin-Human-2	20,091,690
ATAC-hsIFNa2-Human-1	18,784,696	ATAC-hsIFNa2-Human-2	23,287,969
ATAC-hsIFNa2-Bonobo-1	14,149,541	ATAC-hsIFNa2-Bonobo-2	17,088,889

3.5.12 ATAC-seq datasets processing for interspecies dataset

- Read quality was assessed using FastQC (v0.11.5).
- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tpe, tbo, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.
- Mapping was done using HISAT2 (v2.1.0) with options –new-summary –very-sensitive –no-spliced-alignment. The human reference genome hg38 was obtained from `GP/hg38/bigZips/hg38.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY). The bonobo reference genome panPan3 was obtained from `GP/panPan3/bigZips/panPan3.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr3, chr4, chr5, chr6, chr7, chrX, chr8, chr12, chr11, chr2B, chr10, chr9, chr2A, chr13, chr14, chr15, chr17, chr18, chr16, chr20, chr19, chr21, chr22, chrM).
- Converted mapped SAM to BAM files using Samtools (v1.8) view -F 4 to remove unmapped reads.
- Read duplicates were removed using Sambamba (v0.6.6) markdup with options –remove-

duplicates, –overflow-list-size=300000.

- Bedgraph files were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1, –normalizeUsing CPM.
- Peaks were determined using MACS2 (v2.1.1.20160309) callpeak with options –nolambda, –nomodel, –keep-dup all, –call-summits, and filtered out narrowPeaks with a score < 100.
- Peaks were merged across the species datasets using muMerge (v1.1.0) using options –save_sampids, –verbose.
- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the deduplicated BAM files and the muMerged MACS2 peaks.

3.5.13 RNA-seq treatment conditions for interspecies dataset

Each of the 6 animal LCLs were treated with their species-specific IFN-β, or with BSA, for 3 hours prior to the cell lysate step of the RNA-seq libraries preparation. On the day of the treatments, each LCL was moved onto separate 48-well plate wells, each with 135,000 cells/well and left incubating in a total volume of 250 µL after the IFN-β or BSA addition. The human LCLs were treated with human IFN-β at 100 ng/mL (Kingfisher Biotech Ref. RP1788H-100 Lot. KU4428KU, resuspended in 400 µL of 0.1% BSA PBS on 2020/12/23 at 250,000 ng/mL). The gibbon LCLs were treated with gibbon IFN-β at 100 ng/mL (Kingfisher Biotech Ref. RP1791GB-025 Lot. LU4443KU, resuspended in 100 µL of 0.1% BSA PBS on 2020/12/23 at 250,000 ng/mL). The rhesus LCLs were treated with rhesus IFN-β at 500 ng/mL (Kingfisher Biotech Ref. RP1709Y-025 Lot. CU4126BU, resuspended in 100 µL of 0.1% BSA PBS on 2021/05/09 at 250,000 ng/mL). The squirrel monkey LCLs were treated with squirrel monkey IFN-β at 5 ng/mL (Kingfisher Biotech Ref. RP1829SM-025 Lot. BV4549LU, resuspended in 100 µL of 0.1% BSA PBS on 2021/05/09 at 250,000 ng/mL). The cow LCLs were treated with cow IFN-β at 200 ng/mL (Kingfisher Biotech Ref. RP0298B-025

Lot. FO1566FL, resuspended in 100 µL of 0.1% BSA PBS on 2021/05/09 at 250,000 ng/mL). The chicken LCLs were treated with chicken IFN-β at 500 ng/mL (Kingfisher Biotech Ref. RP1786C-025 Lot. KU4407KU, resuspended in 100 µL of 0.1% BSA PBS on 2020/12/23 at 250,000 ng/mL). For the untreated BSA negative controls, each species LCLs were treated with an equal volume of BSA PBS (final 0.00004% similar to the IFN-β treatments). After the 3 hour IFN-β treatment incubations, 900 µL of RNA lysis buffer was added to the 48-well plate wells for a total volume of 1150 µL, and the plates were stored at -70°C until all 3 replicates were ready to be processed together. The 1st replicates were processed on 2021/05/28, the 2nd replicates were processed on 2021/05/29, and the 3rd replicates were processed on 2021/05/30. All cultures and treatments were processed in parallel.

3.5.14 RNA-seq library preparation for interspecies dataset

Total RNA was extracted using the Quick-RNA MiniPrep Plus (Zymo Research Ref. R1058) following the manufacturer's instructions. The RNA purity was determined using a Nanodrop with a 260 nm and 280 nm absorbance ratio ranging from 1.79 to 2.18 for all samples. The RNA-seq libraries were prepared using the KAPA mRNA HyperPrep Kit (Roche Ref. KK8581), KAPA mRNA Capture Kit (Roche Ref. KK8441), and KAPA Pure Beads (Roche Ref. KK8545); following the manufacturer's instructions (KR1352 – v7.21) using 250 ng of total RNA as input with an RNA fragmentation step of 6 minutes at 94°C, and using 12 cycles in the amplification step. The finalized libraries concentrations were obtained using the Qubit dsDNA high sensitivity assay kit (Invitrogen Ref. Q32851). All the 36 RNA-seq libraries were processed in parallel.

3.5.15 RNA-seq sequencing information for interspecies dataset

The 1st, 2nd, and 3rd replicates were pooled together and were sequenced on 2021/07/15 on a NovaSeq 6000 as paired-end 150 bp reads.

Table 3.4 describes the number of reads per RNA-seq library in the IFN interspecies dataset.

3.5.16 RNA-seq datasets processing for interspecies dataset

- Read quality was assessed using FastQC (v0.11.5).

- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r, qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tbo, tpe, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.

- Mapping was done using HISAT2 (v2.1.0) with options –very-sensitive on each species' respective reference genomes. The human reference genome hg38 was obtained from `GP/hg38/bigZips/hg38.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY. The gibbon reference genome nomLeu3 was obtained from `GP/nomLeu3/bigZips/nomLeu3.fa.gz` and was modified so that it only contained the main chromosome contigs chr1a, chr2, chr3, chr4, chr5, chr6, chr7b, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22a, chr23, chr24, chr25, chrX. The rhesus reference genome rheMac10 was obtained from `GP/rheMac10/bigZips/rheMac10.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chrM, chrX, chrY. The squirrel monkey reference genome saiBol1 was obtained from `GP/saiBol1/bigZips/saiBol1.fa.gz` and was modified so that it only contained the contigs numbered from JH378105 to JH378420, which were renamed as chr1 to chr316, respectively. The cow reference genome bosTau9 was obtained from `GP/bosTau9/bigZips/bosTau9.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26,

chr27, chr28, chr29, chrM, chrX. The chicken reference genome galGal6 was obtained from `GP/galGal6/bigZips/galGal6.fa.gz` and was modified so that it only contained the main chromosome contigs chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chr23, chr24, chr25, chr26, chr27, chr28, chr30, chr31, chr32, chr33, chrM, chrW, chrZ.

- SAM to BAM conversion was done using Samtools (v1.8).

- Bigwigs were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1, –normalizeUsing RPKM, –filterRNAstrand forward (for the positive strand file) or reverse (for the negative strand file), –scaleFactor 1 (for the positive strand file) or -1 (for the negative strand file).

- Read counts over genes were obtained using R (v3.6.0) Rsubread featureCounts (v1.32.4) with the options isGTFAnnotationFile=TRUE, useMetaFeatures=TRUE, GTF.featureType="exon", GTF.attrType="gene_id", allowMultiOverlap=TRUE, largestOverlap=TRUE, isPairedEnd=TRUE, strandSpecific=2; using each species GTF annotation file. The human GTF annotation file was obtained from `GP/hg38/bigZips/genes/hg38.ncbiRefSeq.gtf.gz`, the gibbon GTF annotation file was obtained from `GP/nomLeu3/bigZips/genes/nomLeu3.ensGene.gtf.gz`, the rhesus GTF annotation file was obtained from `GP/rheMac10/bigZips/genes/rheMac10.ncbiRefSeq.gtf.gz`, the squirrel monkey GTF annotation file was obtained from `GP/saiBol1/bigZips/genes/saiBol1.ensGene.gtf.gz` and was modified so that the contig names reflect the chrN names just as they were assigned in the genome FASTA file, the cow GTF annotation file was obtained from `GP/bosTau9/bigZips/genes/bosTau9.ncbiRefSeq.gtf.gz`, and the chicken GTF annotation file was obtained from `GP/galGal6/bigZips/genes/galGal6.ncbiRefSeq.gtf.gz`.

- Differential gene expression was done using DESeq2 (v1.26.0). Gene set enrichment was done using the GSEA GUI (v4.3.2) with the Human MSigDB Collections (v7.5.1).

- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the BAM files, and a BED file con-

taining the annotated gene TSSs that was obtained by further processing the above GTF files as follows, using the human hg38 annotation as an example: convert2bed –input=gtf –output=bed –do-not-sort < hg38.ncbiRefSeq.gtf > hg38.ncbiRefSeq.bed grep -w transcript hg38.ncbiRefSeq.bed | grep -v chr[0-9]*_ | cut -f1,2,3,4,5,6 | awk -v OFS='\ t' 'print $1, $2, $3, $4, $5, $6, $3-$2' > hg38.ncbiRefSeq.bed.tmp sort -nk7r hg38.ncbiRefSeq.bed.tmp | sort -u -k4,4 | awk -v OFS='\ t' 'print $1, $2, $3, $4, $5, $6' | sort -k 1,1 -k2,2n > hg38.ncbiRefSeq.oneEntry.bed awk -v OFS='\ t' 'if ($6 == "+") print $1,$2-1500,$2+1500,$4; if ($6 == "-") print $1,$3-1500,$3+1500,$4' hg38.ncbiRefSeq.oneEntry.bed > hg38.ncbiRefSeq.TSS.oneEntry.bed.tmp awk -v OFS='\t' ' if ($2 < 0) print $1,"0",$3,$4; else if ($2 > 0) print $0 ' hg38.ncbiRefSeq.TSS.oneEntry.bed.tmp > hg38.ncbiRefSeq.TSS.oneEntry.bed

3.5.17 B-cell immortalization to make DR LCL for intrahuman dataset

The human B-cell transformation with Epstein-Barr Virus (EBV) to obtain a Lymphoblastoid Cell Line (LCL) was done following a protocol provided by Renata Collard and Angela Rachubinski from the University of Colorado Anschutz Medical Campus, who also provided the cell line GM7404A that produces EBV. Briefly, 20 mL of human blood was drawn and peripheral blood mononuclear cells were separated from the blood plasma and red blood cells using Lymphoprep (Stemcell Technologies Ref. 07851) using the manufacturer's instructions. Next, 15 mL of Lymphoprep was placed on a 50 mL conical tube. 15 mL of whole blood was diluted with 15 mL 2% FBS PBS, carefully poured on top of the 15 mL of Lymphoprep, and centrifuged at 800 x g for 20 minutes at 20°C with brake-off. Removed and discarded the upper plasma layer. Carefully removed the mononuclear cell gray layer (approximately 10 mL), transferred to a new conical tube, and discarded the bottom erythrocyte and granulocyte layer. Washed twice the mononuclear cells by mixing them with 10 mL 2% FBS PBS and centrifuged at 300 x g for 10 minutes at 20°C with break on, and resuspended the pellet in 2 mL of RPMI-1640 (Invitrogen Ref. 23400-062) media. Added 1 mL of the resuspended human cells onto each of two standing T-25 flasks, with 2 mL of 10% FBS and 4 µg/mL of Cyclosporin A (Sigma-Aldrich Ref. C1832-5MG) RPMI, and with 1 mL

of the EBV-containing supernatant of the cell line GM7404A. The 2 T-25 flasks with 4 mL total volume were left undisturbed for 7 days at 37 C in a 5% CO2 incubator, and then added with 3 mL RPMI media made with 10% FBS, 100 U/mL penicillin, 100 µg/mL streptomycin, 250 ng/mL amphotericin B (Gibco Ref. 15240062). Another 7 days later, another 3 mL of the same RPMI mixture was added to the flasks and shaked the flasks. 7 days later 3 more mL of the same RPMI mixture were added. Small clumps reminiscent of LCL clumps started to appear. The flasks kept being shaken every other day. Approximately 1 month after the initial infection, the two T-25 flasks were combined onto a single standing T-75 flask with fresh 3 mL of the same RPMI media. The now transformed DR-LCLs continued to be cultured as the other LCLs. Three independent DR-LCL were generated, but only one used in the intrahuman panel.

3.5.18 Cell lines for the Human-Rhesus cis/trans experiment

Table 3.5 describes the information of the LCLs used to generate the cis vs trans IFN-α2 interspecies dataset; including the date they were received, the species, the ID, and the source.

3.5.19 Cell lines information for intrahuman dataset

Table 3.6 describes the information of the LCLs used to generate IFN intrahuman dataset; including the date they were received, the internal ID used in the lab, the official ID, the country of origin, their ethnicity, their biological sex, and the source.

3.5.20 PRO-seq and RNA-seq growth conditions for intrahuman dataset

The human LCLs were cultured in RPMI-1640 media (Gibco 72400-047) using 15% FBS (Gibco 10437-028) and 100 units/mL Penicillin-Streptomycin (Gibco 15140-122) in vent-cap T-25 flasks (Corning 430639), and kept at a confluency between 400,000 cells/mL to 800,000 cells/mL during cell culture at 37°C with 5% CO_2.

3.5.21 PRO-seq treatment conditions for intrahuman dataset

Each of the 11 human LCLs were treated for 1 hour prior to the nuclei isolation with human IFN-β at 100 ng/mL (Kingfisher Biotech Ref. RP1788H-100 Lot. KU4428KU), or with BSA as negative control for a final concentration 0.00004% BSA, similar to that of the IFN-β treatments. 3 T-25 cultures per LCLs were used per treatment. The 1st replicates were processed on 2020/12/25, 2nd replicates were processed on 2021/03/01, and the 3rd replicates (only Dave, Ethan, and Eric LCLs) were processed on 2022/06/19. All cultures and treatments were processed in parallel.

3.5.22 PRO-seq nuclei extraction for intrahuman dataset

Nuclei isolation was done as described in [44] with some modifications. Briefly, LCL cultures ranging from 3 to 25 million cells were used for each condition. After each culture was treated for 1 hour, the cultures were washed twice with ice-cold PBS. Then, the cell pellets were carefully resuspended in 6 mL of lysis buffer (0.1% DEPC-DI water with 10 mM Tris-HCl pH 7.4, 2 mM $MgCl_2$, 3 mM $CaCl_2$, 0.5% IGEPAL, 10% Glycerol, 1 mM DTT, Invitrogen Ref. AM2696 SUPERase-IN RNAse inhibitor, and with Roche Ref. 11836170001 protease inhibitor cocktail) and centrifuged for 15 minutes at 4°C at 1000 x g. The pellets were carefully resuspended in 1 mL lysis buffer using Finntip wide orifice pipette tips (Thermo Scientific Ref. 9405163), were mixed with 4 mL more of lysis buffer, and centrifuged a second time for 15 minutes at 4°C at 1000 x g. The pellets were carefully resuspended a second time in 1 mL lysis buffer using Finntip wide orifice pipette tips, transferred to low binding 1.7 mL eppendorf tubes (Costar Ref. 3207), and centrifuged for 5 minutes at 4°C at 1000 x g. The pellets were carefully resuspended using Finntip wide orifice pipette tips in 500 µL freezing buffer (0.1% DEPC-DI water with 50 mM Tris-HCl pH 8.0, 5 mM $MgCl_2$, 40% Glycerol, 0.1 mM EDTA pH 8.0, and SUPERase-IN RNAse inhibitor), and centrifuged for 2 minutes at 4°C at 2000 x g. The resulting nuclei pellets were resuspended a final time in 110 µL of freezing buffer using Finntip wide orifice pipette tips. I mixed 10 µL of the resuspended nuclei with 990 µL of PBS for counting the nuclei yield. The remaining 100 µL resuspended nuclei were

snap-frozen in liquid nitrogen and stored at -70°C before being used for the PRO-seq nuclear-run on reactions.

3.5.23 PRO-seq library preparation for intrahuman dataset

PRO-seq datasets were prepared as described in [60], which in turn is a modified protocol from [122]. Briefly, between 2 to 18 million nuclei per dataset were used for the PRO-seq transcription run-on using a mixture of rNTP and Biotin-11-CTP (Biotin-11-CTP at 0.025 mM from PerkinElmer Ref. NEL542001EA; rCTP at 0.025 mM from Promega Ref. E604B, rATP at 0.125 mM Ref. E601B, rGTP at 0.125 mM Ref. E603B, and rUTP at 0.125 mM Ref. E6021). 1% of S2 Drosophila melanogaster nuclei relative to the number of the sample nuclei were added during the run-on reaction as a normalization spike-in. Total RNA was extracted using a phenol/chloroform precipitation. Isolated RNA was fragmented using base hydrolysis with NaOH. Biotinylated fragmented nascent transcripts were isolated using a first streptavidin Dynabeads M-280 (Invitrogen Ref. 11206D) pull down, and the VRA3 RNA adaptor was ligated at their 3' end. A second streptavidin bead pull down was performed, followed by the enzymatic modifications of the RNA fragment 5' ends with a pyrophosphohydrolase and a polynucleotide kinase, and the VRA5 RNA adaptor was ligated at their fixed 5' ends. A third streptavidin bead pull down was performed, followed by the reverse transcription of the resulting adaptor-ligated libraries. The libraries were cleaned up with AMPure XP beads (Beckman Coulter Ref. A63881). Then, the libraries were amplified using 13 PCR cycles, and cleaned up again with another round of AMPure XP beads. The resulting library concentrations were measured with the Qubit dsDNA high sensitivity assay (Invitrogen Ref. Q32851), and their size distributions assessed using the Agilent High Sensitivity D1000 ScreenTape. The 1st replicates were processed on 2021/01/01, the 2nd replicates were processed on 2021/03/02, the 3rd replicates (only Dave, Ethan, and Eric LCLs) were processed on 2022/06/21

3.5.24 PRO-seq sequencing information for intrahuman dataset

The 1st PRO-seq replicates were pooled and sequenced on two consecutive days, 2021/01/11 and 2021/01/12 to get sufficient sequencing depth. The 1st PRO-seq replicates were pooled and sequenced on two consecutive days, 2021/05/10 and 2021/05/11 to get sufficient sequencing depth. Both 1st and 2nd replicates were sequenced using a NextSeq 500. Base calls and demultiplexing was done using Bcl2Fastq2 (v2.2.0). The FASTQ files sequenced on the sequential dates were concatenated. The 3rd PRO-seq replicates for Dave, Ethan, and Eric were sequenced on 2022/10/13 on a NextSeq 2000. All datasets were sequenced as single-end 76 bp long reads.

Table 3.7 describes the number of reads per PRO-seq library in the IFN intrahuman dataset.

3.5.25 PRO-seq datasets processing for intrahuman dataset

- PRO-seq datasets were processed using the Nextflow pipeline found in `https://github.com/Dowell-Lab/Nascent-Flow`.
- Read quality was assessed using FastQC (v0.11.5)
- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r, qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.
- Mapping was done using HISAT2 (v2.1.0) with options –very-sensitive –no-spliced-alignment on each species' respective reference genomes. The human reference genome hs1 was obtained from `GP/hs1/bigZips/hs1.fa.gz`.
- SAM to BAM conversion was done using Samtools (v1.8).
- Bigwigs were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1 –normalizeUsing RPKM –filterRNAstrand reverse (for pos file) or forward (for neg file)

–scaleFactor 1 (for pos file) or -1 (for neg file).

- Bidirectional loci were determined using Tfit and dREG as described in the Nextflow pipeline https://github.com/Dowell-Lab/Bidirectional-Flow. It removes multimapped reads using Samtools (v1.8) view -h -q 1 'bam file' | grep -P '(NH:i:1| ∧@)' | samtools view -h -b. Tfit calls were obtained by first using the Tfit bidir module to call prelim regions. The annotation was used to add 3 kb-wide TSS regions to the prelim file and removed any part of the prelim regions that overlap with the TSS regions. Prelim regions > 10 kb were then fragmented down to equal size regions (< 10kb) with 50% overlap and then coverage filtered to keep prelim regions having > 9 mapped reads. Finally, the adjusted prelim regions were used as regions of interest to the Tfit model module to obtain Tfit calls. dREG calls were filtered as having FDR < 0.05, merged if within 20bp of each other, and having

  >

  9 mapped reads. Bidirectional transcription calls were combined using muMerge (v1.1.0) across experimental conditions.

- Read counts over genes were obtained using R (v3.6.0) Rsubread featureCounts (v1.32.4) with the options isGTFAnnotationFile=FALSE, useMetaFeatures=TRUE, allowMultiOverlap=TRUE, largestOverlap=TRUE, isPairedEnd=FALSE, strandSpecific=1; using the multimapped reads filtered BAM files; and using a custom SAF file that contains the longest annotated entry per gene from the RefSeq annotation, without the initial 25% genic region starting from the 5' end to remove the RNA polymerase pausing region. The specific steps to produce this SAF file are as follows for the human hs1 annotation: wget `GP/hs1/bigZips/genes/hs1.110.20220412.ncbiRefSeq.gtf.gz`, and modified so that the chromosome names were displayed with the UCSC nomenclature (e.g. chr1) and now with its default Genbank nomenclature (e.g. CP068277.2) using the file found in `GP/hs1/bigZips/hs1.chromAlias.txt`. convert2bed –input=gtf –output=bed –do-not-sort < hs1.110.20220412.ncbiRefSeq.gtf >
  hs1.110.20220412.ncbiRefSeq.bed grep -w transcript hs1.110.20220412.ncbiRefSeq.bed | grep

-v chr[0-9]*_ | cut -f1,2,3,4,5,6 | awk -v OFS='\t' 'print $1, $2, $3, $4, $5, $6, $3-$2' > hs1.110.20220412.ncbiRefSeq.bed.tmp sort -nk7r hs1.110.20220412.ncbiRefSeq.bed.tmp | sort -u -k4,4 | awk -v OFS='\t' 'print $1, $2, $3, $4, $5, $6' | sort -k 1,1 -k2,2n > hs1.110.20220412.ncbiRefSeq.oneEntry.bed awk -v OFS='\t' ' print $4, $1, $2, $3, $6' hs1.110.20220412.ncbiRefSeq.oneEntry.bed
> hs1.110.20220412.ncbiRefSeq.oneEntry.saf awk -v OFS='\t' ' if ($5 == "+") printf "%s\t%s\t%.0f\t%.0f\t%s\n", $1, $2, $3+(($4-$3)*0.25), $4, $5; else print $0 ' hs1.110.20220412.ncbiRefSeq.oneEntry.saf >
hs1.110.20220412.ncbiRefSeq.without5prime25.oneEntry.saf.tmp
awk -v OFS='\t' '{ if ($5 == "-") printf "%s\t%s\t%.0f\t%.0f\t%s\n", $1, $2, $3, $4-(($4-$3)*0.25), $5; else print $0 }' hs1.110.20220412.ncbiRefSeq.without5prime25.oneEntry.saf.tmp >
hs1.110.20220412.ncbiRefSeq.without5prime25.oneEntry.saf

- Differential gene expression was done using DESeq2 (v1.26.0). Gene set enrichment was done using the GSEA GUI (v4.3.2) with the Human MSigDB Collections (v7.5.1).

- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the multimapped reads filtered BAM files and the muMerged Tfit or dREG bidirectionals.

3.5.26 RNA-seq treatment conditions for intrahuman dataset

Each of the 11 human LCLs were treated for 3 hours with either human IFN-β (Kingfisher Biotech Ref. RP1788H-025 Lot. KU4427KU) at 100 ng/mL, or as negative controls with 0.00004% BSA similar to the IFN-β treatments. On the day of the treatments, each LCL was moved onto separate 48-well plate wells, each with 125,000 cells/well and left incubating in a total volume of 250 µL after the IFN-β or BSA addition. After the 3 hour IFN-β treatment incubations, 1 mL of RNA lysis buffer was added to the 48-well plate wells for a total volume of 1250 µL, and the plates were stored at -70°C until all 3 replicates were ready to be processed together. The 1st replicates were processed on 2020/12/16, the 2nd replicates were processed on 2020/12/18, and the 3rd replicates were processed on 2020/12/20. All cultures and treatments were processed in parallel.

3.5.27 RNA-seq library preparation for intrahuman dataset

Total RNA was extracted using the Quick-RNA MiniPrep Plus (Zymo Research Ref. R1058) following the manufacturer's instructions, and the RNA concentrations were measured using a Qubit HS RNA kit, yielding concentrations ranging from 2 ng/µL to 12 ng/µL. The RNA-seq libraries were prepared using the KAPA mRNA HyperPrep Kit (Roche Ref. KK8581), KAPA mRNA Capture Kit (Roche Ref. KK8441), and KAPA Pure Beads (Roche Ref. KK8545); following the manufacturer's instructions (KR1352 – v7.21) using 250 ng of total RNA from most samples (though a few with low concentration had only 150-100 ng) as input with an RNA fragmentation step of 6 minutes at 94°C, and using 11 cycles in the amplification step for the samples that had 250 ng of input RNA or 12-14 cycles for those samples with less input RNA. The finalized libraries concentrations were obtained using the Qubit dsDNA high sensitivity assay kit (Invitrogen Ref. Q32851), with final concentrations ranging from 2 ng/µL to 21 ng/µL. The 1st and 2nd replicates were processed in parallel on 2020/12/28, and the 3rd replicates were processed in parallel on 2022/12/29.

3.5.28 RNA-seq sequencing information for intrahuman dataset

The 1st, 2nd, and 3rd replicates were pooled together and were sequenced on 2021/01/26 on a NovaSeq 6000 as paired-end 150 bp long reads.

Table 3.8 describes the number of reads per RNA-seq library in the IFN intrahuman dataset.

3.5.29 RNA-seq datasets processing for intrahuman dataset

- Read quality was assessed using FastQC (v0.11.5).
- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r, qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tbo, tpe, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.

- Mapping was done using HISAT2 (v2.1.0) with options –very-sensitive on the human reference genome hs1, which was obtained from `https://hgdownload.soe.ucsc.edu/goldenPath/hs1/bigZips/hs1.fa.gz`.

- SAM to BAM conversion was done using Samtools (v1.8).

- Bigwigs were obtained using deepTools (v3.0.1) bamCoverage with options –binSize 1, –normalizeUsing RPKM, –filterRNAstrand forward (for the positive strand file) or reverse (for the negative strand file), –scaleFactor 1 (for the positive strand file) or -1 (for the negative strand file).

- Read counts over genes were obtained using R (v3.6.0) Rsubread featureCounts (v1.32.4) with the options isGTFAnnotationFile=TRUE, useMetaFeatures=TRUE, GTF.featureType="exon", GTF.attrType="gene_id", allowMultiOverlap=TRUE, largestOverlap=TRUE, isPairedEnd=TRUE, strandSpecific=2. The human hs1 GTF annotation file was obtained from `GP/hs1/bigZips/genes/hs1.110.20220412.ncbiRefSeq.gtf.gz`. And modified so that the chromosome names were displayed with the UCSC nomenclature (e.g. chr1) and now with its default Genbank nomenclature (e.g. CP068277.2) using the file found in `GP/hs1/bigZips/hs1.chromAlias.txt`.

- Differential gene expression was done using DESeq2 (v1.26.0). Gene set enrichment was done using the GSEA GUI (v4.3.2) with the Human MSigDB Collections (v7.5.1).

- Determined transcription factor activity with TFEA (v1.1.4) using the transcription factor motif database JASPAR2022_CORE_vertebrates_non-redundant using the BAM files, and a BED file containing the annotated gene TSSs that was obtained by further processing the above GTF files as follows: convert2bed –input=gtf –output=bed –do-not-sort < hs1.110.20220412.ncbiRefSeq.gtf > hs1.110.20220412.ncbiRefSeq.bed grep -w transcript hs1.110.20220412.ncbiRefSeq.bed | grep -v chr[0-9]*_ | cut -f1,2,3,4,5,6 | awk -v OFS='\t' ' print $1, $2, $3, $4, $5, $6, $3-$2 '

> hs1.110.20220412.ncbiRefSeq.bed.tmp sort -nk7r hs1.110.20220412.ncbiRefSeq.bed.tmp | sort -u -k4,4 | awk -v OFS='\t' ' print $1, $2, $3, $4, $5, $6 ' | sort -k 1,1 -k2,2n > hs1.110.20220412.ncbiRefSeq.oneEntry.bed awk -v OFS='\t' 'if ($6 == "+") print $1,$2-1500,$2+1500,$4; if ($6 == "-") print $1,$3-1500,$3+1500,$4 ' hs1.110.20220412.ncbiRefSeq.oneEntry.bed > hs1.110.20220412.ncbiRefSeq.TSS.oneEntry.bed.tmp awk -v OFS='\t' ' if ($2 < 0) print $1,"0",$3,$4; else if ($2 > 0) print $0 ' hs1.110.20220412.ncbiRefSeq.TSS.oneEntry.bed.tmp > hs1.110.20220412.ncbiRefSeq.TSS.oneEntry.bed

3.6 Data availability

The sequencing datasets described here were deposited to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database.

The PRO-seq datasets for the human and rhesus LCLs treated with their cognate species-matching (cis) and the other species (trans) IFN-α2 were deposited under the GEO accession number GSE214304.

The PRO-seq and RNA-seq datasets for the interspecies (human, gibbon, rhesus, squirrel monkey, cow, and chicken) LCLs treated with their cognate species-matching IFN-β were deposited under the GEO accession number GSE217402, with the PRO-seq datasets having the series GSE217400, and the RNA-seq datasets having the series GSE217401.

The PRO-seq and RNA-seq datasets for the 8 diverse ethnic humans LCLs treated with human IFN-β were deposited under the GEO accession number GSE217313, with the PRO-seq datasets having the series GSE217294, and the RNA-seq datasets having the series GSE217302.

Table 3.4: Sequencing depth of the IFN RNA-seq interspecies datasets used in chapter 3

Dataset	Read number	Dataset	Read number
RNA-BSA-Human-1	28,346,504	RNA-IFN-Human-1	30,435,526
RNA-BSA-Human-2	24,407,796	RNA-IFN-Human-2	25,358,343
RNA-BSA-Human-3	28,983,129	RNA-IFN-Human-3	25,318,962
RNA-BSA-Gibbon-1	37,448,641	RNA-IFN-Gibbon-1	24,443,450
RNA-BSA-Gibbon-2	30,419,688	RNA-IFN-Gibbon-2	29,256,992
RNA-BSA-Gibbon-3	27,011,639	RNA-IFN-Gibbon-3	31,700,258
RNA-BSA-Rhesus-1	27,540,029	RNA-IFN-Rhesus-1	31,071,451
RNA-BSA-Rhesus-2	28,621,658	RNA-IFN-Rhesus-2	25,343,848
RNA-BSA-Rhesus-3	30,597,594	RNA-IFN-Rhesus-3	30,152,284
RNA-BSA-SquirrelMonkey-1	30,778,491	RNA-IFN-SquirrelMonkey-1	33,219,336
RNA-BSA-SquirrelMonkey-2	28,488,860	RNA-IFN-SquirrelMonkey-2	29,335,812
RNA-BSA-SquirrelMonkey-3	26,761,431	RNA-IFN-SquirrelMonkey-3	32,819,537
RNA-BSA-Cow-1	30,373,475	RNA-IFN-Cow-1	26,840,215
RNA-BSA-Cow-2	27,395,719	RNA-IFN-Cow-2	27,660,767
RNA-BSA-Cow-3	31,288,569	RNA-IFN-Cow-3	29,126,070
RNA-BSA-Chicken-1	28,279,889	RNA-IFN-Chicken-1	27,347,731
RNA-BSA-Chicken-2	28,293,824	RNA-IFN-Chicken-2	27,875,930
RNA-BSA-Chicken-3	30,981,295	RNA-IFN-Chicken-3	34,915,447

Table 3.5: Cell lines information of the IFN cis vs trans dataset used in chapter 3

Received	Species	ID	Source
2019/01/29	Human-F	GM12878	Coriell / NIGMS
2019/07/23	Human-M	HG03077	Coriell / NHGRI
2019/10/04	Rhesus-F	Mm 150-99	Yoav Gilad Lab
2019/10/04	Rhesus-M	Mm 290-96	Yoav Gilad Lab

Table 3.6: Cell lines information of the IFN intrahuman dataset used in chapter 3

Received	Internal ID	Official ID	Country	Ethnicity	Sex	Source
2019/01/29	Ursula	GM12878	United States	Caucasian	F	Coriell / NIGMS
NA	DR	NA	Mexico	NA	M	NA
2019/07/23	Sengbe	HG03077	Sierra Leone	Mende	M	Coriell / NHGRI
2019/07/23	Khaondo	GM19024	Kenya	Luhya	F	Coriell / NHGRI
2019/07/23	Niyilolawa	GM18489	Nigeria	Yoruba	F	Coriell / NHGRI
2019/07/23	Pedro	HG02150	Peru	Peruvian	M	Coriell / NHGRI
2019/07/23	Srivathani	HG03645	Sri Lanka	Tamil	F	Coriell / NHGRI
2019/07/23	ChenChao	GM18530	China	Han	M	Coriell / NHGRI
NA	Eric (D21)	NA	NA	NA	M	Nexus Biobank
NA	Ethan (T21)	NA	NA	NA	M	Nexus Biobank
NA	Dave (D21)	NA	NA	NA	M	Nexus Biobank

Table 3.7: Sequencing depth of the IFN PRO-seq intrahuman datasets used in chapter 3

Dataset	Read number	Dataset	Read number
PRO-BSA-Ursula-1	42,195,917	PRO-IFNB-Ursula-1	41,808,292
PRO-BSA-Ursula-2	46,092,959	PRO-IFNB-Ursula-2	47,377,346
PRO-BSA-DR-1	41,457,652	PRO-IFNB-DR-1	42,705,381
PRO-BSA-DR-2	53,450,928	PRO-IFNB-DR-2	43,533,364
PRO-BSA-Sengbe-1	42,172,382	PRO-IFNB-Sengbe-1	41,928,885
PRO-BSA-Sengbe-2	38,948,852	PRO-IFNB-Sengbe-2	43,193,854
PRO-BSA-Khaondo-1	42,423,753	PRO-IFNB-Khaondo-1	41,437,714
PRO-BSA-Khaondo-2	42,875,523	PRO-IFNB-Khaondo-2	46,542,139
PRO-BSA-Niyilolawa-1	41,631,139	PRO-IFNB-Niyilolawa-1	41,501,879
PRO-BSA-Niyilolawa-2	46,611,255	PRO-IFNB-Niyilolawa-2	40,392,956
PRO-BSA-Pedro-1	41,821,224	PRO-IFNB-Pedro-1	42,071,170
PRO-BSA-Pedro-2	42,975,706	PRO-IFNB-Pedro-2	45,989,703
PRO-BSA-Srivathani-1	41,795,630	PRO-IFNB-Srivathani-1	42,054,804
PRO-BSA-Srivathani-2	47,045,002	PRO-IFNB-Srivathani-2	41,652,259
PRO-BSA-ChenChao-1	42,002,990	PRO-IFNB-ChenChao-1	42,058,600
PRO-BSA-ChenChao-2	46,360,142	PRO-IFNB-ChenChao-2	46,200,586
PRO-BSA-Dave-1	41,194,471	PRO-IFNB-Dave-1	42,453,894
PRO-BSA-Dave-2	44,249,637	PRO-IFNB-Dave-2	40,934,032
PRO-BSA-Dave-3	35,464,522	PRO-IFN-Dave-3	31,737,744
PRO-BSA-Eric-1	39,029,604	PRO-IFNB-Eric-1	42,101,766
PRO-BSA-Eric-2	40,696,392	PRO-IFNB-Eric-2	46,652,910
PRO-BSA-Eric-3	28,547,374	PRO-IFN-Eric-3	28,787,496
PRO-BSA-Ethan-1	41,596,634	PRO-IFNB-Ethan-1	41,620,402
PRO-BSA-Ethan-2	40,623,819	PRO-IFNB-Ethan-2	43,416,781
PRO-BSA-Ethan-3	28,639,578	PRO-IFN-Ethan-3	33,508,006

Table 3.8: Sequencing depth of the IFN RNA-seq intrahuman datasets used in chapter 3

Dataset	Read number	Dataset	Read number
RNA-BSA-Ursula-1	31,909,634	RNA-IFN-Ursula-1	40,162,379
RNA-BSA-Ursula-2	36,552,450	RNA-IFN-Ursula-2	30,464,008
RNA-BSA-Ursula-3	35,459,008	RNA-IFN-Ursula-3	32,786,917
RNA-BSA-DR-1	35,955,569	RNA-IFN-DR-1	31,248,938
RNA-BSA-DR-2	36,194,499	RNA-IFN-DR-2	32,219,991
RNA-BSA-DR-3	29,918,724	RNA-IFN-DR-3	29,311,192
RNA-BSA-Sengbe-1	34,534,303	RNA-IFN-Sengbe-1	36,101,570
RNA-BSA-Sengbe-2	36,240,125	RNA-IFN-Sengbe-2	39,571,943
RNA-BSA-Sengbe-3	26,787,072	RNA-IFN-Sengbe-3	25,737,414
RNA-BSA-Khaondo-1	38,882,432	RNA-IFN-Khaondo-1	33,931,715
RNA-BSA-Khaondo-2	42,343,160	RNA-IFN-Khaondo-2	48,990,204
RNA-BSA-Khaondo-3	25,064,983	RNA-IFN-Khaondo-3	32,175,085
RNA-BSA-Niyilolawa-1	36,792,381	RNA-IFN-Niyilolawa-1	49,718,713
RNA-BSA-Niyilolawa-2	36,034,857	RNA-IFN-Niyilolawa-2	46,966,788
RNA-BSA-Niyilolawa-3	25,716,202	RNA-IFN-Niyilolawa-3	33,621,764
RNA-BSA-Pedro-1	39,598,373	RNA-IFN-Pedro-1	33,245,218
RNA-BSA-Pedro-2	34,300,383	RNA-IFN-Pedro-2	34,119,312
RNA-BSA-Pedro-3	30,474,367	RNA-IFN-Pedro-3	24,924,245
RNA-BSA-Srivathani-1	31,412,960	RNA-IFN-Srivathani-1	42,189,675
RNA-BSA-Srivathani-2	34,607,358	RNA-IFN-Srivathani-2	28,661,148
RNA-BSA-Srivathani-3	33,511,863	RNA-IFN-Srivathani-3	31,859,262
RNA-BSA-ChenChao-1	26,505,929	RNA-IFN-ChenChao-1	37,207,581
RNA-BSA-ChenChao-2	33,966,765	RNA-IFN-ChenChao-2	31,794,771
RNA-BSA-ChenChao-3	30,666,149	RNA-IFN-ChenChao-3	33,286,304
RNA-BSA-Dave-1	36,687,953	RNA-IFN-Dave-1	23,750,002
RNA-BSA-Dave-2	49,676,332	RNA-IFN-Dave-2	34,934,265
RNA-BSA-Dave-3	31,907,175	RNA-IFN-Dave-3	39,546,353
RNA-BSA-Eric-1	34,468,542	RNA-IFN-Eric-1	38,176,754
RNA-BSA-Eric-2	38,251,062	RNA-IFN-Eric-2	39,738,019
RNA-BSA-Eric-3	34,477,884	RNA-IFN-Eric-3	32,169,657
RNA-BSA-Ethan-1	26,222,595	RNA-IFN-Ethan-1	27,211,017
RNA-BSA-Ethan-2	47,751,448	RNA-IFN-Ethan-2	41,846,890
RNA-BSA-Ethan-3	36,767,133	RNA-IFN-Ethan-3	43,914,436

Chapter 4

The role of the microenvironment and mechanosensing on nucleus chromatin

This work complements the efforts published as the two following research articles:

Walker, C.J., Crocini, C., Ramirez, D. *et al.* Nuclear mechanosensing drives chromatin remodelling in persistently activated fibroblasts. *Nature Biomedical Engineering* 5, 1485–1499 (2021). https://doi.org/10.1038/s41551-021-00709-w

Walker, C.J.,Batan, D., Bishop, C.T., Ramirez, D., *et al.* Extracellular matrix stiffness controls cardiac valve myofibroblast activation through epigenetic remodeling. *BioEngineering and Translational Medicine* 22;7(3):e10394 (2022). https://doi.org/10.1002/btm2.10394

4.1 Introduction

Cells are capable of sensing their immediate microenvironment where they live and grow, and they react accordingly. Bacterial cells are able to sense the stiffness of where they form their colonies [25, 24], and also the number of cells in their colonies. In response they modify their gene expression and therefore their physiology and metabolism, with direct consequences in their pathogenicity potential. Eukaryotic cells also have been observed to react to changes in their microenvironment. For example, cancer cells are thought to react to changes in their immediate extracellular matrix as they grow, experiencing also changes in their oxygen availability, and exploit these changes to

transition into blood-vessel circulating metastatic cells [54, 146]. Mammalian heart valve fibroblast cells are also thought to react to changes in the stiffness in their extracellular matrix after muscle injury, causing them to become activated, which in turn produce a stiffer matrix exacerbating the problem, a phenotype that can lead to heart failure [195].

In collaboration with Cierra Walker, Claudia Crocini, *et al* at the University of Colorado Boulder, I found that pig primary valve interstitial myocardial cells (ssVICs) display a marked change in genome-wide DNA accessibility due to differences in chromatin condensation when ssVICs are detached using trypsin from their growth substrate, a microgel with similar stiffness to heart tissue, relative to when cells are left attached. To directly observe genome-wide changes in DNA accessibility I developed a modified bulk ATAC-seq that is performed in situ (see Methods), instead of relying on detaching the cells beforehand as the original protocol instructs [42].

4.2 Experimental system

I decided to further investigate these genome-wide changes in DNA accessibility by testing the trypsin detachment perturbation on three additional conditions. As the original conditions were ssVICs grown in hydrogel, I tested the effects of trypsin-induced detachment on ssVICs grown in lab-standard polystyrene plastic substrate. In addition, I tested the trypsin-induced detachment on human induced pluripotent stem cells (cell line WTC11) to see if a non-pig non-fibroblast cell type would also react similarly. Finally, because trypsin induces dissociation of substrate-attached cells with its proteolytic cleavage of serines found in membrane proteins [148], I assessed the effect of the trypsin proteolysis in a context where there is no cell detachment, namely by exposing to trypsin non-adherent human suspension cells (the lymphoblastoid cell line GM12878).

When comparing trypsinized with non-trypsinized cells with the modified ATAC-seq procedure, another difference is the amount of Tn5 transposase used during the transposition reaction: Detached cells are processed in a smaller volume in an eppendorf tube, whereas attached cells are processed in a bigger volume in their plastic wells. This difference in volume arises because in order to expose a similar amount of cells in both conditions (attached versus detached), a bigger

volume is needed to cover the entirety of the plastic wells, and a bigger volume entails greater amount of transposase is necessary so that the overall concentration remains similar. However, I was concerned that this difference in enzyme amount may have a confounding impact on the accessibility outcomes, and therefore I introduced an additional sample to test this possibility. For the ssVICs, I tested cells attached to their plastic wells using the Well Enzyme Concentration (WEC), trypsin-induced detached cells using Tube Enzyme Concentration (TEC), an in addition a sample with trypsin-induced detached cells using WEC.

To correctly account for genome-wide changes in DNA accessibility through ATAC-seq, which is approximated by a number of DNA-sequencing reads derived from either the pig or human cells, I added a fixed amount of Drosophila melanogaster S2 nuclei to each sample during the transposition reactions. Because the S2 nuclei did not experience the putative effects of trypsin but did get targeted by the transposase, they served as a reliable point of reference to which the proportion of obtained sequence reads from pig or human can be compared to across the detached and attached conditions.

4.3 Results

I relied on two orthogonal approaches to observe changes in genome-wide DNA accessibility due to differences in chromatin condensation between the two experimental conditions, with or without trypsin treatment.

The first approach takes advantage of a common quality metric for ChIP-seq and ATAC-seq datasets that the ENCODE project has proposed called the Fraction of Reads In Peaks (FRIP) score [107]. The FRIP score measures the proportion of mapped reads that are mapped in peak regions, with the peak regions themselves defined through a bioinformatic tool such as the peak caller MACS2. For ATAC-seq datasets, the higher the FRIP score signifies that the transposase accessed the same regions throughout the bulk cell population without introducing “noise” signal (i.e. fewer reads mapped outside well defined peaks); and a low score can be interpreted as the transposase having a greater fraction of the genome accessible to transpose to, yielding spotty reads

throughout the genome.

The second approach uses the spiked-in Drosophila melanogaster S2 nuclei. If two samples are being compared, one very accessible genome-wide and the other very inaccessible, from which an observer is trying to assess differences in accessibility, without a reference spike-in both samples may appear very similar to each other by using ATAC-seq alone even when obtaining similar sequencing depths. On the other hand, when adding a spike-in different genome into both samples, I can assume that the transposase should be equally able to transpose into the added genomes in both reactions, and differences in the ratio between the reads mapped to the sample genome relative to the added genome can be interpreted as differences in the capabilities of the transpose to introduce itself into the sample genomes, which can only be explained by differences in DNA accessibility between the two samples.

After mapping the ATAC-seq reads to either the human hg38 or pig susScr11 reference genomes, I observed that the FRIP scores between the trypsin-treated and untreated samples were different between the ssVICs and WTC11 samples, whereas they remained similar between the two trypsin conditions in the suspension GM12878 cells (Figure 4.1). Moreover, I observed that the fraction of mapped reads to the spike-in Drosophila melanogaster S2 nuclei were also different between the two trypsin conditions, but relatively unchanged in the negative control GM12878 samples (Figure 4.2).

However, the observed patterns seem to be in different directions between the ssVICs and WTC11 samples. For the pig ssVICs samples, having a lower FRIP score and a lower spike-in mapped fraction upon trypsin-induced detachment from their plastic substrate, are both in agreement of their DNA becoming much more accessible genome-wide evidenced by having absorbed a greater proportion of the sequencing reads towards the pig genome but corresponding to less defined loci (i.e. regulatory elements) and being found throughout the genome (Figure 4.3). In contrast, for the human iPSC WTC11 samples, the trypsin-induced detachment is followed by both a greater FRIP score but a decrease in the S2 spike-in mapped fraction, which is harder to interpret. Upon detachment, the WTC11 cells may undergo a chromatin reconfiguration in a

Figure 4.1: FRIP score across the ATAC-seq datasets. Darker colors refer to the samples not treated with trypsin (still in suspension for GM12878 (blue), or attached to plastic for WTC11 (red) and ssVICs (yellow)), lighter colors refer to samples treated with trypsin (still in suspension for GM12878, or detached for WTC11 and ssVICs)

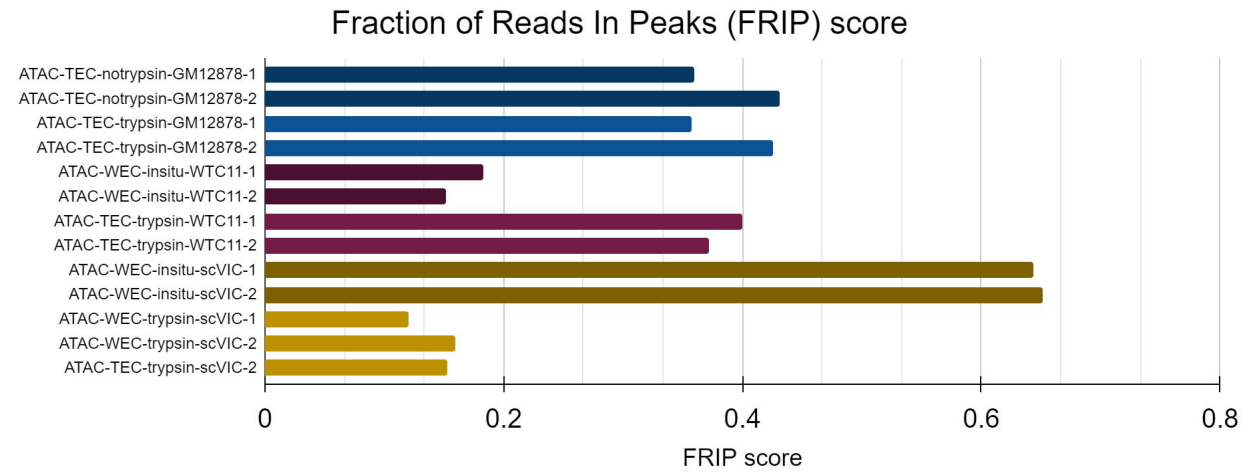
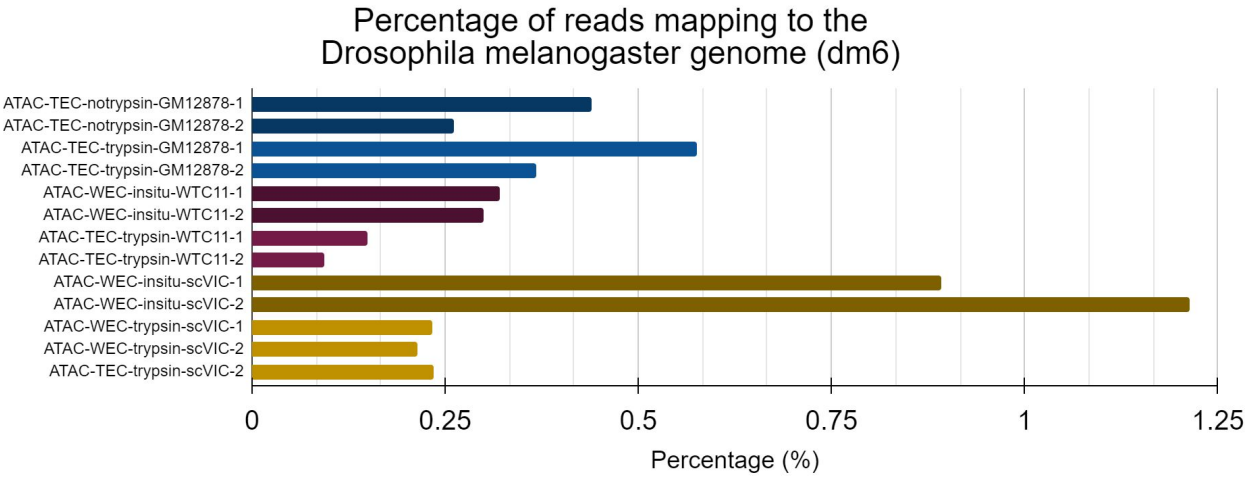
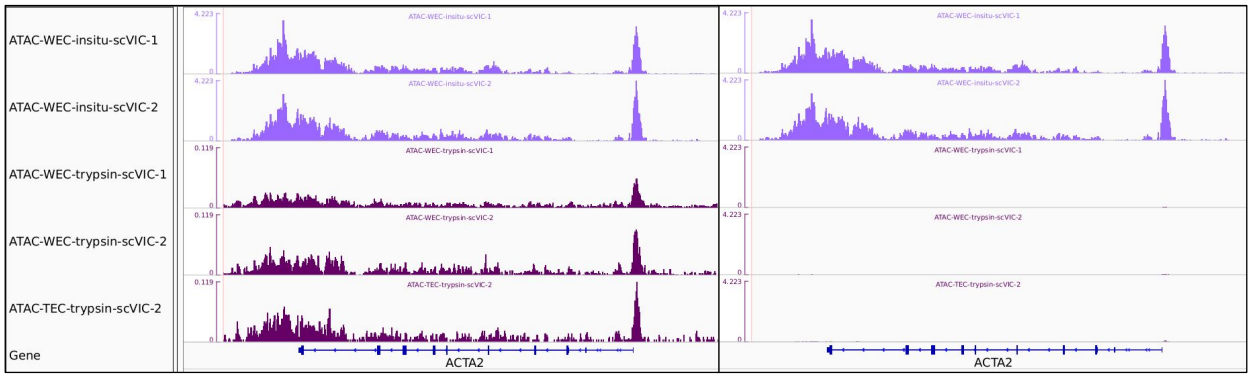


Figure 4.2: Percentage of total reads per dataset that mapped to the Drosophila melanogaster dm6 reference genome. Darker colors refer to the samples not treated with trypsin (still suspended for GM12878 (blue), or attached to plastic for WTC11(red) and ssVICs(yellow)), lighter colors refer to samples treated with trypsin (still in suspension for GM12878, or detached for WTC11 and ssVICs).



way that becomes more accessible (and therefore fewer reads are pooled from the S2 nuclei), but the chromatin decondensation is not aleatory throughout the genome, but localized in or nearby preexisting peaks in the cell population. These changes can be appreciated by looking at the genomic read coverage (Figure 4.4 top) in that the trypsinized samples show much less noise signal relative to high peaks at promoter regions.

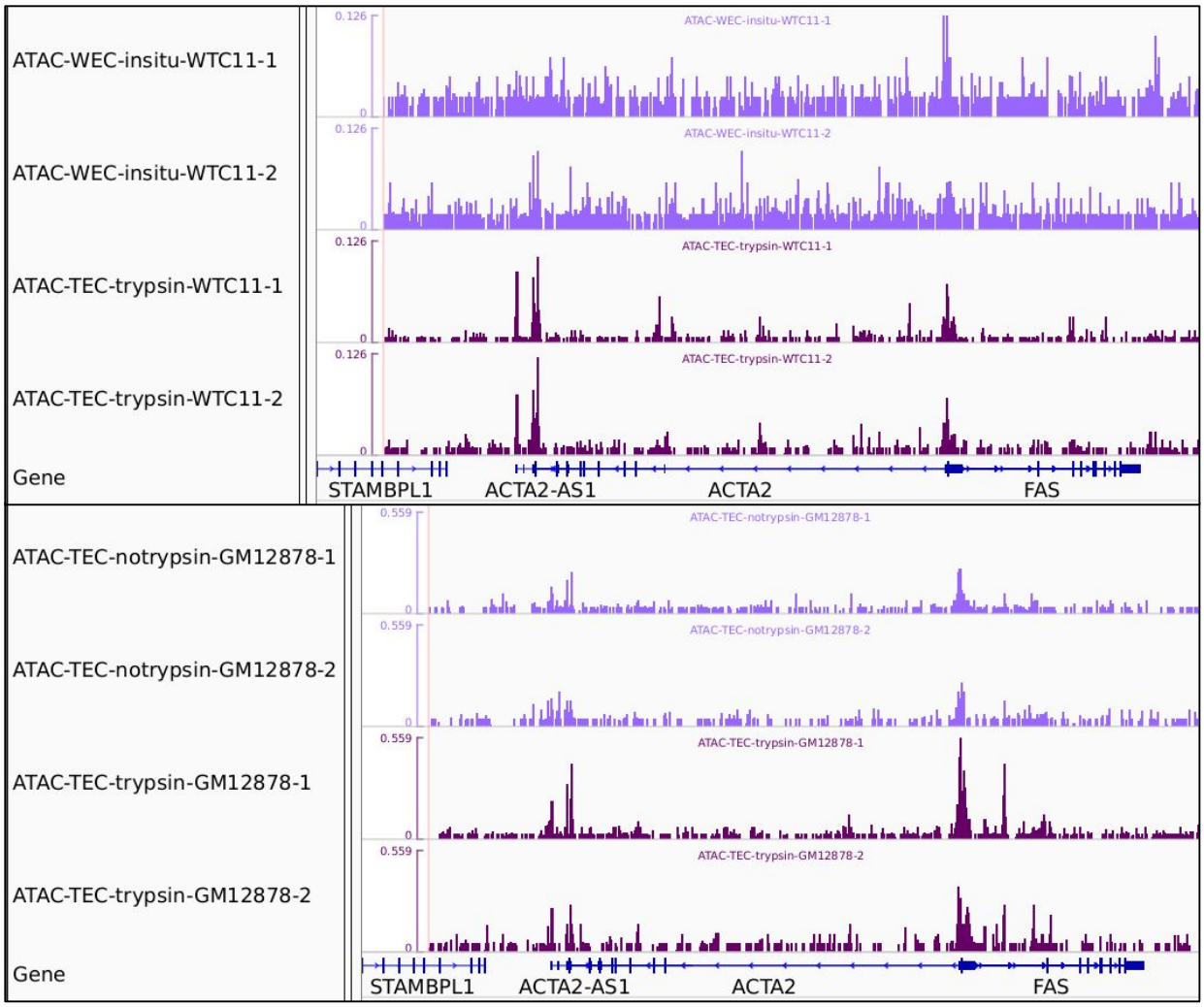
Figure 4.3: IGV genome browser displaying the ATAC-seq signal coverage over the ACTA2 gene for the ssVICs samples. In light purple are ssVICs attached without trypsin, in dark purple are ssVICs detached after trypsin treatment. Left: the non-trypsin and trypsin-treated samples are scaled across each group. Right: all five samples are scaled together.



With respect to the trypsin-induced detachment of ssVICs with the two Tn5 transposase ratios (TEC and WEC), I observe no discernible differences for either the FRIP score or the spike-in S2 nuclei mapped fractions, nor in the overall visual signal coverage. This result suggests that the difference in the amount of transposase to which the cells are exposed to does not seem to translate in differences in the ability of the transposase to find and intercalate in DNA-accessible regions.

All in all, these experiments provide us with insights into the different chromatin condensation dynamics that occur after cells experience perturbations in their immediate growth microenvironments (i.e. rough detachment from their substrate). These findings open up interesting research questions and follow-up experiments to dissect the biological processes that transmit biomechanical cues of cellular attachment and their ensuing changes in gene expression for the cell to content to

Figure 4.4: IGV genome browser displaying the ATAC-seq signal coverage over the ACTA2 gene for the WTC11 (top) and GM12878 (bottom) samples. In light purple are cells not treated with trypsin, in dark purple are cells after trypsin treatment. All four tracks are scaled together per cell line.



such disturbances.

4.4 Limitations

Though the aforementioned experiments suggest interesting biological phenomena, additional non-genomic approaches to validate any chromatin condensation or decondensation are needed, such as DAPI DNA staining or immunofluorescence markers. In addition, the fact that the in situ transposition reactions were performed with longer incubation times compared to the standard "in tube" reactions (i.e. 50 minutes versus 30 minutes) should be tested to rule out that this difference is the defining variable explaining the phenotype. Finally, other cell types should be tested as well, as I already observed non-agreeable observations with only two cell types: the pig ssVICs and the human iPSCs WTC11.

4.5 Methods

4.5.1 Cell lines information

Table 4.1 describes the information of the cells used to generate the ATAC-seq datasets; including their species, the cell type and ID, and their source.

Table 4.1: Information on the cell lines used in chapter 4.

Species	Cell type and ID	Source
Homo sapiens	LCLs, GM12878	Coriell / NIGMS
Homo sapiens	iPSCs, WTC11	Claudia Crocini
Sus scrofa	Primary valve interstitial myocardial cells	Cierra Walker

4.5.2 Growth conditions for ATAC-seq datasets

The pig VIC cells were grown and handled by Cierra Walker from the Anseth Lab. The human iPSC were grown and handled by Claudia Crocini from the Leinwand Lab. Both the pig VICs and the human iPSCs were transferred to 24-well plates with plastic as their substrate the day

before the experiment. The human LCLs were cultured in RPMI-1640 media (Gibco 72400-047) using 15% FBS (Gibco 10437-028) and 100 units/mL Penicillin-Streptomycin (Gibco 15140-122) in vent-cap T-25 flasks (Corning 430639), and kept at a confluency between 400,000 cells/mL to 800,000 cells/mL during cell culture at 37°C with 5% CO_2. On the day of the experiment, Cierra and Claudia set up the cells, which entailed either leaving the cells undisturbed in their 24-well plate wells with the pig VICs having ∼50,000 cells per well and the human iPSCs having ∼250,000 cells per well, or trypsinizing the cells by exposing them to 0.25% Trypsin-EDTA (Gibco Ref. 25200056) for approximately 10 minutes, before neutralizing the Trypsin by adding 3 volumes of culture media. In the case of the LCLs, ∼50,000 cells were spun down and their grown media replaced with trypsin for the same amount of time. The trypsinized cells were then washed once with PBS and transferred to 1.8 mL eppendorf tubes before proceeding with the ATAC-seq protocol. Each sample was prepared in duplicates.

4.5.3 ATAC-seq libraries preparation

The ATAC-seq libraries were prepared following two alternative procedures. In both variations, 2,500 of Drosophila melanogaster S2 nuclei (aiming for 5% relative to the cell number) were added in the transposition reactions. 1) For the datasets defined as "in situ", the procedure entailed doing the transposition reaction on the pig VICs and human iPSCs still attached to the 24-well plate plastic substrate instead of detaching the cells with trypsin beforehand, as described in [195] which in turn is a modified protocol from [42]. Briefly, the culture media was carefully aspirated from the 24-well plate wells. 500 µL of ice-cold lysis buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL, 0.1% Tween-20, 0.01% Digitonin Promega Ref. G944A) were added to the wells and let incubate at 4°C for 3 minutes. Carefully removed the supernatant, added 1 mL of wash buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20), and let incubate at 4°C for 10 minutes. Carefully removed the supernatant, 300 µL of the transposition mix (referred to as Well Enzyme Concentration or WEC) were added to all the samples still attached to the 24-well plate wells (150 µL Tagment DNA Buffer

Illumina Ref. 15027866, 7.5 µL Tagment DNA Enzyme 1 Illumina Ref. 15027865, 3 µL Digitonin diluted 1:1 with water, 30 µL Tween-20, 10.5 µL water, 89 µL PBS, and 10 µL of Drosophila S2 nuclei in at 250 nuclei/µL in PBS), and let incubate for 50 minutes at 37°C in a shaker incubator at 100 RPM. After the transposition reaction, 50 µL of EB buffer (Qiagen Ref. 19086) were added to the wells to obtain a volume of 350 µL, and 350 µL of UltraPure Phenol:Chloroform:Isoamyl Alcohol 25:24:1 (Invitrogen Ref. 15593-031) were added to the wells. The cells were detached from the 24-well plate plastic substrate using a cell lifter (Celltreat Ref. 229305), and their DNA extracted doing a standard phenol-chloroform precipitation, finishing by resuspending the DNA in 20 µL of EB buffer. Afterwards, a PCR pre-amplification was done using NEBNext Ultra II Q5 Master Mix (NEB Ref. M0544S) using 5 cycles. Then, a qPCR was done using NEBNext Ultra II Q5 Master Mix, SYBR Gold (Life Tech Ref. S11494), and 5 µL of the pre-amplified sample, and the results used to determine the additional number of extra PCR cycles using Nextera DNA CD Indices (Illumina Ref. 20015882). This number of cycles ranged from 2 to 7, and were chosen so that all samples reached the same concentration starting from different amounts of DNA. The post-amplified ATAC-libraries were cleaned-up with the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014). The libraries were size-selected to remove DNA fragments greater than 1000 bp with a Sage Science BluePippin. The ATAC-seq libraries were quantified with Qubit HS DNA assay and their fragment size-distributions determined with Agilent HS D5000 ScreenTape. 2) The alternative procedure was done following the [42] protocol with no modifications. Briefly, starting with the trypsinized samples residing in 1.8 mL eppendorf tubes with PBS, the cells were centrifuged at 500 x g for 5 minutes at 4°C, the supernatant carefully removed and replaced with 50 µL of ice-cold lysis buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL, 0.1% Tween-20, 0.01% Digitonin), the cells were resuspended 3 times pipetting up and down, and let incubate on ice for 3 minutes. Then, added 1 mL of wash buffer (water with 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20) and the tubes inverted 3 times to mix. The tubes were centrifuged at 500 x g for 10 minutes at 4°C and the supernatant was carefully removed without disturbing the small cell pellet. The pellets were then carefully resuspended by

pipetting 6 times with 50 µL of the transposition mix (referred to as Tube Enzyme Concentration or TEC) (25 µL Tagment DNA Buffer Illumina Ref. 15027866, 2.5 µL Tagment DNA Enzyme 1 Illumina Ref. 15027865, 0.5 µL Digitonin diluted 1:1 with water, 5 µL Tween-20, 0.5 µL water, 6.5 µL PBS, and 10 µL of Drosophila S2 nuclei in at 250 nuclei/µL in PBS). One of the trypsinized pig VICs samples, however, was mixed with a transposition mix that contained three times as much enzyme, and is referred to as WEC (25 µL Tagment DNA Buffer Illumina Ref. 15027866, 7.5 µL Tagment DNA Enzyme 1 Illumina Ref. 15027865, 0.5 µL Digitonin diluted 1:1 with water, 5 µL Tween-20, 2 µL PBS, and 10 µL of Drosophila S2 nuclei in at 250 nuclei/µL in PBS). After all the eppendorf tubes were properly mixed, they were incubated for 30 minutes in a heat block at 37°C, flicking the tube often. Afterwards, the samples were cleaned using the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014) following the manufacturer's instructions, and eluted in 21 µL elution buffer. Then, a PCR pre-amplification was done using NEBNext Ultra II Q5 Master Mix (NEB Ref. M0544S) using 5 cycles. Then, a qPCR was done using NEBNext Ultra II Q5 Master Mix, SYBR Gold (Life Tech Ref. S11494), and 5 µL of the pre-amplified sample, and the results used to determine the additional number of extra PCR cycles using Nextera DNA CD Indices (Illumina Ref. 20015882). This number of cycles ranged from 2 to 8, and were chosen so that all samples reached the same concentration starting from different amounts of DNA. The post-amplified ATAC-libraries were cleaned-up with the DNA Clean and Concentrator-5 Kit (Zymo Research Ref. D4014). The libraries were size-selected to remove DNA fragments greater than 1000 bp with a Sage Science BluePippin. The ATAC-seq libraries were quantified with Qubit HS DNA assay and their fragment size-distributions determined with Agilent HS D5000 ScreenTape. All samples were processed on the same day, first with the "in situ" reactions, followed by the tube reactions, and all samples processed together in the PCR amplifications, final clean-up and quantification. One of the replicates for the trypsinized pig VICs with WEC failed to amplify and was discarded.

4.5.4 ATAC-seq datasets sequencing information

All the ATAC-seq libraries were pooled together and sequenced on 2020/07/29 on a NextSeq 500 as paired-end 37 bp long reads. Base calls and demultiplexing was done using Bcl2Fastq2 (v2.2.0).

Table 4.2 describes the number of reads per ATAC-seq library.

Table 4.2: Sequencing depth of the ATAC-seq datasets used in chapter 4.

Dataset	Read number
ATAC-WEC-insitu-ssVIC-1	32,498,437
ATAC-WEC-insitu-ssVIC-2	26,546,502
ATAC-TEC-trypsin-ssVIC-2	20,953,629
ATAC-WEC-trypsin-ssVIC-1	87,840,676
ATAC-WEC-trypsin-ssVIC-2	48,296,378
ATAC-WEC-insitu-WTC11-1	22,511,684
ATAC-WEC-insitu-WTC11-2	28,302,110
ATAC-TEC-trypsin-WTC11-1	35,529,817
ATAC-TEC-trypsin-WTC11-2	29,222,625
ATAC-TEC-notrypsin-GM12878-1	24,947,759
ATAC-TEC-notrypsin-GM12878-2	21,661,461
ATAC-TEC-trypsin-GM12878-1	34,711,274
ATAC-TEC-trypsin-GM12878-2	26,328,795

4.5.5 ATAC-seq datasets processing

- Read quality was assessed using FastQC (v0.11.5).
- Read quality and adapter trimming was done using BBMap (v38.05) bbduk with options ktrim=r qtrim=10, k=23, mink=11, hdist=1, maq=10, minlen=25, tpe, tbo, literal=AAAAAAAAAAAAAAAAAAAAAA, ref= The FASTA file containing common adapters found in `https://github.com/Dowell-Lab/Nascent-Flow/blob/master/bin/adapters.fa`.
- Mapping was done using HISAT2 (v2.1.0) with options 'new-summary', 'very-sensitive',

'no-spliced-alignment'. The human reference genome hg38 was obtained from `https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrM, chrX, chrY). The pig reference genome susScr11 was obtained from `https://hgdownload.soe.ucsc.edu/goldenPath/susScr11/bigZips/susScr11.fa.gz`, and modified so that it only contained the main chromosome contigs (chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chrM, chrX, chrY). All samples were also mapped to the Drosophila reference genome dm6, which obtained from `https://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/dm6.fa.gz`, so that the fraction of reads mapped to the spiked-in S2 nuclei could be measured to infer genome-wide changes in DNA accessibility.

- Converted mapped SAM to BAM files using Samtools (v1.8) view -F 4 to remove unmapped reads.

- Read duplicates were removed using Sambamba (v0.6.6) markdup with options 'remove-duplicates', 'overflow-list-size=300000'.

- Bedgraph files were obtained using deepTools (v3.0.1) bamCoverage with options 'binSize 1', 'normalizeUsing CPM'.

4.5.6 Determining FRIP score from ATAC-seq datasets

The Fraction of Reads In Peaks (FRIP) score was obtained by counting the number of mapped reads assigned to open chromatin regions (e.g. peaks) relative to the total number of mapped reads. The peaks were determined using MACS2 (v2.1.1.20160309) callpeak with options 'nolambda', 'nomodel', 'keep-dup all', 'call-summits', and filtered out narrowPeaks with a score ¡ 100. Then, the narrowPeaks were merged across the cell types (e.g. all WTC11 samples together) using muMerge (v1.1.0) using options 'save_sampids', 'verbose'. The muMerge BED output file was modified

so that it contains 12 columns as follows: awk -v OFS='\t' ' print $1, $2, $3, "peak"NR, NR, ".", $7=$2, $8=$3, $9="0", $10=1, $11=$3-$2, $12=0 ' scVIC.susScr11-macs2_MUMERGE.bed > scVIC.susScr11-macs2_MUMERGE.12.bed. Finally, split_bam.py (v3.0.0) from the RSeQC package was used to determine the number of reads mapping to the merged peak regions or to the rest of the genome to obtain FRIP score, by simply dividing the number of reads in peaks over the total number of mapped reads as displayed in the log file from split_bam.py.

4.5.7 Obtaining scaling factors that correct for Drosophila spike-ins for ATAC-seq datasets

The total number of mapped reads to the Drosophila melanogaster dm6 genome was determined by summing the numbers from the HISAT2 summary stats file corresponding to the two lines "Aligned concordantly 1 time" and "Aligned concordantly ¿1 times" after mapping the sambamba-deduplicated BAM files to to the dm6 reference genome. The dm6 mapped reads were then expressed as a percentage relative to the total reads from the datasets, and averaged across replicates. Finally, a scaling factor ratio was calculated by dividing the non-trypsinized average from each cell type by either of the non-trypsinized or trypsinized average fractions, as shown in the table below. The bedgraphs were then adjusted with these scaling factors by dividing the 4th column with these factors as follows: awk -v OFS='\t' 'print $1, $2, $3, $4/4.618' ATAC-WEC-trypsin-scVIC-1.susScr11.bedgraph > ATAC-WEC-trypsin-scVIC-1.susScr11.scaled.bedgraph. The rescaled bedgraph files were then converted to bigwig files for visualization purposes using the UCSC Genome Browser kentUtils bedGraphToBigWig (v4.0.0) function.

Table 4.3 describes the ration (or scaling factor) used to rescale each of the ATAC-seq datasets.

Table 4.3: Normalized scaling factor estimated from dm6 reads used in chapter 4.

Dataset	Total reads	dm6 reads	% dm6	Average % dm6	Ratio
ATAC-TEC-notrypsin-GM12878-1	24,947,759	109,862	0.4404	0.3506	1.000
ATAC-TEC-notrypsin-GM12878-2	21,661,461	56,512	0.2609	0.3506	1.000
ATAC-TEC-trypsin-GM12878-1	34,711,274	199,852	0.5758	0.4716	0.744
ATAC-TEC-trypsin-GM12878-2	26,328,795	96,735	0.3674	0.4716	0.744
ATAC-WEC-insitu-WTC11-1	22,511,684	72,108	0.3203	0.3098	1.000
ATAC-WEC-insitu-WTC11-2	28,302,110	84,713	0.2993	0.3098	1.000
ATAC-TEC-trypsin-WTC11-1	35,529,817	52,909	0.1489	0.1211	2.557
ATAC-TEC-trypsin-WTC11-2	29,222,625	27,289	0.0934	0.1211	2.557
ATAC-WEC-insitu-scVIC-1	32,498,437	290,105	0.8927	1.0534	1.000
ATAC-WEC-insitu-scVIC-2	26,546,502	322,316	1.2142	1.0534	1.000
ATAC-WEC-trypsin-scVIC-1	87,840,676	205,040	0.2334	0.2281	4.618
ATAC-WEC-trypsin-scVIC-2	48,296,378	103,790	0.2149	0.2281	4.618
ATAC-TEC-trypsin-scVIC-2	20,953,629	49,447	0.2360	0.2281	4.618

Chapter 5

Conclusions and future work

In this work, I present the study of the evolution of gene transcriptional regulation spanning almost 300 million years of the metazoan branch of the terrestrial tree of life. I explored how eukaryotic cells have evolved, and continue to evolve, to withstand hazards to their subsistence that are thrown at them from multiple fronts by our hostile universe. From one side, I examined how primates react to potential damage to their precious genome, in the form of the activation of the p53, the guardian of the genome. On another side, I asked how animals contend with foreign pathogens by deploying varying intracellular defenses in the form of the cell-intrinsic innate immune system controlled by interferon. And I ended with some examples of how drastic changes in the microenvironment outside of cells is transduced inside the nucleus with potentially catastrophic effects on the chromatinized genome and the subsequent capabilities of the cells to react. All of this, of course, was explored only from the context of gene transcriptional regulation, and there is so much more yet to work on to fully understand these phenomena.

In the case of both the p53 and interferon gene transcriptional responses described here, it is crucial to test if any changes observed across species are still noticeable further down in the genes' expression cycle; at the protein levels, and ultimately at the organismal level. For evolution by means of natural selection works with changes in the fitness of organisms (and of course, the role of neutral selection). If a gene is observed to be transcribed more, but this phenotype has no effect on the ability of the host organism to pass on its genes to the next generation, then it is hard to posit that such a change in gene transcription was shaped by directional selection.

In a more immediate and feasible manner, I propose the following steps to complement the results shown in this thesis. PRO-seq is a powerful tool to observe both immediate gene transcription, but also the transcriptional regulatory elements that control such genes. However, there are two big hurdles that the author of this text had difficulties overcoming: 1) Assigning what are the target genes of the putative enhancers [141, 65]. 2) Finding the ortholog sequences of a given putative enhancer in other species. For the first point, the use of existing chromatin interaction techniques is recommended, or the many available datasets already publicly available. For the second point, the use of relatively new multiple sequence alignments [41]. In addition, I recommend the validation of the observed transcriptional changes with the usage of primary cells, perhaps just with relatively readily available subjects such as humans and rhesus macaques.

Further, the evidence of acquisition or loss of regulatory elements should be accompanied by the testing of a few of these sequence changes across orthologous loci. For example, if a gene is shown to be added into the p53-responsive network in hominoids, and the putative regulatory element is zeroed in, then such an element should be either perturbed with a DNA-editing tool or the changes tested in the context of a plasmid (or both).

In the case of the variation of the IFN response in diverse human ethnicities, I propose the exploitation of the abundance of DNA sequencing information that has been made available by the 1000 genomes project [40] and similar consortia. Here, I only sampled 8 individuals, and one from each ethnic group. But the future results that these datasets can provide can be validated by observing their fixation or lack thereof in the many other individuals from the same ethnic backgrounds.

Finally, the datasets provided here from the p53 and IFN transcriptional responses across primates are primed to be compared against each other. The immune system is under a strong evolutionary pressure to diversify constantly to withstand the numerous pathogens hosts interact with throughout their lifetimes [189, 58, 170]. Will the IFN transcriptional response be more diversified than the one from p53?

In conclusion, here I put forward an analysis of the p53 transcriptional response in primates,

as well as datasets for the study of the IFN transcriptional response among human populations, and across animals. I also present observations on the effect of cell detachment on the genome-wide chromatin state. I hope these contributions will inspire future researchers to follow their curiosity and obtain more answers that help us explain this strange and marvelous spectacle that is life on our little blue planet floating in the void of space.

Bibliography

[1] . Microbiology by numbers. Nat. Rev. Microbiol., 9:628, 2011.

[2] B. Alberts, A. Johnson, J. Lewis, et al. Molecular biology of the cell. Garland Science, 2014.

[3] M.A. Allen, Z. Andrysik, V.L. Dengler, et al. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. eLife, 3:e02200, 2014.

[4] R. Andersson and A. Sandelin. Determinants of enhancer and promoter activities of regulatory elements. Nat. Rev. Genet., 21:71—-87, 2020.

[5] Z. Andrysik, M.D. Galbraith, A.L. Guarnieri, et al. Identification of a core tp53 transcriptional program with highly distributed tumor suppressive activity. Genome Research, 27:1645–1657, 2017.

[6] N.T. Arndt and E.G. Nisbet. Processes on the young earth and the habitats of early life. Annual Review of Earth and Planetary Sciences, 40:521–549, 2012.

[7] P.R. Arnold, A.D. Wells, and X.C. Li. Diversity and emerging roles of enhancer rna in regulation of gene expression and cell fate. Front. Cell. Dev. Biol., 7:377, 2019.

[8] B. Aubrey, G. Kelly, A. Janic, et al. How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression? Cell Death Differ., 25:104—-113, 2018.

[9] C. Banks, A. Joshi, and T. Michoel. Functional transcription factor target discovery via compendia of binding and expression profiles. Sci. Rep., 6:20649, 2016.

[10] A. Bannister and T. Kouzarides. Regulation of chromatin by histone modifications. Cell Res., 21:381—-395, 2011.

[11] F. Baralle and J. Giudice. Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol., 18:437—-451, 2017.

[12] L.B. Barreiro, J.C. Marioni, R. Blekhman, et al. Functional comparison of innate immune signaling pathways in primates. PLoS Genetics, 6:e1001249, 2010.

[13] L.W. Barrett, S. Fletcher, and S.D. Wilton. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell. Mol. Life Sci., 69:3613—-3634, 2012.

[14] A. Battle, Z. Khan, S.H. Wang, et al. Impact of regulatory variation from rna to protein. Science, 347:664–667, 2015.

[15] E. Baugh, H. Ke, A. Levine, et al. Why are there hotspot mutations in the tp53 gene in human cancers? Cell Death Differ., 25:154—-160, 2018.

[16] I. Belda, J. Ruiz, A. Santors, et al. Saccharomyces cerevisiae. Trends in Genetics, 35:956–957, 2019.

[17] E.A. Bell, P. Boehnke, T.M. Harrison, et al. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. Proceedings of the National Academy of Sciences, 47:14518–14521, 2015.

[18] A. Bellantuono, D. Bridge, and D.E. Martinez. Hydra as a tractable, long-lived model system for senescence. Invertebrate Reproduction & Development, 59:39–44, 2015.

[19] V.A. Belyi, P. Ak, E. Markert, et al. The origins and evolution of the p53 family of genes. Cold Spring Harbor Perspectives in Biology, 2:a001198, 2009.

[20] T. Boehm and B.J. Swann. Origin and evolution of adaptive immunity. Annual Review of Animal Biosciences, 2:259—-283, 2014.

[21] G. Bourque. Transposable elements in gene regulation and in the evolution of vertebrate genomes. Current Opinion in Genetics & Development, 19:607–612, 2009.

[22] A. Breschi, T. Gingeras, and R. Guigó. Comparative transcriptomics in human and mouse. Nat. Rev. Genet., 18:425—440, 2017.

[23] C.S. Britton, T.R. Sorrells, and A.D. Johnson. Protein-coding changes preceded cis-regulatory gains in a newly evolved transcription circuit. Science, 367:96—100, 2020.

[24] G.N. Bruni and J.M. Kralj. Membrane voltage dysregulation driven by metabolic dysfunction underlies bactericidal activity of aminoglycosides. eLife, 9:e58706, 2020.

[25] G.N. Bruni, R.A. Weekly, B.J.T. Dodd, et al. Voltage-gated calcium flux mediates escherichia coli mechanosensation. Proceedings of the National Academy of Sciences, 114:9445–9450, 2017.

[26] C. Buccitelli and M. Selbach. mRNAs, proteins and the emerging principles of gene expression control. Nat. Rev. Genet., 21:630—-644, 2020.

[27] D.W. Burt. The cattle genome reveals its secrets. J. Biol., 8:36, 2009.

[28] C.A. Buttler and E.B. Chuong. Emerging roles for endogenous retroviruses in immune epigenetic regulation. Immunological Reviews, 305:165—-178, 2021.

[29] E. Cannavo, P. Khoueiry, D.A. Garfield, et al. Shadow enhancers are pervasive features of developmental regulatory networks. Current Biology, 26:38–51, 2016.

[30] L. Carbone, R. Alan Harris, S. Gnerre, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature, 513:195—-201, 2014.

[31] S.B. Carroll. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. Cell, 134:25–36, 2008.

[32] J.A. Castro-Mondragon, R. Riudavets-Puig, I. Rauluseviciute, et al. Jaspar 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Research, 50:D165—D173, 2022.

[33] N. Chatterjee and G.C. Walker. Mechanisms of dna damage, repair, and mutagenesis. Environ. Mol. Mutagen., 58:235–263, 2017.

[34] F.X. Chen and E.R Smith A. Shilatifard. Born to run: control of transcription elongation by rna polymerase ii. Nat. Rev. Mol. Cell Biol., 19:464—478, 2018.

[35] J. Choi and L.A. Donehower. p53 in embryonic development: maintaining a fine balance. Cellular and Molecular Life Sciences, 55:38–47, 1999.

[36] M. Chorev and L. Carmel. The function of introns. Frontiers in Genetics, 3:55, 2012.

[37] E. Chuong, N. Elde, and C. Feschotte. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet., 18:71—-86, 2017.

[38] B.J. Clavijo, L. Venturini, C. Schuduma, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Research, 27:885–896, 2017.

[39] R. Cojocaru and P.J. Unrau. Origin of life: Transitioning to dna genomes in an rna world. eLife, 6:e32330, 2017.

[40] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 526:68–74, 2015.

[41] Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. Nature, 587:240—245, 2020.

[42] M. Corces, A. Trevino, E. Hamilton, et al. An improved atac-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods, 14:959—-962, 2017.

[43] L. Core and K. Adelman. Promoter-proximal pausing of rna polymerase ii: a nexus of gene regulation. Genes and Development, 33:960–982, 2019.

[44] L.J. Core, J.J. Waterfall, and J.T. Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322:1845—1848, 2008.

[45] J. Cotney, J. Leng, J. Yin, et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell, 154:185–196, 2013.

[46] F. Crick. Central dogma of molecular biology. Nature, 227:561—-563, 1970.

[47] F. Crick, L. Barnett, S. Brenner, et al. General nature of the genetic code for proteins. Nature, 192:1227—-1232, 1961.

[48] Francis Crick. What Mad Pursuit: A Personal View of Scientific Discovery. Basic Books, 1988.

[49] D. D. Panne, T. Maniatis, and S.C. Harrison. An atomic model of the interferon-β enhanceosome. Cell, 129:1111–1123, 2007.

[50] C.G. Danko, L.A. Choate, B.A. Marks, et al. Dynamic evolution of regulatory element ensembles in primate cd4+ t cells. Nat. Ecol. Evol., 2:537—548, 2018.

[51] L. Dao, A. Galindo-Albarrán, J. Castro-Mondragon, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat. Genet., 49:1073—-1081, 2017.

[52] C. Darwin. On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life. J. Murray, 1859.

[53] M.D. Daugherty and H.S. Malik. Rules of engagement: Molecular insights from host-virus arms races. Annual Review of Genetics, 46:677–700, 2012.

[54] C. Delloye-Bourgeois, L. Bertin, K. Thoinet, et al. Microenvironment-driven shift of cohesion/detachment balance within tumors induces a switch toward metastasis in neuroblastoma. Cancer Cell, 32:427–443, 2017.

[55] P. Dettmer. Immune: A journey into the mysterious system that keeps you alive. Random House, 2021.

[56] E. Dolgin. The most popular genes in the human genome. Nature, 551:427—431, 2017.

[57] R.D. Dowell. Transcription factor binding variation in the evolution of gene regulation. Trends in Genetics, 26:468–475, 2010.

[58] D. Enard, L. Cai, C. Gwennap, et al. Viruses are a dominant driver of protein adaptation in mammals. eLife, 5:e12469, 2016.

[59] J. Falo-Sanjuan, N.C. Lammers, H.G. Garcia, et al. Enhancer priming enables fast and sustained transcriptional responses to notch signaling. Developmental Cell, 50:411–425, 2019.

[60] C.B. Fant, C.B. Levandowski, K. Gupta, et al. Tfiid enables rna polymerase ii promoter-proximal pausing. Molecular Cell, 78:785—793, 2020.

[61] N.V. Fedoroff. Transposable elements, epigenetics, and genome evolution. Science, 338:758–767, 2012.

[62] M. Fischer. Conservation and divergence of the p53 gene regulatory network between mice and humans. Oncogene, 38:4095—-4109, 2019.

[63] M. Flajnik and M. Kasahara. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nat. Rev. Genet., 11:47—59, 2010.

[64] N. Frankel, G. Davis, D. Vargas, et al. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature, 466:490—-493, 2010.

[65] M. Gasperini, J.M. Tome, and J. Shendure. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet., 21:292—310, 2020.

[66] N.H. Gehring and J. Roignant. Anything but ordinary – emerging splicing mechanisms in eukaryotic gene regulation. Trends in Genetics, 37:355–372, 2021.

[67] M.B. Gerstein, C. Bruce, J.S. Rozowsky, et al. What is a gene, post-encode? history and updated definition. Genome Research, 17:669–681, 2007.

[68] M.V.C. Greenberg and D. Bourc'his. The diverse roles of dna methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol., 20:590—607, 2019.

[69] R. Hakem. Dna-damage repair; the good, the bad, and the ugly. EMBO J., 27:589–605, 2008.

[70] S. Harris and A. Levine. The p53 pathway: positive and negative feedback loops. Oncogene, 24:2899—2908, 2005.

[71] G.F. Harrison, J. Sanz, J. Boulais, et al. Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nat. Ecol. Evol., 3:1253—-1264, 2019.

[72] H. Hasegawa, Y. Yamada, H. Iha, et al. Activation of p53 by nutlin-3a, an antagonist of mdm2, induces apoptosis and cellular senescence in adult t-cell leukemia cells. Leukemia, 23:2090—2101, 2009.

[73] M.S. Hill, P.V. Zande, and P.J. Wittkopp. Molecular and evolutionary processes generating variation in gene expression. Nat. Rev. Genet., 22:203—215, 2021.

[74] A. Hof, P. Campagne, D. Rigden, et al. The industrial melanism mutation in british peppered moths is a transposable element. Nature, 534:102—-105, 2016.

[75] H.H. Hoffmann, W.M. Schneider, and C.M. Rice. Interferons and viruses: an evolutionary arms race of molecular interactions. Trends in Immunology, 36:124–138, 2015.

[76] M.M. Horvath, X. Wang, M.A. Resnick, et al. Divergent evolution of human p53 binding sites: Cell cycle versus apoptosis. PLoS Genet., 3:e127, 2007.

[77] Pestka Biomedical Laboratories Inc. Cynomolgus IFN-alpha 2 (Ile 16) mammalian, Catalog No. 16105, Lot No. 6987r.

[78] Proteintech Group Inc. Humankine recombinant human IFN alpha 2a protein, Catalog No. hz-1066, Lot No. 0615-01.

[79] F. Inoue and N. Ahituv. Decoding enhancers using massively parallel reporter assays. Genomics, 106:159–164, 2015.

[80] A.G. Jegga, A. Inga, D. Menendez, et al. Functional evolution of the p53 regulatory network through its target response elements. Proceedings of the National Academy of Sciences, 105:944–949, 2008.

[81] E.R. Jerison, S. Kryazhimskiy, J.K. Mitchell, et al. Genetic variation in adaptability and pleiotropy in budding yeast. eLife, 6:e27167, 2017.

[82] Y. Jin, U. Eser, K. Struhl, et al. The ground state and evolution of promoter region directionality. Cell, 170:889–898, 2017.

[83] M. Jinek, K. Chylinski, I. Fonfara, et al. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. Science, 337:816–821, 2012.

[84] F. Jones, M. Grabherr, Y. Chan, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature, 484:55—61, 2012.

[85] I. Jonkers and J. Lis. Getting up to speed with transcription elongation by rna polymerase ii. Nat. Rev. Mol. Cell Biol., 16:167—-177, 2015.

[86] E.N. Judd, A.R. Gilchrist, N.R. N.R. Meyerson, et al. Positive natural selection in primate genes of the type i interferon response. BMC Ecol. Evol., 21:65, 2021.

[87] M.S. Kang and E. Kieff. Epstein–barr virus latent genes. Exp. Mol. Med., 47:e131, 2015.

[88] K. Karakostis and R. Fåhraeus. Shaping the regulation of the p53 mrna tumour suppressor: the co-evolution of genetic signatures. BMC Cancer, 19:915, 2019.

[89] E.R. Kastenhuber and S.W. Lowe. Putting p53 in context. Cell, 170:1062–1078, 2017.

[90] S. Kato, S. Han, W. Liu, et al. Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proceedings of the National Academy of Sciences, 100:8424–8429, 2003.

[91] B. Kempkes and E.S. Robertson. Epstein-barr virus latency: current and future perspectives. Current Opinion in Virology, 14:138–144, 2015.

[92] W.J. Kent, R. Baertsch, A. Hinrichs, et al. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proceedings of the National Academy of Sciences, 100:11484–11489, 2003.

[93] R. Khamsi. Chickens join the genome club. Nature, 2004.

[94] Z. Khan, M.J. Ford, D.A. Cusanovich, et al. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science, 342:1100–1104, 2013.

[95] J.T. King and A. Shakya. Phase separation of dna: From past to present. Biophysical Journal, 120:1139–1149, 2021.

[96] M.C. King and A.C. Wilson. Evolution at two levels in humans and chimpanzees. Science, 188:107–116, 1975.

[97] T. Klann, J. Black, M. Chellappan, et al. Crispr–cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat. Biotechnol., 35:561—-568, 2017.

[98] J.C. Klein, A. Keith, V. Agarwal, et al. Functional characterization of enhancer evolution in the primate lineage. Genome Biol., 19:99, 2018.

[99] S.L. Klemm, Z. Shipony, and W.J. Greenleaf. Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet., 20:207—220, 2019.

[100] A. Klug. Rosalind franklin and the discovery of the structure of dna. Nature, 219:808—810, 1968.

[101] J. Klunk, T.P. Vilgalys, C.E. Demeure, et al. Evolution of immune genes is associated with the black death. Nature, 2022.

[102] E.V. Koonin and K.S. Makarova. Origins and evolution of crispr-cas systems. Philosophical Transactions of the Royal Society B, 374, 2019.

[103] C.D. Krause and S. Pestka. Cut, copy, move, delete: The study of human interferon genes reveal multiple mechanisms underlying their evolution in amniotes. Cytokine, 76:480–495, 2015.

[104] E.Z. Kvon, R. Waymack, M. Gad, et al. Enhancer redundancy in development and disease. Nat. Rev. Genet., 22:324—336, 2021.

[105] M.T.Y. Lam, W. Li, M.G. Rosenfeld, et al. Enhancer rnas and regulated transcriptional programs. Trends in Biochemical Sciences, 39:170–182, 2014.

[106] S.A. Lambert, A. Jolma, L.F. Campitelli, et al. The human transcription factors. Cell, 172:650–665, 2018.

[107] S.G. Landt, G.K. Marinov, A. Kundaje, et al. Chip-seq guidelines and practices of the encode and modencode consortia. Genome Res., 22:1813–1831, 2012.

[108] A.J.M. Larsson, P. Johnsson, M. Hagemann-Jensen, et al. Genomic encoding of transcriptional burst kinetics. Nature, 565:251—254, 2019.

[109] M. Lawrence, S. Daujat, and R. Schneider. Lateral thinking: How histone modifications regulate gene expression. Trends in Genetics, 32:42–56, 2016.

[110] T.I. Lee and R.A. Young. Transcriptional regulation and its misregulation in disease. Cell, 152:1237–1251, 2013.

[111] A.J. Levine. p53: 800 million years of evolution and 40 years of discovery. Nat. Rev. Cancer, 20:471—480, 2020.

[112] J.J. Lewis, R.C. Geltman, P.C. Pollak, et al. Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. Proceedings of the National Academy of Sciences, 116:24174–24183, 2019.

[113] D. Li and M. Wu. Pattern recognition receptors in health and diseases. Sig. Transduct. Target Ther., 6:291, 2021.

[114] S. Li, E.Z. Kvon, A. Visel, et al. Stable enhancers are active in development, and fragile enhancers are associated with evolutionary adaptation. Genome Biol., 20:140, 2019.

[115] A. Liston, S. Humblet-Baron, D. Duffy, et al. Human immune diversity: from evolution to modernity. Nat. Immunol., 22:1479—-1489, 2021.

[116] Y. Liu, A. Beyer, and R. Aebersold. On the dependency of cellular protein levels on mrna abundance. Cell, 165:535–550, 2016.

[117] H.K. Long, S.L. Prescott, and J. Wysocka. Ever-changing landscapes: Transcriptional enhancers in development and evolution. Cell, 167:1170–1187, 2016.

[118] I. Lopes, G. Altab, P. Raina, et al. Gene size matters: An analysis of gene length in the human genome. Frontiers in Genetics, 12:559998, 2021.

[119] M.I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol., 15:550, 2014.

[120] K. Luger, M. Dechassa, and D. Tremethick. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? Nat. Rev. Mol. Cell Biol., 13:436–447, 2012.

[121] B. Maddox. The double helix and the 'wronged heroine'. Nature, 421:407—-408, 2003.

[122] D. Mahat, H. Kwak, G. Booth, et al. Base-pair-resolution genome-wide mapping of active rna polymerases using precision nuclear run-on (pro-seq). Nat. Protoc., 11:1455—1476, 2016.

[123] S. Martire and L.A. Banaszynski. The roles of histone variants in fine-tuning chromatin organization and function. Nat. Rev. Mol. Cell Biol., 21:522—-541, 2020.

[124] E. Masy, E. Adriaenssens, C. Montpellier, et al. Human monocytic cell lines transformed in vitro by epstein-barr virus display a type ii latency and lmp-1-dependent proliferation. J. Virol., 76:6460—-6472, 2002.

[125] K. Mattioli, W. Oliveros, and C. Gerhardinger et al. Cis and trans effects differentially contribute to the evolution of promoters and enhancers. Genome Biol., 21:210, 2020.

[126] F. McNab, K. Mayer-Barber, A. Sher, et al. Type i interferons in infectious disease. Nat. Rev. Immunol., 15:87—-103, 2015.

[127] M. Min and S.L. Spencer. Spontaneously slow-cycling subpopulations of human cells originate from activation of stress-response pathways. PLOS Biology, 17:e3000178, 2019.

[128] S. Mitschka and C. Mayr. Context-specific regulation and function of mrna alternative polyadenylation. Nat Rev Mol Cell Biol, 2022.

[129] G. Monaco, B. Lee, W. Xu, et al. Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types. Cell Rep., 26:1627–1640, 2019.

[130] J.M. Moon, J.A. Capra, P. Abbot, et al. Signatures of recent positive selection in enhancers across 41 human tissues. G3, 9:2761–2774, 2019.

[131] S. Mostafavi, H. Yoshida, D. Moodley, et al. Parsing the interferon transcriptional network and its disease associations. Cell, 164:564—-578, 2016.

[132] N. Mouraret, E. Marcos, S. Abid, et al. Activation of lung p53 by nutlin-3a prevents and reverses experimental pulmonary hypertension. Circulation, 127:1664–1676, 2013.

[133] J. Mouw, G. Ou, and V. Weaver. Extracellular matrix assembly: a multiscale deconstruction. Nat. Rev. Mol. Cell Biol., 15:771—-785, 2014.

[134] M. Mumbach, A. Rubin, R. Flynn, et al. Hichip: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods, 13:919—922, 2016.

[135] W.E. Müller, B. Blumbach, and I.M. Müller. Evolution of the innate and adaptive immune systems: Relationships between potential immune molecules in the lowest metazoan phylum (porifera) and those in vertebrates. Transplantation, 68:1215—-1227, 1999.

[136] N. Nakazawa, A.R. Sathe, G.V. Shivashankar, et al. Matrix mechanics controls fhl2 movement to the nucleus to activate p21 expression. Proceedings of the National Academy of Sciences, 113:6813–6822, 2016.

[137] M. Nei, P. Xu, and G. Glazko. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proceedings of the National Academy of Sciences, 98:2497—2502, 2001.

[138] S. Nurk, S. Koren, A. Rhie, et al. The complete sequence of a human genome. Science, 376:44–53, 2022.

[139] M. Osterwalder, I. Barozzi, V. V. Tissières, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature, 554:239—243, 2018.

[140] S.R. Paludan, T. Pradeu, S.L. Masters, et al. Constitutive immune mechanisms: mediators of host defence and immune regulation. Nat. Rev. Immunol., 21:137—-150, 2021.

[141] A. Panigrahi and B.W. O'Malley. Mechanisms of enhancer action: the known and the unknown. Genome Biol., 22:108, 2021.

[142] E.C. Partridge, S.B. Chhetri, J.W. Prokop, et al. Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature, 583:720—-728, 2020.

[143] C.N. Passow, A.M. Bronikowski, H. Blackmon, et al. Contrasting patterns of rapid molecular evolution within the p53 network across mammal and sauropsid lineages. Genome Biology and Evolution, 11:629—643, 2019.

[144] S.L. Pereira, R.A. Grayling, R. Lurz, et al. Archaeal nucleosomes. Proceedings of the National Academy of Sciences, 94:12633–12637, 1997.

[145] P. Perelman, W.E. Johnson, C. Roos, et al. A molecular phylogeny of living primates. PLOS Genetics, 7:e1001342, 2011.

[146] V. Petrova, M. Annicchiarico-Petruzzelli, G. Melino, et al. The hypoxic tumour microenvironment. Oncogenesis, 7:10, 2018.

[147] J. Piehler, C. Thomas, K.C. Garcia, et al. Structural and dynamic determinants of type i interferon receptor assembly and their functional interpretation. Immunol. Rev., 250:317—-334, 2012.

[148] L. Polgár. The catalytic triad of serine peptidases. Cell. Mol. Life Sci., 62:2161—-2172, 2005.

[149] O. Porrua and D. Libri. Transcription termination and the control of the transcriptome: why, where and how to stop. Nat. Rev. Mol. Cell Biol., 16:190—202, 2015.

[150] P. Portin and A. Wilkins. The evolving definition of the term 'gene'. Genetics, 205:1353––1364, 2017.

[151] S.L. Prescott, R. Srinivasan, M.C. Marchetto, et al. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. Cell, 163:68–83, 2015.

[152] N.J. Proudfoot. Transcriptional termination in mammals: Stopping the rna polymerase ii juggernaut. Science, 352:aad9926, 2016.

[153] L. Przybyla, J.M. Muncie, and V.M. Weaver. Mechanical control of epithelial-to-mesenchymal transitions in development and cancer. Annual Review of Cell and Developmental Biology, 32:527–554, 2016.

[154] A. Ramanathan, G.B. Robb, and S. Chan. mrna capping: biological functions and applications. Nucleic Acids Research, 44:7511—7526, 2016.

[155] F. Randow, J.D. MacMicking, and L.C. James. Cellular self-defense: How cell-autonomous immunity protects against pathogens. Science, 340:701—706, 2013.

[156] M. Rebeiz and M. Tsiantis. Enhancer evolution and the origins of morphological novelty. Current Opinion in Genetics & Development, 45:115–123, 2017.

[157] S.K. Reilly, J. Yin, A.E. Ayoub, et al. Evolutionary changes in promoter and enhancer activity during human corticogenesis. Science, 347:1155–1159, 2015.

[158] W.F. Richter, S. Nayak, J. Iwasa, et al. The mediator complex as a master regulator of transcription by rna polymerase ii. Nat. Rev. Mol. Cell. Biol., 23:732—-749, 2022.

[159] I. Romero, I. Ruvinsky, and Y. Gilad. Comparative studies of gene expression and the evolution of gene regulation. Nat. Rev. Genet., 13:505—-516, 2012.

[160] M.J. Rowley and V.G. Corces. Organizational principles of 3d genome architecture. Nat. Rev. Genet., 19:789—-800, 2018.

[161] J.D. Rubin, J.T. Stanley, R.F. Sigauke, et al. Transcription factor enrichment analysis (tfea) quantifies the activity of multiple transcription factors from a single experiment. Commun. Biol., 4:661, 2021.

[162] L. Sagan. On the origin of mitosing cells. J. Theor. Biol., 14:255–274, 1967.

[163] V. Sartorelli and S.M. Lauberth. Enhancer rnas are an important regulatory layer of the epigenome. Nat. Struct. Mol. Biol., 27:521—528, 2020.

[164] D. Schmidt, M.D. Wilson, B. Ballester, et al. Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. Science, 328:1036–1040, 2010.

[165] D. Schoenberg and L. Maquat. Regulation of cytoplasmic mrna decay. Nat. Rev. Genet., 13:246—259, 2012.

[166] S. Schoenfelder and P. Fraser. Long-range enhancer–promoter contacts in gene expression control. Nat. Rev. Genet., 20:437—-455, 2019.

[167] G. Schreiber. The molecular basis for differential type i interferon signaling. J. Biol. Chem., 292:7285—7294, 2017.

[168] A.E. Shaw, J. Hughes, Q. Gu, et al. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type i interferon responses. PLOS Biology, 15:e2004086, 2017.

[169] B.A. Shen and R. Landick. Transcription of bacterial chromatin. Journal of Molecular Biology, 431:4040–4066, 2019.

[170] A.J. Shultz and T.B. Sackton. Immune genes are hotspots of shared positive selection across birds and mammals. eLife, 8:e41815, 2019.

[171] A. Siepel, G. Bejerano, J.S. Pedersen, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15:1034–1050, 2005.

[172] S.A. Signor and S.V. Nuzhdin. The evolution of gene expression in cis and trans. Trends in Genetics, 34:532–544, 2018.

[173] P. Speight, M. Kofler, K. Szászi, et al. Context-dependent switch in chemo/mechanotransduction via multilevel crosstalk among cytoskeleton-regulated mrtf and taz and tgfβ-regulated smad3. Nat. Commun., 7:11642, 2016.

[174] F. Spitz and E. Furlong. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet., 13:613—626, 2012.

[175] M. Spivakov. Spurious transcription factor binding: Non-functional or genetically redundant? Bioessays, 36:798–806, 2014.

[176] R. Stark, M. Grzelak, and J. Hadfield. Rna sequencing: the teenage years. Nat. Rev. Genet., 20:631—656, 2019.

[177] T. Stephan, S.M. Burgess, H. Cheng, et al. Darwinian genomics and diversity in the tree of life. eLife, 119:e2115644119, 2021.

[178] N. Stevens, E. Seiffert, P. O'Connor, et al. Palaeontological evidence for an oligocene divergence between old world monkeys and apes. Nature, 497:611—-614, 2013.

[179] A. Subramanian, P. Tamayo, V.K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102:15545–15550, 2005.

[180] M. Sulak, L. Fong, K. Mika, et al. Tp53 copy number expansion is associated with the evolution of increased body size and an enhanced dna damage response in elephants. eLife, 5:e11994, 2016.

[181] K. Sullivan, M. Galbraith, Z. Andrysik, et al. Mechanisms of transcriptional regulation by p53. Cell Death Differ., 25:133—-143, 2018.

[182] M. Sullivan and D. Morgan. Finishing mitosis, one step at a time. Nat. Rev. Mol. Cell Biol., 8:894—903, 2007.

[183] E.D. Tarbell and T. Liu. Hmmratac: a hidden markov modeler for atac-seq. Nucleic Acids Research, 47:e91, 2019.

[184] P. Trojet and D. Reinberg. Facultative heterochromatin: Is there a distinctive molecular signature? Molecular Cell, 28:1–13, 2007.

[185] E. Tunnacliffe and J.R. Chubb. What is a transcriptional burst? Trends in Genetics, 36:288–297, 2020.

[186] P. Turelli, C. Playfoot, D. Grun, et al. Primate-restricted krab zinc finger proteins and target retrotransposons control gene expression in human neurons. Science, 6:eaba3200, 2020.

[187] C. Uhler and G. Shivashankar. Regulation of genome organization and gene expression by nuclear mechanotransduction. Nat. Rev. Mol. Cell Biol., 18:717—-727, 2017.

[188] V. V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell. Biol., 19:621—-637, 2018.

[189] E.J. Vallender and B.T. Lahn. Positive selection on the human genome. Human Molecular Genetics, 13:245–254, 2004.

[190] B. van Steensel and E.E.M. Furlong. The role of transcription in shaping the spatial organization of the genome. Nat. Rev. Mol. Cell Biol., 20:327—337, 2019.

[191] J. Vaquerizas, S. Kummerfeld, S. Teichmann, et al. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet., 10:252—263, 2009.

[192] M. Vermunt, S. Tan, B. Castelijns, et al. Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. Nat. Neurosci., 19:494—503, 2016.

[193] J. Vierstra, J. Lazar, R. Sandstrom, et al. Global reference mapping of human transcription factor footprints. Nature, 583:729—736, 2020.

[194] D. Villar, C. Berthelot, S. Aldridge, et al. Enhancer evolution across 20 mammalian species. Cell, 160:554–566, 2015.

[195] C.J. Walker, C. Crocini, D. Ramirez, et al. Nuclear mechanosensing drives chromatin remodelling in persistently activated fibroblasts. Nature Biomedical Engineering, 5:1485—-1499, 2021.

[196] T. Wang, J. Zeng, C.B. Lowe, et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proceedings of the National Academy of Sciences, 104:18613–18618, 2017.

[197] E.A. Warman, D. Forrest, T. Guest, et al. Widespread divergent transcription from bacterial and archaeal promoters is a consequence of dna-sequence symmetry. Nat. Microbiol., 6:746––756, 2021.

[198] J. Watson and F. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. Nature, 171:737—-738, 1953.

[199] E.M. Wissink, A. Vihervaara, N.D. Tippens, et al. Nascent rna analyses: tracking transcription and its regulation. Nat. Rev. Genet., 20:705—723, 2019.

[200] P. Wittkopp, B. Haerum, and A. Clark. Evolutionary changes in cis and trans gene regulation. Nature, 430:85—88, 2004.

[201] C. Woese. The universal ancestor. Proc. Natl Acad. Sci., 95:6854—-6859, 1998.

[202] C. Woese, O. Kandler, and M. Wheelis. Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. Proc. Natl. Acad. Sci., 87:4576—-4579, 1990.

[203] G. Wolf, A. de Iaco, M. Sun, et al. Krab-zinc finger protein gene expansion in response to active retrotransposons in the murine lineage. eLife, 9:e56337, 2020.

[204] G. Wolf, D. Greenberg, and T.S. Macfarlan. Spotting the enemy within: Targeted silencing of foreign dna in mammalian genomes by the krüppel-associated box zinc finger protein family. Mobile DNA, 6:17, 2015.

[205] A. Wurmser and S. Basu. Enhancer-promoter communication: It's not just about contact. Front. Mol. Biosci., 9:867303, 2022.

[206] C. Xu, C. Fan, and X. Wang. Regulation of mdm2 protein stability and the p53 response by nedd4-1 e3 ligase. Oncogene, 34:281—289, 2015.

[207] C. Yang, C. Tian, T.E. Hoffman, et al. Melanoma subpopulations that rapidly escape mapk pathway inhibition incur dna damage and rely on stress signalling. Nat. Commun., 12:1747, 2021.

[208] L. Yao, J. Liang, A. Ozer, et al. A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nat. Biotechnol., 40:1056—-1065, 2022.

[209] F. Yue, Y. Cheng, A. Breschi, et al. A comparative encyclopedia of dna elements in the mouse genome. Nature, 515:355—364, 2014.

[210] K.S. Zaret and J.S. Caroll. Pioneer transcription factors: establishing competence for gene expression. Genes and Development, 25:2227–2241, 2011.

[211] A. Zemach and D. Zilberman. Evolution of eukaryotic dna methylation and the pursuit of safer sex. Current Biology, 20:780–785, 2010.

[212] Y. Zhang, T. Liu, C.A. Meyer, et al. Model-based analysis of chip-seq (macs). Genome Biol., 9:R137, 2008.

[213] Carl Zimmer. Life's Edge: Searching for What It Means to Be Alive. Dutton, 2021.

[214] E. Åberg, F. Saccoccia, M. Grabherr, et al. Evolution of the p53-mdm2 pathway. BMC Evol. Biol., 17:177, 2017.

[215] M. Çalışkan, D.A. Cusanovich, C. Ober, et al. The effects of ebv transformation on gene expression levels and methylation profiles. Human Molecular Genetics, 20:1643—1652, 2011.