Identifying Gene Regulatory Activity Divergence in *Cis* and *Trans* with ATAC-STARR-seq

By

Tyler John Hansen

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry

May 12th, 2023

Nashville, Tennessee

Approved:

Vito Quaranta, M.D.

Scott Hiebert, Ph.D.

Manuel Ascano Jr., Ph. D.

Emily Hodges, Ph.D.

John Anthony Capra, Ph.D.

To my brilliant and loving wife, Emili

To my wonderful kids

and

To my amazing parents

Thank you for everything

ACKNOWLEDGEMENTS

Outside of family, it is hard to come up with a more important relationship than that of a trainee and their PhD advisor. I want to thank my advisor, Dr. Emily Hodges for an enjoyable, exciting, and truly inspiring experience in her lab. I have learned so much from Emily about conducting science, communicating science, and selling yourself and your project to others. There are so many lessons that Emily has taught me that have made me a much better scientist and will help me succeed in the future. “If it can work a little then it can work a lot” and “we don’t look for a needle in the haystack, we define the haystack” are words of wisdom that I will always live by and carry with me for the rest of my career. Emily was also very gracious with her patience and care towards my personal life, which, with getting married and having two kids, always seemed to change. This change, while exciting, led to a lot of trial and error on figuring out how balance my PhD work with all the important new responsibilities I acquired at home. I’d like to say I’ve found a good balance, and this would not have been possible without Emily’s kindness, understanding, and patience.

I want to also thank all members of the Hodges lab for being great people and always making lab a great place to be. Special thank you to Tim Scott and Lindsey Guerin for all the coffee runs, lunch trips, and great conversations we had over the years; grad school was tough, and I could not have made it through without you two. I also want to thank other current and former Hodges Lab members—Kelly Barnett, Kritika Singh, Verda Agan, Adam Miranda, Jonathan Attalla, Elizabeth Dorans, Ali Wilt, and Doug Shaw—for all of their great ideas over the years. You guys inspire me to great science, thank you.

Thank you to Sarah Fong and Tony Capra, who collaborated with me on Chapter III of this dissertation. Thank you for being great scientists, good people, and giving this project your all. It was a truly complicated project and would not be anywhere near as impactful as it is today without your intellectual and analytical contributions. Even from two time zones away, our weekly meetings were often the highlight of my week and I hope we have another opportunity to work together sometime.

Thank you to the core group of friends I've made during my time in graduate school: James Held, Payam Fathi, Hillary Layden, Helen Parrington, and Logan Richards. Spending time with you guys has always been fun. We have made a lot of great memories that I will cherish forever. Thank you in particular to James helping me in so many ways during Grad school. From talking to police after my house was broken into (just three days before my wedding day) to taking care of Henry at 3am so that Emili and I could go to the hospital and have our second baby. You are one of the best people I know, thank you for everything.

Part of what made my PhD experience special was the fantastic community of the Biochemistry Department. Thank you to the faculty leadership and all current and former members of the Biochemistry Student Association for working so hard to make this department a wonderful place to do science. Thank you to Sam Lisy, Lindsay Redman Rivera, Robert Mann, Caroline Wiser, Stephen Clark, Yelena Perevalova, Sarah Arcos, Katie Rothamel, Kavi Mehta, Tata Kavlashvili, Juan Carvajal-Garcia, Kaitlyn Browning, Anna Johnson, Ronan Bracken, Monica Bomber, Vincent Yao, Kate Clowes, Nicky Eleuteri, Colby Tubbs and so many other new colleagues I made over the years. You all are so genuinely kind and talented individuals. I hope I can maintain contact with you all throughout my career.

The scientific aspect of this project would not have been possible without the help and guidance of several individuals. Thank you to my committee—Vito Quaranta, Scott Hiebert, Manny Ascano, Tony Capra, and Emily Hodges—for always challenging me and keeping me on the right path. Thank you also to Amanda Lea for providing feedback on the *cis/trans* project (Chapter III) and for her kindness and generosity in many other respects—I'm not sure I would have gotten my dream postdoc without her. Other faculty I'd like to thank are Yi Ren, Breann Brown, Dave Cortez, and Bill Tansey. You are great scientists and good people—know that I look up to you as role models and hope to be like you one day.

Thank you to all of my previous teachers and mentors since my scientific journey began: Angie Midthun-Hensen, Ryan Olson, Barbara Bielec, Judith Kimble, Aaron Kershner, Amy Groth, Andy Golden, Harold Smith, and Aimee Jaramillo-Lambert. You all played such an important role in my early career, and it was your kindness, generosity, and patience that allowed me to discover my passion for science—thank you so much.

Last and most importantly, I want to thank my family for their love, support, and patience during my time in graduate school. The life of an academic is not easy on family—there are long nights, limited options of cities to live in, and constant distractions that make it easy to lose focus on non-work things—thank you for your flexibility and understanding as I tried to figure out how to balance my work life with my family life. Thank you to my parents for your constant love and support and always motivating me to follow my dreams, even though that took me out of Wisconsin. Thank you to my brother, Ben, and my sister, Jenna, for your love and all the fun we have had over the years. I enjoy hearing about all the amazing things you two do. Thank you to my kids for being the absolute joy of my life. Thank you to my wife, Emili, for her unconditional love and her ability to always be optimistic, no matter the circumstance. You inspire me to be a better person every day and I'm so lucky to have you in my life.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

DNA Regulatory Elements in the Human Genome

With few exceptions, all life on earth uses the same fundamental mechanism to express genes: DNA is transcribed into messenger RNA (mRNA), which is then translated into proteins. This process, called gene expression, controls the uniqueness of species and how hundreds of specialized cell types can be made from a single genome. The initial regulatory steps of transcription are critically important in tightly controlling gene expression programs in the cell. In particular, non-coding *cis*-regulatory DNA sequences in the genome, called DNA regulatory elements, control which genes are transcribed and how many mRNA transcripts are made (Andersson & Sandelin, 2019; Haberle & Stark, 2018; Heinz, Romanoski, Benner, & Glass, 2015; Long, Prescott, & Wysocka, 2016; Rickels & Shilatifard, 2018). To actively regulate transcription, DNA regulatory elements are bound by transcription factors (TFs), which, through complex biochemical interactions and biophysical phase separation mechanisms, recruit regulatory complexes that affect transcription of nearby target genes (Figure 1A) (Haberle & Stark, 2018; Peng, Li, & Xu, 2020; Ptashne, 1967, 1986). Because the activity of DNA regulatory elements control which and how much genes are transcribed, DNA regulatory element activity dictates cellular identity, including both maintenance of cell identity and determination of cell fate during differentiation (Corces et al., 2016; Heinz et al., 2015). For this reason, global DNA regulatory element dysfunction can drive diseases where cells adopt new identities, fail to differentiate completely, or differentiate incorrectly. For example, the mis-regulation of DNA regulatory elements can drive neoplastic transformation and cause a variety of cancers (Bradner, Hnisz, & Young, 2017; Smith & Shilatifard, 2014).

Promoters & Enhancers

DNA regulatory elements can be classified into many categories. Promoters and enhancers are two types of positive DNA regulatory elements that, when active, drive transcription of their target genes. Promoters and enhancers are functionally similar overall (Andersson, Sandelin, & Danko, 2015), but they differ in a few key ways. First, promoters exist directly upstream of the genes they regulate while enhancers are located distal to their target genes and can exist in any orientation because they can be brought into close proximity via 3D chromatin looping mechanisms (Figure 1A) (Andersson & Sandelin, 2019; Panigrahi & O'Malley, 2021; Schaffner, 2015). Because they are unrestricted by target gene distance and because they lack consensus DNA sequence motifs, such as a TATA box or initiator sequence, it is challenging to identify enhancers in the human genome (Long et al., 2016). Nonetheless, significant efforts from ENCODE, FANTOM, and other consortia using different enhancer identification methods have discovered over 1 million putative enhancers in the human genome for hundreds of tissues and cell types (Abugessaisa et al., 2021; The ENCODE Project Consortium et al., 2020; Wang et al., 2019).

Promoters and enhancers also differ in the types of genes they regulate in the human genome. In general, genes regulated only by promoters are constitutively expressed and are associated with housekeeping functions in the cell, whereas enhancers regulate genes associated with cell-type specific functions (Long et al., 2016). In this way, enhancers control cell fate and identity to a much greater extent than promoters. Alterations to enhancer function can cause disease and are thought to drive the development of complex, polygenic diseases (Herz, 2016;

Maurano et al., 2012; Smith & Shilatifard, 2014). Understanding where enhancers are in the human genome and how they function is vital for identifying the genetic causes of many common debilitating complex, polygenic diseases like Crohn's disease.

Silencers

Another class of DNA regulatory elements called silencers actively repress gene expression. In principle, they act similar to promoters and enhancers, but they instead are bound by TFs that repress transcription of nearby target genes (Pang, van Weerd, Hamoen, & Snyder, 2022). Like enhancers, silencers are challenging to identify because they lack consensus DNA sequence motifs and can be located distal from the genes they regulate. Furthermore, it is difficult to discern whether a gene is inactive due to the activity of silencers or other silencing mechanisms because silencers are just one mechanism for repressing gene expression. This makes silencers much more difficult to identify than enhancers, so they remain a critically understudied component in the transcriptional regulation field. While recent efforts have been made to identify silencers in the human genome, these studies have not been performed at consortia-like scales, so a comprehensive evaluation of silencers in the human genome is also lacking. Some assays that identify putative enhancers may also identify silencers, but the lack of quantitative measures on regulatory activity—*i.e.* how much target gene transcription they drive—makes it difficult to discern whether these regions have positive or negative effects on gene transcription.

Epigenetic Control of DNA Regulatory Elements

DNA regulatory elements themselves can be regulated. In eukaryotes, genomic DNA is tightly bound to nucleosomes, which are octameric complexes of histone proteins (Klemm, Shipony, & Greenleaf, 2019). The majority of DNA is tightly wound up by nucleosomes into structures called heterochromatin, which, by steric hindrance, prevents most proteins from binding. In any given cell type, only ~2% of genomic DNA is accessible to TFs, so most human DNA regulatory elements are not active (Klemm et al., 2019). Furthermore, accessibility is only one layer of control, other processes prevent DNA regulatory elements from being bound by TFs and driving transcription of their target genes, so DNA regulatory elements exist in several distinct functional chromatin states (Figure 1B) (Atlasi & Stunnenberg, 2017).

As the first layer of regulatory control, accessibility is mediated by pioneer transcription factor binding; pioneer TFs bind short DNA sequences exposed on the outside of nucleosomes (Cirillo et al., 2002; Soufi et al., 2015). Pioneer TFs "open" DNA by recruiting chromatin remodelling complexes, such as SWI/SNF, which eject nucleosomes from the locus substantially reducing the level of steric hinderance imposed by the bound nucleosomes (Wolf et al., 2023).

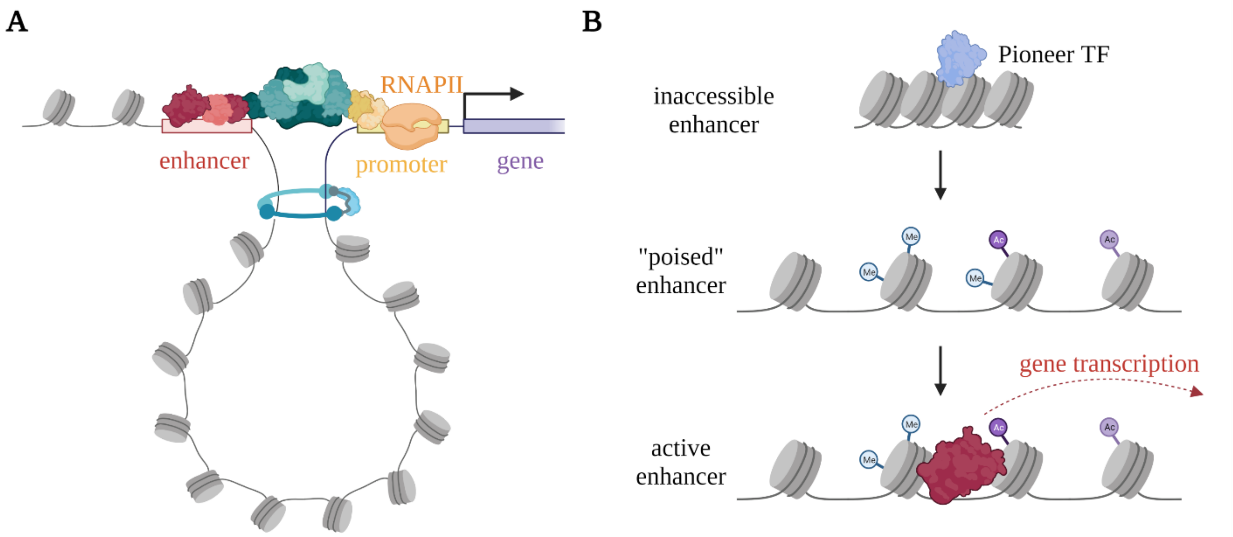

Figure 1: Schematic of enhancer concepts. (A) Enhancers interact with their target genes via chromatin looping. The TFs that bind enhancers interact with coactivators which regulate transcription initiation and RNAPII pause release. (B) Enhancers exist in three broad chromatin states: inaccessible, poised, and active. Only active enhancers, by binding TFs, drive transcription of their target genes. Poised enhancers are accessible but not active. They can contain several epigenetic modifications that are associated with active enhancers.

Once accessible, epigenetic features of DNA regulatory elements can be modified by a variety of epigenetic modifiers; this includes demethylation of DNA, post transcriptional modifications to histone tails, and many other biochemical modifications (Atlasi & Stunnenberg, 2017; Gasperini, Tome, & Shendure, 2020). Whether these epigenetic features are required for regulatory element activity remains a controversial question in the field (Morgan & Shilatifard, 2020), and recent work suggests many of them are generally dispensable for activity (Barnett et al., 2020; Dorighi et al., 2017; Douillet et al., 2020; Kreibich, Kleinendorst, Barzaghi, Kaspar, & Krebs, 2023; Rickels et al., 2017; Zhang, Zhang, Dong, Xiong, & Zhu, 2020). The distinct overlap of these marks can be used to place DNA regulatory elements into over a dozen different functional states (Ernst & Kellis, 2012; The ENCODE Project Consortium et al., 2020). While this many different chromatin states nicely demonstrates the vast complexity of this process and can be useful in other contexts, this level of detail can be overwhelming. For simplification purposes, inactive DNA regulatory elements that are accessible and contain various combinations of these epigenetic features are generally designated as "poised" regulatory elements.

In contrast, "active" DNA regulatory elements are both accessible and actively drive transcription of their target genes. Compared to the other functional chromatin states, active DNA regulatory elements play the most direct role on gene expression, so identifying them is important for understanding which genomic loci regulate a given cell identity. While several assays can profile which DNA regulatory elements are accessible, discerning whether accessible elements are "active" is a vitally important distinction to make when identifying which non-coding DNA sequences dictate cell identity and cause specific diseases when dysfunctional.

Profiling DNA Regulatory Element Activity

Epigenetic Annotations

There are several approaches to identify DNA regulatory elements in the human genome. Promoters can be easily identified based on their fixed, proximal distance to the genes they regulate, but enhancers are much more difficult to identify. One of the first approaches to identify enhancers used a comparative genomics framework to identify regions of high sequence conservation (Lindblad-Toh et al., 2011). This approach can yield many genomic regions with functional constraint, but they cannot determine which DNA regulatory elements are used by a particular cell type or whether the DNA regulatory region has an activating, silencing, or neutral effect on gene expression. Therefore, assays that functionally characterize the genome, collectively called "functional genomic methods" are required to identify the enhancers driving gene expression programs that yield a given cell type.

One of the most common functional genomic methods to identify putative enhancers is to profile the genome for epigenetic marks associated with enhancer activity. This approach uses the presence of distinct combinations of epigenetic marks to infer DNA regulatory element activity and function. For example, a common approach is to profile the genome for the three histone modifications H3K27ac, H3K4me1, and H3K4me3 (Rada-Iglesias et al., 2011). The relative abundance of these epigenetic marks classifies DNA regulatory elements into active enhancers, active promoters, poised enhancers, and poised promoters (Creyghton et al., 2010; Villar et al., 2015). Another approach classifies enhancers as non-promoter regions of bidirectional transcription using either PRO-seq or CAGE-seq to demarcate sites of RNA polymerase II (RNAPII) engagement (Abugessaisa et al., 2021; Andersson et al., 2014). Approaches like these are convenient, because they reflect the endogenous state of the DNA regulatory element and leverage well-defined methods. However, they do not directly measure enhancer activity, but rather infer it by association with these marks. Furthermore, there is disagreement on which epigenetic marks most accurately identify active enhancers, so many enhancer studies define enhancers differently. This is problematic because use of different epigenetic marks yields enhancer sets that do not agree with each other in terms of their identity, the number of active enhancers called overall, and their evolutionary and functional characteristics (Benton, Talipineni, Kostka, & Capra, 2019). While these approaches are unquestionably useful for identifying candidate enhancers, the lack of a direct functional measure of enhancer activity may cause them to be misleading. This approach may be substantially more accurate in the future when studies into the biochemical mechanisms of enhancer-mediated transcription uncover better molecular identifiers of active DNA regulatory elements.

Massively Parallel Reporter Assays

The initial characterization of enhancers described them as DNA sequences capable of driving transcription of a plasmid reporter gene irrespective of its orientation or distance to the gene (Banerji, Olson, & Schaffner, 1983). This description led to the development of reporter assays as tools to identify enhancers in the human genome (Sadowski, Ma, Triezenberg, & Ptashne, 1988; Triezenberg, LaMarco, & McKnight, 1988). In these assays, a single candidate test sequence is cloned into a transcriptional reporter plasmid and assayed for its ability to drive transcription of the reporter gene. This approach, however, is low throughput since only one DNA sequence is tested at a time. For this reason, Massively Parallel Reporter Assays (MPRAs), which leverage next-generation sequencing, were developed to test the regulatory activity of thousands of DNA sequences at once in order to identify putative DNA regulatory elements in a high-throughput manner (Melnikov et al., 2012; Patwardhan et al., 2009; Santiago-Algarra, Dao, Pradel, Espana, & Spicuglia, 2017). One advantage of MPRAs is that, unlike the epigenic profiling methods described above, they directly measure regulatory activity. To do this, MPRAs clone candidate DNA sequences into a reporter plasmid on a massively parallel scale. This resulting plasmid library is transfected into cells and all candidate sequences are assayed for regulatory activity at once. MPRAs vary in design, but all follow the same overall logic. The test sequences that have regulatory activity drive transcription of the reporter gene and yield "reporter RNAs" which are sequenced and matched to their associated test sequence. Their abundance in the reporter RNA pool is normalized to their abundance in the plasmid DNA input, and the greater reporter RNA to plasmid DNA ratio, the more active the test sequence (Santiago-Algarra et al., 2017).

While MPRAs directly measure the ability of a DNA sequence to drive transcription, they too have limitations. By removing DNA sequence from their native environment and placing them within a plasmid reporter, MPRAs are exogenous assays and it is hard to know for certain if their activity on a plasmid recapitulates their activity at their endogenous locus. To circumvent this issue, several groups have conducted "lenti-MPRAs" which insert reporter constructs into chromatin using lentivirus particles (Inoue et al., 2017). While some differences in regulatory activity were observed between integrated and non-integrated MPRAs, overall, the results are largely similar. Furthermore, the lenti-MPRA approach is confounded in that the insertion site is a different chromatin landscape from the endogenous locus. Moreover, some applications are not amenable to lentiviral integration and for the ones that are, they require additional experimental steps that can prove to be tricky. It therefore does not appear that the improvements in accuracy outweigh the technical challenges when considering a lentiviral versus episomal approach. In most cases, it seems that an episomal MPRA is sufficient.

There are many types of episomal MPRAs, each tailored to a specific purpose. Self-transcribing active regulatory region sequencing (STARR-seq) is one type of MPRA that is uniquely designed to assay an entire genome for regulatory activity (Arnold et al., 2013; Inoue et al., 2017; Kircher et al., 2019; Maricque, Dougherty, & Cohen, 2017; Melnikov et al., 2012; Muerdter et al., 2018; Patwardhan et al., 2012). STARR-seq quantifies regulatory activity genome-wide by cloning randomly fragmented genomic DNA into the 3'UTR of the reporter plasmid. Because the test sequence is contained within the 3'UTR of the reporter RNAs that are produced, active DNA regulatory elements will drive transcription of themselves, so activity is quantified by the abundance of DNA regulatory element sequences within the reporter RNA pool.

STARR-seq, which was developed for use in the *Drosophila melanogaster* genome, is a tremendously elegant approach, but has major limitations when applied to the human genome. Because the human genome is about 20 times larger than the *Drosophila* genome, it is technically challenging to accommodate all 3 billion base pairs of human DNA in one assay. Whole human genome STARR-seq requires large-scale cloning procedures and can only produce shallow sequencing coverage of human regulatory elements (Johnson et al., 2018). In addition, STARR-seq assays both accessible and inaccessible chromatin. Because only 2% of the human genome is accessible in any given cell type, ~98% of all regions assayed in whole human genome STARR-seq are inaccessible and therefore inactive endogenously (Klemm et al., 2019). In this way, nearly all of the regulatory information provided by this approach is for regions that would not drive transcription of nearby genes in their endogenous chromatin state. Therefore, most reads map to inaccessible chromatin so that whole human genome STARR-seq is overall inefficient.

Because of this limitation, recent methods have been developed to reduce the scope of the assay to accessible chromatin. These approaches, HiDRA and FAIRE-STARR-seq, accomplished this by combining STARR-seq with techniques that capture accessible chromatin to specifically test the regulatory potential of accessible DNA (Chaudhri, Dienger-Stambaugh, Wu, Shrestha, & Singh, 2020; Glaser et al., 2021; X. Wang et al., 2018). As a result, these methods only sample 2% of the human genome while assaying nearly all regulatory elements capable of driving transcription endogenously. This approach enables deeper sequencing coverage of all biologically relevant regulatory regions.

Although HiDRA and FAIRE-STARR-seq have been performed previously, they have not been characterized in-depth or developed to their full capability. Because these assays combine accessibility and regulatory activity methods, they have the potential to reveal multiple levels of gene regulatory information simultaneously, but this potential has not been explored. Additionally, important parameters of these methods, such as effects from query sequence length, effects from orientation of insert on plasmid, and the development of an optimal data analysis strategy have not been investigated. These methods, like most other MPRA approaches, have also largely ignored detection of silencing activity, even though, in theory, they could identify silencers. Altogether, these approaches require better characterization and an expansion of their capabilities to address difficult questions in transcriptional regulation, such as profiling of gene regulatory divergence between species.

Gene Regulatory Divergence Between Species

Humans are among the most complex organisms on earth, yet we have the same number of genes as *Caenorhabditis elegans*, a substantially less complex organism that is only made up of 959 total somatic cells (Kimble & Hirsh, 1979; Sulston & Horvitz, 1977; Sulston, Schierenberg, White, & Thomson, 1983). So how are humans so different from *C. elegans* if the number of genes is the same? In general, the complexity of the human species arises from drastically different levels of gene regulation. At every step from gene to protein, humans have extremely complex and expansive gene regulatory mechanisms when compared to simpler species like *C. elegans* (Ledford, 2008). This increased regulatory complexity, due in large part to an increased size of the non-coding genome, allows humans to generate over a million different proteins from the ~20,000 genes encoded in the human genome (Aebersold et al., 2018). Furthermore, these regulatory processes fine-tune the amounts of those proteins, so that over 200 different cell types can be produced from the same genome (Heinz et al., 2015). In addition, this increased complexity, allows for highly specialized processes like adaptive immunity to exist. As species become more closely related, such as in the case of humans and non-human primates, the complexity of their gene regulatory mechanisms is much more similar. In this case, the phenotypic differences are determined largely by divergent use of the same gene regulatory processes rather than in the proteins and the regulatory pathways that make them up.

Differences in gene expression control phenotypic changes between humans and non-human primates

In 1971, Britten and Davidson proposed that phenotypic changes between organisms may be driven primarily by changes in expression of gene rather than changes in their identity (Britten & Davidson, 1971; King & Wilson, 1975). Four years later, Mary-Claire King and Allan Wilson famously extended this hypothesis to phenotypic differences between humans and chimpanzees. They noted that there are not enough changes in protein sequences to explain the differences in phenotype, and therefore proposed that alterations to gene regulation is what drove evolution of humans from our most recent common chimpanzee ancestor over 8 million years ago. This hypothesis is attractive because many of the genes involved in human-specific phenotypes, particularly those related to morphology, are pleiotropic and important for several other biological processes, including development. By changing the regulation of genes, one can alter their expression in only one cell type. Whereas changing the sequence of the gene itself will alter its overall function in all cell types. In other words, changes in gene regulation provide a better mechanism for fine-tuning gene expression to adapt and evolve phenotypes that improve overall fitness than changing protein coding sequences (Reilly & Noonan, 2016).

Since 1975, this gene regulatory hypothesis has largely been validated by several studies. Most convincingly, a survey of gene expression levels between similar cell types across primate species reveals phylogenetic relationships that accurately reflect the true evolutionary relationships between species (Brawand et al., 2011). Interestingly, the degree of gene expression divergence between two closely related species varies widely across different cell types and tissues, which is likely due to different functional constraints on the biological function of each tissue (Reilly & Noonan, 2016). In addition, gene expression levels between species of the same cell type are much more similar than gene expression levels between cell types of the same species indicating that cell-type specific gene expression changes are favored over pleiotropic changes that effect many cell types (Brawand et al., 2011). Critically, most gene expression does not need to change in order to produce a phenotypic outcome, as only ~10-39% of genes display divergent expression depending on the cell types and species that are being compared (Brawand et al., 2011; Reilly & Noonan, 2016).

Differential Enhancer Activity Drives Gene Expression Divergence Between Species

The changes in gene expression across species are ultimately caused by changes in gene regulation, which are primarily mediated by changes in DNA regulatory element activity. Several studies have compared epigenetic profiles of DNA regulatory elements between similar cell types across species. In general, changes in epigenetic modifications correlate with changes in gene expression (Cain, Blekhman, Marioni, & Gilad, 2011; Zhou et al., 2014), with one study finding that divergent epigenetic profiles explain ~42% of the gene expression variance between humans and chimpanzee lymphoblastoid cell lines (Zhou et al., 2014). Because gene expression changes between primate species are cell-type specific, regulatory activity divergence more often occurs at enhancers than promoters (Villar et al., 2015). Altogether, differential regulation of enhancers drives differences in gene expression that produce species-specific phenotypes.

Gene Regulatory Divergence in *Cis* and *Trans*

Changes in gene regulation can occur in either *cis* or in *trans*. How these two terms are defined depend on what the unit of measure is, but broadly, *cis* changes are local substitutions to the nearby DNA sequence, whereas *trans* changes are global, cell environment changes to diffusible products, like transcription factors (Hill, Vande Zande, & Wittkopp, 2020; Signor & Nuzhdin, 2018; Vande Zande, Hill, & Wittkopp, 2022). These two modes of change are very different approaches of altering gene regulation to adapt to a given environmental stress. Because *cis* changes are local and typically only affect the expression of one gene, their effects are precise but overall small, making them less likely to be overall deleterious. However, changes to many genes may be required if evolutionary pressures are strong and require rapid adaptation. Because *cis* changes only affect one gene, many changes would have to occur, which may be too slow. By contrast, a single *trans* change affects the entire cellular environment so it can alter the expression of many genes at once. Therefore, *trans* changes can allow the species to more rapidly adapt to a given evolutionary pressure than changes in *cis.* However *trans* changes are more likely to be pleiotropic and deleterious to overall fitness (Hill et al., 2020). In some way, *trans* effects are encoded in the genome, so *cis* changes could ultimately cause *trans* changes. It is important to note that the terms *cis* and *trans* are respective to the unit of divergence. While a *trans* change must arise from a genetic difference between species, this genetic difference has a different effect on gene regulation than a mutation that affects divergence only in *cis,* so they are under different selective pressures during evolution.

Parsing *cis* and *trans* effects on gene regulatory divergence has been primarily investigated at the level of gene expression, as measured by differences in mRNA transcript levels between closely related species. Because DNA sequence changes and cellular environment changes are inherently linked within an endogenous setting, these studies leverage unique methods that can directly and exclusively test divergence in either *cis* or *trans.* One approach is to measure allele-specific expression differences within a common cellular environment so that changes in *cis* are compared in a common *trans*-regulatory setting. A common way to do this involves mating two closely related species so that they generate F1 hybrids; gene expression for each allele is measured in the hybrids, which represents a common cellular environment, and this is then compared to the expression in the parental environments. This approach has been applied to many different taxa including *Drosophila,* yeast*,* plants, and mice (Coolon, McManus, Stevenson, Graveley, & Wittkopp, 2014; Emerson et al., 2010; Goncalves et al., 2012; McManus et al., 2010; Osada, Miyagi, & Takahashi, 2017; Shi et al., 2012; Tirosh, Reikhav, Levy, & Barkai, 2009; Wittkopp, Haerum, & Clark, 2004, 2008). Overall, these studies have yielded widely different measures of the relative abundance of *cis* and *trans* effects on gene expression, however, they commonly find that the proportion of *cis* effects increases with increased evolutionary divergence (Signor & Nuzhdin, 2018).

Performing F1 hybrid studies is impractical when investigating *cis* and *trans* divergence between primate species. One recent study circumvented this limitation by measuring allele-specific expression in a fused human-chimpanzee tetraploid iPSC cell line (Agoglia et al., 2021). By comparing gene expression in the fused cell line—the common hybrid environment—to expression in the native cell lines, this approach is similar to the experimental logic of F1 hybrid studies. The authors of this study found that ~39% of differentially expressed genes can be explained by divergence in *cis.* Taking this a step further, another study, using this same tetraploid cell line, generated embryoid bodies, performed single-cell RNA-seq, and measured allele-specific expression of each cell type within the embryoid body (Barr, 2022). They found that, on average, ~70% of inter-species differences in gene expression could not be explained by changes in *trans*. Common cellular environment studies like these are powerful because they are well-controlled and the local differences of each allele on target gene expression can be easily identified.

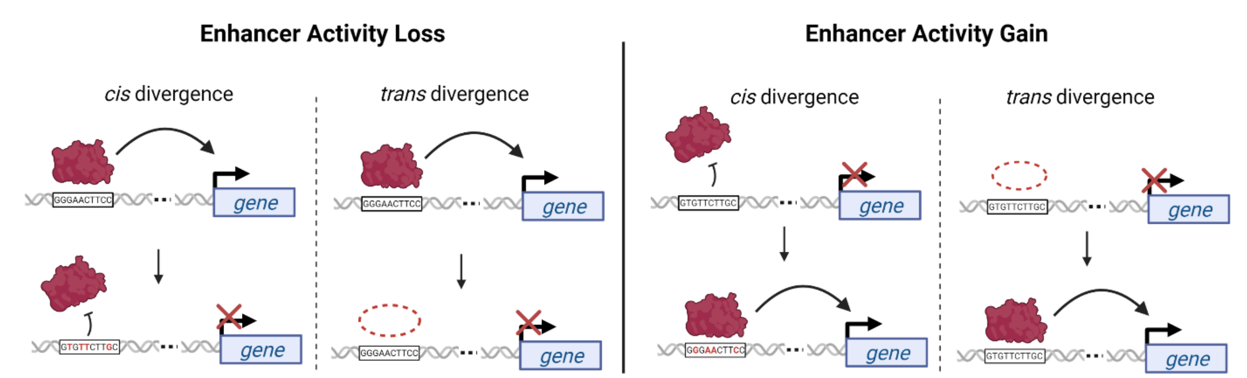

Figure 2: *Cis* and *trans* modes of divergence in gene regulatory element activity for both gains and losses in activity. *Cis* changes alter the DNA sequence of the enhancer, and these changes affect its own activity. *Trans* changes affect the cellular environment, and these changes affect the activity of the enhancer and likely many other enhancers. See also Figure 30D-E.

DNA Regulatory Element Activity Divergence in Cis and Trans

While most studies have focused on differences in mRNA levels as a way to measure the effect of *cis* and *trans* changes, these expression differences are ultimately mediated by divergent DNA regulatory element function. However, only a handful of studies have investigated *cis* and *trans* divergence directly on DNA regulatory element activity. With DNA regulatory element activity as the unit of measure, *cis* divergence is simply changes to the underlying sequence of the DNA regulatory element that alters its own function, so these changes only alter the activity of one regulatory element (Figure 2). On the other hand, *trans* divergence is changes to the cellular environment that affect regulatory element activity of many DNA regulatory elements at one time (Figure 2). For example, the differential abundance of a TF can alter all regulatory elements that bind that TF.

Like gene expression studies, investigations into *cis* and *trans* effects on DNA regulatory element activity are similarly challenged by an inherent link between genome and cellular environment, and they too must adopt unique methods that allow for the parsing of these two effects. To do this, researchers leverage MPRAs, which allow DNA sequences to be removed and tested outside of their native environment. One MPRA-based approach compares regulatory activity measures of homologous sequences between closely related species within a common cellular environment (Arnold et al., 2014; Klein, Keith, Agarwal, Durham, & Shendure, 2018; Uebbing et al., 2021; Weiss et al., 2021). By controlling the cellular environment, any effects from *trans*-regulatory differences between species are negated when assessing DNA regulatory element activity. While this approach allows direct identification of regulatory divergence in *cis*, it lacks direct assessment of regulatory activity changes in *trans*.

Another, more direct, approach compares regulatory activity of the same sequences across species-specific cellular environments (Gordon & Ruvinsky, 2012; Mattioli et al., 2020; Whalen et al., 2023). This MPRA-based approach allows direct identification of DNA regulatory element activity divergence in both *cis* and *trans*. In this way, these approaches can be leveraged to investigate the relative proportions of *cis* and *trans* divergence on regulatory element activity. The largest scale study to date using such an approach analyzed differential activity of ~1,600 homologous regulatory elements between human and mouse embryonic stem cells (Mattioli et al., 2020). They selected the ~1,600 regions to test, 268 of which are enhancers, from the FANTOM consortium, which uses a single biochemical feature—enhancer RNAs—to identify enhancers (Abugessaisa et al., 2021). They found 660 elements with regulatory divergence in *cis* and 293 elements with regulatory divergence in *trans.* This study, along with two other much smaller scale studies suggest that *cis* divergence primarily drives species-specific regulatory element activity between closely related species (Gordon & Ruvinsky, 2012; Mattioli et al., 2020; Whalen et al., 2023). However, these studies considered small, pre-selected subsets of regulatory elements, so their results represent only a small and selection-biased portion of the genome. Evolution over millions of years acts at genomic scale, so these studies lack a global view of how the *cis* and *trans* modes of divergence on regulatory activity drove existing gene regulatory differences between closely related species. Therefore, the field requires a comprehensive and unbiased survey of *cis* and *trans* contributions to global gene regulatory divergence to better understand the mechanisms driving gene regulatory evolution. This gap in knowledge is largely due to limitations of current technologies to profile DNA regulatory elements in the human genome.

Scope of Dissertation

In this dissertation I present two projects. In Chapter II, I present a substantially improved and well-characterized ATAC-STARR-seq method that allows simultaneous profiling of chromatin accessibility, TF occupancy, and DNA regulatory activity. When I began my PhD, ATAC-STARR-seq was a completely novel idea and had not been developed yet. Since then, two versions of ATAC-STARR-seq—HiDRA and FAIRE-STARR-seq—were published before ours. These methods applied slightly different techniques to accomplish the same overall goal of performing STARR-seq on accessible chromatin sequences, but they did not explore key parameters of the assay and realize the full potential of these methods. We present our own version of ATAC-STARR-seq as a substantial improvement to these techniques and unlike the previous methods, we also provided the field with computational and technical support to increase accessibility of the method to others. The overall goal of developing ATAC-STARR-seq was to create a method that allowed us and others to investigate exciting biological questions that were not previously possible. At its core, ATAC-STARR-seq now allows researchers to identify all biologically relevant DNA regulatory elements in the human genome with an assay that directly quantifies regulatory activity. Given how much non-coding DNA sequences play a role in human disease, the value of this method cannot be understated.

In Chapter III, I use ATAC-STARR-seq to investigate the respective contributions of *cis* and *trans* changes on DNA regulatory element activity between human and rhesus macaque on a genome-wide scale to understand the preferred mode of DNA regulatory element activity evolution between closely related species. This question was limited by the available technologies, and we could not have investigated this second story without first developing ATAC-STARR-seq. Overall, we observe a greater role for *trans*-regulatory mechanisms driving primate evolution than previously appreciated and identify that changes in both *cis* and *trans* affect most divergent active regulatory regions.

CHAPTER II[1]

ATAC-STARR-SEQ REVEALS TRANSCRIPTION FACTOR-BOUND ACTIVATORS AND SILENCERS ACROSS THE CHROMATIN ACCESSIBLE HUMAN GENOME

Introduction

Transcription is regulated by transcription factors (TFs) and the DNA sequences they bind, called *cis*-regulatory elements. Enhancers, which are a class of *cis*-regulatory elements, are distally located from the genes they target and serve as key drivers of cell-type specific gene expression (Heinz et al., 2015). Because enhancers require TF binding, they are largely dependent on chromatin accessibility to elicit transcriptional activity. Therefore, chromatin accessibility is a vital regulator of enhancer function, and this is evidenced by the observation that ~94% of all ENCODE TF ChIP-seq peaks fall within accessible chromatin (Klemm et al., 2019). In any given cell type, only a small fraction (~2%) of the genome is accessible to TF binding (Klemm et al., 2019; Thurman et al., 2012). In this way, most enhancers are inaccessible and are less likely to drive transcription endogenously.

Enhancers are difficult to identify and validate because they lack uniform features and are less constrained by gene proximity than promoters (Gasperini et al., 2020). Massively parallel reporter assays (MPRAs) were developed to test the regulatory potential of thousands to millions of DNA sequences in parallel, providing high-throughput identification of putative enhancers.

[1] This chapter is adapted from "ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome" published in Genome Research and has been reproduced with the permission of the publisher and my co-author Emily Hodges, Ph.D. | Citation: "Hansen, T. J., & Hodges, E. (2022). ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome. Genome Research, 32, 1529-1541. doi:10.1101/gr.276766.122"

Overall, MPRAs test the regulatory potential of genomic regions by cloning them *en masse* into a reporter plasmid and leveraging high-throughput sequencing to quantify regulatory activity (Santiago-Algarra et al., 2017). Among the variety of different vector backbones and assay designs applied to MPRAs, Self-Transcribing Active Regulatory Region sequencing (STARR-seq) is uniquely designed to assay an entire genome for regulatory activity (Arnold et al., 2013; Inoue et al., 2017; Kircher et al., 2019; Maricque et al., 2017; Melnikov et al., 2012; Muerdter et al., 2018; Patwardhan et al., 2012). STARR-seq quantifies regulatory activity genome-wide by cloning randomly fragmented genomic DNA into the 3'UTR of the reporter plasmid. Thus, active enhancers drive transcription of themselves, and activity is quantified by the abundance of its own sequence in the transcript pool, removing the need for barcodes that some MPRAs employ. One major limitation of STARR-seq is that it is technically challenging to accommodate the massive size of the human genome; it requires large-scale cloning procedures and produces shallow sequencing coverage of human regulatory elements (Johnson et al., 2018). In addition, STARR-seq assays both accessible and inaccessible chromatin. Thus, many assayed regions are derived from heterochromatin and are less likely to be transcriptionally active in the cell type in question.

To narrow the scope of the assay, recent methods have combined STARR-seq with techniques that capture accessible chromatin to specifically test the regulatory potential of accessible DNA (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013; Chaudhri et al., 2020; Glaser et al., 2021; X. Wang et al., 2018). As a result, these methods only sample a fraction of the human genome (~2%) while assaying nearly all regulatory elements capable of driving transcription endogenously, because they are derived from open chromatin. This approach remains comprehensive while enabling deeper sequencing coverage of biologically relevant genomic regions. Furthermore, integrated approaches have recently been described that combine measurements of chromatin accessibility with analysis of transcription and other epigenomic features from a single population of cells (Barnett et al., 2020; Chen et al., 2022; Clark et al., 2018; Kelly et al., 2012). Similarly, ATAC-STARR-seq has the potential to reveal multiple levels of gene regulatory information simultaneously, but this potential has not been explored. In addition, a complete understanding of gene regulatory activity is lacking with most MPRA approaches because silencing activity is largely overlooked, with a few recent exceptions (Doni Jayavelu, Jajodia, Mishra, & Hawkins, 2020; Y. S. Kim et al., 2021; Pang & Snyder, 2020); this is potentially due to technical caveats of distinguishing silencers from either that of missing data or interference from head-on transcriptional conflicts or post-transcriptional silencing mechanisms.

Here, we demonstrate a new workflow that substantially expands the capabilities of ATAC-STARR-seq to extract and measure gene regulatory information. Using this approach, we aimed to identify both activators and silencers, as well as to simultaneously profile chromatin accessibility, and perform TF footprinting. From a single ATAC-STARR-seq dataset, a multi-layered, integrated view of the human genome can be captured—a feature that has not been explored previously. We provide a protocol and code repository so that this new ATAC-STARR-seq workflow may be easily used and adopted by the field.

Results

ATAC-STARR-seq Experimental Design

The ATAC-STARR-seq approach is divided into the three main parts: 1) ATAC-STARR-seq plasmid library generation, 2) reporter assay, and 3) data analysis (Figure 3A). To generate ATAC-STARR-seq plasmid libraries, nuclei are isolated from a cell type of interest and exposed to Tn5, the cut-and-paste transposase used in the ATAC-seq method (Buenrostro et al., 2013). Tn5 simultaneously cleaves DNA fragments within accessible chromatin and attaches customizable sequence adapters to their 5’ ends. ATAC-STARR-seq adapters are designed to serve as homology arms for direct Gibson cloning into the STARR-seq reporter plasmid, which enables cloning of accessible DNA fragments *en masse*. The resulting ATAC-STARR-seq plasmid library consists of millions of unique plasmids each harboring their own unique open chromatin-derived DNA fragment.

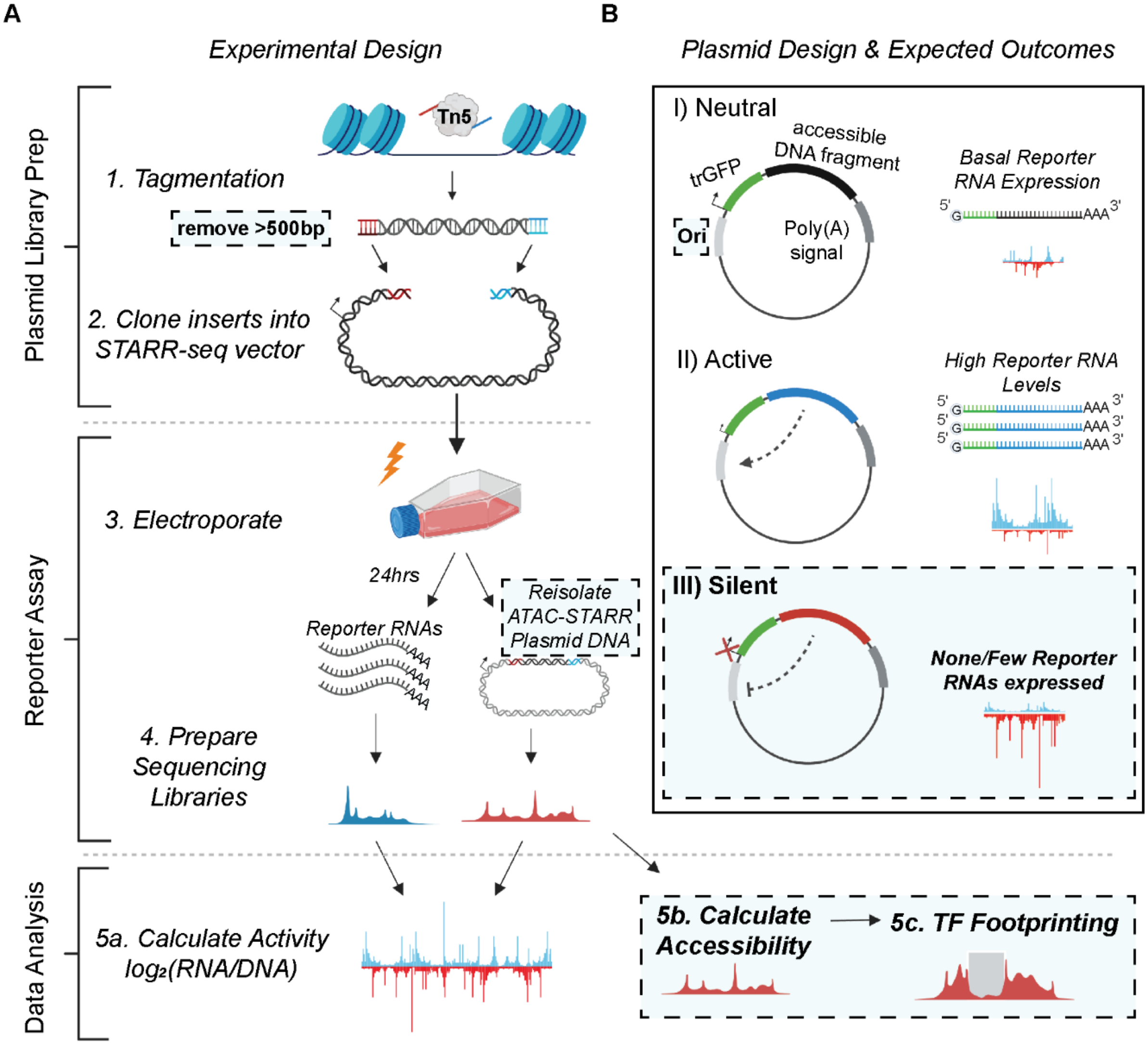
A
Experimental Design
Plasmid Library Prep
Tn5
1. Tagmentation
remove >500bp
2. Clone inserts into STARR-seq vector
Reporter Assay
3. Electroporate
24hrs
Reporter RNAs
Reisolate ATAC-STARR Plasmid DNA
4. Prepare Sequencing Libraries
Data Analysis
5a. Calculate Activity log2(RNA/DNA)
B
Plasmid Design & Expected Outcomes
I) Neutral
accessible DNA fragment
trGFP
Ori
Poly(A) signal
Basal Reporter RNA Expression
II) Active
High Reporter RNA Levels
III) Silent
None/Few Reporter RNAs expressed
5b. Calculate Accessibility
5c. TF Footprinting

Figure 3: Schematic of the ATAC-STARR-seq methodology. (A) The experimental design of ATAC-STARR-seq consists of three parts: plasmid library generation, reporter assay, and data analysis. Open chromatin is isolated from cells with the cut and paste transposase Tn5 and only large DNA fragments (>500bp) are removed. The open chromatin fragments are cloned into a reporter plasmid and the resulting clones—called an ATAC-STARR-seq plasmid library—are electroporated into cells. 24 hours later, both reporter RNAs (blue)—which are transcribed directly off the ATAC-STARR-seq plasmid—and ATAC-STARR-seq plasmid DNA (red) are harvested, and Illumina-sequencing libraries are prepared and sequenced. The resulting ATAC-STARR-seq sequence data is analyzed to extract regulatory activity, chromatin accessibility, and transcription factor footprints. (B) Reporter plasmid design and the expected outcomes for neutral, active, and silent regulatory elements. Each ATAC-STARR-seq plasmid within a library contains a truncated GFP (trGFP) coding sequence, a poly-adenylation signal sequence, an origin of replication (Ori) (which moonlights as a minimal core promoter), and the unique open chromatin fragment being assayed. Since the accessible region is contained in the 3' UTR, the abundance of itself in the transcript pool reflects its activity. In this way, neutral elements do not affect the system and reporter RNAs are expressed at a basal expression level dictated by the minimal core promoter, the Ori. Accessible chromatin fragments that are active express reporter RNAs at a higher level than the basal expression level, while silent elements repress the Ori and reporter RNAs are expressed at a lower level than basal expression. Dashed boxes represent new components of the ATAC-STARR-seq assay design and workflow.

In our updated ATAC-STARR-seq workflow, we employ the STARR-seq Ori backbone, where the origin of replication (Ori) functions as the minimal promoter (Muerdter et al., 2018) (Table 1). Each plasmid in the ATAC-STARR-seq plasmid library contains a truncated GFP (trGFP) coding sequence, a poly-adenylation signal sequence, the Ori, and the unique accessible DNA fragment being assayed (Figure 3B). Critically, the accessible region is cloned into the 3' UTR, so if the accessible region is active, it interacts with the Ori to drive self-transcription. Thus, an accessible region's level of activity is reflected by its own level of expression. Transcripts from ATAC-STARR-seq plasmids, termed "reporter RNAs", are expressed at basal levels from the activity of the Ori itself. This allows detection of silencing activity—the inhibition of the basal expression—in this assay.

Following its creation, the ATAC-STARR-seq plasmid library is transfected via electroporation into a given cell line. From the same flask of cells, both reporter RNAs and plasmid DNA are harvested 24 hours later, then prepared as Illumina sequencing libraries and sequenced. Activity is calculated as the $\log_2$ ratio between normalized read counts from the reporter RNA and plasmid DNA datasets. The re-isolation of plasmid DNA recovers only the ATAC-STARR-seq plasmids that were successfully transfected, thus providing a more accurate representation of the "input" sample than sequencing without transfection. Table 1 provides a comparison of experimental and analytical features as well as reported data metrics for the current ATAC-STARR design and previously reported approaches (Chaudhri et al., 2020; X. Wang et al., 2018).

Table 1: A comparison of experimental differences and result metrics between accessible chromatin coupled to STARR-seq techniques.

Type	***Description***	**ATAC-STARR-seq (Hansen & Hodges)**	**HiDRA (Wang et al. 2018)**	**FAIRE-STARR-seq (Chaudhri et al. 2020)**
Experimental Differences	*Cell type*	GM12878	GM12878	Purified murine splenic B cells
	Accessible chromatin extraction process	ATAC-seq (Tn5-tagmentation)	ATAC-seq (Tn5-tagmentation)	FAIRE-seq (crosslinking-based)
	mtDNA removal process	Omni-ATAC (detergent-based)	CRISPR against mtDNA gRNAs	none
	Size selection	0-500bp	150-500bp	300-700bp
	Reporter plasmid promoter	Bacterial origin of replication (ORI)	Super Core Promoter 1	Super Core Promoter 1
	Manner of plasmid library sequence library preparation	Reisolated after electroporation (in parallel with reporter RNAs)	Sequenced as-is, no reisolation	Not sequenced
	Analysis	Sliding windows & DESeq2	Fragment groups & DESeq2	Homer *findPeaks*, no normalization to DNA
Result metrics	*Library Complexity*	~50 million	9.7 million	Not reported directly, ~81% coverage of input
	Number of active regions called	30,078 active regions	66,254 active HiDRA regions	11,809 STARR-positive regions
	Number of silent regions called	21,125 silent regions	*None reported*	*None reported*
	Number of accessible chromatin peaks called	101,904 peaks	*None reported*	55,133 peaks (from FAIRE-seq not the plasmid library)
	Number of TFs footprinted	746 TFs	*None reported*	*None reported*
	Number of SHARPER-RE driver elements identified	*None reported*	~13,000	*None reported*

The GM12878 ATAC-STARR-seq plasmid library is highly complex

Following the experimental design outlined above, we tagmented GM12878 cells and generated an ATAC-STARR-seq plasmid library. A successful ATAC-STARR-seq experiment is predicated on maintaining complexity at all stages of the protocol. We estimated the initial complexity of our ATAC-STARR-seq plasmid library by sequencing the library at low depth and estimating the number of unique reads with the Preseq software package (Daley & Smith, 2013) (Figure 4A). The GM12878 ATAC-STARR-seq plasmid library contains a maximum complexity of about 50 million unique accessible DNA fragments, providing ample coverage of accessible loci.

24hrs post-transfection is the optimal time to harvest ATAC-STARR-seq reporter RNAs

The introduction of plasmid DNA into cells produces an interferon-stimulated gene response that can confound the isolation of biologically relevant regulatory activity (Muerdter et al., 2018). To minimize this interference in our data, we determined the optimal incubation time between electroporation and harvest. Two factors play an important role in determining when to harvest RNA: global reporter RNA expression levels and the timing of interferon stimulated gene response to STARR-seq reporter plasmid DNA. To investigate both factors, we electroporated ATAC-STARR-seq plasmid DNA, isolated poly-adenylated RNA at several time points after transfection, quantified RNA expression with qPCR, and compared to an untransfected sample (Figure 4B). An increase in reporter RNA expression is observed at 3 hours (the earliest timepoint) and remains stable at later time points. We measured expression of *IFNB1*, *IFIT2*, and *ISG15* to

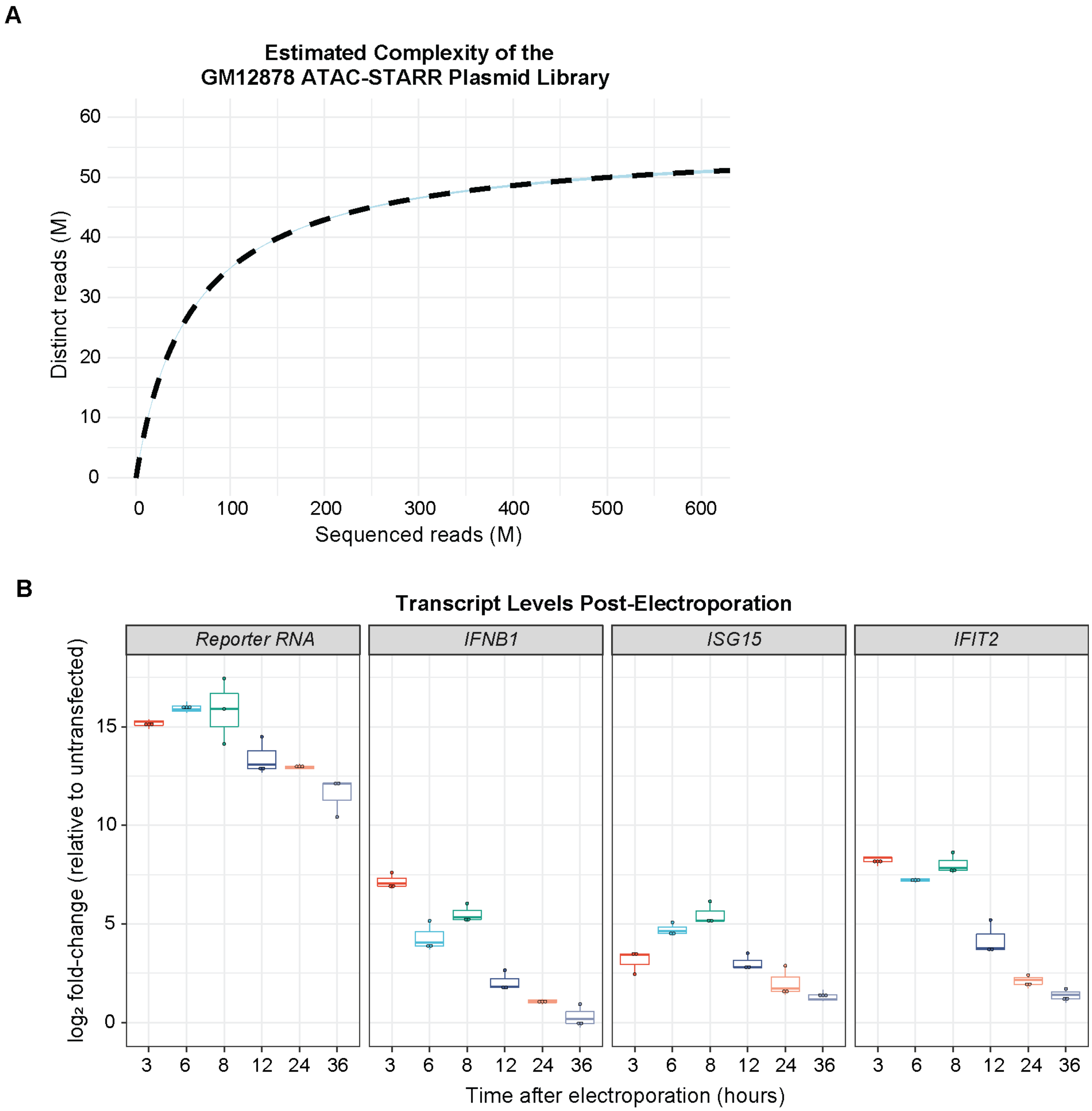

Figure 4: ATAC-STARR Optimization. (A) Estimated complexity curve for the GM12878 ATAC-STARR plasmid library. Dashed lines represent predicted values from Preseq's lc-extrap. The associated ribbon plots (light blue) represent the 95% confidence interval reported with the predicted value. (B) Relative expression of reporter RNAs and three interferon-stimulated genes (IFNB1, IFIT2, and ISG15) at varying timepoints between 0- and 36-hours post-electroporation. For each analysis, fold-change values are relative to the untransfected condition. Three replicates were isolated and quantified for each timepoint.

characterize the interferon stimulated gene response in our system. RNA expression for all three genes increases initially but returns to baseline by 24 hours. Given the persistent level of reporter

RNAs and the attenuated interferon stimulated gene response in our system, we decided to harvest 24 hours after electroporation. Together, this allows us to capture reporter RNAs that reflect steady-state regulatory properties of GM12878 accessible regions without sacrificing reporter RNA recovery.

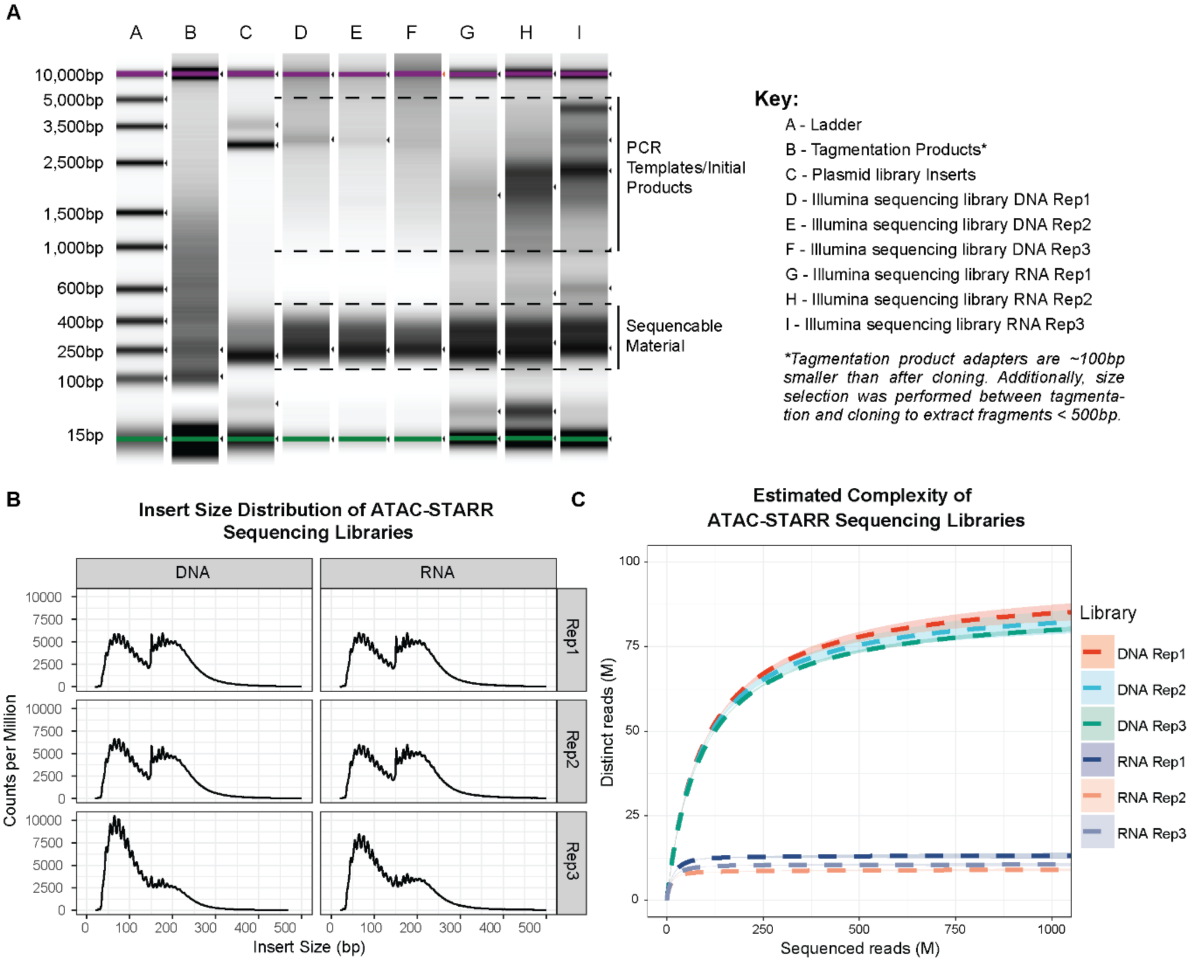

Figure 5: Characterization of ATAC-STARR sequencing libraries. (A) Agilent Tapestation results for relevant steps of ATAC-STARR, this includes the following: tagmented products, plasmid library inserts, and Illumina sequencing libraries for all three replicates of DNA and RNA. Tagmented products lack the full Illumina adapter and therefore are about 100bp smaller than their later-stage counterparts. They also include larger fragments which were removed via selection before the cloning step. The Illumina-ready libraries were amplified using a minimal PCR cycle number and therefore the plasmid or cDNA template as well as the first and second round products can be seen as larger material—this material is not sequenceable as it lacks at least one of the adapters required for cluster amplification. (B) Insert size distribution of ATAC-STARR-seq reads, as quantified by Picard's CalculateInsertSizeMetrics. (C) Estimated complexity curves for ATAC-STARR sequencing libraries. Dashed lines represent predicted values from Preseq's lc-extrap. The associated ribbon plots (light blue) represent the 95% confidence interval reported with the predicted value.

ATAC-STARR-seq maintains nucleosome profiles of Tn5 selected DNA fragments

For a total of three replicates, we then transfected the library into GM12878 cells and harvested both reporter RNAs and plasmid DNA from the same flask of cells 24 hours later. Using the captured reporter RNAs and plasmid DNA, we prepared Illumina sequencing libraries for each replicate and submitted for sequencing. The size distribution of the accessible DNA fragments remained consistent throughout the ATAC-STARR-seq procedure and displayed the characteristic nucleosome banding and DNA pitch typified by ATAC-seq fragment libraries (Figure 5A,B). Analysis of library complexity between replicates revealed an average maximum complexity of 90 million unique fragments for input DNA, and 10 million unique fragments for reporter RNAs (Figure 5C). The difference between RNA and DNA complexities is likely due to higher duplication rates in the RNA samples (Table 2) driven by both the expression of multiple

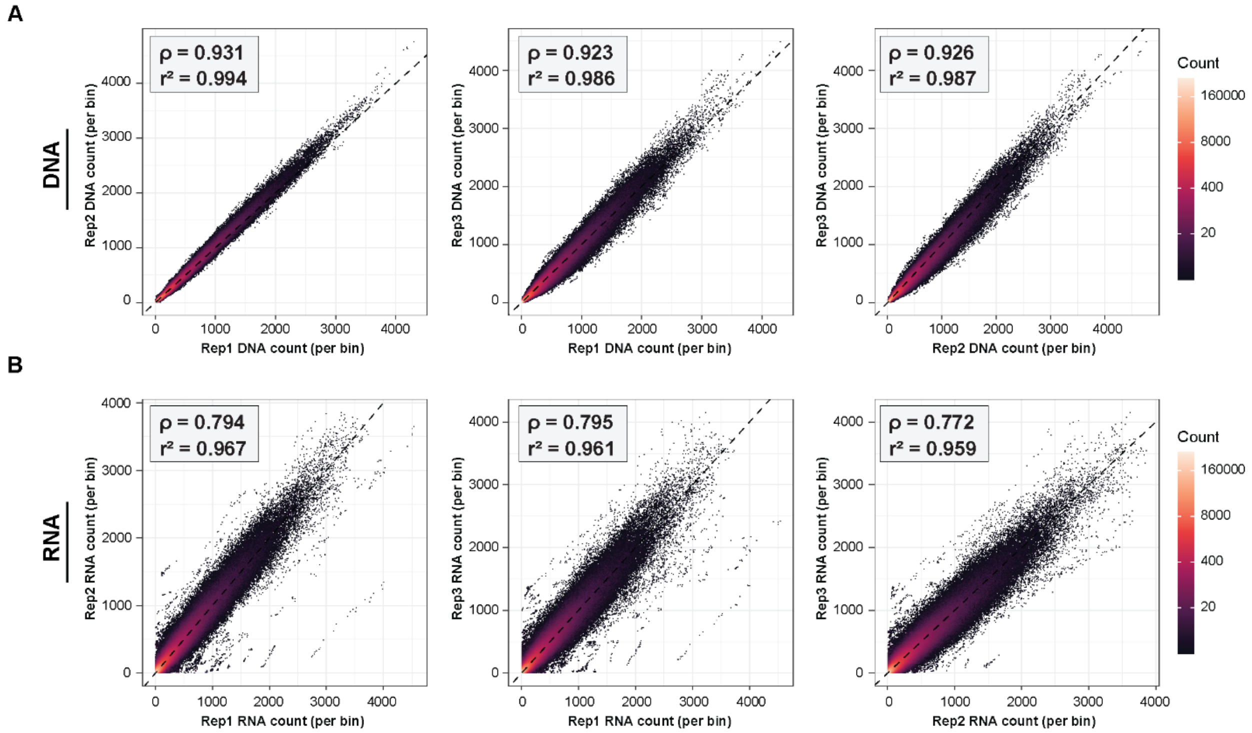

Figure 6: Correlation between ATAC-STARR-seq replicates. Scatter plots of DESeq2-normalized read counts per bin between replicates for both (A) DNA and (B) RNA samples. Pearson (r2) and spearman (ρ) correlation coefficients are indicated in the top left corner for each pairwise comparison.

transcripts per plasmid and more PCR cycles required for the RNA samples. In addition, for both RNA and DNA samples, replicates displayed high Pearson (r^2: 0.96-0.99) and Spearman's (ρ: 0.77-0.93) correlation coefficients indicating strong agreement among the three replicates assayed (Figure 6). Altogether the ATAC-STARR-seq sequence libraries demonstrated the necessary quality and complexity for downstream analysis.

Table 2: ATAC-STARR-seq sequencing summary statistics. Plasmid library column represents data from the library complexity check.

Metric	Plasmid Library	DNA Rep 1	DNA Rep 2	DNA Rep 3	RNA Rep 1	RNA Rep 2	RNA Rep 3
Total read count (paired end)	113,978,542	55,453,364	47,609,989	81,350,911	101,163,327	122,274,760	103,410,392
Filtered read count (paired end)	66,730,249	30,803,098	26,530,451	44,046,983	56,307,716	67,956,476	56,098,454
Filtered & deduplicated read count (paired end)	29,482,015	22,626,181	20,015,687	28,369,114	11,385,851	8,122,462	9,285,796
Trimming Rate	79.7%	76%	79%	82%	76%	76%	76%
Mapping Rate (>30MAPQ)	73%	61%	61%	59%	61%	61%	59%
% mtDNA reads	19.13%	8.6%	8.7%	8.6%	8.6%	8.6%	8.3%
% ENCODE blacklist reads	0.147%	0.05%	0.05%	0.05%	0.05%	0.05%	0.05%
Duplication rate	56%	27%	25%	35.6%	80%	88%	83%
Number of PCR Cycles	10	8	8	8	13	13	12
FastQC fields failed	Per base sequence content, Sequence Duplication Levels						

Table 3: Genrich peak counts for varying FDR thresholds. Underlined values indicate the peak sets that were analyzed further.

Sample	FDR < 0.01	FDR < 0.001	FDR < 0.0001	FDR < 0.00001
Corces	133,007	89,829	66,471	50,784
ATAC-STARR	162,877	124,612	101,904	85,668

ATAC-STARR-seq faithfully captures chromatin accessibility with high signal-to-noise

The use of Tn5 on native chromatin to selectively clone chromatin accessible DNA fragments provides the opportunity to quantify not only reporter activity, but also chromatin accessibility simultaneously from the same plasmid library. This is because the same DNA fragments sequenced in a typical ATAC-seq workflow are contained in the ATAC-STARR-seq plasmids. Given the insert fragments from reisolated plasmids are sequenced, we asked if the resulting peak profiles recapitulate native ATAC-seq to measure chromatin accessibility. This is important because, in contrast to a typical ATAC-seq procedure, ATAC-STARR-seq involves several additional steps including cloning, transfection and reisolation, which could distort the content of the library such that it no longer represents its native profile in the genome. Specifically, mapped sequence reads derived from inserts of reisolated plasmids are counted at a given locus and this estimate infers the accessibility of the region at the time of tagmentation. This also reflects the number of plasmids that represent a given region within the reisolated ATAC-STARR-seq plasmid library. To test this, we processed the reisolated plasmid DNA as an Omni-ATAC-seq dataset and benchmarked against the GM12878 Omni-ATAC-seq dataset from Corces *et al*. 2017. Raw sequences obtained for both datasets were processed through identical workflows (see Methods). After collapsing read duplicates, we called peaks for each dataset using a variety of false-discovery rates (FDRs) (Table 3). To closely match the number of peaks previously reported by Corces *et al*. 2017 (~108,433), we chose two separate FDR thresholds—0.0001 for ATAC-STARR-seq and 0.001 for the Corces data—yielding 101,904 and 89,829 accessible chromatin peaks respectively (Corces et al., 2017). The ATAC-STARR-seq and Corces *et al*. peak sets represent 2.22% and 2.11% of the genome, respectively, which agrees with previous reports

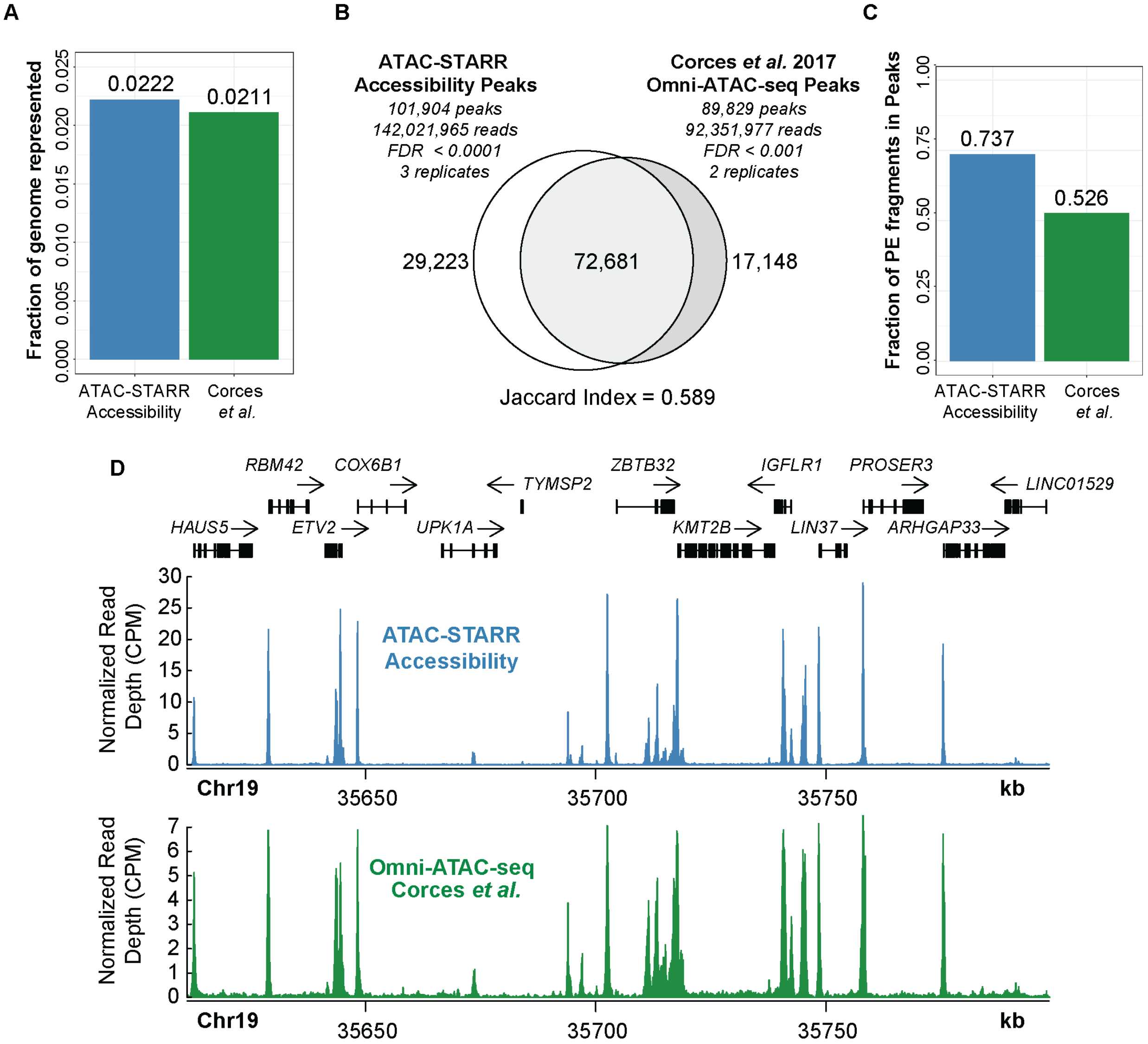

Figure 7: ATAC-STARR-seq accurately quantifies chromatin accessibility. ATAC-seq data from Corces et al. 2017 is compared with ATAC-STARR-seq plasmid DNA data. (A) Fraction of the human genome represented by each peak set. (B) Venn diagram of peak overlap between the two datasets and the associated Jaccard Index. (C) Fraction of paired-end (PE) fragments in peaks—FRiP scores—for both samples. (D) Signal tracks comparing counts per million (CPM) normalized read count at a representative locus.

(Figure 7A, (Klemm et al., 2019; Thurman et al., 2012)). Overall, 71% of ATAC-STARR-seq peaks are reproduced in the Corces *et al*. dataset, while 81% of Corces *et al*. peaks overlap the ATAC-STARR-seq dataset (Figure 7B; Jaccard index = 0.589), indicating strong agreement between these data despite substantial differences in ATAC-STARR DNA sample preparation.

Furthermore, the fraction of reads in peaks score (FRiP), an ENCODE ATAC-seq standard measure of noise, is considerably higher for both ATAC-STARR-seq (0.74) and Corces *et al.* (0.526) than the ENCODE accepted standard (>0.2, Figure 7C), indicating minimal background in our dataset. The high signal-to-noise is also evident when looking at normalized read pileups at a representative locus (Figure 7D), where the signal mirrors the Corces *et al.* accessibility signal patterns. Based on these results, we conclude that ATAC-STARR-seq can accurately retain chromatin accessible peaks in the human genome with high signal-to-noise.

A sliding windows approach increases activity region calling sensitivity

ATAC-STARR-seq tests regulatory activity in DNA enriched for accessible chromatin. Unlike whole genome STARR-seq or other MPRAs, where the genomic DNA fragment distribution is relatively constant, read coverage varies substantially from peak-to-peak in ATAC-STARR-seq. In this way, ATAC-STARR-seq requires an analysis strategy that calls active and silent regulatory regions within accessibility peaks. To address this “peaks-within-peaks” problem, we developed an analytical approach using DESeq2 to normalize reporter RNA read counts to reisolated plasmid DNA read counts. DESeq2 additionally performs an independent filtering step which removes low count data confounders that can influence ratios and result in false positive peak calls (Love, Huber, & Anders, 2014).

We tested two different approaches for regulatory activity analysis. The two approaches differ in how genomic regions are defined prior to differential analysis with DESeq2. Our “sliding window” method, defines regions by slicing accessible peaks into 50bp sliding bins with a 10bp step size (Figure 8A). Alternatively, the “fragment group” method, which is the approach used in

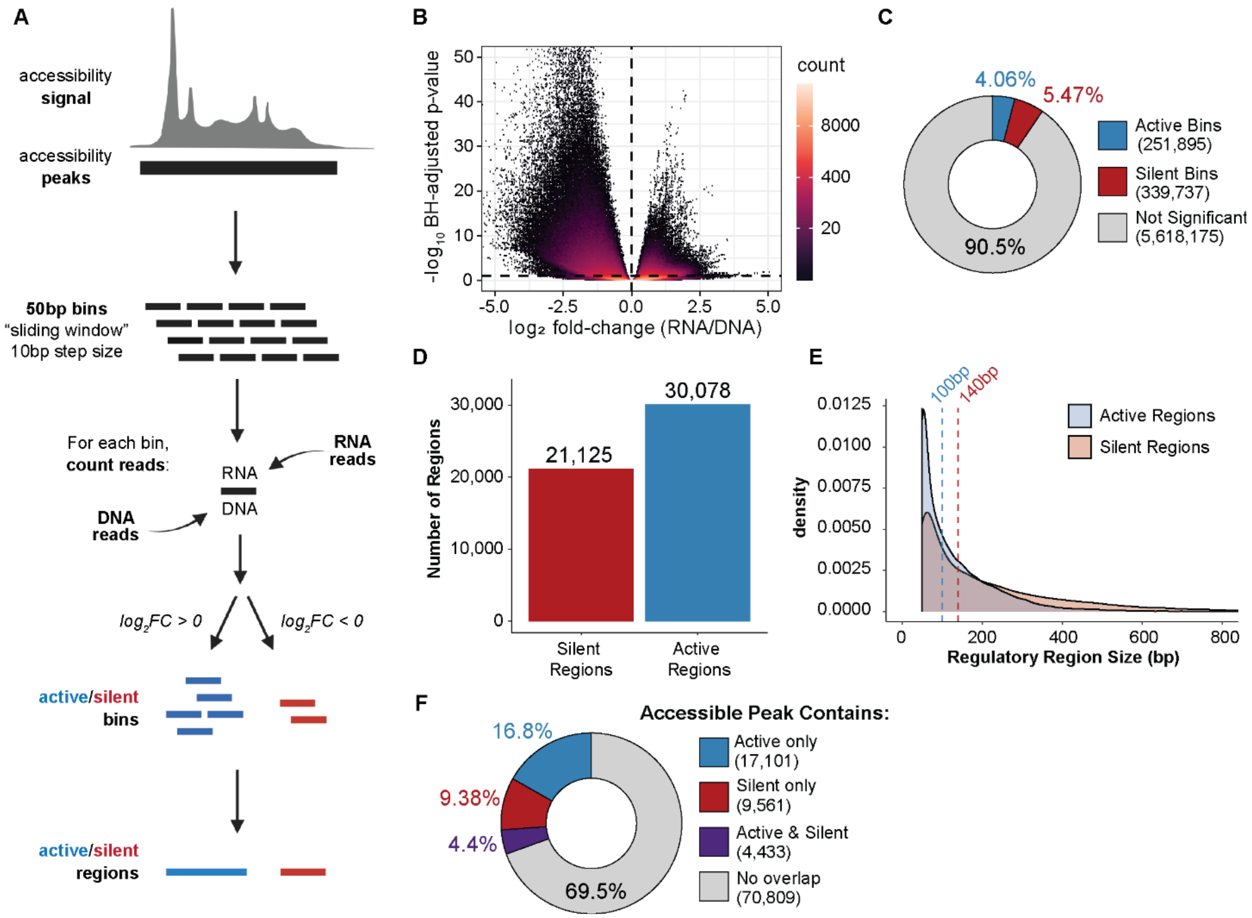

Figure 8: ATAC-STARR-seq quantifies regulatory activity within accessible chromatin. (A) Schematic of the sliding window peak calling method. Accessibility peaks are chopped into 50bp bins at a 10bp step size with the BEDTools makewindows function (options -w 50, -s 10). For each window, RNA and DNA reads are counted using Subread's featureCounts function. Differential analysis comparing RNA and DNA read count is performed with DESeq2. Significant bins are called at an Benjamini-Hochberg (BH) adjusted p-value < 0.1 and parsed into active or silent depending on $\log_2$ fold-change (FC) value (+/- zero). Finally, bins are collapsed into regions using the BEDTools merge function. Log_2FC scores are averaged across merged bins. (B) Volcano plot of $\log_2$FC scores against $-\log_{10}$-transformed BH adjusted p-value from DESeq2 for all bins analyzed. (C) The proportion of bins called as active or silent. (D) The number of regions defined as either active or silent. (E) Overlapping density plots of active and silent regulatory region size; dashed lines represent the medians in each case. (F) The proportion of accessible peaks that overlap an active or silent region, or both.

Wang *et al*. 2018, synthesizes regions by grouping paired-end sequencing fragments by 75% or greater overlap (Figure 9A). Using a different set of genomic regions, both methods assign and count overlapping RNA and DNA reads to each genomic region and, using DESeq2, identify regions where the RNA count is statistically different from the DNA count at a Benjamini-

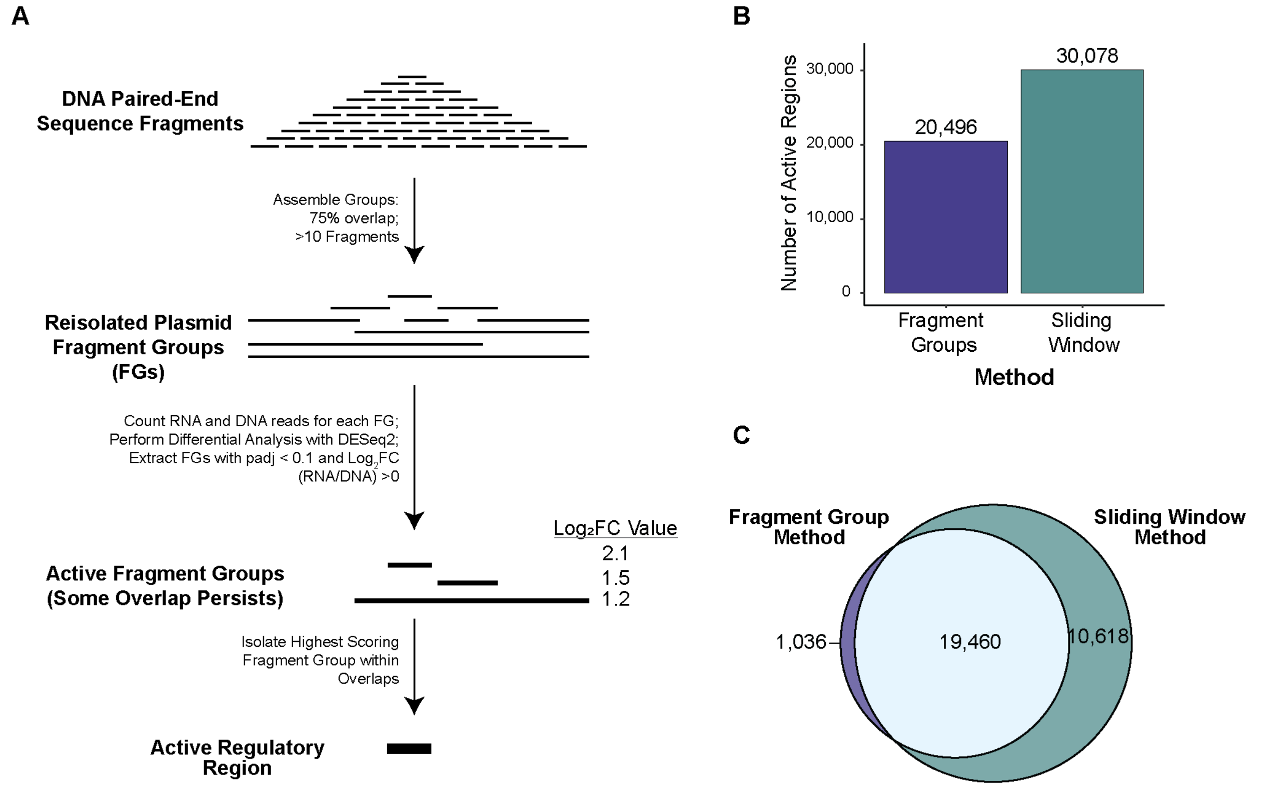

Figure 9: Comparison between the sliding window and the fragment group active region calling methods. (A) Diagram of the fragment group region calling scheme. Paired-end fragments from the DNA samples are first assembled into "fragment groups" (FGs) which represent groups of more than 10 paired-end fragments with each fragment overlapping another fragment by at least 75%. Like the sliding window method, reads from RNA and DNA samples are then assigned to each FG and active FGs are identified using differential analysis with DESeq. The same padj (<0.05) and log2fold-change (>0) filters are applied. For FGs that overlap, the FG with the largest activity score is isolated. (B) The number of active regions called with either method. (C) Euler plot comparing the region overlap between the two methods.

Hochberg (BH) adjusted p-value < 0.1. The "sliding window" method yielded ~30,000 distinct active regions, while the "fragment groups" method yielded ~20,000 distinct active regions (Figure 9B). In addition, nearly all active regions defined using the fragment group method (95%) are also captured in the sliding window method regions (Figure 9C). Given this overlap and a 50% greater recovery with the sliding windows approach, we used the sliding windows method to call active ATAC-STARR-seq regulatory regions.

Investigating the influence of replicates on region calls

We note that the Wang *et al*. 2018 study reported twice the number of active regions reported herein. This discrepancy may be explained in part by using the super core promoter in their assay, but another major difference between the two studies is replicate number (five replicates versus three replicates). To determine if the difference in active region count is driven by replicate number, we downloaded and analyzed raw sequencing data from Wang *et al.* 2018 using our pipeline and analysis methods. We then assigned reads to the bins we analyzed and called active regions using either three or five replicates (Figure 10A). With five replicates, we also captured ~66,000 active regions; however, we identified ~39,000 regions with only three replicates. This is much closer to the number we report (~30,000) and suspect the extra 9,000 regions may be the result of experimental differences, such as the promoter employed. Altogether the number of called active regions increases with more replicates.

To further investigate the effect of replicate number on region calling sensitivity in our data, we merged and split our three ATAC-STARR-seq replicates into five randomly sampled

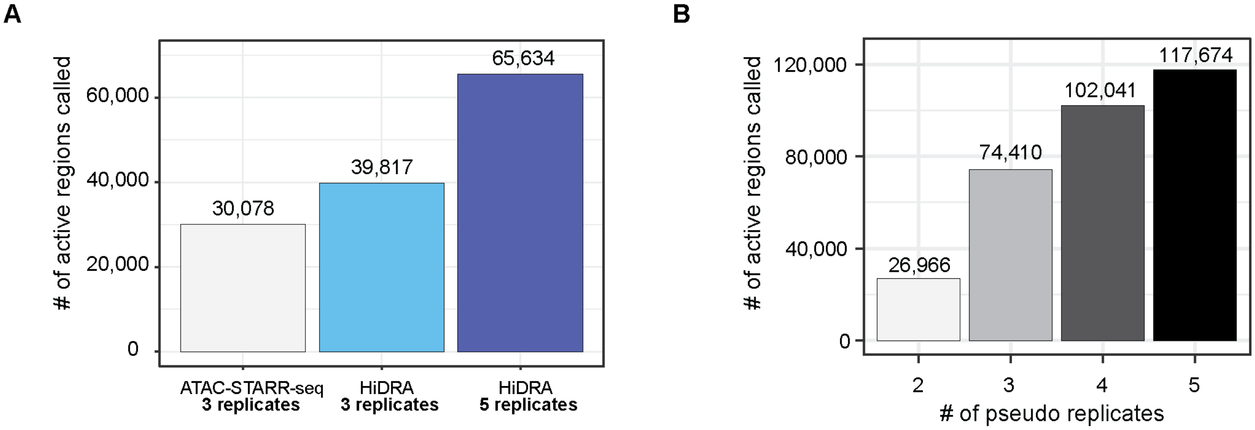

Figure 10: Analysis of replicate count on region calling sensitivity. (A) Number of active regions called using HiDRA data with either 3 or 5 replicates. Current ATAC-STARR-seq active region number is plotted for comparison. (B) Number of active regions called when 2, 3, 4, or 5 pseudoreplicates are provided. To generate pseudoreplicates, replicates were merged and then split into 5 separate files.

“pseudo-replicates”. We then called active regions using two, three, four, or five pseudo-replicates (Figure 10B). We find the largest increase in region count going from two to three replicates. Thus, the three replicate condition seems to yield the best value, while additional replicates may be needed to detect more weakly active regulatory regions. However, it is also very important to note that studies investigating the relationship between replicate number, sensitivity, and accuracy for RNA-seq data have demonstrated that performing more replicates yields more differentially expressed genes, but this is concomitant with an increase in false positive rate (Lamarre et al., 2018; Schurch et al., 2016). Therefore, the additional regions that are called with increasing replicate counts may represent a disproportionate number of false positives and may affect the outcomes of certain accuracy-sensitive applications like computational modelling.

Duplicate removal hinders region calling sensitivity

A question that often arises when determining biological signals from sequence read count data is whether to collapse read duplicates, as duplicates can arise both technically (PCR duplicates) and biologically (active regions generate multiple transcripts of themselves). To understand their contribution to data interpretation, we analysed our data with and without duplicates and compared the output. Removal of duplicates produces modest improvements to correlation coefficients between replicates, although both conditions had correlations indicative of satisfactory reproducibility (Figure 6, Figure 11A-B). However, excluding duplicates produced many fewer active regions called than including duplicates (~21,000 fewer regions) (Figure 11C). Together, this indicates that removing duplicates modestly improves reproducibility but

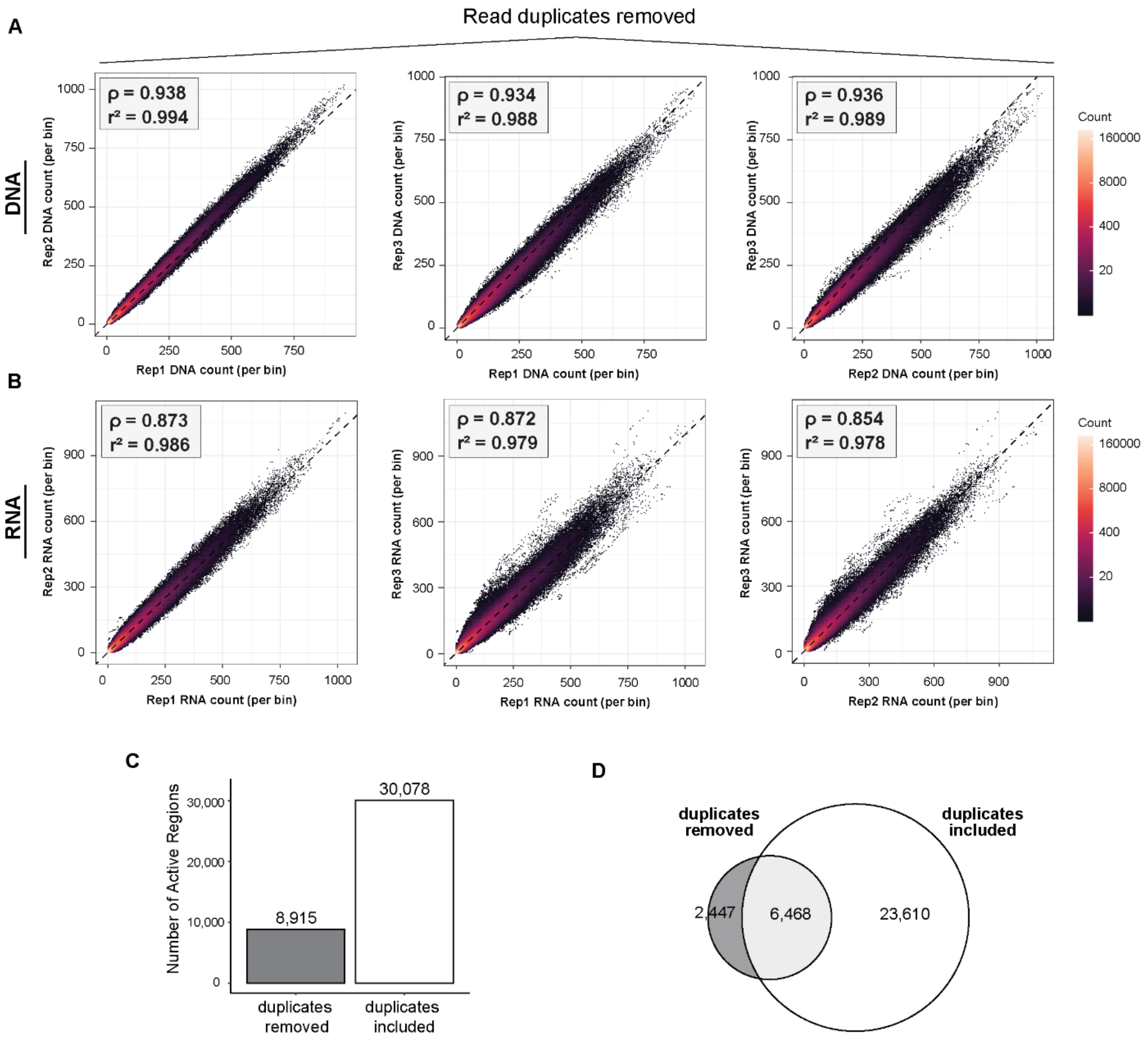

Figure 11: Comparison between keeping duplicates and removing duplicates to call active regions. (A-B) Scatter plots of DESeq2-normalized read counts per bin between replicates for both (A) DNA and (B) RNA samples when duplicates are removed. Pearson (r2) and spearman (ρ) correlation coefficients are indicated in the top left corner for each pairwise comparison. (C) The number of active regions called with or without duplicates. (D) Euler plot comparing the region overlap between the two methods.

significantly sacrifices sensitivity. Furthermore, most of the regions called without duplicates are also called when duplicates are included, indicating that, for the most part, duplicate removal affects sensitivity and not accuracy (Figure 11D). Because the with-duplicate analysis yielded many more additional regions and is reproducible between replicates, we included duplicates in our activity analysis moving forward. Importantly, because our approach filters by significance, reproducibility is required when calling active and silent regions. Therefore, identified active and silent regions are of high confidence when including duplicates.

ATAC-STARR-seq quantifies regulatory activity of open chromatin

In the sliding window approach, bins are classified as active or silent depending on whether RNA is enriched or depleted, respectively, and then like-bins are merged to collapse overlaps (Figure 8A). Using this approach, we identified ~590,000 bins where RNA and DNA counts were significantly different (Figure 8B). More specifically, this analysis identified 251,895 (4.1%) active bins and 339,737 (5.5%) silent bins from the ~5.6 million total bins measured (Figure 8C). Overlapping bins were merged into 30,078 active and 21,125 silent regulatory regions (Figure 8D). It is important to note that more silent than active bins are called; however, because silent regions are generally larger (Figure 8E), merging overlapping bins results in fewer silent regions than active. Collectively, the active and silent bins represent ~9.5% of all bins measured, indicating that the majority of accessible DNA is transcriptionally neutral. Moreover, most accessible peaks do not have an active or silent region contained within them (69.5%), suggesting that most accessible regions are neutral regulatory regions according to our assay (Figure 8F). This suggests that the majority of accessible DNA has no regulatory potential in this cellular context or, alternatively, that ATAC-STARR-seq is not sensitive enough to measure weakly active or weakly silent regions. A recent study in mouse embryonic stem cells made the same observation using an orthogonal approach, suggesting this phenomenon is present in other mammalian species (Glaser et al., 2021). We note that a small percentage of accessible peaks (4.4%) contain both active and silent regions, demonstrating that there can be competing regulatory regions within the same accessible peak.

Both short and long DNA fragments are required for comprehensive region calling

Because DNA fragment synthesis for MPRAs is limited to 200 bp including the adapters and barcode, a significant advantage of ATAC-STARR-seq and other capture-based MPRAs is the ability to measure activity of longer DNA sequences (Santiago-Algarra et al., 2017). To investigate the effect of fragment length on regulatory region calls, we divided mapped reads into short (>125bp) and long (<125bp) fragments and independently called active and silent regulatory regions; 125bp was chosen as it bisects the bimodal peak distribution displayed by RNA and DNA libraries (Figure 5B). Overall, read counts were similar for each sample after splitting into short and long groups (Figure 12A). Two to three times as many active and silent regions were called in the long fragment group compared to the short group (20,833 versus 10,789 for active and 16,872 versus 6,213 for silent). Nonetheless, a substantial number of regions are called within the short fragment group, although both fell short of the number of active and silent regions called when both long and short were used (Figure 12B). The regulatory regions called using long DNA fragments are larger than those called with short fragments, as expected (Figure 12C); however, they display little difference in TSS distance, indicating these groups are not comprised of different genomic annotations (Figure 12D). A critical observation is that only 23% of active regions called using short reads overlap active regions called using longer reads, revealing the two groups identify different regulatory regions in the genome (Figure 12E); this is also true for the silent regulatory regions, although to a lesser extent. Altogether this analysis reveals that short and long DNA fragments identify different regulatory region sets both in number and similarity. Therefore, to be

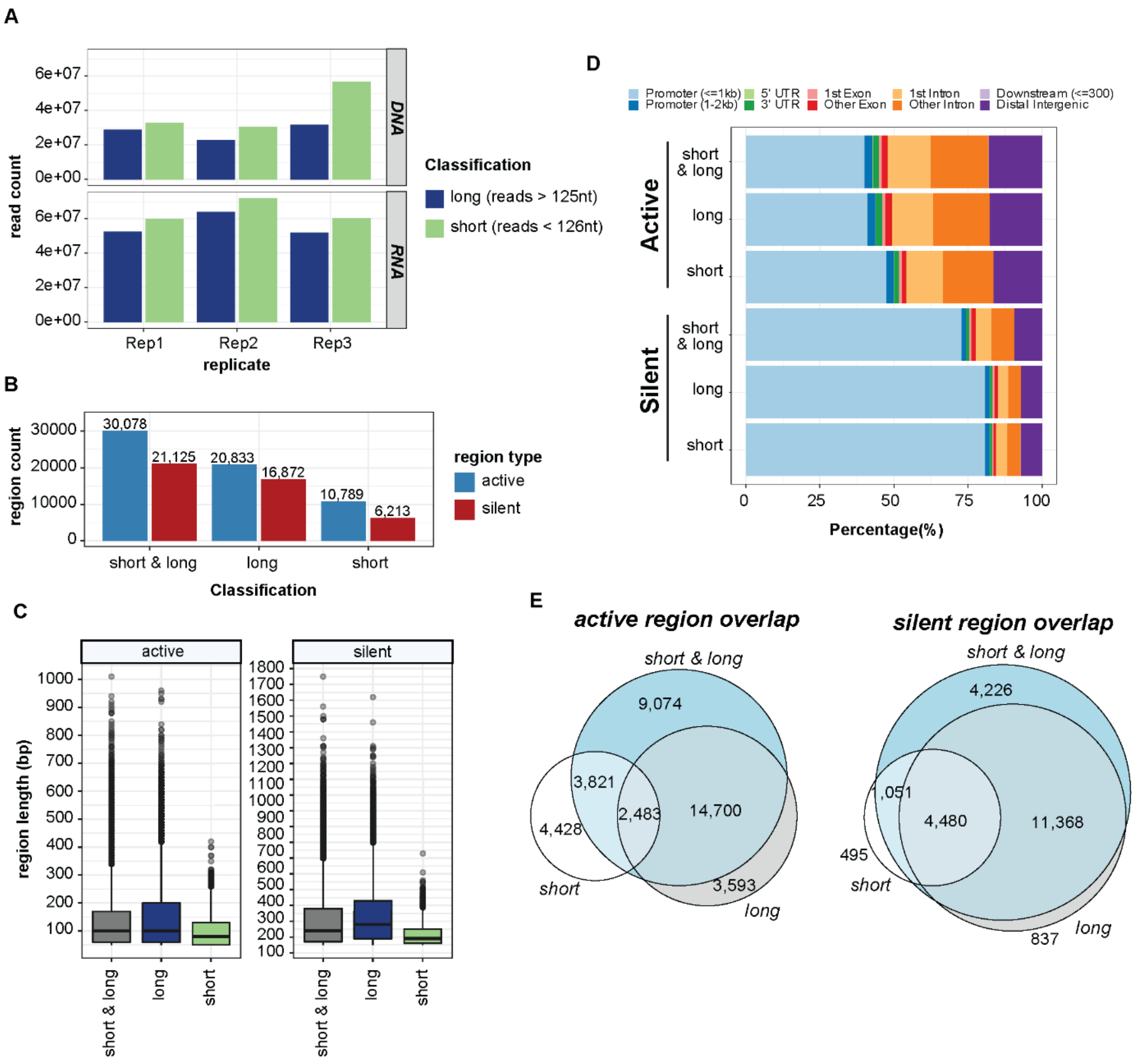

Figure 12: Effect of fragment length on regulatory region calls. ATAC-STARR-seq fragments were parsed into "long" and "short" files based on whether they were greater than or less than or equal to 125nt. (A) read counts of each fragment length classification for each replicate for both plasmid DNA and reporter RNA samples. (B) Active and silent region counts using only long fragments, only short fragments, or both. (C) Boxplots of basepair (bp) length for the active and silent region sets called for each fragment length classification. (D) Annotation of regulatory regions relative to the transcriptional start site (TSS). The promoter is defined as 2kb upstream and 1 kb downstream of the TSS. (E) Venn diagrams representing the amount of active or silent region overlap between the region sets called from each fragment length classification.

as comprehensive as possible, STARR-seq assays should be designed to include both short and long DNA fragments rather than impose a size selection to remove smaller fragments.

Active and silent ATAC-STARR-seq regions represent both proximal and distal cis-regulatory elements and lie within functional chromatin states

To gain insight into the regulatory features of active regions, we annotated both active and silent regions according to genomic location. Active regions are found in both promoter proximal and distal areas of the genome, with a majority occurring in intronic and intergenic sites (~55%), whereas silent regions coincide primarily with promoters (~75%) (Figure 13A). Functional classification of active and silent regions by the 18-state ChromHMM model (Roadmap Epigenomics Consortium et al., 2015) revealed that active regions consist of TSS active, TSS flanking upstream, and Enhancer Active 1 chromatin states and are devoid of repressive states like Repressed Polycomb Weak and Quiescent (Figure 13B). By contrast, silent regions are slightly enriched for bivalent chromatin states (TSSBiv, EnhBiv), consistent with the observation that they are accessible but not active. Most silent regions also coincide with TSS Active and TSSFlnk ChromHMM states, which corroborates their promoter proximal locations; however, their designation as "active" by ChromHMM is somewhat puzzling considering these DNA fragments do not drive transcription in our assay. One explanation is that silent regulatory activity, as measured by episomal-based reporter assays, does not fully copy regulatory activity as predicted by ChromHMM. Alternatively, active promoters may confound the reporter assay by initiating transcription from the 3'UTR of the plasmid causing conflicts with active transcription from the Ori.

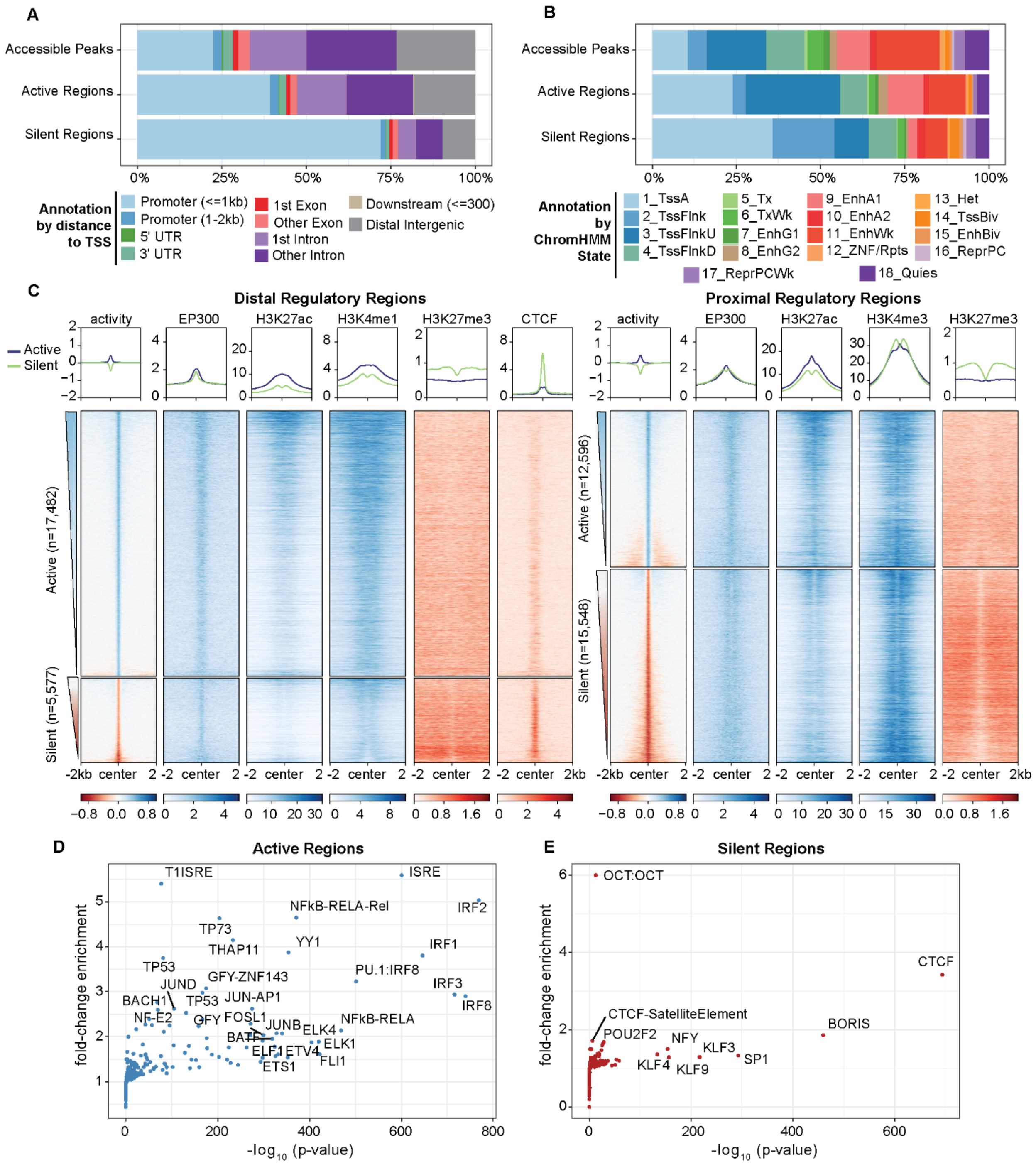

Figure 13: Regulatory regions defined by ATAC-STARR exhibit annotations, histone modifications, and TFs characteristic of their function. (A) Annotation of regulatory regions relative to the transcriptional start site (TSS). The promoter is defined as 2kb upstream and 1kb down-stream of the TSS. (B) Annotation of regulatory regions by the ChromHMM 18-state model for GM12878 cells. (C) Heatmaps of GM12878 ENCODE ChIP-seq signal and regulatory activity for proximal and distal ATAC-STARR-defined regulatory regions. Proximal regions were classified as within 2kb upstream and 1kb downstream of a TSS; all other regions were annotated as distal. Active and silent regions were ranked by mean activity signal for both proximal and distal regions. (D-E) Transcription factor motif enrichment analysis as quantified by HOMER. Fold-change values are relative to the default background calculated by HOMER.

To further investigate if silent regions are a result of 3'UTR transcription initiation, we considered if an orientation bias existed in reporter RNAs levels. If 3'UTR transcription conflicts exist, we would expect many fewer reporter RNAs when transcription results in head-on conflicts rather than occurring in the same direction as the Ori. We therefore subset reads based on whether they arose from an insert cloned in a 3' to 5' direction or in a 5' to 3' direction (Figure 14A). We then assigned read counts to all bins analyzed (Figure 14B-C), the bins called active (Figure 14D-E), or the bins called silent (Figure 148F-G). Because this is expected to be a promoter-specific effect, we also split bins into proximal and distal based on location to the nearest transcription start site. In all cases, more than 95% of the bins do not display an orientation bias, which we defined as a normalized read count difference greater than five between orientations (Figure 14H). Moreover, we observe high Pearson and Spearman's correlation coefficients between orientations for all conditions (r^2: 0.80-0.91 and ρ: 0.73-0.90) and the minimal contribution of orientation bias to silent regions is in agreement with a previous report (Klein et al., 2020). For the <5% of regions that do display orientation bias, proximal bins are more affected than distal bins, as expected. Altogether, ATAC-STARR-seq does not display a significant orientation bias and most of the 21,000 silent regions we observe result from legitimate silencing activity or another source.

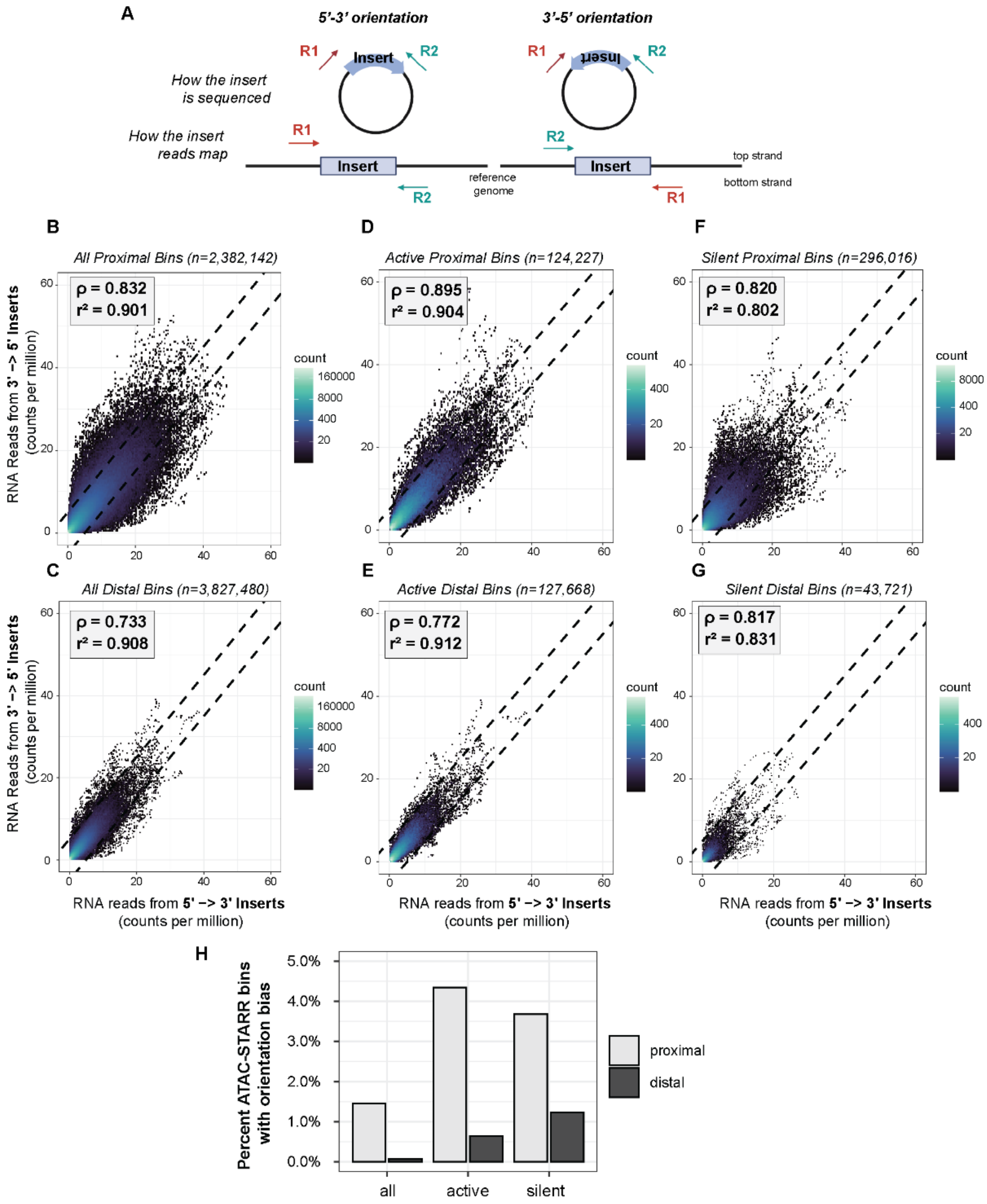

Figure 14: Assessment of potential orientation bias in ATAC-STARR-seq data. (A) Schematic of the method for separating reads based on insert orientation. Read 1 and Read 2 are sequenced from the same position regardless of insert orientation on the plasmid and reporter RNA samples. Therefore, insert orientation can be specified based on how the read pair map to the genome. 5'-3' inserts have R1 on the top strand, while 3' -5' inserts have R1 on the bottom strand. (B-G) Scatter plots of counts per million normalized reporter RNA read counts between 5' to 3' inserts and 3' to 5' inserts for (B) all proximal bins analyzed, (C) all distal bins analyzed, (D) active proximal bins only, (E) active distal bins only, (F) silent proximal bins only or (G) silent distal bins only. Pearson (r2) and spearman (ρ) correlation coefficients are indicated in the top left corner for each pairwise comparison. Proximal bins were defined as within 2kb upstream and 1kb downstream of a transcription start site, while distal bins were defined as everything else. Dashed lines indicate +/- 5 counts from the expectation (y=x). The percentage of bins that lie outside of these lines are denoted in (H).

Active and silent ATAC-STARR-seq regions are distinct functional classes and are enriched for specific histone modifications and TF motifs

To further investigate the chromatin landscape of the active and silent regions, we plotted ENCODE GM12878 ChIP-seq signal (The ENCODE Project Consortium et al., 2020) for EP300, CTCF, and histone modifications associated with active and repressed chromatin states (Figure 13C). As expected, active regions contain EP300 at their center with histone 3 lysine 27 acetylation (H3K27ac) more broadly distributed across the center; histone 3 lysine 4 mono-methylation (H3K4me1) is also present at distal regions, while histone 3 lysine 4 tri-methylation (H3K4me3) is at proximal regions. In addition, histone 3 lysine 27 tri-methylation (H3K27me3)—a bivalent repressive mark—is largely absent from active regions. Proximal silent regions, on the other hand, are enriched for H3K27me3 and H3K4me3. This suggests many of the proximal silent regions are accessible bivalent regulatory elements in lymphoblastoid cells. To support their designation as silent calls, we compared histone modification signal at accessible peaks that contain either a silent region, an active region, both a silent and active region, or neither, which we define as neutral accessible peaks (Figure 15A). Consistent with the observations above, silent accessible peaks contain more H3K27me3 signal and are devoid of H3K27ac signal relative to the other accessible peak types.

It is important to note that silent regions are distinct from neutral regions, which are defined as regions failing to reach significance in the RNA-DNA differential analysis. Overall, neutral regions exhibit baseline levels of histone modifications and distribution in genomic annotations like that of all accessible peaks (Figure 15B-C, Figure 13A-B). While neutral regions represent the

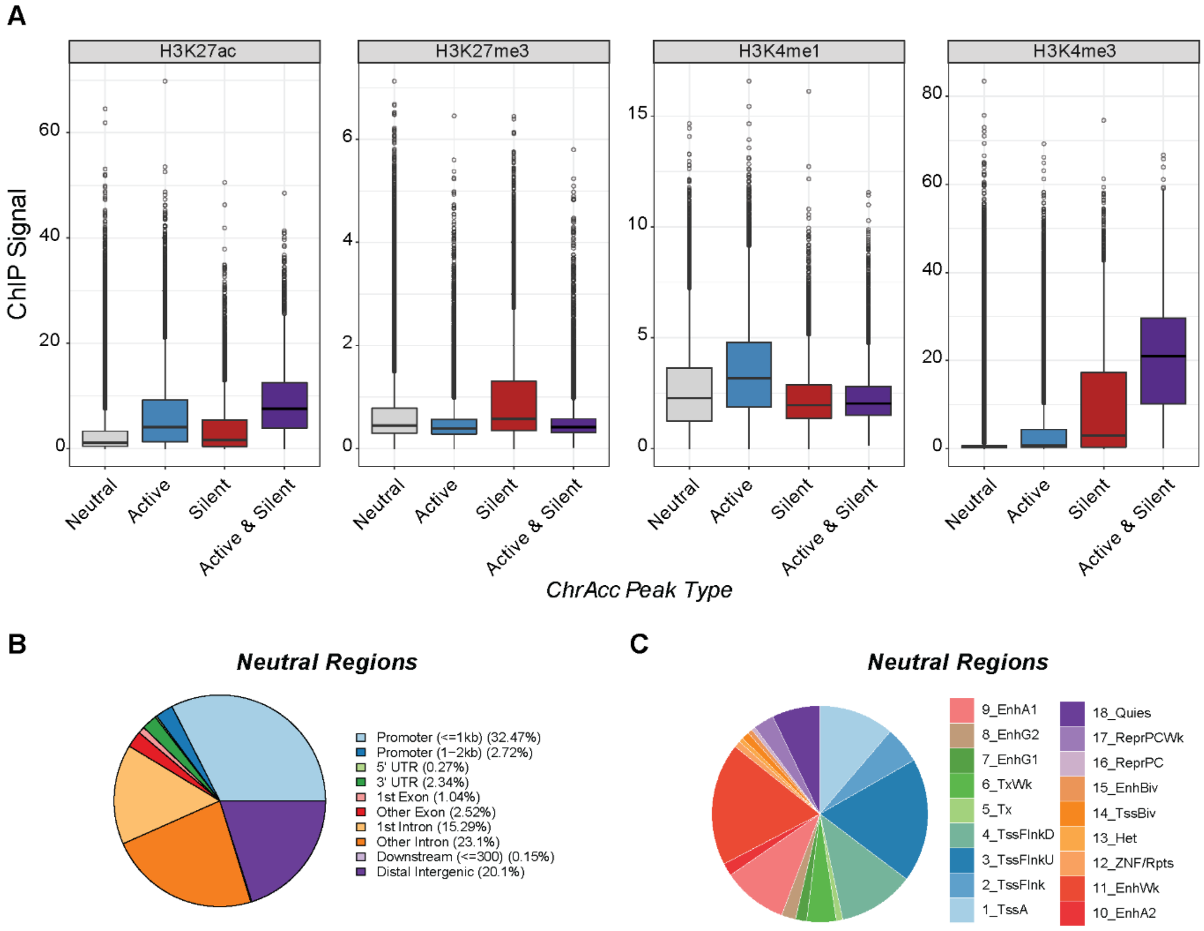

Figure 15: Additional characterization of ATAC-STARR-seq regulatory regions. (A) Histone modification ChIP-seq signal at accessible chromatin peaks. Boxplot of the distribution of histone modification ChIP-seq signal for accessible chromatin peaks (ChrAcc) that contain an active region, a silent region, both an active and silent region, or neither (neutral). Values represents the average fold change over control signal per region for each histone modification. (B) Annotation of regulatory regions relative to the transcriptional start site (TSS). The promoter is defined as 2kb upstream and 1 kb downstream of the TSS. (C) Annotation of regulatory regions by the ChromHMM 18-state model for GM12878 cells.

majority of accessible peaks, it is possible that a subset are weak enhancers as indicated by overlap with ChromHMM states, or regulatory elements that display activity in a different cellular context.

Our analysis of TF motifs within active and silent regions revealed prominent differences in motif enrichment. Distal silent regions are strongly enriched for CTCF and its counterpart BORIS, which is associated with diverse functions including gene repression and insulator activity

(Figure 13D-E)(S. Kim, Yu, & Kaang, 2015). In addition, we found enrichment for the SP/KLF family several of which (Cao, Sun, Icli, Wara, & Feinberg, 2010) are known to be transcriptional repressors. By contrast, the most enriched TFs in active regions were the IRF family, the ETS family, subunits of the NF-kB complex, and general promoter TFs such as THAP11 and YY1. These data are consistent with our current understanding of immune gene regulation and regulatory element function, which together corroborates the quantification of regulatory activity with ATAC-STARR-seq.

ATAC-STARR-seq retains the ability to map in vivo TF binding

An inherent advantage of an ATAC-seq based approach is the ability to perform TF footprinting. Computational footprinting methods identify Tn5 cleavage events or "cut sites" from ATAC-seq data and, when combined with motif analysis, can identify TF binding sites with high accuracy (Bentsen et al., 2020; Yan, Powell, Curtis, & Wong, 2020). Since ATAC-STARR-seq produces similar high-quality chromatin accessibility peak profiles as standard ATAC-seq, we explored whether TF footprints were also preserved. We generated Tn5-bias corrected cut site signal files for both Corces *et al*. 2017 and ATAC-STARR-seq accessibility datasets and plotted cut site signal at all accessible CTCF motifs (Figure 16A) (Bentsen et al., 2020). As a control, we also plotted GM12878 CTCF ChIP-seq signal from ENCODE and ranked region order by highest mean ChIP-seq signal. We observed consistent depletion of Tn5 cut-sites for both Corces *et al*. 2017 and ATAC-STARR-seq accessibility at CTCF sites. Moreover, we only observe footprints at motifs with CTCF ChIP-signal, demonstrating the utility of TF footprinting to determine motifs

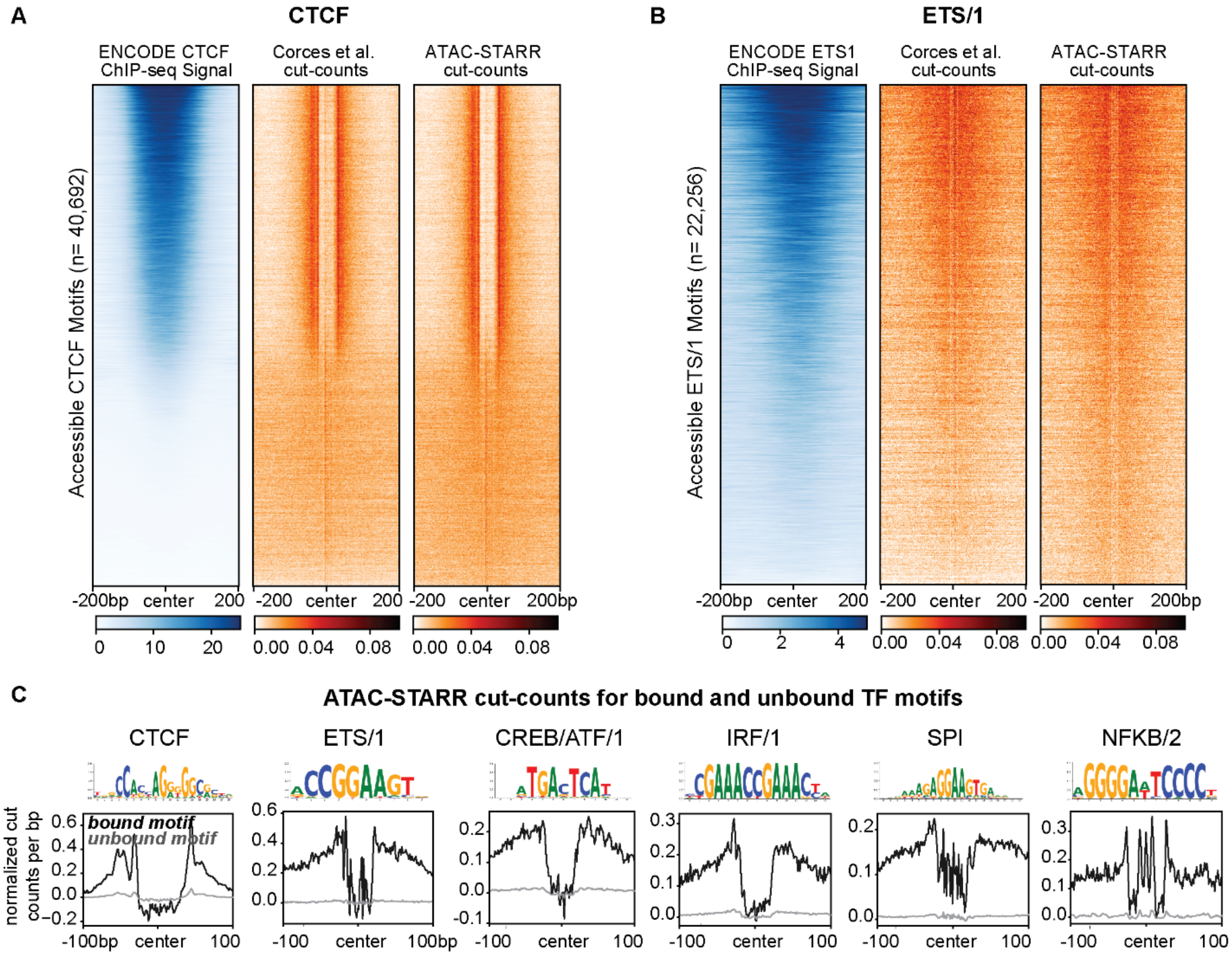

Figure 16: ATAC-STARR-seq identifies transcription factor footprints. (A) Comparison of ENCODE CTCF ChIP-seq signal to Corces et al. and ATAC-STARR-seq cut count signal for all accessible CTCF motifs. (B) Comparison of ENCODE ETS1 ChIP-seq signal to Corces et al. and ATAC-STARR-seq cut count signal for all accessible motifs with the ETS/1 motif archetype. For both, regions were ranked by largest mean ChIP-seq signal. (C) Aggregate plots representing mean signal for the TOBIAS-defined bound and unbound motif archetypes: CTCF, ETS/1, CREB/ATF/1, IRF/1, SPI, NFKB/2.

that are bound or unbound by TFs. Given the importance of TFs in driving enhancer function, this distinction is vital when dissecting transcriptional regulation in human cells.

TF motif enrichment analysis pointed to multiple ETS family members, including ETS1 which is an important immune cell regulator (Garrett-Sinha, 2013) (Figure 13D). So, we asked whether ETS1 footprints are also present in our data. Unlike CTCF, ETS1 shares its motif with many other transcription factors, such as ETV4; therefore, footprinting cannot distinguish ETS1 and ETV4 binding sites. For this reason, we refer to TFs using their ENCODE-defined "archetypes", which reflects the group of TFs that share the same motif (Vierstra et al., 2020). For each archetype, we performed footprinting against one of the TFs within an archetype to infer motifs bound by members of the group, such as ETS1 for the ETS/1 archetype. To assess the extent to which ETS1 footprints can be explained by ETS1 binding, we plotted GM12878 ETS1 ChIP-seq signal from ENCODE within both Corces *et al*. 2017 and ATAC-STARR-seq cut sites (Figure 16B). Indeed, ETS1 ChIP-seq signal explains the majority but not all the ETS/1 footprints present. We observe similar cut-site signal to Corces *et al*. 2017, further indicating that ATAC-STARR-seq can detect *in vivo* binding of transcription factors despite the additional cloning and transfection steps involved in producing ATAC-STARR-seq DNA libraries.

We performed footprinting for several more immune related TF archetypes to identify bound or unbound TF motifs (Figure 16C). For all TFs, bound motifs display substantially larger footprint depth than unbound motifs. Together, this indicates that ATAC-STARR-seq, when combined with footprinting, can identify regions of the genome where TFs are bound. This additional level of information can be leveraged in conjunction with accessibility and activity to understand the context of TF binding while circumventing the need to perform individual chromatin immunoprecipitations.

Collective profiling of accessibility, in vivo TF binding, and activity with ATAC-STARR-seq reveals distinct networks of gene regulation

Interrogating chromatin accessibility, TF binding, and regulatory activity together can be used to interpret locus-specific gene regulatory mechanisms. For example, active regulatory elements surrounding the B cell-specific expressed gene *ZBTB32* overlap IRF8 and NFKB1 footprints suggesting these regions are regulated by IRF8 and NFKB1 binding (Figure 17A). We also observe SP1 and KLF3 footprints overlapping a silent region at the *ETV2* locus, a gene lowly expressed in B cells, according to the Human Protein Atlas (Uhlen et al., 2015; Uhlen et al., 2019). Together this indicates that active and silent regions can, in part, be explained by the occupancy of these TFs.

To demonstrate the power of integrating TF footprints and regulatory regions on a global scale, we clustered active and silent regions based on the presence or absence of several TF footprints (Figure 17B-C). Footprints were selected based on top hits from the previous motif enrichment analysis (Figure 13D-E). Regulatory activity may be driven by one or multiple TF binding events that defines the cluster and is representative of a gene regulatory network in the genome. Indeed, we find that the putative target genes regulated by each unique group are enriched for distinct gene regulatory pathways and are often related to the TFs in the cluster (Figure 17D-E). For example, cluster C is primarily defined by the presence of IRF/1 and is enriched for interferon alpha/beta signalling. It is interesting that active clusters tend to be more associated with B cell function than silent clusters, which are more associated with general, non-B cell related pathways.

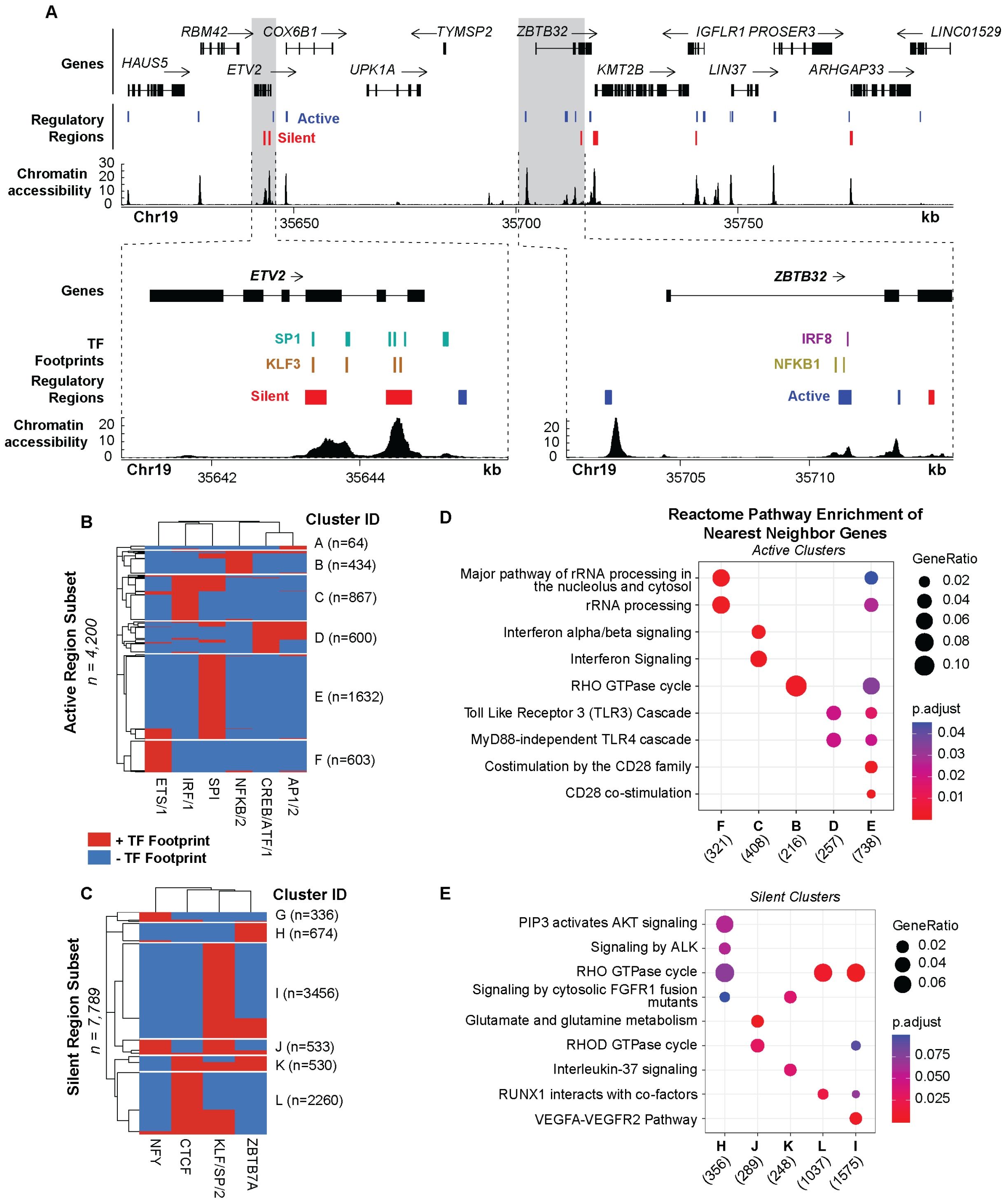

Figure 17: TF footprints stratify ATAC-STARR-defined regulatory regions into gene regulatory networks. (A) ATAC-STARR-defined chromatin accessibility, TF footprints, and regulatory regions at Chr19:35,611,232-35,798,446 (hg38). Signal tracks represent counts per million normalized read depth of chromatin accessibility. Zooms into *ETV2* and *ZBTB32* show that some regulatory regions are occupied by a SP1, KLF3, IRF8, or NFKB1 footprint. (B-C) Heatmaps of clustered (B) active and (C) silent regions based on presence or absence of footprints for select TF motif archetypes. (D-E) Reactome pathway enrichment analysis for nearest-neighbor gene sets for each of the clusters. Genes counts for each cluster are displayed below their group identifier.

Altogether, these distinct gene regulatory networks provide an additional layer of insight into the mechanisms that control gene expression and showcase how integration of the multiple layers of gene regulatory information provided by ATAC-STARR-seq can narrow the focus of gene targets for active and silent regions. We envision such an analysis could be used to interpret the functional consequences of a dysregulated transcription factor or disease-associated genetic variants. We provide this level of detail from a single dataset, which further demonstrates the strong potential of our workflow to reveal distinct functional layers of human gene regulation. The resolution we achieve here would not be possible without all three levels of regulatory information provided by ATAC-STARR-seq.

Discussion

Genome-wide approaches that integrate measurements of multiple layers of gene regulation are needed to better understand enhancer function. By combining ATAC-seq with STARR-seq, ATAC-STARR-seq assays regulatory activity only within the context of accessible chromatin. This allows deeper coverage of regulatory elements by narrowing scope but remaining inclusive of nearly all active regulatory elements. In this report, we substantially expand the capabilities of ATAC-STARR-seq and present an improved workflow which uniquely permits simultaneous profiling of accessibility, TF occupancy, and regulatory activity from a single DNA fragment source. Specifically, we implement key experimental and analytical improvements and present data rationalizing the decisions we make. Experimentally, we adapt a modified tagmentation protocol (Omni-ATAC) to remove mitochondrial DNA from the DNA fragment pool. We also utilize the Ori as the minimal promoter on the STARR-seq backbone which improves reporter RNA expression, recovery, and dynamic range over the super core promoter

(SCP1) backbone (Klein et al., 2020; Muerdter et al., 2018). Furthermore, we reisolate the transfected plasmid DNA to capture only the DNA that is available to cells, which is a more accurate measure of the input than sequencing prior to transfection. Reisolating plasmid DNA drives a greater degree of variance between samples and better reflects a true experimental replicate than sequencing the same DNA sample for each RNA replicate. Finally, we show that replicate number and inclusion of long and short fragment sizes are critical for comprehensive region calling.

Critically, we developed and tested a simple and sensitive region calling strategy that improves detection of regulatory regions including silencers. We also quantify chromatin accessibility and identify TF footprints, which is surprising given the added processing steps in ATAC-STARR-seq including cloning, transfection, and recapture of DNA libraries that can dull or degrade footprint signal. This enabled us to stratify the active and silent regulatory regions into distinct gene regulatory networks defined by the presence of one or multiple TF footprints. Such an analysis typically requires multiple genomic sequencing assays, but we do this using a single dataset.

With this improved workflow, we identified 30,078 active regions and 21,125 silent regions in lymphoblastoid cells. Most active regions were distal to transcription start sites, enriched for functional active ChromHMM states, and were enriched for known B cell regulating-TF motifs such as IRF8 and NFkB. By contrast, the silencers are proximal to transcription start sites and enriched for CTCF and the SP/KLF TF family. Silent regions are also enriched for the bivalent marks H3K27me3 and H3K4me3 and may represent regulatory regions that are poised, particularly at promoters. Because our plasmid design places regulatory regions within the 3'UTR of the truncated reporter gene, it is possible that the lack of observed reporter RNAs at silent regions are a result of head-on transcriptional conflicts that arise from antisense transcription initiation from the 3'UTR. However, we show this minimally occurs in our system and the silent regions reflect true silencing activity or another source that has yet to be identified. While further studies may be needed to validate these silent regions, this work confirms that the silencers are a distinct class of regulatory element with distinct properties compared to active and neutral regions and warrant further investigation. Even with an increasing number of studies targeted at identifying silencers in the human genome, silencing regulatory regions remain an under-studied aspect of gene regulation and our approach provides a new strategy to investigate these elements on a global scale (Doni Jayavelu et al., 2020; Y. S. Kim et al., 2021; Pang & Snyder, 2020).

ATAC-STARR-seq data has several distinct attributes that require a tailored analysis strategy. Current MPRA bioinformatic tools and pipelines are not tractable for these data, because in ATAC-STARR-seq the input itself is enriched for accessible chromatin and the read pileup varies considerably within these loci. In this way, the analysis of our data required calling essentially "peaks within peaks". For this reason, it was critical to 1) normalize RNA to DNA and 2) avoid regions of low count data, which is why we adapted approaches using DESeq2. We also showed that including PCR duplicates was preferred over collapsing duplicates. In the future it would be beneficial to introduce a unique molecular identifier to the system—such as the strategy employed by UMI-STARR-seq (Neumayr, Pagani, Stark, & Arnold, 2019)—to collapse only the duplicates arising from PCR. While we show comparisons of analysis strategies here, we believe that more information could be extracted from this and future ATAC-STARR-seq datasets with improved analysis strategies. In recent years we have seen the development of tailormade peak callers for whole genome STARR-seq, such as CRADLE (Y. S. Kim et al., 2021) and

STARRPeaker (D. Lee et al., 2020); a similarly tailored ATAC-STARR-seq peak caller could further improve the capabilities of the method.

While this study was limited to one condition, there are many potential applications of ATAC-STARR-seq. With the ability to subset enhancers by TF occupancy, ATAC-STARR-seq could be leveraged to investigate enhancer grammar by pairing measurable regulatory activity with multiple TF footprints that drive enhancer function. This approach also has the potential to identify dysfunctional gene regulatory networks in diseases like cancer where neoplastic transformation can be driven by the dysfunction of a specific TF. Additionally, an ATAC-STARR-seq plasmid library may be generated from one cell-type and tested in another. This flexibility could be used as a tool to determine context dependent activity or investigate enhancer and TF usage patterns during a differentiation time course.

In this study, we demonstrated that our improved ATAC-STARR-seq workflow is a powerful approach enabling joint quantification of chromatin accessibility, transcription factor occupancy, and regulatory activity. We further demonstrate how this single assay can characterize the human genome at many functional levels from chromatin accessibility to distinct gene regulatory networks. This method provides a state-of-the-art approach to deeply investigate transcriptional regulation of the human genome. We provide a detailed protocol, a well-documented code repository, and guidelines for quality control (below) so that ATAC-STARR-seq may be easily used and adapted by the field.

Guidelines for ATAC-STARR-seq quality control

Generate highly complex ATAC-STARR-seq plasmid libraries. Library complexity is the most important consideration when generating an ATAC-STARR-seq plasmid library. Library complexity is defined by the number of unique DNA fragments analyzed in the library, i.e., the number of unique plasmid inserts, and the more complex a plasmid library, the more DNA sequences that are tested. Greater library complexity translates to greater coverage of the genome. While we have not experimented directly with different library complexities, less complex libraries would likely result in a reduction in sensitivity and fewer regions being called active and silent. To estimate library complexity, we suggest performing low-depth sequencing of the plasmid library prior to conducting the reporter assay portion of ATAC-STARR (see methods). In this report we find our library complexity is roughly 50 million unique sequences. We made critical choices in procedure and reagents used to ensure this high library complexity; therefore, we strongly discourage replacement of key procedures with faster, cheaper, or simpler alternatives. For the human genome, we recommend library complexities of at least 20 million.

Perform minimal PCR cycles to keep PCR duplication rates low. As mentioned previously, duplicates should not be collapsed when calling active and silent regions, because they can arise both technically (PCR duplicates) and biologically (active regions generate multiple transcripts of themselves). Due to this issue, it is important to minimize PCR duplicates when preparing sequencing libraries. To achieve this, we try to obtain just enough sequence-able material using the fewest number of PCR cycles. We recommend a duplication rate < 90% for Reporter RNA samples and < 50% for plasmid DNA samples.

Reads should pass general quality filters. The sequenced Reporter RNA and plasmid DNA libraries should be analyzed for quality using FastQC. Both should pass all FastQC quality filters except *per base sequence content* (Tn5 has a bias) and *sequence duplication levels* (inherent quality of ATAC-STARR-seq). Mapping rate should be high (>80%) for most cell lines. For GM12878 cells, at least in our hands, ~20% of reads map to the Epstein-Barr Virus genome which causes our mapping rates to be low (~60-70%). This phenomenon is unique to viral-transformed cell lines like GM12878.

Replicates should be reproducible. We recommend calculating Spearman's correlation values between ATAC-STARR-seq replicates (see methods). In STARR-seq-based methods, Spearman's correlation values > 0.7 are typically sufficient for downstream analysis (Arnold et al., 2013; Barakat et al., 2018; Chaudhri et al., 2020; Glaser et al., 2021; X. Wang et al., 2018). Importantly, our analytical pipeline does not identify non-replicating regions as active or silent. Therefore, data for regions that are not reproducible should not manifest as false positives in our system. Less reproducibility, however, will lead to drop out and a greater false negative rate.

Assessment of Batch Effects. While correlation scores are one measure of assessing batch effects between replicates, principal component analyses (PCA) can also provide critical insights into batch effects, particularly when several conditions are compared to each other. If batch effects are minimal, samples should cluster together only by condition and not by the batch in which they were processed. In our system, batch effects could contribute to false negatives, rather than false positives, as reproducibility is required for active and silent region calling to reach the necessary statistical significance. If needed, we recommend correcting for batch effects by including replicate number in the DESeq2 formula, i.e., ~ replicate + condition, as described in the DESeq2 vignette: (http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html).

Plasmid DNA data should meet general ATAC-seq standards. Because plasmid DNA samples reflect ATAC-seq libraries, they should generally meet ATAC-seq quality thresholds, such as a FRiP score > 0.2. Importantly, a stringent q-value should be applied to yield between 50,000-110,000 ChrAcc peaks that represent about 2% of the human genome. The fragment size distribution should be bimodal with two peaks representing nucleosome free DNA fragments

(>100bp) and mono-nucleosomal DNA fragments (~200bp). This should be determined prior to sequencing via tapesation (Supplemental Figure S2A) and during the analysis phase (Supplemental Figure S2B). We do not see the di-, tri-, quad-, etc. nucleosomal bands due to removal of large fragments via SPRI bead size selection in the plasmid library generation process.

Materials & Methods

Cell Culture

GM12878 cells were obtained from Coriell and cultured with RPMI 1640 Media containing 15% fetal bovine serum, 2mM GlutaMAX, 100 units/mL penicillin and 100 µg/mL streptomycin. Cells were cultured at 37°C, 80% relative humidity, and 5% CO_2. Cell density was maintained between 0.2×10^6 and 1×10^6 cells/mL with a 50% media change every 2-4 days. All cell lines were regularly screened for mycoplasma contamination using the MycoAlert kit (Lonza).

Plasmids

The hSTARR-seq_ORI plasmid vector was a gift from Alexander Stark (Addgene plasmid #99296) and the pcDNA3-EGFP was a gift from Doug Golenbock (Addgene plasmid #13031). The bacterial stabs from Addgene were spread onto an LB agar plate containing 100µg/mL ampicillin and incubated at 37°C overnight. For each, a single colony was picked and grown in 50mL LB containing 100µg/mL ampicillin overnight at 37°C while shaking at 225rpm. Plasmid DNA was extracted using the ZymoPURE II Plasmid Midiprep kit (Zymo Research, #D4200).

The linear vector used in the ATAC-STARR-seq gibson cloning step was generated by a single 50µL PCR reaction using NEBNext® Ultra™ II Q5® Master Mix (NEB, #M0544S). While not necessary for this study, primers were designed to add the i5 barcode to the linearized vector;

this allows for different ATAC-STARR-seq plasmid libraries to be pooled and tracked. Following this approach, a universal forward primer (Fwd_universal_STARR) and a reverse primer (Rev_N504_STARR) designed to add the N504 barcode were used (primer sequences are provided in Table 4). PCR Products were purified with the Zymo Research DNA Clean & Concentrator-5 kit. DNA yield was determined by Nanodrop, and purity was analysed by gel electrophoresis; the linearized vector was the only product observed on the gel.

Tagmentation

A total of eight tagmentation reactions were performed on 50,000 GM12878 cells each. We followed a slightly modified version of the Omni-ATAC approach used in Corces *et al.* 2017 (Corces et al., 2017). Specifically, twice as much Tn5 than described in the protocol was used. Standard Tn5 transposase was prepared in-house following the method described in Picelli *et al.* 2014 (Picelli et al., 2014). Standard Tn5 transposome was assembled as described in Barnett *et al.* 2020 (Barnett et al., 2020) with the following oligos: Tn5_1, Tn5_2_ME_comp, and TN5MEREV. Tagmented products were pooled together and purified with the Zymo Research DNA Clean & Concentrator-5 kit (#D4013). The entire elution was split and amplified via five-10μL PCR reactions. We used NEBNext® High-Fidelity 2× PCR Master Mix (#M0541S)—which is not a hot-start formulation—to first extend tagments before the initial denaturation step of PCR via the following cycling parameters: 72°C 5 min, 98°C 30s; 4 cycles of 98°C 10s, 62°C 30s, 72°C 60s; final extension 72°C 2 min; hold at 10°C. Forward and reverse primer sequences, Fwd_atac-starr_tag and Rev_atac-starr_tag, are provided in Table 4. Amplified products were purified with the Zymo Research DNA Clean & Concentrator-5 kit and then analyzed for concentration and size distribution with a HSD5000 screentape (Agilent, #5067) on an Agilent 4150 TapeStation system.

After amplification, we selected PCR products less than 500bp using SPRISelect beads (Beckman-Coulter, #B23317) at a 0.6× volume ratio of beads:sample. Selection was verified using a HSD5000 screentape.

Table 4: Oligo sequences used in ATAC-STARR-seq and qPCR. Red denotes i7 or i5 barcode.

Name	Sequence (5' to 3')	Description
Fwd_universal_STARR	CATCTCCGAGCCCACGAGACTCGACGAATTCGGCCGG	Primer to linearize vector with PCR and add homology arms for gibson cloning.
Rev_N504_STARR	ACATCTGACGCTGCCGACGATCTACTCTTGCATGCTCTAGATCAATCTAATTC	Primer to linearize vector with PCR, add a nextera i5 barcode to the linear vector, and add homology arms for gibson cloning.
Tn5_1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG	Used to assemble Tn5 transposome. Nextera adapter A. Hybridize with TN5MEREV
Tn5_2_ME_comp	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG	Used to assemble Tn5 transposome. Nextera adapter B. Hybridize with TN5MEREV
TN5MEREV	/5Phos/CTGTCTCTTATACACATCT	Used to assemble Tn5 transposome. Nextera adapter complementary sequence to Tn5_1 and Tn5_2_ME_comp
Fwd_atac-starr_tag	TCGTCGGCAGCGTCAGATG	used to extend and amplify tagments prior to cloning
Rev_atac-starr_tag	GTCTCGTGGGCTCGGAGATG	used to extend and amplify tagments prior to cloning
STARR_GSP	CTCATCAATGTATCTTATCATGTCTG	Gene Specific Primer used in Reverse Transcription (from Muerdter et al. 2018)
Nextera Index N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG	i7 Barcode primer #1 for Nextera library preps (8bp barcode)
Nextera Index N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG	i7 Barcode primer #2 for Nextera library preps (8bp barcode)
Nextera Index N504	AATGATACGGCGACCACCGAGATCTACACAGAGTAGATCGTCGGCAGCGTC	i5 Barcode primer #4 for Nextera library preps (8bp barcode)
Nextera Index N505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGTCGTCGGCAGCGTC	i5 Barcode primer #5 for Nextera library preps (8bp barcode)
Fwd_qPCR_STARR-ORI-reporter-RNA	CACTGGGCAGGTGTCC	qPCR primer
R_qPCR_hSTARR_ORI_reporter_RNA	GTCTCTTATACACATCTGACGC	qPCR primer
GAPDH_fwd	AAATCAAGTGGGGCGATGCT	qPCR primer
GAPDH_rev	CAAATGAGCCCCAGCCTTCT	qPCR primer
ACTB_fwd	GTTGTCGACGACGAGCG	qPCR primer
ACTB_rev	GCACAGAGCCTCGCCTT	qPCR primer
IFIT2_fwd	AAGGGTGGACACGGTTAAAG	qPCR primer
IFIT2_rev	GGTACTGGTTGTCAGGATTCAG	qPCR primer
ISG15_fwd	AGCATCTTCACCGTCAGGTC	qPCR primer
ISG15_rev	GCGAACTCATCTTTGCCAGT	qPCR primer
IFNB1_fwd	GTTTCGGAGGTAACCTGTAAGT	qPCR primer
IFNB1_rev	GAACCTCCTGGCTAATGTCTATC	qPCR primer

Massively Parallel Cloning

Four 10µL gibson cloning reactions were performed with NEBuilder® HiFi DNA Assembly Master Mix at a vector:insert molar ratio of 1:2. As a negative control, we performed one cloning reaction substituting tagments with nuclease-free water. Gibson products were pooled and purified via ethanol precipitation as previously described in Sambrook & Russell (Sambrook & Russell, 2006); we used glycoblue (150µg/mL) as a co-precipitant. Purified gibson products were electroporated into MegaX DH10B T1R Electrocomp™ Cells (Invitrogen, # C640003) using a Bio-Rad Gene Pulser. Three electroporations for the ATAC-STARR-seq sample (and 1 for the control) were performed with the following parameters: exponential decay pulse type, 2kV, 200Ω, 25µF, and 0.1cm gap distance. Pre-warmed SOC media (1mL) was added immediately following electroporation; the three reactions were pooled and incubated at 37°C for 1 hour. We confirmed cloning success by plating a dilution series—using a small aliquot from the ATAC-STARR-seq and negative control samples—onto pre-warmed LB agar plates containing 100µg/mL ampicillin and visualizing colonies 24 hours later. The remaining ATAC-STARR-seq transformation was added directly to a 1L LB liquid culture with 100µg/mL ampicillin and grown at 37°C while shaking at 225rpm overnight. The next day, plasmid DNA was harvested from the 1L culture using the ZymoPURE II Plasmid Gigaprep (Zymo Research, #D4204). Before prepping, we recorded a 1.633 optical density.

Electroporation

GM12878 cells were cultured so that cell density was between 400,000 and 800,000 cells/mL on day of transfection. Three replicates were performed on separate days. For each replicate, a total of 20 electroporation reactions was performed using the Neon™ Transfection

System 100 µL Kit (Invitrogen, #MPK10025) and the associated Neon™ Transfection System (Invitrogen, #MPK5000). 121 million GM12878 cells were collected, washed with 45mL PBS, and resuspended in 2178µL Buffer R. For each reaction, 5 million cells (in 90µL Buffer R) were electroporated with 5µg of ATAC-STARR-seq plasmid DNA (in 10µL nuclease-free water) in a total volume of 100µL with the following parameters: 1100V, 30ms, and 2 pulses. Electroporated cells were dispensed immediately into a pre-warmed T-75 flask containing 50mL of RPMI 1640 with 20% fetal bovine serum and 2mM GlutaMAX.

Cell Harvest

24 hours after transfection, the 50mL ATAC-STARR-seq flask was divided into two equal volumes; plasmid DNA was extracted from one volume, while reporter RNAs were extracted from the other. Plasmid DNA was isolated with the ZymoPURE II Plasmid Midiprep kit (#D4200) and eluted in 50µL 10mM Tris-HCL pH 8.0. Prior to lysis, cells were washed with 25mL PBS to remove any extracellular plasmid DNA. Reporter RNAs were extracted in a stepwise process. First, total RNA was isolated from the second volume of transfected cells using the TRIzol™ Reagent and Phasemaker™ Tubes Complete System (Invitrogen™, #A33251). Specifically, 5mL TRIzol was added to homogenize the washed and pelleted cells. Next, polyadenylated RNA was isolated from total RNA using oligo(dT)25 Magnetic Beads (NEB, #S1419S) at a 1µg Total RNA to 10µg beads ratio. We performed this step at 4°C and eluted into 50µL10mM Tris-HCl pH 7.5. The extracted poly(A)$^+$ RNA was treated with DNase I (NEB, #M0303S). This reaction was cleaned up using the Zymo Research RNA Clean & Concentrator-25 kit (Zymo Research, #R1018).

First-strand cDNA synthesis

For each sample, ten 50µL reverse transcription reactions were carried out using PrimeScript™ Reverse Transcriptase (Takara, #2680) and a gene specific primer (STARR_GSP) as described by Muerdter *et al.* 2018 (Muerdter et al., 2018). Single-stranded cDNA was treated with RNase A at a concentration of 20µg/mL in low salt concentrations and cleaned up with a Zymo Research DNA Clean & Concentrator-5 kit.

Illumina Sequencing Library Preparation

For reisolated plasmid and reporter RNA samples, Illumina-compatible libraries were generated using NEBNext® Ultra™ II Q5® Master Mix and a unique combination of the following Nextera indexes: N504-N505 (i5) and N701-N702 (i7), see Table 4 for primer sequences. DNA samples were amplified for 8 PCR cycles, while RNA was amplified for 12-13 cycles. In both cases, products were purified with the Zymo Research DNA Clean & Concentrator-5 kit and analyzed for concentration and size distribution using a HSD5000 screentape. Purified products were sequenced on an Illumina NovaSeq, PE150, at a requested read depth of 50 or 75 million reads, for DNA and RNA samples, respectively, on an Illumina NovaSeq 6000 machine through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core.

Read Processing

FASTQ files for the two Omni-ATAC-seq replicates from Corces *et al*. 2017 and all five HiDRA replicates from Wang *et al.* 2018 were downloaded from the NCBI sequence read archive (run codes: SRR5427886- SRR5427887 and SRR6050484-SRR6050523, respectively) and were processed using the same pipeline as ATAC-STARR-seq (Corces et al., 2017; X. Wang et al., 2018). For this publicly available data and our own, FASTQ files were trimmed and analysed for quality with Trim Galore! (version 0.6.7) using the --fastqc and --paired parameters. Trimmed reads were mapped to hg38 with bowtie2 (version 2.3.5.1) using the following parameters: -X 500 --sensitive --no-discordant --no-mixed (Langmead & Salzberg, 2012). Mapped reads were filtered to remove reads with MAPQ < 30, reads mapping to mitochondrial DNA, and reads mapping to ENCODE blacklist regions using a variety of functions from the Samtools software package (version 1.13) (H. Li et al., 2009). When desired, duplicates were removed with the *markDuplicates* function from Picard (version 2.26.3) (https://broadinstitute.github.io/picard/). Read count was determined using the *flagstat* function from Samtools. Read counts for each step are provided in Supplemental Table S1. We also provide a python script on our GitHub repository (Hansen & Hodges, 2022b) that performs the processing steps above. Complexity was estimated using the *lc-extrap* function from the Preseq package (version 2.0.0) (Daley & Smith, 2013) and insert size was determined using the *CollectInsertSizeMetrics* function from Picard. Complexity curves were plotted in R with ggplot2.

Accessibility Analysis

Peak Calling. We called accessibility peaks with the Genrich software package (version 0.5, https://github.com/jsh58/Genrich), using deduplicated bam files. For ATAC-STARR-seq, we used all three replicates of reisolated plasmid samples. For Corces data, we used the two available replicates. For both, we set a false-discovery rate of 0.0001 and the -j parameter, which specifies ATAC-seq mode.

Peak Comparisons. Peaks between Corces and ATAC-STARR-seq plasmid DNA were compared using the *jaccard* function from the BEDTools package (version 2.30.0) (Quinlan &

Hall, 2010). FRiP scores (the fraction of reads in peaks) and the genomic fraction represented by each peak set was calculated using custom code available on our GitHub repository. Euler plots were made in R with the eulerr package (version 6.1.0) (Larsson, 2021) and bar charts were made in R with ggplot2.

Signal Tracks. Accessibility signal tracks were generated with the *bamCoverage* function from the deepTools package (version 3.5.1) (Ramirez et al., 2016) using the following parameters: -bs 10 --normalizeUsing CPM -e --centerReads. Signal was plotted using the Sushi package (version 1. 30.0) (Phanstiel, Boyle, Araya, & Snyder, 2014) in R.

Active and Silent Region Calling

We called active and silent regions using the sliding window and fragment groups methods. In both cases, except where specified, mapped read files containing duplicates were used for region calling. Overlap between the two active region sets identified by each method was determined using BEDTools jaccard. Methods for each are listed below.

Sliding Window. Within ATAC-STARR-defined open chromatin regions, we generated 50 bp genomic, sliding window bins with a 10bp step size using the *makewindows* function and -s 10 -w 50 parameters from the BEDTools software package. Bins smaller than 50bp were removed from the analysis and reads were counted per bin for each replicate using the *featureCounts* function from the Subread package with the following parameters: -p -B -O --minOverlap 1 (Liao, Smyth, & Shi, 2014). The resulting counts matrix was pre-filtered to remove bins with zero counts and then analyzed with the DESeq2 software package (version 1.32.0) in R to identify active and silent bins (Love et al., 2014). Bins with an Benjamini–Hochberg (BH) adjusted p-value < 0.1 and $\log_2$ fold-change (RNA/DNA) > 0 were defined as active, whereas silent had a BH adjusted p-value < 0.1 and $\log_2$ fold-change (RNA/DNA) < 0. Overlapping and book-ended bins were merged with the *merge* function from BEDTools (using default parameters), resulting in active and silent regions. A python script for region calling is available on our GitHub repository. For the sliding window strategy, we also performed the analysis with or without duplicates in order to compare the results. For the without-duplicate analysis, deduplicated bam files were used at the *featureCounts* step, otherwise all parameters were the same. Active regions were compared using the *jaccard* function from the BEDTools package. Scatter plots and correlation coefficients for replicate-to-replicate comparisons were generated by first extracting DESeq-normalized counts, using the *counts(normalized=TRUE)* function, plotted using ggplot2, and compared using the *cor.test()* function in R using both Spearman's and pearson correlation methods.

Fragment Groups. We generated fragment groups using custom code based on the method described in Wang *et al* 2018 (X. Wang et al., 2018). Paired-end mapped reads were converted from bam to bed format using the *bamtobed* function from the BEDTools software package with option -bedpe and a custom *awk* function. Overlapping paired-end fragments were grouped using the *bedmap* function from the BEDOPS software package (version 2.4.28) (Neph et al., 2012) using the following parameters: --count --echo-map-range --fraction-both 0.75. Importantly, only fragment groups made up of 10 or more reads were used for downstream analysis. Reads were counted per fragment group for each replicate bam file using the *featureCounts* function from the Subread package (version 2.0.1) with the following parameters: -p -B -O --minOverlap 1. The resulting counts matrix was pre-filtered to remove bins with zero counts and then analyzed with the DESeq2 software package in R to identify active fragment groups. Fragment groups with an adjusted p-value < 0.1 and $\log_2$ fold-change (RNA/DNA) > 0 were defined as active. This method resulted in many fragment groups that overlapped each other, so we isolated the most active region within each overlap using a custom function available on our GitHub repository; the resulting, non-redundant regions were defined as active peaks.

Replicate Count Effects

HiDRA replicate count comparison. Raw HiDRA sequencing data was downloaded and processed as described in the read processing section above. Using the same bins generated and analyzed in the active and silent region calling section, reads from all five HiDRA replicates were counted per bin using the *featureCounts* function from the Subread package and the following parameters: -p -B -O --minOverlap 1. Active and regions were called in the same manner as described in the active and silent region calling section using either three or five replicates. Region counts for each condition were plotted using ggplot2.

Pseudo-replicate analysis. To create pseudo-replicates, all three replicate bam files of our ATAC-STARR data were merged using Samtools *merge.* Merged reads were split into five separate files using the Samtools *view* command with the *-s* options set to $rep.2, where .2 represents 20% of the reads and $rep represents the seed number for random sampling. In this way, each pseudo-replicate was sampled with a unique seed number and should, therefore differ from the other pseudo-replicates. Using the same bins analyzed in the active and silent region calling section, reads from all five pseudo-replicates were counted per bin and active regions were called in the same manner as described in the active and silent region calling section using two, three, four, or five pseudo-replicates. Region counts for each condition were plotted using ggplot2.

Short vs. Long DNA Fragment Analysis

Reads were split from filtered bam files (read duplicates included) into short and long groups using samtools *view* piped to an awk command that filters paired end fragments shorter/equal to 125nts (awk ‘*substr($0,1,1)=="@"* || *($9<= 125 && $9>=0)* || *($9>= -125 && $9<=0)’*) or longer than 125nts (awk *‘substr($0,1,1)=="@"* || *($9> 125)* || *($9<-125)’*). Read counts were performed with samtools *flagstat*. Active and silent regulatory regions were called in the same manner as described above using the “sliding windows” approach. Overlaps were calculated using bedtools *jaccard* (default parameters). Region size was calculated in R and annotation was perfomed using the ChIPSeeker package (version 1.28.3) (Yu, Wang, & He, 2015); promoters were defined as 2kb upstream and 1kb downstream of a TSS. All plots were made using ggplot2 in R.

Orientation Analysis

Replicate bam files were merged using Samtools *merge*. Reads were split by orientation using Samtools *view -f*, which selects reads based on their SAM flags. Reads with flags 99 and 147 were assigned to the 5’-3’ bam file, while reads with flags 83 and 163 were assigned to the 3’-5’ bam file. The same bins generated and analyzed for region calling were used. Bins designated as active and silent were used for the active only and silent only analysis, respectively. The three bin sets were further subset into proximal and distal based on distance to the nearest TSS using the ChIPSeeker software package; proximal bins were defined as 2kb upstream and 1kb downstream of a TSS while distal was everything else. For each subset of bins, reads were counted per bin for the orientation-specific bam files using the *featureCounts* function from the Subread package with the following parameters: -p -B -O --minOverlap 1. Scatter plots of counts per million normalize read count were generated with ggplot2 and both Spearman's and pearson correlation coefficients were determined with the *cor.test()* function in R. Bins with a greater than 5 read count difference between insert orientations were considered to be biased; we based this threshold on the all distal bins scatterplot with the assumption that distal bins should not display an orientation bias. The percentage biased was plotted with ggplot2.

Active and Silent Peak Characterization

Annotation. Active and silent peak sets were annotated relative to transcription start site (TSS) locations and plotted in R using the ChIPSeeker package (version 1.28.3) (Yu et al., 2015); promoters were defined as 2kb upstream and 1kb downstream of a TSS. ChromHMM state was assigned to each peak using the BEDTools *intersect* function and -u parameter; the list of hg38 18-state ChromHMM regions (Roadmap Epigenomics et al., 2015) (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/E116_18_core_K27ac_hg38lift_mnemonics.bed.gz) were intersected against the regions sets of interest and the proportion was plotted with ggplot2.

Heatmaps. The activity bigwig was generated with the deepTools package. Merged bam files for RNA and DNA were converted to counts per million normalized bedGraph files using the *bamCoverage* function and the following parameters: -bs 10 --normalizeUsing CPM. The resulting RNA bigwig was normalized to the DNA bigwig to generate a signal file of $\log_2$(RNA/DNA) ratio using the *bigwigCompare* function and the following parameters: -bs 1 --operation log2 --pseudocount 1 –skipZeroOverZero. Heatmaps were generated using the deepTools package. Activity signal was plotted at distal and proximal regions and region order was ranked by maximum mean signal. GM12878 ChIP-seq bigwig files were downloaded from the ENCODE consortium (The ENCODE Project Consortium et al., 2020) and plotted. The matrix was made using the *computeMatrix* function, with the following parameters: -a 2000 -b 2000 --referencePoint center -bs 10 --missingDataAsZero. The matrix was plotted using the *plotHeatmap* function with the following key parameters: --sortUsing mean --sortUsingSamples 1.

Histone Signal Boxplots. We intersected silent and active regions with our accessible peaks file using the *intersect* function from the BEDTools software package to get peaks that contain an active region, a silent region, both an active and silent region, or neither. Using the *slop* function from BEDTools we then extended ChrAcc peaks by 1kb on either side and then used the bigwigCompare function from the DeepTools package to determine H3K4me1/H3K4me3/H3K27ac/H3Kme3 GM12878 ChIP-seq bigwig signal distributions for each for the ChrAcc peak types. The same ENCODE files used in the heatmap analysis above, were also used here. The plotted values represent the average *fold-change over control* for each ChrAcc peak +/- 1kb. Plots were made with ggplot2.

Motif enrichment. We performed motif enrichment on the active and silent peak sets using the findMotiftsGenome.pl script from the HOMER package (version 4.10, http://homer.ucsd.edu/) (Duttke, Chang, Heinz, & Benner, 2019) using the following parameters: -size given -mset vertebrates. Plots were made with ggplot2.

Neutral region calling. Neutral regions were called in the exact same manner as active or silent except for one critical difference: only bins with padj > 0.1 were selected. Annotation of distance to nearest TSS and ChromHMM were performed as described for the active and silent regions above.

TF footprinting

Computational footprinting. Transcription factor footprinting was performed using the TOBIAS software package (version 0.12.12) (Bentsen et al., 2020). Deduplicated mapped reads were used to generate Tn5-bias corrected bigwig signal files using the *ATACorrect* function. Using the corrected signal files, TF binding was calculated with the *ScoreBigWig* function and footprints for individual TFs were called for all core non-redundant vertebrate JASPAR motifs (Fornes et al., 2020) using the *BINDetect* function. Motifs with a footprint were classified as "bound", while motifs without a footprint were classified as "unbound". The "archetype" for each TF was assigned by cross-referencing the motif annotations table from Viestra *et al*. 2020 (Vierstra et al., 2020).

Data Visualization. Heatmaps were generated using the deepTools package. GM12878 ChIP-seq bigwig files were downloaded from ENCODE (www.encodeproject.org) (The ENCODE Project Consortium et al., 2020) and plotted with Tn5-corrected signal at all accessible CTCF and ETS/1 motifs (defined as the "all" bed file for CTCF or ETS1 from BINDetect) using the *computeMatrix reference-point* function with the following key parameters: -a 200 -b 200 --referencePoint center --missingDataAsZero -bs1. The resulting matrix was plotted using the *plotHeatmap* function and the following key parameters: --sortUsing mean --sortUsingSamples 1. Aggregate plots were also generated using the deepTools package. Tn5-corrected signal was measured at bound and unbound sites for each TF archetype using the *computeMatrix reference-point* function with the following key parameters: -a 75 -b 75 --referencePoint center --missingDataAsZero -bs 1. The resulting matrix was plotted using the *plotProfile* function.

Integration of Regulatory Activity, Chromatin Accessibility, and TF footprinting

Signal and regions were visualized at the given locus using the Sushi package in R. To determine the presence or absence of a TF footprint, we intersected TF footprints with the active and silent regions bed file and reported +/- for presence of the footprint using custom code available on our GitHub repository. Footprints were selected based on top hits from the motif enrichment analysis above. Active and silent regions without a footprint for the queried TFs were removed from the analysis. We clustered the region subsets with the pheatmap package (version 1.0.12, https://github.com/raivokolde/pheatmap), using the clustering_distance_row/columns = "binary" parameter; we cut the tree into 6 clusters for active and silent. We extracted the regions from each cluster and then, using the ChIPSeeker package, assigned the nearest neighbor gene. Using ClusterProfiler (Yu, Wang, Han, & He, 2012) and ReactomePA (Jassal et al., 2020), we then performed reactome pathway enrichment analysis on the nearest neighbor gene sets. We applied a 0.05 and 0.1 p-value cut-off for active and silent clusters, respectively.

Determination of Harvest Time with Quantitative PCR

GM12878 cells were cultured so that cell density was between 400,000 and 800,000 cells/mL on day of transfection. Three replicates were performed on separate days. For each sample, 5 million GM12878 cells were electroporated with 5µg ATAC-STARR-seq plasmid DNA using the Neon™ Transfection System 100 µL Kit (Invitrogen, #MPK10025) and the associated Neon™ Transfection System (Invitrogen, #MPK5000) in Buffer R with the following parameters: 1100V, 30ms, and 2 pulses. Electroporated cells were dispensed immediately into pre-warmed T-12.5 flasks containing 6.25mL of RPMI 1640 with 20% fetal bovine serum and 2mM GlutaMAX.

Total RNA was harvested at various time points—3hr, 6hr, 12hr, 24hr, and 36hr—using the TRIzol™ Reagent and Phasemaker™ Tubes Complete System (Invitrogen™, #A33251). For each sample, 0.75mL TRIzol was added to cell pellets. First-strand cDNA synthesis was performed using an Oligo $(dT)_{25}$ primer and the SuperScript™ IV First-Strand Synthesis System (Invitrogen™, #18091050). cDNA was treated with RNase H to remove RNA from RNA-DNA dimers. For each replicate, 10µL quantitative PCR reactions were performed in technical triplicate using PowerUp™ SYBR™ Green Master Mix (Applied Biosystems™, #A25742) on a StepOnePlus™ Real-Time PCR System (Applied Biosystems™, #4376600). For each reaction, 1µL of the reverse-transcribed product was added and gene-specific primers were supplied at a final concentration of 500nM (see Supplemental Table S4 for primer sequences). Fold-change was calculated with the ΔΔCt method, using either GAPDH or ACTB as the housekeeping gene for reporter RNA or ISG targets, respectively. Plots were made with ggplot2 (version 3.3.5) (Wickham, 2016) in R (version 4.1.1).

Plasmid Library Complexity Estimation

Plasmid inserts were amplified via PCR for 10 cycles from 3.75µg ATAC-STARR-seq plasmid library using NEBNext® Ultra™ II Q5® Master Mix and the Nextera indexes, N505 and N701, see Supplemental Table S3 for primer sequences. Products were purified with the Zymo Research DNA Clean & Concentrator-5 kit (#D4013) and analyzed for concentration and size distribution using a HSD5000 screentape. Purified products were sequenced on an Illumina NovaSeq, PE150, at a requested read depth of 25 million reads through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core.

Transfection efficiency estimation

Transfection efficiency is a critical ATAC-STARR-seq bottleneck, particularly for difficult to transfect cells like GM12878. In parallel with ATAC-STARR-seq, we electroporated GM12878 cells with a pcDNA3.1-eGFP plasmid and estimated transfection efficiency as the percentage of GFP positive cells when measured by flow cytometry 24 hours later. Specifically, GM12878 cells were electroporated following same conditions as above with either purified pcDNA3.1-eGFP plasmid or nuclease-free water and then prepared for flow cytometry 24 hours later at a concentration of 1.25×10^6 cells/mL in 1xPBS solution containing 1% BSA. We halved both GFP and water samples and stained one half of each with propidium iodide (Sigma-Aldrich, #P4864). Unstained cells (water/PI-) were used in conjunction with compensation control cells (GFP/PI- or water/PI+) to quantify the percentage of living GFP positive cells in the experimental condition (GFP/PI+) via flow cytometry; this percentage was the reported transfection efficiency. When performed in parallel to ATAC-STARR-seq plasmid library transfection, we consistently achieve around 10-20% efficiency (data not shown).

CHAPTER III[2]

HUMAN GENE REGULATORY EVOLUTION IS DRIVEN BY THE DIVERGENCE OF REGULATORY ELEMENT FUNCTION IN BOTH CIS AND TRANS

Introduction

Phenotypic divergence between closely related species is driven primarily by non-coding mutations that alter gene expression, rather than protein structure or function (Brawand et al., 2011; Britten & Davidson, 1969, 1971; Franchini & Pollard, 2017; King & Wilson, 1975; Reilly & Noonan, 2016; Sholtis & Noonan, 2010). Gene expression changes can result from divergence in 1) *cis*, where DNA mutations alter local regulatory element activity, or 2) *trans*, where changes alter the abundance or activity of transcriptional regulators (Hill et al., 2020; Signor & Nuzhdin, 2018). These two modes of change have different mechanisms and scopes of effects on gene expression outputs. Each *cis* change influences a single regulatory element and its immediate local targets, while a *trans* change globally influences many regulatory elements and their gene targets. Thus, determining the respective contributions of *cis* versus *trans* changes to between-species gene expression differences is key to understanding the mechanisms that generate phenotypic divergence. Furthermore, because gene regulatory variants in humans are often associated with

[2] This chapter is adapted from "Human gene regulatory evolution is driven by the divergence of regulatory element function in both *cis* and *trans*" published in bioRxiv and has been reproduced with the permission of the publisher and my co-authors Sarah Fong, Ph.D., John A. Capra, Ph.D., and Emily Hodges, Ph.D. | Citation: "Hansen, T. J., Fong, S., Capra, J. A., & Hodges, E. (2023). Human gene regulatory evolution is driven by the divergence of regulatory element function in both *cis* and *trans*. bioRxiv, 2023.02.14.528376; doi: https://doi.org/10.1101/2023.02.14.528376"

disease phenotypes, understanding these mechanisms will facilitate interpretation of genetic variation on disease.

Cis and *trans* changes are difficult to study independently because cellular environment and genomic sequence are inherently linked within endogenous settings. Previous studies have developed different approaches largely focused on gene expression levels to attempt to disentangle *cis* and *trans* mechanisms of gene regulatory evolution (Agoglia et al., 2021; Barr, 2022; Consortium, 2020; Coolon et al., 2014; Emerson et al., 2010; Goncalves et al., 2012; Graze, McIntyre, Main, Wayne, & Nuzhdin, 2009; Hill et al., 2020; X. C. Li & Fay, 2017; X. Liu, Li, & Pritchard, 2019; McManus et al., 2010; Meiklejohn, Coolon, Hartl, & Wittkopp, 2014; B. P. H. Metzger, Wittkopp, & Coolon, 2017; Osada et al., 2017; Shi et al., 2012; Takahasi, Matsuo, & Takano-Shimizu-Kouno, 2011; Tirosh et al., 2009; Vosa et al., 2021; Wittkopp et al., 2004, 2008). Overall, these studies have yielded a complex picture of the roles of *cis* and *trans* changes in different settings, but they generally argue that *cis* changes drive most divergence in gene expression between closely related species.

Gene expression is driven by regulatory element activity; thus, to gain a better understanding of the molecular mechanisms underlying gene regulatory evolution, it is necessary to investigate *cis* and *trans* changes at the regulatory element level. To directly identify *cis* differences, several recent studies have compared the regulatory activity of homologous sequences between closely related species within a common cellular environment (Arnold et al., 2014; Klein et al., 2018; Uebbing et al., 2021; Weiss et al., 2021). By controlling the cellular environment, the regulatory element activity differences identified by these studies must be the result of changes in *cis* (i.e., sequence).

In contrast, only a handful of studies have directly tested the contributions of *trans* changes to regulatory element activity between species by comparing regulatory activity of the same sequences across species-specific cellular environments (Gordon & Ruvinsky, 2012; Mattioli et al., 2020; Whalen et al., 2023). Collectively, these studies conclude that *trans* changes to regulatory element function occur less frequently than *cis* changes and suggest that *cis*-variation primarily drives divergent regulatory element activity between closely related species (Irene Gallego Romero & Lea, 2022). One recent study comparing regulatory element activity in human and mouse embryonic stem cells reported ~70% of activity differences were due to changes in *cis* (Mattioli et al., 2020). However, this study considered small (~1,600), pre-selected subsets of regulatory elements, and as a result, a comprehensive and unbiased survey of *cis* and *trans* contributions to global gene regulatory divergence remains a key gap in understanding mechanisms of gene regulatory evolution.

In this study, we develop a comparative ATAC-STARR-seq framework to comprehensively dissect *cis* and *trans* contributions to regulatory element divergence between species. ATAC-STARR-seq captures almost all chromatin accessible DNA fragments and assays them for regulatory activity. Because we create a reporter plasmid library separate from performing the reporter assay, our approach decouples sequence from cellular environment. Thus, sequences from a species of interest can be tested for activity within any chosen cellular environment. This allows us to systematically measure the effect of homologous sequence differences while controlling the cellular environment and *vice versa*.

Our approach expands the scope of analysis from a few thousand regulatory elements to ~100,000 regulatory elements genome-wide without the need for prior knowledge of regulatory potential (Hansen & Hodges, 2022a; X. Wang et al., 2018). Applying ATAC-STARR-seq to human and rhesus macaque lymphoblastoid cell lines (LCLs), we discover that *cis* and *trans* changes contribute to regulatory elements with divergent activity at similar frequencies, which contrasts with previous smaller studies that found *cis* changes drive most gene regulatory variation between species. We show that *cis* divergent elements are enriched for accelerated substitution rates and variants that influence gene expression in human populations, while *trans* divergent elements are enriched for footprints of differentially expressed transcription factors (TFs) that affect multiple gene regulatory loci. Furthermore, we find that the activity of most species-specific regulatory elements diverged in both *cis* and *trans* between human and macaque LCLs. These *cis* & *trans* regions are characterized by enrichment for specific transposable element sub-families harboring distinct TF binding footprints in humans. Finally, we illustrate how knowledge of mechanisms of regulatory divergence enriches interpretation of human variation and gene regulatory networks. By leveraging new technology to evaluate mechanisms of regulatory element divergence genome-wide, our study highlights the interplay between *cis* and *trans* changes on gene regulation and reveals a central role for *trans*-regulatory divergence in driving gene regulatory evolution.

Results

Comparative ATAC-STARR-seq produces a multi-layered view of human and macaque gene regulatory divergence

We applied ATAC-STARR-seq (Hansen & Hodges, 2022a) to assay the regulatory landscape of LCLs between humans and macaques (International HapMap, 2003; Rangan, Martin, Bozelka, Wang, & Gormus, 1986; Tosato & Cohen, 2007) (GM12878 vs. LCL8664; Figure 1A,B). ATAC-STARR-seq enables genome-wide measurement of chromatin accessibility, TF occupancy,

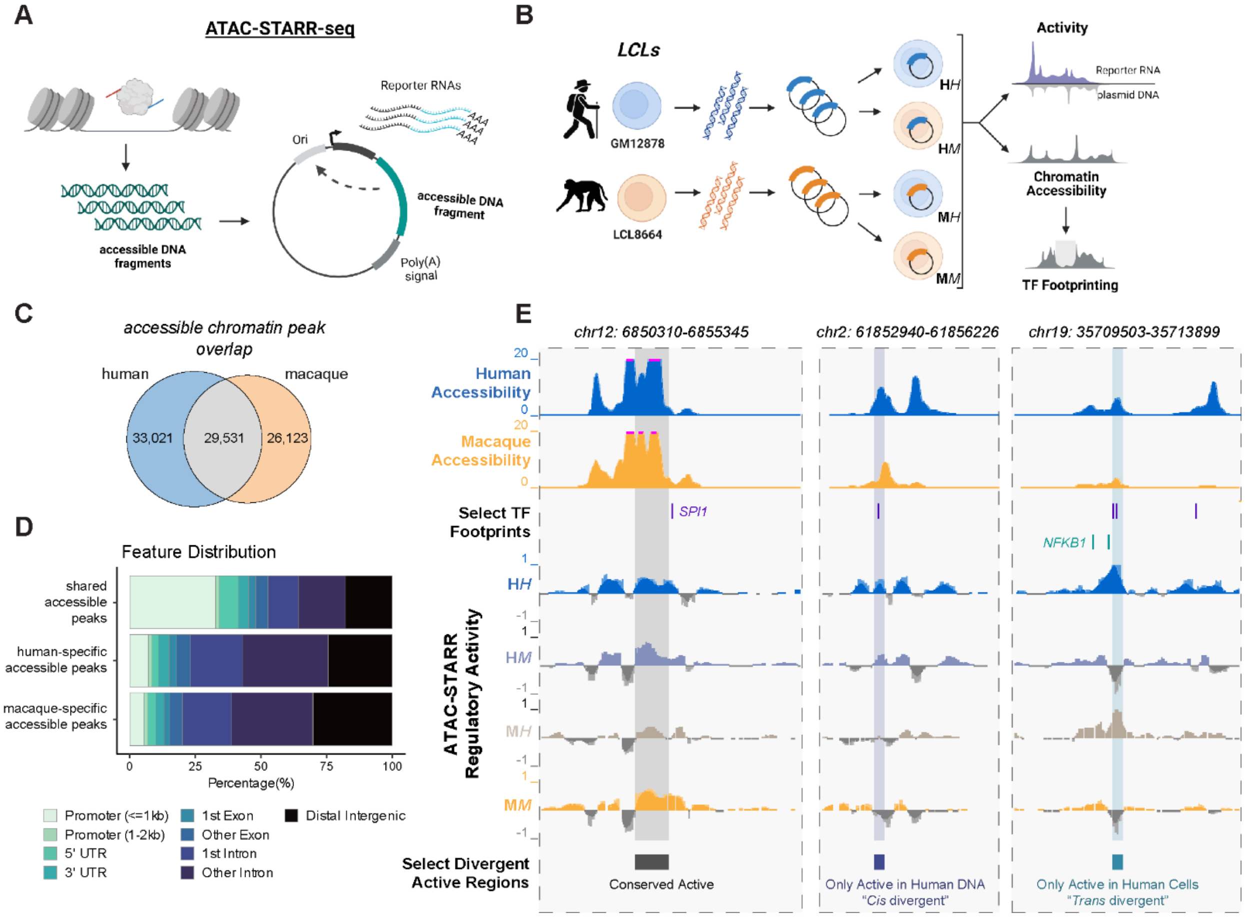

Figure 18: Comparative ATAC-STARR-seq produces a multi-layered view of human and macaque gene regulatory divergence. (A) A schematic of the ATAC-STARR-seq methodology. Accessible DNA fragments are isolated from cells and subsequently cloned into a self-transcribing reporter vector plasmid, which are then electroporated into cells and assayed for regulatory activity by harvesting and sequencing Reporter RNAs and input plasmid DNA. (B) Our comparative ATAC-STARR-seq strategy to assay human and macaque genomes in both cellular environments. ATAC-STARR-seq plasmid libraries were independently generated for GM12878 and LCL8664 cell lines and then assayed separately in either cellular context. Our comparative approach provides measures in chromatin accessibility and transcription factor (TF) footprinting for both genomes as well as regulatory activity for the four experimental conditions: human DNA in human cells (HH), human DNA in macaque cells (HM), macaque DNA in human cells (MH) and macaque DNA in macaque cells (MM). (C) Euler plot representing the number of species-specific and shared accessibility peaks identified from ATAC-STARR-seq data. (D) Distribution of genomic annotations for species-specific and shared accessibility peaks based on the distance to nearest transcription start site. (E) Select genomic loci at hg38 coordinates representing conserved or differentially active regions of the two genomes. Tracks represent human and rhesus macaque accessibility, TF footprints for SPI1 and NFKB1, and regulatory activity measures for HH, HM, MH, MM.

and regulatory element activity, which is the ability of a DNA sequence to drive transcription (Figure 18,19). For each experimental condition, we performed three replicates and obtained both reporter RNA and successfully transfected plasmid DNA samples for each replicate. In all conditions, DNA input libraries were highly complex with estimated sizes ranging between 31-54

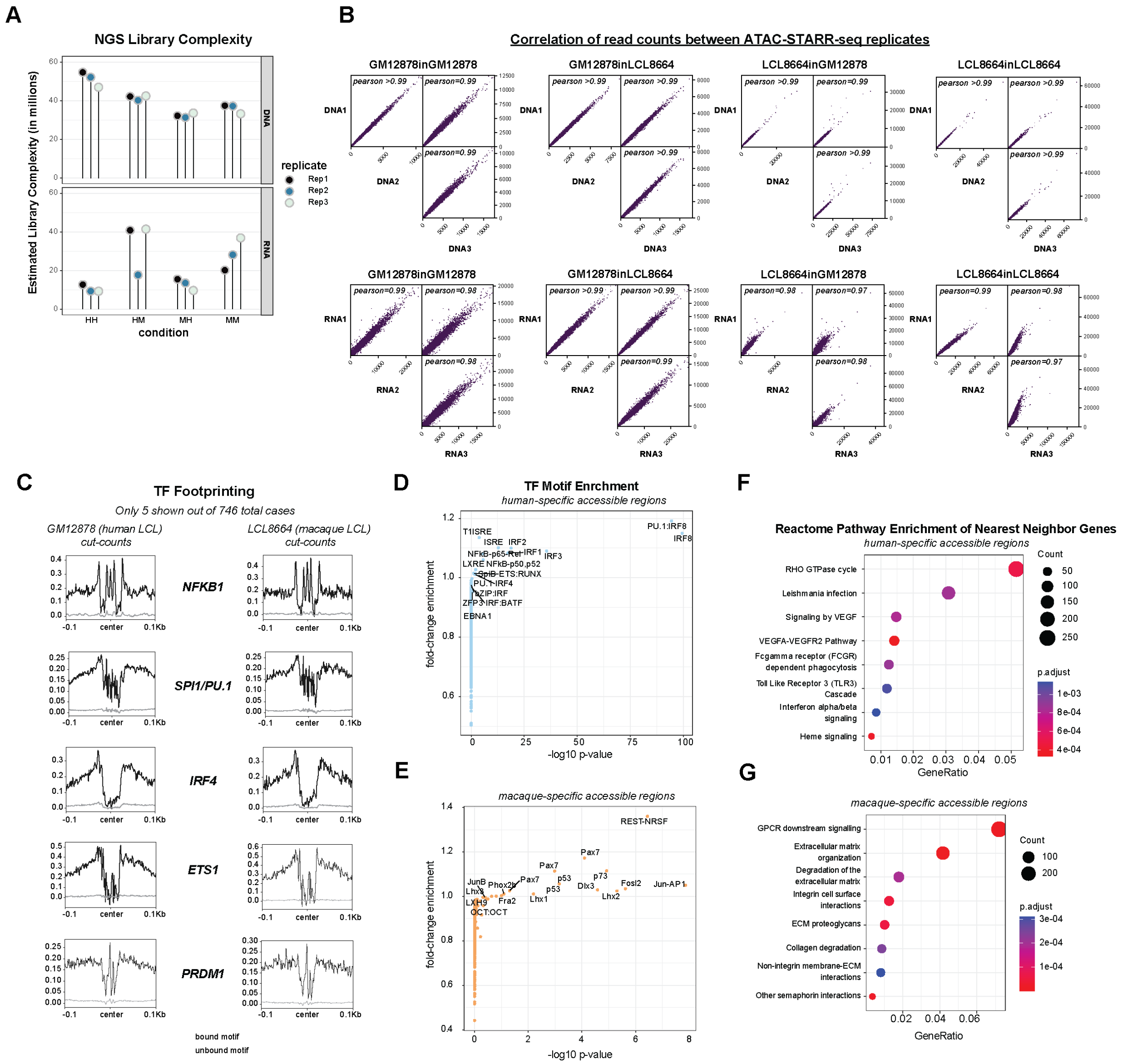

Figure 19: Differential accessibility analysis, TF footprinting, and ATAC-STARR-seq quality control. (A) Estimated sequence library complexities from Picard for each replicate of each condition. This represents the total number of non-redundant sequences contained within the library. (B) Pearson correlation plots between replicates for both RNA and DNA samples for each condition. (C) 5 representative examples of TF footprinting in human and macaque LCLs from ATAC-STARR-seq data. A total of 746 JASPAR motifs were analyzed to identify bound (black line) and unbound (grey line) motifs classified by Tn5 cut-count distributions at the motifs. Bound motifs are also called footprints. (D-E) TF motif enrichment analysis results for either (D) human-specific or (E) macaque-specific accessible regions. (F-G) Reactome pathway enrichment analysis of nearest neighbor genes for either (F) human-specific or (G) macaque-specific accessible regions. Only the top 8 terms are displayed.

million DNA sequences (Figure 19A). Both reporter RNA and plasmid DNA sequencing data were reproducible across the three replicates (Figure 19B; Pearson r^2: 0.97-0.99).

We first determined accessibility peaks using the sequence reads obtained from the input DNA libraries, as previously described.(Hansen & Hodges, 2022a) Previous studies have investigated regions of differential chromatin accessibility in primate LCLs and other tissues,(Edsall et al., 2019; Garcia-Perez et al., 2021; Shibata et al., 2012; Yao et al., 2022) and consistent with these results, most chromatin accessibility peaks identified between the human and macaque genomes (59,144, 67%) is species-specific, while 29,531 (33%) peaks had shared accessibility between species (Figure 18C). As expected, we find that divergent accessibility peaks are distally located and enriched for cell-type relevant functions (Figure 18D, 19C-G). Pinpointing the mechanisms underlying divergent activity requires that regulatory element DNA be captured from and tested in both species. Therefore, we analyzed shared accessible chromatin peaks so that both the human and macaque homologs were assayed. We quantified regulatory activity in four conditions: human DNA in human cells (HH), human DNA in macaque cells (HM), macaque DNA in human cells (MH), and macaque DNA in macaque cells (MM) (Figure 18B). By comparing activity levels of orthologous sequences in these four settings, we can dissect whether *cis* changes, *trans* changes, or both have occurred in every single element tested. Altogether, this produces an integrated, high-resolution quantification of accessibility, TF occupancy, and regulatory activity at both conserved and divergent regulatory elements between human and macaque LCLs (Figure 18E).

Unlike in differential RNA expression analysis, it was necessary to both identify regions of interest and estimate their activity prior to any condition-specific comparison. To do this, we divided the 29,531 shared accessible peaks into sliding bins and retained bins with 1:1 orthology between human and macaque. We called activity for each bin using replicates to determine p-values for activity in each condition and collapsed overlapping bins with consistent activity. This yielded a set of robust active regions for each condition (Figure 20A,B, Methods). Next, we directly compared active regions between the four conditions. We used a rank-based comparison scheme to account for power differences that would affect significance thresholds, assuming that each condition has similar numbers of active regions within shared accessible chromatin. We compared results at several rank thresholds corresponding to different false discovery rate (FDR) thresholds and we observed similar patterns in the divergent activity calls between conditions at all thresholds considered (Figure 20C,D). Thus, we focus in the main text on a rank threshold of 10,000 active regions per condition corresponding to an FDR range of 0.026-0.11. The condition-specific regions were similarly distributed across the genome, with marginal differences in genomic feature content (Figure 21A).

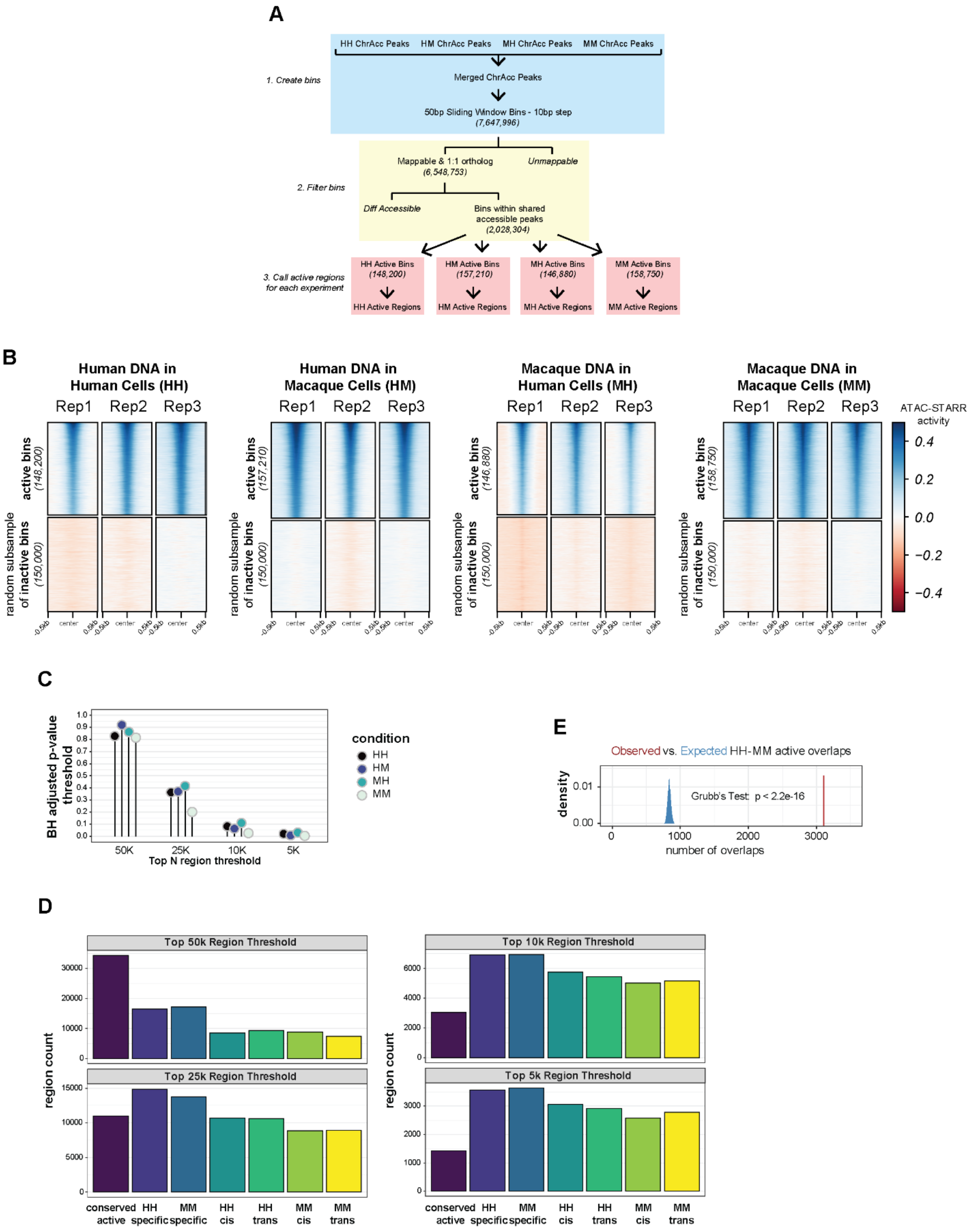

Figure 20: Support of differential activity calls. (A) A schematic of the activity calling approach. Exact bin counts are provided to show how many bins were lost due to filtering steps. (B) Comparison of ATAC-STARR-seq activity values for each replicate of each condition for both all bins called active and for a random subsample of inactive bins. (C) Lollipop chart representing the Benjamini-Hochberg adjusted p-values applied to obtain the various number of regions for each condition. (D) The number of regions classified into each region set based on the number of active regions called per condition. (E) Observed vs. expected analysis of overlaps between the region sets compared in Figure 2B. Red line represents the observed, while blue density plot represents the expected distribution of overlaps for 1000 random shuffles within shared accessible chromatin.

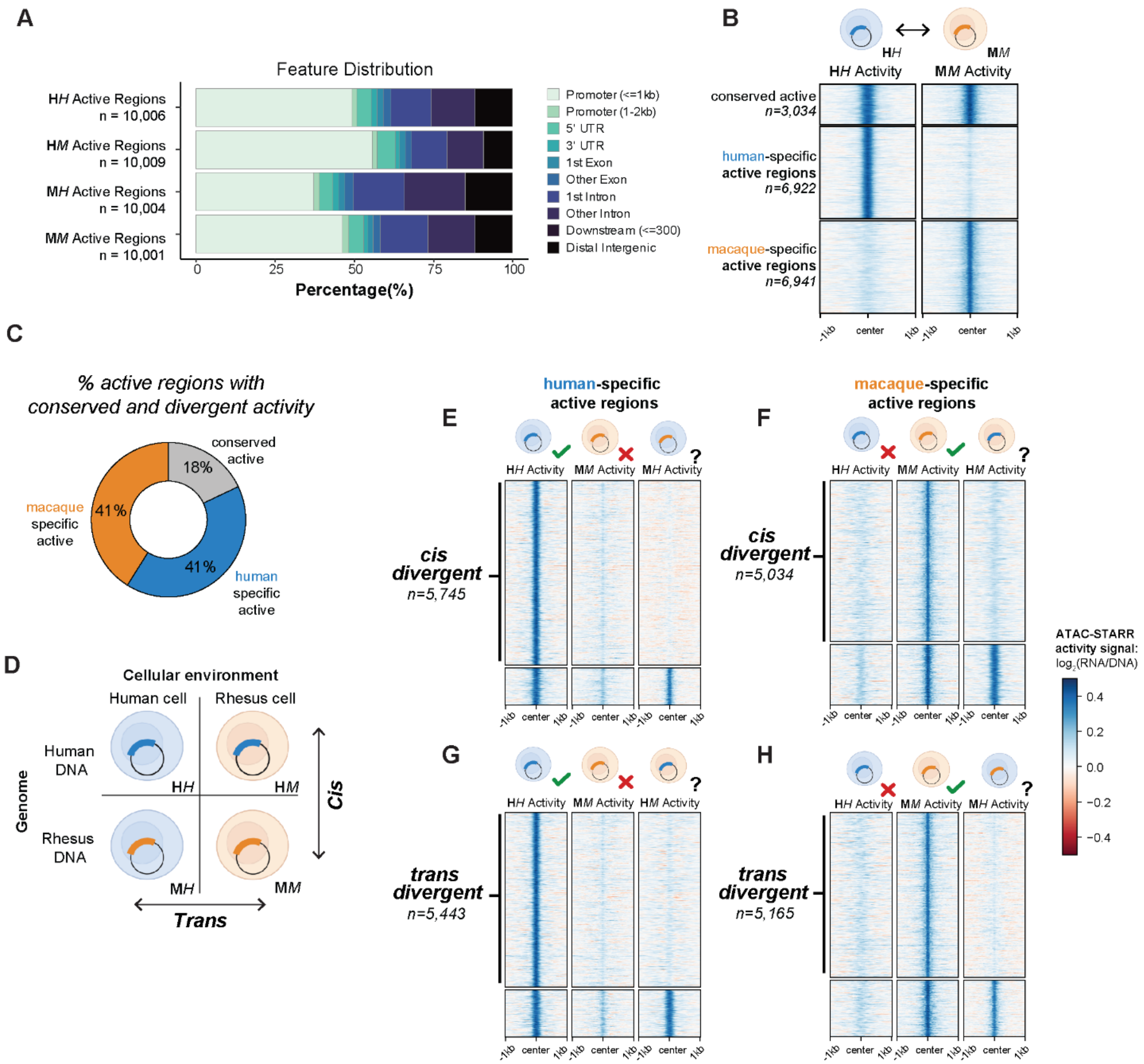

Figure 21: *Cis* and *trans* gene regulatory divergence occur at similar frequencies. (A) Distribution of genomic annotations for the ~10,000 active regions called in each condition based on the distance to nearest transcription start site. (B) Comparison between the human and macaque native states to reveal conserved and species-specific active regions. (C) The percentage of active regions with conserved and divergent activity. (D) Cartoon depicting the four conditions tested and how they are compared to identify *cis* and *trans* divergent regions. (E) Human-specific *cis* divergent regions determined by comparing human-specific active regions with the MH condition. Regions without MH activity were called *cis* divergent regions. (F) Macaque-specific *cis* divergent regions determined by comparing human-specific active regions with the HM condition. (G) Human-specific *trans* divergent regions determined by comparing human-specific active regions with the HM condition. (H) Macaque-specific *trans* divergent regions determined by comparing human-specific active regions with the HM condition. The heatmaps display ATAC-STARR-seq activity values for the specified region sets and experimental conditions.

Cis and trans gene regulatory divergence occur at similar frequencies

We first tested the conservation of regulatory activity between "native states" by comparing human DNA in human cells (HH) and macaque DNA in macaque cells (MM) (Figure 2B). Of the top ~10,000 regions considered, 3,034 (18%) regions have conserved activity, 6,922 (41%) regions were active only in the HH state and 6,941 (41%) were active only in the MM state (Figure 21B,C). The overlap between HH and MM active regions was significantly greater than expected (Figure 20E; p < 2.2e-16), and the divergent activity calls are supported by clear differences in ATAC-STARR-seq regulatory activity signal between HH and MM (Figure 21B). This indicates that many active regulatory sequences with shared accessibility have divergent activity, challenging the widely held assumption that conserved chromatin accessibility signifies conserved regulatory activity.

To determine the contribution of *cis* and *trans* changes to the differentially active regulatory regions, we compared their native activity to the corresponding non-native contexts—i.e., human DNA in the rhesus cellular environment (HM) and rhesus DNA in the human cellular environment (MH) (Figure 21D). We define *cis* changes as cases when sequence orthologs are tested in the same cellular environment but result in activity differences, implying that DNA variation contributes to regulatory activity differences. Conversely, we define *trans* changes as cases when a single sequence tested in different cellular environments results in activity differences, suggesting cellular environment changes contribute to the activity difference.

As expected, *cis* changes contributed to a large proportion of human-specific active regions (83%; 5,745). For these regulatory elements, the human DNA sequence was active in the human cellular environment, but the macaque sequence was inactive in both the macaque and human cells (Figure

21E). Likewise, 73% of macaque-specific active regions (5,034) diverged due to changes in *cis* (Figure 21F).

Surprisingly, similar proportions of human-specific active regions (79%; 5,443) were differentially active due to changes in *trans*, i.e., their DNA sequences were not active when assayed in the macaque cellular environment (Figure 21G). Likewise, 74% of macaque-specific active regions (5,165) were differentially active due to *trans* changes (Figure 21H). This was unexpected based on findings from previous smaller-scale studies that *cis* changes contribute to a greater number of differentially active regions than *trans* changes.(Gordon & Ruvinsky, 2012; Mattioli et al., 2020; Whalen et al., 2023)

Collectively, these data demonstrate that *trans* changes to regulatory element activity occur as frequently as *cis* changes between human and macaque LCLs, indicating that *trans* changes in cellular environments have widespread impact on species-specific gene regulatory activity. These classifications are supported by clear qualitative differences in ATAC-STARR-seq regulatory activity signal between conditions (Figure 21E-H). We also observe equivalent proportions of *cis* and *trans* differences in activity when we vary our threshold for calling activity, indicating the relative abundance of *cis* and *trans* divergence is not sensitive to the threshold used (Figure 20C,D).

Most species-specific regulatory differences are driven by changes in both cis and trans

Because *cis* changes and *trans* changes each contribute to the differential activity of many divergent active regulatory regions, we quantified how often they occur together in the same DNA regulatory element. Unexpectedly, we found that 70% of the human specific active regions (4,631) and 64% of the macaque specific active regions (3,994) displayed both *cis* and *trans* divergence (Figure 22A-D). Accordingly, we classified these regulatory regions as *cis & trans*, and regions only divergent in *cis* or *trans* as *cis only* and *trans only*, respectively. With these definitions, the *cis & trans* class accounts for 67.5% of all divergent active regions (human and macaque combined), whereas *cis only* and *trans only* represent about 17% and 15.5%, respectively. Thus, the regions with divergent regulatory activity between humans and macaques predominantly exhibit functional changes in both sequence and cellular environment, suggesting that *cis* and *trans* mechanisms jointly contributed to the evolution of individual gene regulatory elements.

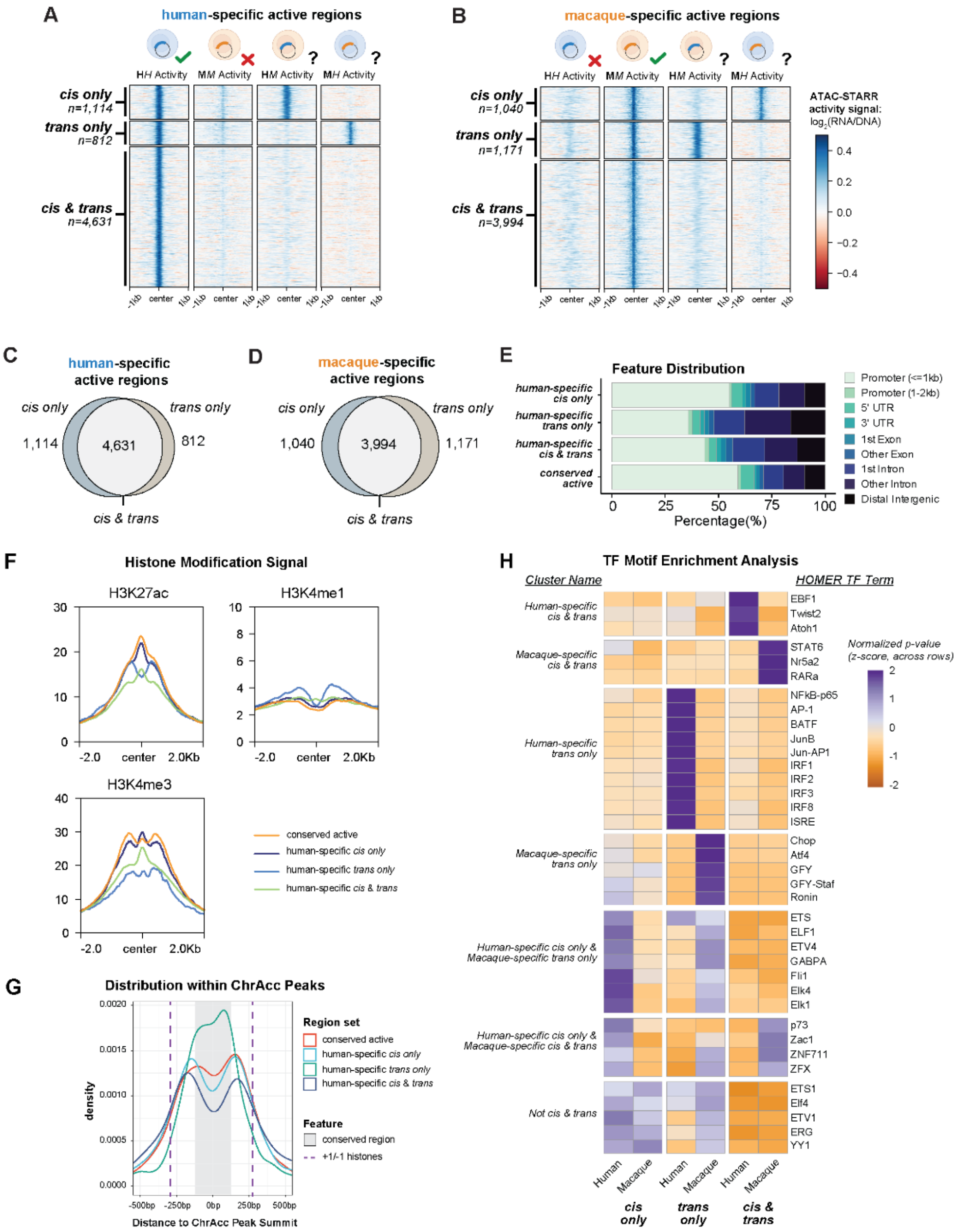

Figure 22: Most species-specific regulatory differences are driven by changes in both *cis* and *trans*. (A,B) Comparison of ATAC-STARR-seq activity values across all conditions for (A) human-specific and (B) macaque-specific *cis* and *trans* divergent regions. *Cis only*, *trans only*, and *cis & trans* regions display activity signals consistent with their calls. (C,D) Euler plots of the *cis only*, *trans only*, and cis & *trans* classifications for (C) human-specific and (D) macaque-specific active regions. (E) Distribution of genomic annotations for human-specific *cis only*, *trans only*, *cis & trans*, and conserved active regions. (F) Profile plots of ENCODE GM12878 ChIP-seq signal for H3K27ac, H3K4me1, and H3K4me3 histone modifications for the human-specific region classes. (G) Density plot of the distances between region center and accessible chromatin (ChrAcc) peak summits for human-specific *cis only*, *trans only*, *cis & trans*, and conserved active regions. The +1 and -1 histones are estimated with purple dashed lines by the ENCODE GM12878 H3K27ac signal summits and the conserved portion of the ChrAcc peaks is estimated with a grey box by the 17-way PhyloP score, see Figure 23C,D. (H) Clustered heatmap of TF motif enrichments for the combined or species separated *cis only*, *trans only*, *cis & trans* regions. Values are the z-score distributions of p-values, normalized across rows. Only the top 15 motifs for each region set are plotted.

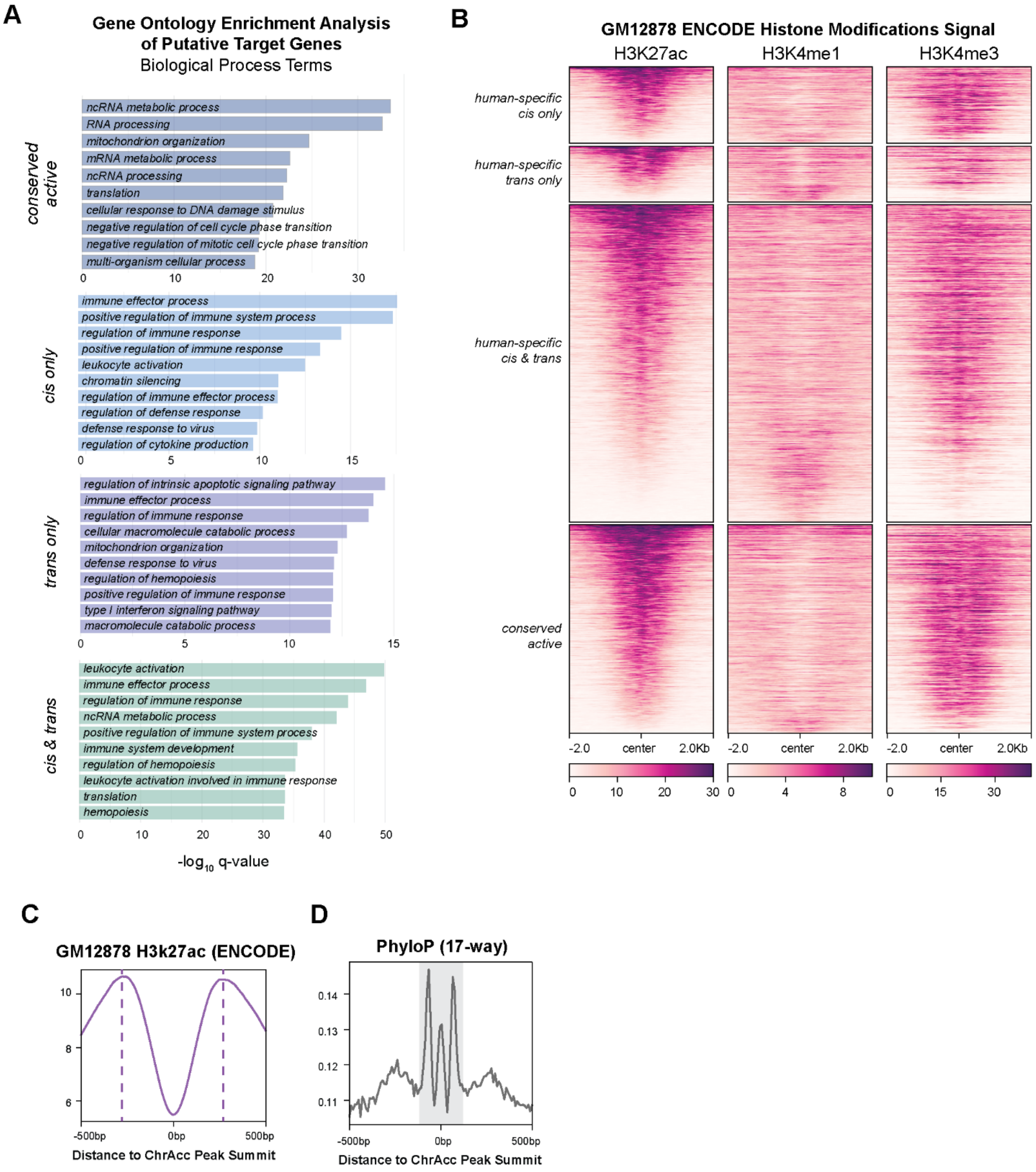

Figure 23: Additional functional characteristics of *cis* only, *trans* only, *cis & trans*, and conserved active region sets. (A) Gene ontology (GO) enrichments for the putative target genes of conserved active, *cis* only, *trans* only, and *cis* & *trans* regions. Only the top 10 terms are shown for each. (B) Heatmaps of ENCODE GM12878 ChIP-seq signal for H3K27ac, H3K4me1, and H3K4me3 histone modifications for each human-specific region class. This is summarized by the profile plots in Figure 3F. (C) H3K27ac and (D) PhyloP signal distributions from accessible chromatin peak centers to define the +1/-1 nucleosomes and conserved region shown in Figure 22G.

Different mechanisms of regulatory divergence exhibit different TF motifs and locations within nucleosome-free regions

Given the prevalence of these distinct modes of regulatory divergence, we investigated the genomic context and functional annotations of the divergent region classes (*cis only*, *trans only*, *cis* & *trans*, and *conserved active*). Functional genomic data for the human GM12878 cell line is readily available, so we focused on the human-specific active regions unless otherwise specified. While all three divergent classes consisted of more promoter-distal regions than the conserved active class, a substantially higher proportion of *trans only* regions overlapped promoter-distal annotations than either *cis only* or *cis* & *trans* regions (Figure 22E), consistent with recent results on *trans* changes between human and mouse.(Mattioli et al., 2020) Gene ontology annotations of genes near each region class revealed that all three *cis*/*trans* region classes were enriched for genes involved in cell-type specific pathways such as *immune effector process* and *regulation of immune response*. However, several terms distinguished the three divergent region classes, such as *type I interferon signaling* for the *trans only* regions and *chromatin silencing* for the *cis only* regions (Figure 23A). Conserved active regions were enriched for nearby genes involved in housekeeping pathways, such as *RNA processing* and *translation*. Together, this indicates that genes involved in different functional pathways may be prone to different kinds of regulatory divergence.

Human-specific *cis only*, *trans only*, and *cis & trans* regions also displayed different patterns of histone modifications, including histone H3 lysine 27 acetylation (H3K27ac), histone H3 lysine 4 monomethylation (H3K4me1), and histone H3 lysine 4 trimethylation (H3K4me3) (Figure 22F, 23B). *Trans only* regions showed greater H3K4me1 signal and less H3K4me3 signal than the other classes, and this is likely explained by the human-specific region class annotations, since the *trans only* class is more enriched for promoter-distal annotations than the *cis only* or *cis*

& *trans* classes (Figure 22E). We also observed a bimodal distribution of histone signal for *trans only* regions but not the others. This suggests that *trans only* elements are generally located within the center of the nucleosome free region (NFR), while the others are more common on the NFR periphery. To test this, we plotted the distance between region centers and the NFR center—the summit of the accessible chromatin peak (Figure 22G). We used GM12878 H3K27ac ChIP-seq signal to map the -1 and +1 nucleosomes (Figure 23C) and phyloP signal to identify the most conserved portion of the NFR (Figure 23D). As predicted, *trans only* regions are more often at the center of the NFR, while the *cis only* and *cis & trans* regions are more frequently located at the edges of the NFR. This means that *trans only* changes are more likely to occur at the center of NFRs, where there is stronger evolutionary constraint. Thus, evolutionary constraint at NFR centers may prevent *cis* changes, so *trans* changes could be required to drive differential activity of these elements.

TF binding differences likely drive activity differences between *cis*, *tran*s, and *cis* & *trans* region classes. TF motif enrichment analysis revealed distinct TF motifs that distinguish regulatory regions both by the mechanism of gene regulatory divergence and species-specificity (Figure 3H). For example, human-specific *trans only* regions are enriched for IRF family TFs while macaque-specific *trans only* regions are enriched for the ATF4 TF, among others. Furthermore, IRF TFs are not enriched in human-specific *cis* & *trans* regions, suggesting the TFs that drive *trans* divergence for *trans only* regions are different from those that drive the *cis* & *trans* regions.

Key immune-related transcriptional regulators are differentially expressed between human and macaque LCLs

Trans regulatory divergence results from differences in the cellular environment, including differences in gene expression. To explore the mechanisms underlying the striking number of *trans* divergent regions (10,611 *trans only* and *cis* & *trans* combined), we performed RNA sequencing (RNA-seq) on both GM12878 and LCL8664 cell lines. The human and macaque LCL expression profiles cluster together and away from other human and macaque tissues (Figure 24A). Both LCLs also cluster closely with expression profiles from bulk, naïve, and memory B cells to the exclusion of other hematopoietic lineages (Figure 24B), suggesting they are transcriptionally similar to one another and to primary B cells.(Calderon et al., 2019) We also confirmed that waiting 24 hours after transfection to collect data resulted in minimal, if any, detection of plasmid-induced interferon-stimulated gene expression (Figure 24C-E). Thus, the human and macaque LCLs closely reflect primary B cells, and their transcriptional differences are likely the result of regulatory divergence between human and macaque.

We identified 2,975 differentially expressed genes with 1,505 upregulated in human and 1,470 upregulated in macaque (Figure 25A; human-specific $\log_2$(fold-change) > 2; macaque-specific $\log_2$(fold-change) < -2; both $p_{adj} < 0.001$). The human-specific genes were enriched for immune pathways, like *interferon signaling* and *interleukin-10 signaling*; while macaque-specific genes were enriched for extracellular matrix pathways, like *collagen formation* (Figure 4B). This indicates that, although these cell lines have broadly similar expression profiles (Spearman's ρ = 0.85; Figure 24F), they display specific expression differences that could drive the *trans*-regulatory environment effects we observe. Moreover, these gene expression differences are likely due to species differences, and not cell line immortalization (Figure 24B) or plasmid-induced interferon-stimulated gene expression (Figure 24C-E) artifacts.

Trans only regions are bound by differentially expressed TFs

The differential enrichment of IRF family motifs in human-specific *trans only* regions (Figure 22H) as well as the enrichment of interferon signaling genes in human-specific differentially expressed genes (Figure 25B) suggests a potential link between these differentially expressed TFs and the observed *trans*-divergent regions. To explore this hypothesis, we used TF footprints determined from ATAC-STARR-seq (Figure 19C) to test for TF footprint enrichment in the human-specific *trans only* and macaque-specific *trans only* regions. Indeed, we identified many TFs that are both significantly differentially expressed and enriched for binding in species-specific *trans only* regions; we define these TFs as "putative *trans* regulators" (Figure 25C, 24G). These putative *trans* regulators include several members of the IRF family (IRF4/7/8) that are markedly upregulated in human compared to macaque cells and are enriched for footprints in human-specific *trans only* regions (Figure 25C,D). Moreover, 18.7% of human-specific *trans only* regions were found to contain a TF footprint for one of these IRF family members that are canonically involved in innate immune responses (Fitzgerald & Kagan, 2020) (Figure 4D).

In total, the putative *trans* regulators we identified bind 37.1% of human specific *trans only* regions and 11.5% of macaque specific *trans only* regions. This highlights how changes to the expression of a few TFs can affect activity at a substantial number of the divergent DNA regulatory elements in a cell (Figure 25D,24H). The remaining *trans only* regions may be explained by TFs that did not meet our *putative trans regulator* criteria, which included stringent significance thresholds and a 1:1 ortholog requirement in the comparative RNA-seq workflow. It is also likely that other mechanisms contribute to differences in the *trans*-regulatory environment, such as previously described species-specific differences in post-transcriptional and post-translational regulation of TFs (Lin et al., 2010; Mittleman et al., 2021). Notwithstanding, this data argues that the differential expression of only a handful of transcription factors drives a substantial amount of the *trans*-regulatory divergence observed.

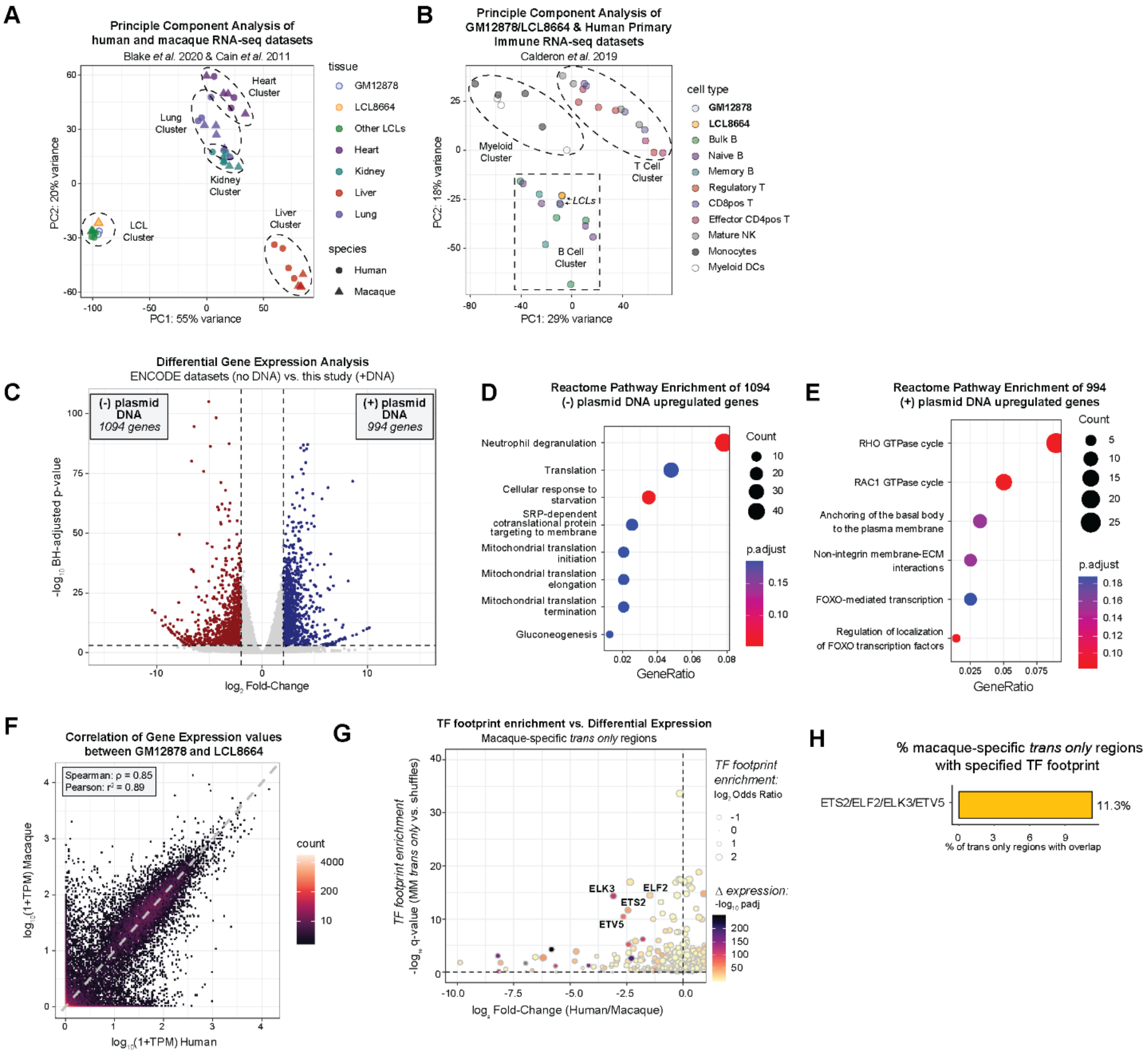

Figure 24: GM12878 and LCL8664 cells are transcriptionally similar to each other and primary B cells. (A) Principal component analysis (PCA) comparing our data with publicly available human and macaque RNA-seq datasets for heart, liver, lung, kidney and LCL tissue types. (B) PCA of our data with publicly available human primary immune cell RNA-seq datasets. (C) Volcano plot of differential expression analysis between GM12878 RNA-seq datasets with and without transfection of plasmid DNA 24hrs before collection; without plasmid DNA samples are from ENCODE. Point color represents genes more expressed in the with-plasmid condition (blue) or without-plasmid condition (red). Thresholds were $\log_2$ fold-change > | 2 | and padj < 0.001. (D-E) Reactome pathway enrichment of differentially expressed gene sets, either (D) without DNA enriched or (E) with DNA enriched. (F) Correlation plot of $\log_{10}$ transformed transcript per million (TPM) values for orthologous genes between GM12878 and LCL8664 cell lines. A pseudo count of 1 was added to TPM before log transforming. Correlation values were calculated on the untransformed TPM counts. (G-H) Macaque versions of Figure 25C-D. (G) Enrichment of macaque-specific *trans* only regions for TF footprints stratified by the differential expression of the TF. Text is only shown for the most differentially expressed and enriched TFs. (H) Percentage of macaque-specific *trans* only regions that overlap a given footprint. TFs within the same motif archetype were merged before determining the number of overlaps.

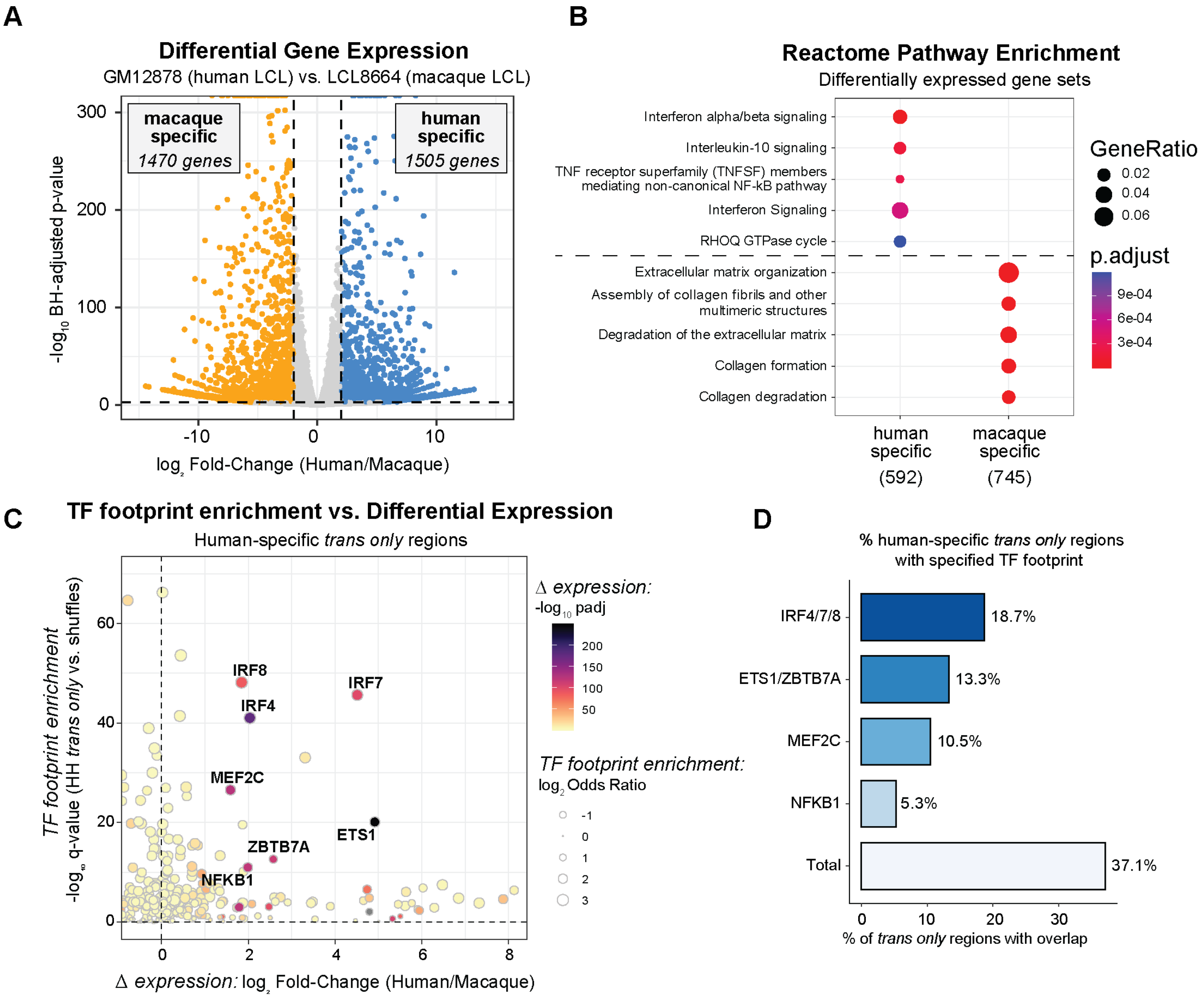

Figure 25: *Trans only* regions are bound by differentially expressed TFs. (A) Volcano plot of differential expression analysis between GM12878 (human) and LCL8664 (macaque) cell lines. Point color represents genes upregulated in human (blue) or macaque (orange). Thresholds were $\log_2$ fold-change > | 2 | and padj < 0.001. (B) Enrichments of differentially expressed gene sets for Reactome pathways. Only the top 5 terms in each were plotted. (C) Enrichment of human-specific *trans only* regions for TF footprints stratified by the differential expression of the TF. Text is only shown for the most differentially expressed and enriched TFs. See Figure 24G for macaque *trans only* results. (D) Percentage of human-specific *trans only* regions that overlap a given footprint. TFs within the same motif archetype were merged before determining the number of overlaps. See Figure S4H for macaque *trans only* results.

Trans **only regions are more conserved than *cis* only regions**

Because *trans* changes result from differences in the cellular environment, while *cis* changes result from functional sequence differences, we hypothesized that DNA sequences in *trans only* regions would be more conserved than sequences in *cis only* regions. Supporting this hypothesis, both *trans only* and *cis only* regions are enriched for primate PhastCons conserved elements compared to expected background distributions (p = 1.4e-11 and 9.1e-4, respectively), but *trans only* regions are more enriched than *cis only* regions (Figure 26A; *trans only* odds ratio (OR) = 1.5; *cis only* OR = 1.2). In contrast, *cis* & *trans* regions are significantly depleted of conserved elements (Figure 26A; OR = 0.67, p = 1.1e-30). As expected, regulatory sequences with conserved activity between human and macaque had the strongest enrichment for conserved elements (Figure 27A; p = 8.1e-157, OR = 3.1).

Accelerated substitution rates compared to neutral expectations can indicate shifts in sequence constraint, possibly resulting from positive selection (Capra, Erwin, McKinsey, Rubenstein, & Pollard, 2013; Hubisz & Pollard, 2014; Pollard, Hubisz, Rosenbloom, & Siepel, 2010). Both *cis only* and *trans only* elements are significantly enriched for elements with higher-than-expected substitution rates (Figure 26B; 27B; *cis only* p=4.9e-3; *trans only* p=4.7e-2), but as expected from their sequence-based mechanism of divergence, *cis only* regions are more enriched than *trans only* regions (*cis only* OR=1.4; *trans only* OR=1.3). *Cis* & *trans* elements showed no significant difference in substitution rates compared to background expectation (p=0.3). Overall sequence identity was similar across *cis/trans* groups, ruling out the possibility of systematic differences in the substitution rates of these regions underlying activity differences (Figure 27C).

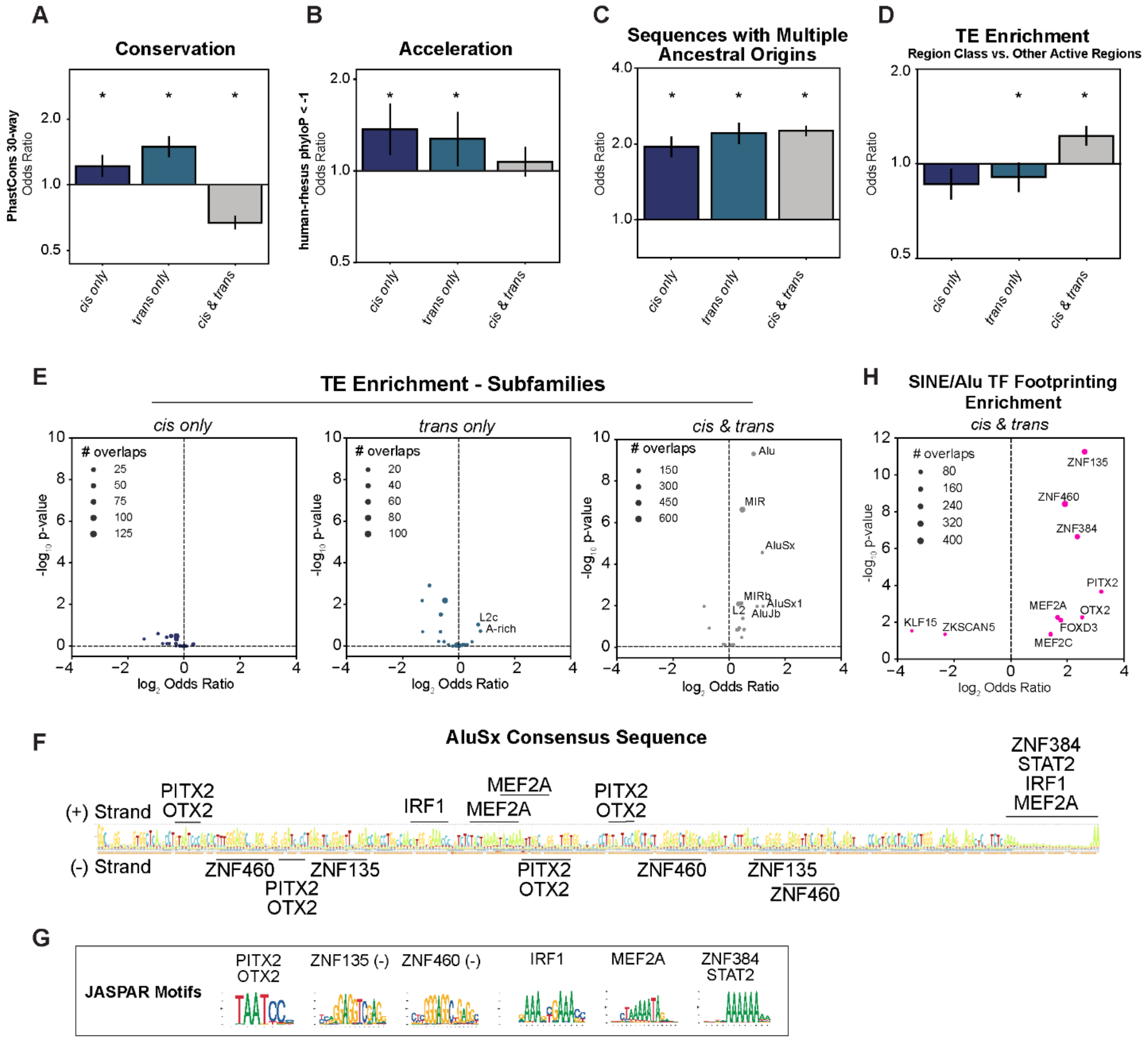

Figure 26: *Cis only*, *trans only*, and *cis & trans* regions have different degrees of conservation, acceleration, and transposable element enrichment. (A-C) Enrichments of *cis only*, *trans only*, and *cis & trans* regions for (A) 30-way PhastCons elements, (B) human accelerated elements (defined as human-rhesus PhyloP < -1), and (C) sequences with multiple ancestral origins compared to an expected background. (D) Enrichment of divergent regions for transposable element (TE) overlap compared to other active regions. For all bar charts, the Fisher's Exact Test odds ratio (OR) is plotted with 95% confidence intervals, which were estimated from 10,000 bootstraps. Windows were log2-scaled. Asterisks indicate a 5% FDR p-value < 0.05. (E) Enrichments of *cis only*, *trans only*, and *cis & trans* regions for subfamilies of TEs compared to an expected background. (F) The AluSx consensus sequence with TF binding sites for the TFs with enriched footprints. (G) Jaspar motifs of the relevant TFs. (H) Enrichments of SINE/Alu overlapping *cis & trans* regions for human TF footprints compared to an expected background. For the scatter plots, text is only shown for the most enriched subfamilies/TFs and point size represents the number of overlaps observed.

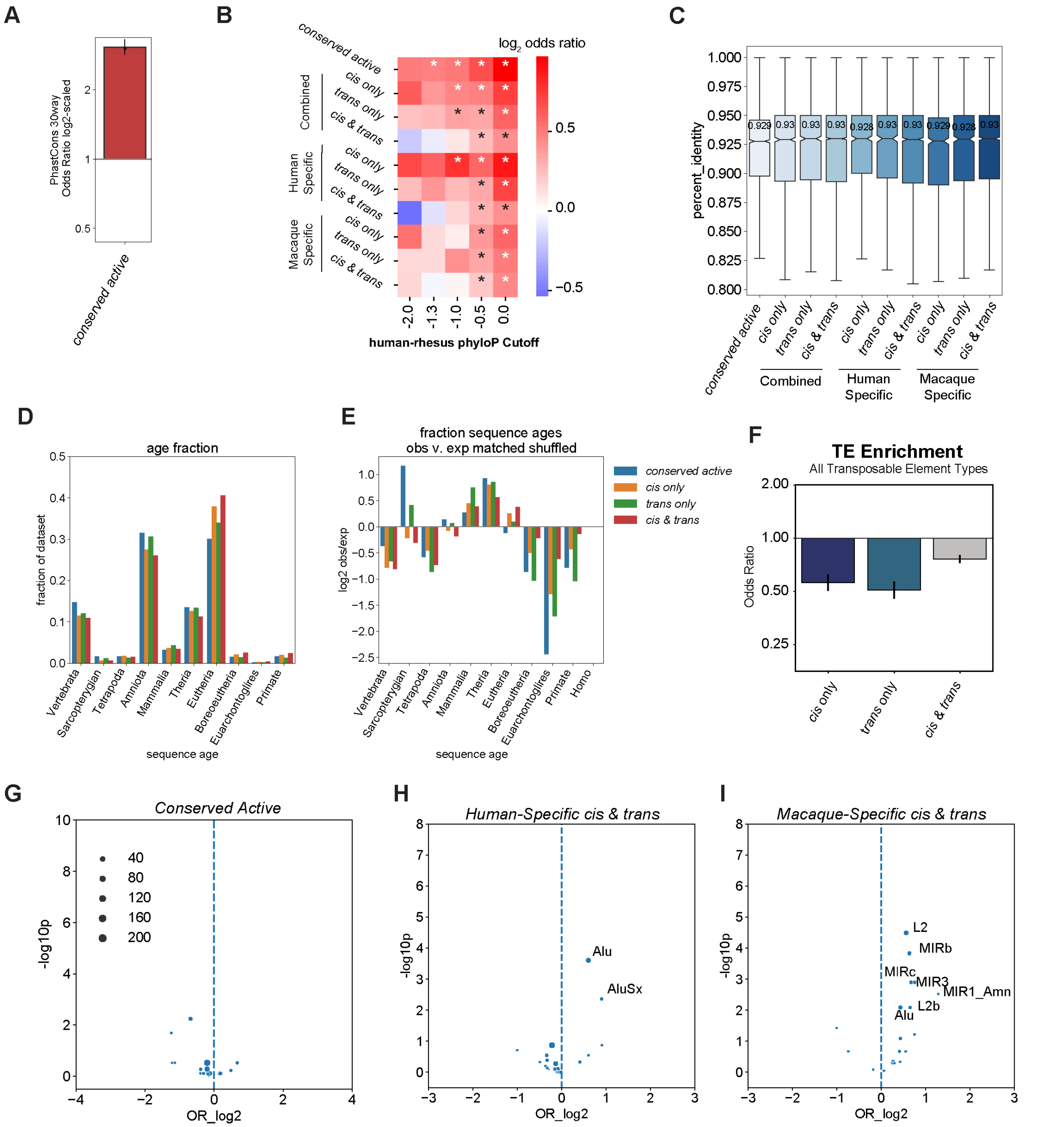

Figure 27: Additional evolutionary analysis of *cis* only, *trans* only, *cis* & *trans* and conserved active regions. (A) Enrichments of *conserved active* regions for 30-way PhastCons elements. For the bar chart, the Fisher's Exact Test odds ratio is plotted with 95% confidence intervals, which were estimated from 10,000 bootstraps. Windows were log2-scaled. Asterisks indicate p-value < 0.05. (B) Enrichments of *cis* only, *trans* only, *cis* & *trans*, and *conserved active* regions for human accelerated elements for multiple human-rhesus PhyloP thresholds. (C) Boxplots of the percent sequence identity for each region. (D) Fraction of each region set assigned to a given sequence age. (E) The observed vs. expected values of each region set for a given sequence age. (F) Enrichments of *cis only*, *trans only*, and *cis & trans* regions for all transposable elements (TEs) compared to an expected background. The Fisher's Exact Test odds ratio (OR) is plotted with 95% confidence intervals, which were estimated from 10,000 bootstraps. Windows were log2-scaled. (G) Enrichments of conserved active regions for subfamilies of TEs compared to an expected background. (H-I) Enrichments of (H) human-specific *cis & trans* regions and (I) macaque-specific *cis & trans* regions for subfamilies of TEs compared to an expected background.

Next, we investigated evolutionary origins of the regions in the divergent classes (Fong & Capra, 2021, 2022). All region sets are enriched for ancient sequences—from the placental common ancestor and older—so it is unlikely that differences in conservation are due to differences in sequence age (Figure 27D-E). Each region set is enriched for sequences with multiple ancestral origins, and *cis & trans* regions are the most significantly enriched (Figure 26C; conserved active p =3.6e-27; *cis only* p =7.9e-43; *trans only* p = 1.3e-56; *cis* & *trans* p = 4.6e-233).

Altogether, *cis only* and *trans only* regions both exhibit extremes of sequence conservation, divergence, and origin, as expected for sets of functional sequences in which some are experiencing negative selection and others positive selection. However, the sequences with *cis only* changes have more evidence of high substitution rates while *trans only* sequences are more enriched for conservation. This is consistent with their respective modes of divergence—sequence vs. cell environment. The fact that elements with *cis & trans* changes show substantially less evidence for selection suggests that they may arise from alternative mechanisms and have different functional roles.

Cis & trans regions are enriched for SINE/Alu transposable elements

Transposable element-derived sequence (TEDS) insertions are a source of raw sequence that often develops novel, species-specific regulatory functions (Chuong, Elde, & Feschotte, 2016; Chuong, Rumi, Soares, & Baker, 2013; Elbarbary, Lucas, & Maquat, 2016; Lynch et al., 2015; Trizzino et al., 2017). Thus, we investigated whether TEDS contribute to the divergent regulatory region classes, specifically in the less-conserved *cis & trans* elements. Overall, each class is depleted of TEDSs compared with genome-wide expectation (Figure 27F), consistent with previous findings that all gene regulatory sequences are depleted of TEDS (Fong & Capra, 2021; Simonti, Pavlicev, & Capra, 2017). However, comparing within the regulatory element classes, *cis & trans* regions were enriched for TEDS compared to the other categories (Figure 5D; *cis & trans* OR = 1.14, p =9.7e-4; *trans* only OR = 0.86, p= 0.02; *cis* only OR = 0.91, p=0.08) suggesting that *cis & trans* elements more frequently originate from TEDS. Several TEDS families were uniquely enriched in *cis & trans* regions, most notably SINE/Alu and MIR derived sequences (Figure 26E, 27G-I). Additionally, SINE/Alu elements were more enriched in human-specific *cis & trans* regions compared to macaque-specific *cis & trans* regions (Figure 27H-I), suggesting that SINE/Alu derived sequence activity is more favorable in the human cellular environment.

SINE/Alu elements are a common source for new DNA regulatory elements (Su, Han, Boyd-Kirkup, Yu, & Han, 2014; Sundaram et al., 2014; Sundaram & Wysocka, 2020). These sequences might have provided *proto-enhancers* in the last common ancestor of humans and rhesus macaques, developing over time into species-specific regulatory elements that experienced both *cis & trans* changes to obtain activity. The consensus AluSx sequence contains several sequences with high similarity to known TF binding sites (Figure 26F,G). Furthermore, TF footprinting analysis of *cis & trans* SINE/Alu elements (Figure 26H) provides strong evidence for the presence of TF binding, including the zinc-finger TFs, ZNF135, ZNF460, ZNF384, and PITX2, FOXD2, OTX2, RARG, and MEF2A. This demonstrates *cis & trans* regions are enriched for sequences derived from SINE/Alu elements and identifies several TFs that likely contributed to species-specific regulatory divergence.

Cis only regions are enriched for human variants associated with gene expression

Next, we explored the effects of genetic variation within human populations in the different regulatory divergence classes. First, we quantified enrichment for expression quantitative trait loci (eQTL) in regions with divergent activity, hypothesizing that variation in *cis only* and *cis & trans* regions would be more likely to associate with variable gene expression within humans.

Cis only elements were significantly enriched for *cis*-eQTLs in EBV-transformed B cells from the GTEx consortium, while the other classes were not enriched for *cis*-eQTLs (Figure 28A; 1.6x fold-change, empirical p-value = 1e-4). Focusing on human-specific active elements, the difference between *cis only* and *trans only* regions is even more extreme (Figure 28A inset). This suggests that regulatory elements that experienced sequence-based evolutionary divergence between human and macaques are more likely to harbor variants that modulate gene expression among humans, while *trans only* regions are less likely to tolerate functional variants.

We also evaluated enrichment for human genome-wide association study (GWAS) variants in divergent region classes. We selected immune and inflammatory traits from the UK Biobank (UKBB) where heritability had previously been observed in B cell gene regulatory loci (Calderon et al., 2019). After removing HLA-overlapping peaks, we observed modest enrichment in all region classes for GWAS variants across 17 inflammatory and autoimmune traits with few differences between the classes (Figure 29A,B; empirical p-value <0.05).

We were particularly interested to explore variants associated with viral hepatitis C, because humans and chimpanzees, but not macaques or other Old-World Monkeys, are susceptible (Sandmann & Ploss, 2013). Human-specific *trans only* regions are significantly and specifically enriched for viral hepatitis C GWAS variants, while macaque-specific regions are not (Figure 29B). This suggests that *trans*-regulatory changes contributed to the ape-specific susceptibility to hepatitis C and that human genetic variants in the regions bound by these *trans* factors modulate susceptibility to infection.

A human accelerated cis only element regulates NLRP1 expression and downstream trans changes

Our approach can identify the causes of evolutionary divergence at regulatory elements and quantify the resulting phenotypic outcomes at both the molecular and organismal levels. To illustrate this, we analyzed a GTEx *cis*-eQTL (rs1805264) associated with *NLRP1*, *MIS12*, *SCIMP*, *RABEP1*, *RPAIN*, *DERL2* expression variation across multiple tissues (Figure 28B, 29C) (Consortium, 2020). This locus overlaps a *cis only* region on chromosome 17 in the *MIS12* promoter that shows accelerated evolution between human and macaque (99th percentile of human acceleration scores; phyloP = -2.89) suggesting the locus experienced positive selection (Figure 28C,D). To understand how variation in this *cis only* region evolved to produce human-specific regulation, we evaluated differential TF footprinting between the human and rhesus macaque homologs. Human substitutions influenced binding site affinities for ZFX, ZNF460, NR2C2, EGR1, NRF1, and KLF15 transcription factors, which exclusively bind in human LCLs, as evidenced by differential footprinting (Figure 28E). Together, this indicates that human substitutions at this element created human-specific TF binding sites and human-specific *cis only* regulatory activity. This human-specific regulatory activity is then modulated by the *cis*-eQTL.

Of the genes influenced by genetic variation in this locus, *NLRP1* shows the highest human-specific differential expression between the two LCLs (Figure 28F). NLRP1 is a viral sensor, including for SARS-CoV-2 (Planes et al., 2022), and a core component of the pro-inflammatory signaling pathway. Thus, we hypothesize that variable *NLRP1* expression may have substantial downstream effects on pro-inflammatory signaling that affects the *trans*-regulatory cellular environment (Bauernfried & Hornung, 2022; Bauernfried, Scherr, Pichlmair, Duderstadt, & Hornung, 2021; Chavarria-Smith, Mitchell, Ho, Daugherty, & Vance, 2016; Fenini, Karakaya, Hennig, Di Filippo, & Beer, 2020). Indeed, the eQTL (rs1805264) is associated with human immune traits including higher platelet count and lymphocyte blood counts (Figure 28G). Together, this locus provides a key example of how a positively selected *cis only* region can affect expression of a target gene with potential to create substantial *trans* changes downstream, and, in turn, influence human-specific trait variation.

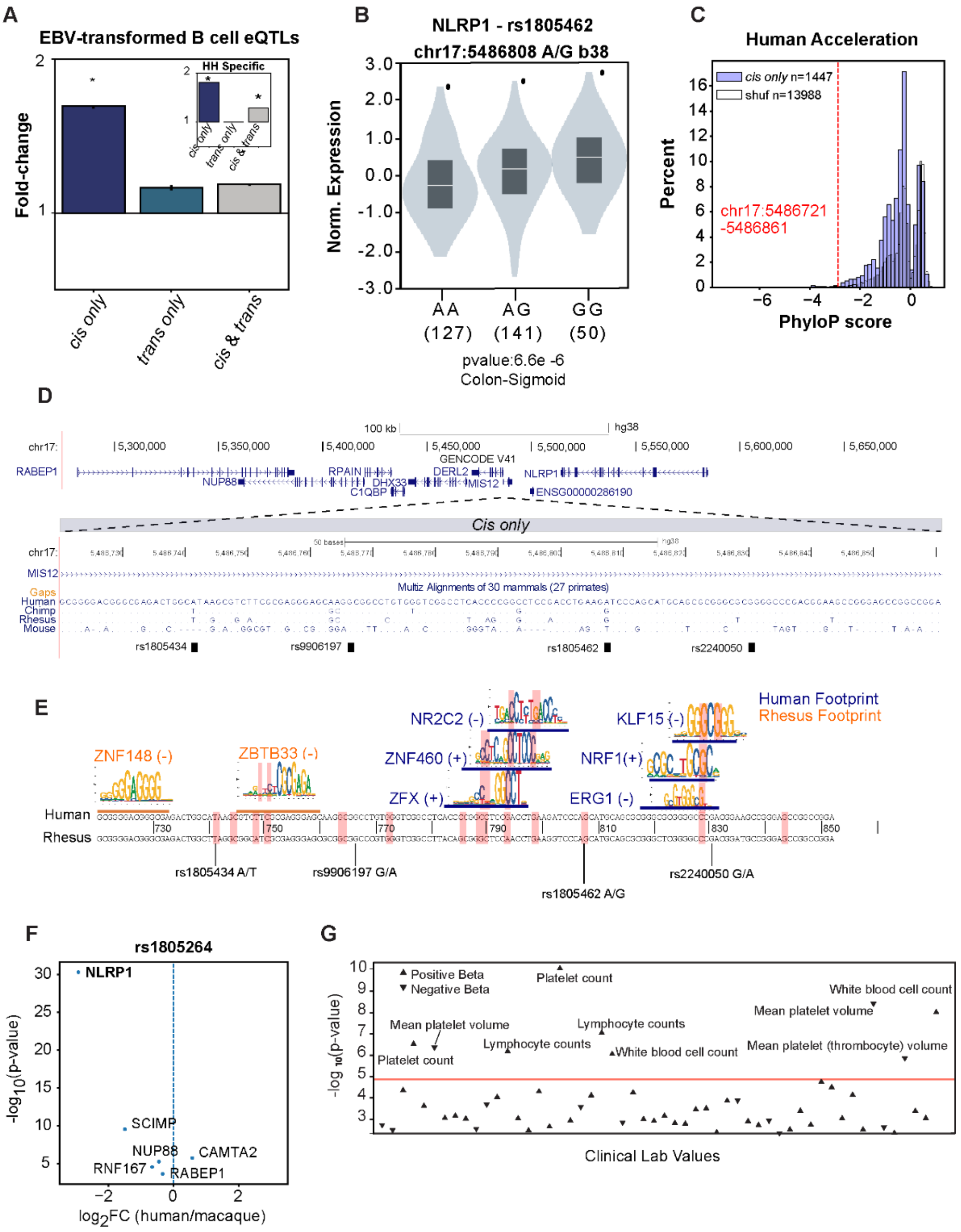

Figure 28: A human accelerated *cis only* element regulates *NLRP1* expression. (A) Enrichments of *cis only*, *trans only*, and *cis & trans* regions for EBV-transformed B cell eQTLs. The median fold-change compared to the expected background is plotted with 95% confidence intervals, which were estimated from 10,000 bootstraps. The inset in represents EBV-transformed B cell eQTLs enrichments for human-specific *cis only*, *trans only*, *cis & trans* regions. (B) Normalized expression scores of *NLRP1* for the three possible genotypes of rs1805264. (C) PhyloP score distribution for *cis only* and expected shuffled regions compared to the PhyloP score of the chr17: 5,486,721-5,486,861 locus (red dotted line). (D) Genomic locus on Chr17 with a zoomed-in view of a multi-way sequence alignment for a highly accelerated human-specific *cis only* element. (E) Differential TF footprints between human and macaque coincide with human-accelerated substitutions. (F) Differential expression of rs1805462-associated eQTL genes between human and macaque LCLs. (G) PheWAS associations for rs1805462 with variation in quantitative blood traits.

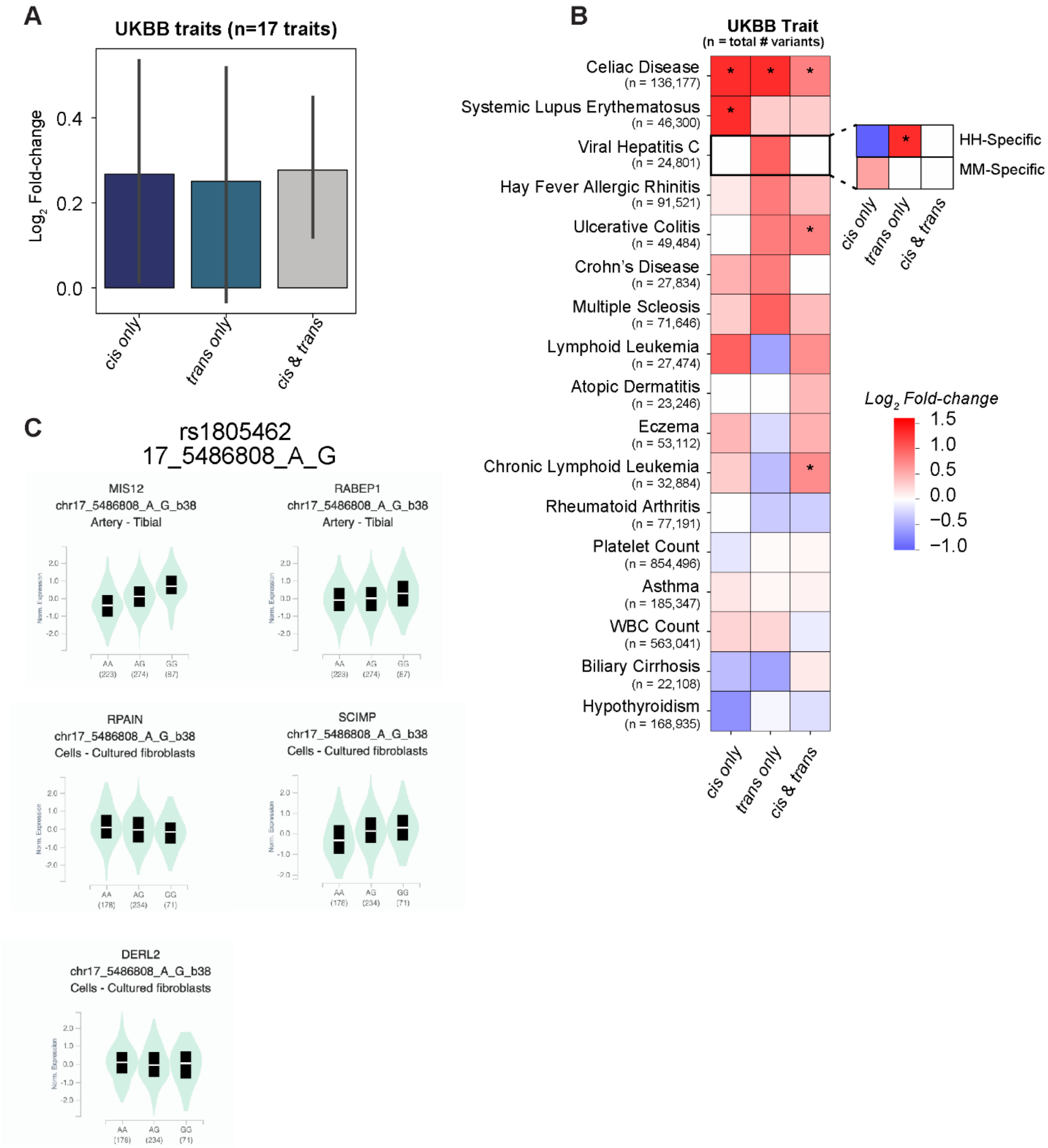

Figure 29: *Cis* only, *trans* only, and *cis* & *trans* regions are similarly enriched for genetic variation associated with UKBB traits (A) Enrichments of *cis* only, *trans* only, and *cis* & *trans* regions for 17 UK biobank traits compared to an expected background. The median fold-change is plotted with 95% confidence intervals, which were estimated from 10,000 bootstraps. (B) Heatmap of *cis* only, *trans* only, and *cis* & *trans* enrichment scores for each of the 17 UK biobank traits. The scores for the human-specific and macaque-specific groups are displayed for Viral Hepatitis C. Asterisk represents p-value < 0.05. (C) Versions of Figure 6B for all other associated genes.

A single substitution may drive differential expression of ETS1 by perturbing RUNX3 binding in macaques

We demonstrate that differential expression of a small number of TFs can explain a substantial portion of the human-specific *trans only* regions observed (Figure 25), and that *cis only* regions can be a potent source of gene expression variation (Figure 28). These observations suggest that a small number of *cis* changes may ultimately lead to substantial *trans* changes if they act on genes, like TFs, that alter the cellular environment (Hill et al., 2020; Signor & Nuzhdin, 2018). To illustrate the ability of our approach to enable inference of these regulatory cascades, we identified a human-specific *cis only* region at a putative enhancer for *ETS1*, a *trans* regulator that is substantially more expressed in human LCLs and binds to >13% of human-specific *trans only* regions (Figure 25C,E and Figure 30A-C). The activity of this putative enhancer is supported by GM12878 H3K27ac signal and human B cell DNA hypomethylation (Hodges et al., 2011; Moore et al., 2020). Furthermore, *ETS1* is the closet gene to the DNA regulatory element and is contained within the same topologically associated domain (TAD) according to GM12878 Hi-C data (Figure 30C) (Y. Wang et al., 2018), so *ETS1* is the likely target gene. Within this human-specific *cis only* region, we identified a macaque-specific substitution (T→C) that disrupts a RUNX3 motif, which is corroborated by the presence of a RUNX3 footprint detected in human but not macaque (Figure 30A). A GM12878 RUNX3 ChIP-seq peak also supports human TF binding at this locus (Moore et al., 2020). Furthermore, the functional relevance of this element is supported by two nearby SNPs, rs4262739 and rs4245080, which are eQTLs for *ETS1* and have been associated with human trait variation including lymphocyte percentage (Mountjoy et al., 2021; Vuckovic et al., 2020). The *ETS1* enhancer provides a powerful example of how a nucleotide substitution impacting the function of a single regulatory element leads to widespread changes in the activity of hundreds of regulatory elements across the genome. Altogether, these examples lead us to a model of how individual *cis* changes can ultimately generate substantial *trans*-divergent regulatory activity between species (Figure 30D).

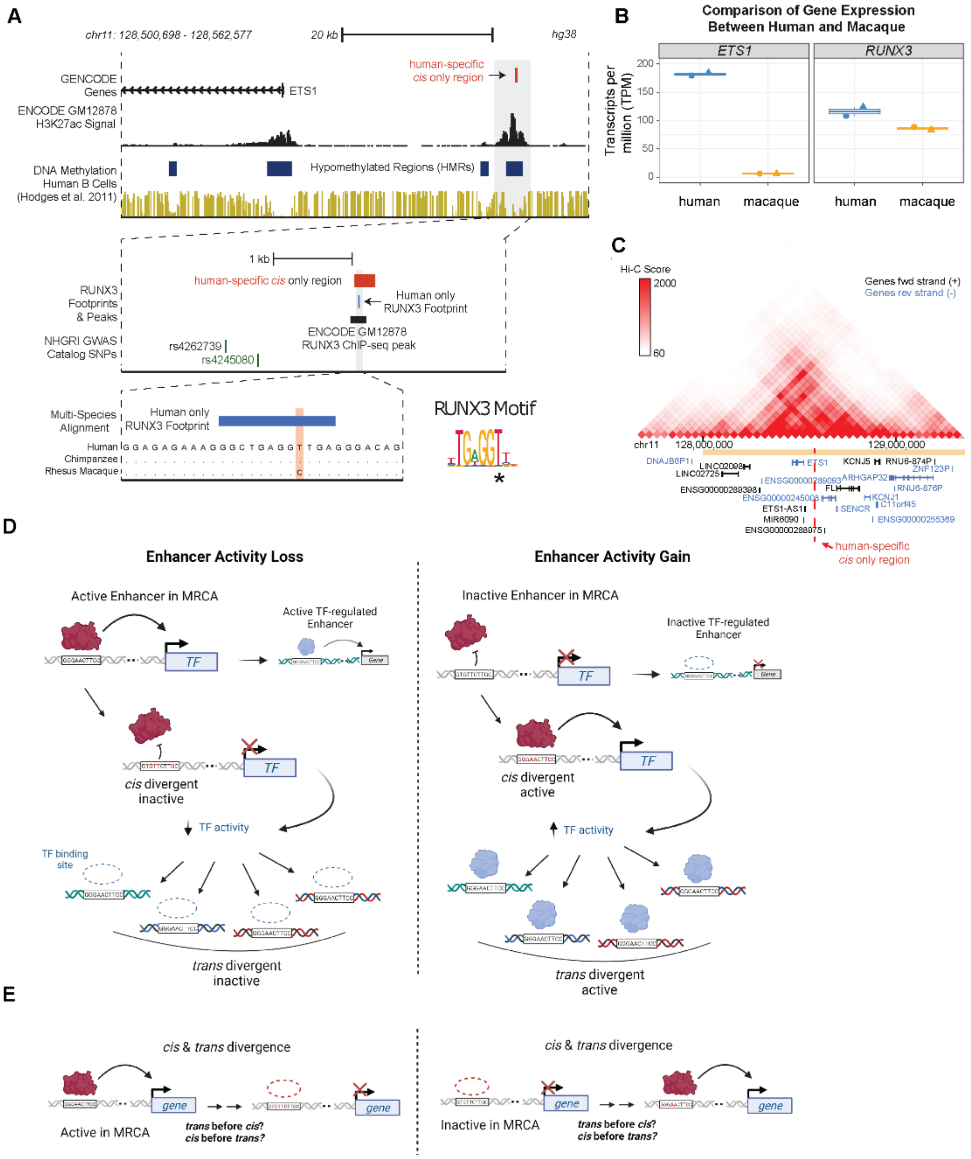

Figure 30: A single substitution may drive differential expression of *ETS1* by perturbing RUNX3 binding in macaques. (A) Genomic locus of a human-specific *cis* only regions within a putative *ETS1* enhancer. Public tracks for GM12878 H3K27ac and Human B cell DNA methylation corroborate this region as a putative enhancer. The first zoomed-in view of the locus shows a RUNX3 footprint present in human cells but not macaque cells. Nearby SNPs, rs4262739 and rs4245080, are associated with human trait variation. A further zoomed-in view of the footprint with a multi-species sequence alignment between human, chimpanzee, and macaque to reveal a macaque-specific substitution that perturbs an important nucleotide of the RUNX3 binding motif. (B) *ETS1* and *RUNX3* transcript per million (TPM) values for each replicate in human and macaque cells. (C) Hi-C data browser view of the *ETS1* locus in GM12878 cells. Vertical dashed line represents the relative location of the putative *ETS1* enhancer. (D) Model of how *cis* changes can become *trans* changes for other loci via TF expression/activity changes. First, *cis* changes alter the DNA sequence of a regulatory element to alter the affinity of TFs to the locus. This causes either enhancer activity loss or gain, based on the ancestral activity state of the enhancer. Alteration of enhancer activity, in turn, modifies the expression of target genes. If the target gene is a transcriptional regulator, the *cis* change would, therefore, also alter the cellular environment and become a *trans* change for other regulatory regions. (E) Model of how regions divergent in both *cis* & *trans* jointly drive differential regulatory element activity.

Discussion

Here, we used a comparative ATAC-STARR-seq framework to directly identify differentially active DNA regulatory elements between human and rhesus macaque and to characterize their mechanisms of divergence—changes in *cis*, in *trans*, or in both *cis* & *trans*. We observe that *trans*-regulatory divergence is common, despite previous work suggesting that *cis* changes drive most gene regulatory divergence between species. Moreover, we find that most divergent elements have both *cis* and *trans* differences in activity, indicating that divergent gene regulatory elements are often shaped by changes in both the homologous DNA sequence and the cellular environment.

Cis only, trans only, and cis & trans region classes display unique characteristics

We identify three classes of regulatory elements based on their mode of divergence: *cis & trans*, *cis only*, and *trans only*. We discovered unique functional and evolutionary characteristics that define these region classes. In summary, *cis only* regions are more enriched for high substitution rates than *trans only* regions, while *trans only* regions are more enriched for evolutionarily conserved sequences, which is consistent with the fact that mutations within the regulatory regions are necessary for divergent activity in *cis*, but not in *trans*. In contrast, *cis & trans* regions show less sequence constraint, but are enriched for complex genomic rearrangements and transposable element derived sequences (SINE/Alu elements, in particular) compared to *cis only* and *trans only* regions, indicating that many arose from mutations to transposable element sequences that were present in the last common ancestor of humans and rhesus macaques. We also identified distinct TF motif enrichments for each region class, which highlights how differential activity, and its mode of divergence depends on unique TFs. Altogether our characterization of the divergent region classes provides insight into the relationship between mode of regulatory divergence and the gene regulatory networks they act on, which remains a key gap in the field (Hill et al., 2020).

Trans-regulatory divergence is more extensive than previously recognized

In this study, we discovered more *trans*-regulatory divergence than previously reported (Irene Gallego Romero & Lea, 2022; Gordon & Ruvinsky, 2012; Hill et al., 2020; Mattioli et al., 2020; Signor & Nuzhdin, 2018; Whalen et al., 2023). Several differences in study design, experimental system, and scale may explain this apparent discordance. First previous work largely focused on gene expression rather than regulatory element activity as the functional output. Second, many previous studies have not been able to directly test for *trans* changes, and thus assumed that elements without *cis* changes were driven by *trans* changes. Thus, they would miss a large number of elements with evidence of both types of change. Third, the two recent studies that did directly evaluate *cis* and *trans* changes on regulatory element activity focused on more limited, pre-selected sets of regions (Mattioli et al., 2020; Whalen et al., 2023). Whalen *et al.* reported that nearly all of 159 tested human accelerated regions (HARs) diverged in *cis*. This is concordant with our findings that many *cis* divergent elements have accelerated substitution rates and are more likely to have accelerated substitution rates than other elements. Furthermore, they focus on HARs, rare elements with extreme evolutionary pressures that do not represent most regulatory loci. Mattioli *et al.* compared human and mouse regulatory element homologs and discovered that more regions were divergent due to changes in *cis* (n=660) than changes in *trans* (n=293). The difference in the *cis:trans* ratio may be due to different sampling of the elements tested, but the longer evolutionary divergence between human and mouse compared to human and macaque may also contribute. As previously mentioned, *cis* changes have been proposed to increase with evolutionary divergence (Coolon et al., 2014; Hill et al., 2020; B. P. H. Metzger et al., 2017), so we would expect to detect more *cis* changes at further evolutionary distances. More work is needed to determine the modes of gene regulatory divergence over both longer and shorter evolutionary distances, as well as different cellular contexts.

Putative trans regulators drive a substantial amount of trans-regulatory divergence in our system

To identify potential drivers of the *trans* regulatory divergence we observe, we defined "putative *trans* regulators" as a TF class that both display expression differences between species and bind to *trans only* regions as determined by TF footprinting. This revealed that a small number of key immune regulators, including ETS1, drive a substantial fraction of the human *trans* divergence we observed. This suggests that the differential expression of only a handful of transcription factors can drive a substantial amount of the *trans*-regulatory divergence.

We further showed that one of the putative *trans* regulators, ETS1, is likely regulated by a human-specific *cis only* region and discovered a key substitution in macaques that perturbs a RUNX3 TF motif. This is evidence of how a single substitution might influence the differential activity of a whole network of gene regulatory elements and species-specific immune-related traits, like Hepatitis C susceptibility in humans but not rhesus macaques. Indeed, we observed that only the human-specific *trans only* regions were highly enriched for Viral Hepatitis C associated variants. Altogether, our data will enable further characterization of putative *trans* regulators and identification of specific loci like the ETS1 regulatory element that may contribute to human-specific phenotypes.

A model of how cis and trans changes jointly drive divergent regulatory element activity

Cis & *trans* divergent regions acquired a change in *cis* and a change in *trans* during their evolution from the most recent common ancestor (MRCA) between humans and rhesus macaques (Figure 30E). We speculate that perturbations in *trans* are often likely to occur prior to *cis.* Once the relevant *trans* factors no longer bind, some elements will accumulate enough sequence variation to result in *cis* changes as well. Several lines of evidence from previous reports and our study support this hypothesis. For example, *cis* changes have been proposed to accumulate with greater evolutionary divergence whereas *trans* changes are favored short-term (Coolon et al., 2014; Hill et al., 2020; B. P. H. Metzger et al., 2017). This is likely because *trans* changes can change many regulatory region activities at once but may be more deleterious than *cis* changes (Vande Zande et al., 2022). In this way, more significant phenotypic changes may be driven by changes to the *trans*-regulatory environment, but with a long-term fitness cost that can be ameliorated by local and precise *cis* changes to DNA regulatory elements.

Limitations of the Study

Several limitations of our study must be considered when interpreting our results. First, we only directly assay one genotype per species and infer evolutionary divergence from these models. While it would be ideal to evaluate additional genotypes for each species (Kelley & Gilad, 2020), this approach was necessary for several reasons. First, there are few non-human primate cell lines available to assay. Second, the comprehensive design of our comparative ATAC-STARR-seq approach is prohibitive for testing and interpreting activity variation across multiple genotypes and across multiple cellular environments.

Second, for experimental reasons, we leverage immortalized cell lines, whose cellular biology may not completely mirror the biology of primary B cells. The immortalization strategies differ for human and rhesus B cells. Specifically, the human B cell line was immortalized using Epstein-Barr Virus (EBV) (International HapMap, 2003; Tosato & Cohen, 2007); whereas the rhesus cell line was immortalized *in vivo* by a rhesus lymphocryptovirus (rhLCV) related to EBV—so-called Rhesus Epstein-Barr Virus (RheEBV) (Cho, Gordadze, Ling, & Wang, 1999; Muhe & Wang, 2015; Rangan et al., 1986). Although the viral EBNA2 gene, which drives transcription of many gene targets in EBV-infected cells (Wu, Kalpana, Goff, & Schubach, 1996), is homologous between EBV and rhLCV, host-restriction and co-evolutionary pressures may exaggerate many of our results. We envision that this could be avoided in future studies by using primate induced Pluripotent Stem Cell (iPSC) lines (I. Gallego Romero et al., 2015). Beyond these possible confounders, our analysis of publicly available RNA-seq datasets shows that, at least transcriptionally, the two cell lines are highly similar both to each other and to human primary B cells (Figure 24A,B).

Despite the greater scale of the assay, ATAC-STARR-seq lacks the within-sample reproducibility of synthetic MPRA approaches that take dozens of measurements for each sequence assayed (Santiago-Algarra et al., 2017). For this reason, we cannot reliably compare effect sizes of activity. Instead, we binarize activity measures by applying significance thresholds to call active regions, which we then compare between conditions. Future analytical approaches may incorporate strategies that enable direct comparisons of activity. This would allow investigation of additional hypotheses, including proposed *cis/trans* compensation mechanisms on regulatory elements (Mattioli et al., 2020). In this way, we interpret *cis* & *trans* regions as individual regulatory regions where both species-specific DNA and species-specific environment are necessary to observe regulatory activity. We caution against interpreting compensatory or directional mechanisms on individual regulatory element activity from our data. However, while we did not explore how multiple regulatory elements control gene expression in a directional or compensatory fashion, this would be possible with our data, but validation studies that place gene regulatory elements in their endogenous context would be needed.

Concluding Remarks

We find that *trans* changes contribute to DNA regulatory element activity divergence between human and macaque nearly as often as *cis* changes. Moreover, we observed that both *cis* and *trans* changes affect most divergent regulatory elements. These findings enabled by our comparative ATAC-STARR-seq framework highlight an underappreciated role for the cellular environment in driving gene regulatory changes. We envision that our comparative strategy will be useful in future studies for mapping gene regulatory divergence between different species and across different cell types within the same species to agnostically determine the locations and roles of *cis* and *trans* divergence on gene regulatory function.

Materials & Methods

Experimental Model and Subject Details

Cell Lines

One human lymphoblastoid cell line (GM12878) and one rhesus macaque lymphoblastoid cell line (LCL8664) were used in this study (International HapMap, 2003; Rangan et al., 1986; Tosato & Cohen, 2007). GM12878 is female, while LCL8664 is male. GM12878 and LCL8664 were purchased directly from Coriell and ATCC (CRL-1805), respectively. We cultured both cell lines with RPMI 1640 Media containing 15% fetal bovine serum, 2mM GlutaMAX, 100 units/mL penicillin and 100 µg/mL streptomycin. Cells were cultured at 37°C, 80% relative humidity, and 5% CO_2. Cell density was maintained between 0.2×10^6 and 1.5×10^6 cells/mL with a 50% media change every 2-4 days. All cell lines were regularly screened for mycoplasma contamination.

ATAC-STARR-seq

We performed four ATAC-STARR-seq experiments following the method as described in Hansen & Hodges 2022 (Hansen & Hodges, 2022a). We created two ATAC-STARR-seq plasmid libraries, one for the GM12878 accessible genome and another for the LCL8664 accessible genome. For a total of four experiments, we electroporated each ATAC-STARR-seq plasmid library into both GM12878 and LCL8664 cells, resulting in the following conditions: GM12878 Library in GM12878 Cells (referred to as HH in text), GM12878 Library in LCL8664 Cells (HM), LCL8664 Library in GM12878 Cells (MH), and LCL8664 Library in LCL8664 Cells (MM). For HH and MH, we used Buffer R, whereas, for HM and MM, we used Buffer T from the Neon™ Transfection System 100 µL Kit (Invitrogen, #MPK10025). Both plasmid DNA and reporter RNAs were harvested from the same flask of cells and processed into llumina sequencing libraries. We repeated the electroporation, harvest, and sequencing library preparation steps for a total for three replicates; replicates were performed on separate days. The plasmid DNA and reporter RNA sequencing libraries for each replicate of each condition was sequenced on an Illumina NovaSeq 6000 machine, PE150, at a requested read depth of 50 or 75 million reads, for DNA and RNA samples, respectively, through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core. The GM12878 Library in GM12878 Cells was previously analyzed (Hansen & Hodges, 2022a) but in a different manner (GEO accession: GSE181317).

RNA-sequencing

Before RNA isolation, we electroporated hSTARR-seq_ORI plasmid (Addgene #99296) into GM12878 and LCL8664 and matched the experimental conditions performed for the ATAC-STARR-seq plasmid library transfections, but on a smaller scale. Instead of twenty 100µL electroporation reactions, we performed a single 100µL reaction for each replicate and kept the cell count:DNA ratio ($3x10^6$ cells and 3µg plasmid DNA per reaction) and electroporation conditions the same. We performed two replicates each for GM12878 and LCL8664 cell lines.

24 hours later, we harvested total RNA using the TRIzol™ Reagent and Phasemaker™ Tubes Complete System (Invitrogen™, #A33251) and prepared Illumina-ready RNA-sequencing libraries using the SMARTer® Stranded Total RNA Sample Prep Kit - HI Mammalian (Takara Bio, #634874). Libraries were analyzed for quality and submitted for sequencing on an Illumina NovaSeq 6000 machine, PE150, at a requested read depth of 50 million reads through the Vanderbilt Technology for Advanced Genomics (VANTAGE) sequencing core.

Quantification and Statistical Analysis

ATAC-STARR-seq Read Processing

FASTQ files were trimmed and analyzed for quality with Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) using the --fastqc and --paired parameters. Trimmed reads were mapped to hg38 with bowtie2 using the following parameters: -X 500 --sensitive --no-discordant --no-mixed (Langmead & Salzberg, 2012). Mapped reads were filtered to remove reads with MAPQ < 30, reads mapping to mitochondrial DNA, and reads mapping to ENCODE blacklist regions using a variety of functions from the Samtools software package (H. Li et al., 2009). When desired, duplicates were removed with the *markDuplicates* function from Picard (https://broadinstitute.github.io/picard/). Read count was determined using the *flagstat* function from Samtools. Library complexity was measured using the *EstimateLibraryComplexity* function from Picard and plotted with ggplot2 in R (Wickham, 2016). Correlation plots were generated with the deepTools package (Ramirez et al., 2016). Read counts for 1kb genomic windows were compared between the filtered, with-duplicates bam files using the *multiBamSummary bins* function and the following parameters: -e and --binSize 1000. Plots were generated using the *plotCorrelation* function and the following parameters: --skipZeros --corMethod pearson.

Chromatin Accessibility Peak Calling and Filtering

Accessible chromatin (ChrAcc) peaks were called in all four conditions (GM12878inGM12878, LCL8664inLCL8664, GM12878inLCL8664, LCL8664inGM12878) using Genrich with the -j parameter, which specifies ATAC-seq mode (https://github.com/jsh58/Genrich). For each condition, de-duplicated bam files for the three plasmid DNA replicates were provided to the peak caller; as part of peak calling, Genrich collapses replicates to yield one peak set for the given condition and uses variance between replicates to assign q-values. Peaks were filtered by q-value so that the genomic coverage of the entire peak set for a given condition was ~1.8% (q-value thresholds ranged between 1.1e-7 and 4.3e-6). The purpose of filtering for genomic coverage of each peak set was to account for data quality differences between the samples. This allows us to compare the most accessible 1.8% of the respective genomes rather than regions defined by a significance threshold. We compared several different genome coverages but qualitatively determined 1.8% best reflected true accessible peaks when looking at read pileup in a genome browser. We subsequently removed XY chromosomes since LCL8664 is male and GM12878 is female. Together, this yielded between 58,000-63,000 peaks for each of the four experiments. Peaks called in rheMac10 coordinates (LCL8664inGM12878 and LCL8664inLCL8664) were converted to hg38 coordinates using liftOver with -minMatch set to 0.9.

Differential Accessibility Analysis

We intersected the filtered ChrAcc peaks from each experiment using the default parameters of BEDTools *intersect*(Quinlan & Hall, 2010) to isolate ChrAcc regions shared across all four contexts—this resulted in 29,531 shared ChrAcc peaks. To obtain specific-specific accessible regions, we intersected only the GM12878inGM12878 and LCL8664inLCL8664 ChrAcc peaksets and wrote non-overlaps using the -v parameter. We performed motif enrichment using the *findMotiftsGenome.pl* script from the HOMER package (http://homer.ucsd.edu/) (Duttke et al., 2019) using the following parameters: -size given -mset vertebrates. We used ChIPSeeker to annotate differential accessible regions based on their distance to the nearest TSS (annotatePeak, *level = gene & tssRegion = -2000/+1000*), assign nearest neighbor genes, and perform Reactome pathway enrichment analysis using the assigned genes (Jassal et al., 2020; Yu et al., 2015). For the annotation plotting, we removed the *Downstream (<=300)* term from the legend to simplify, since we did not observe assignments to that term.

Genome Browser

The respective genome browser tracks were viewed in the hg38 build using the UCSC genome browser (C. M. Lee et al., 2020) and a combination of custom and public tracks. PDFs of these views were downloaded and further annotated in illustrator; positions of the tracks did not change during illustrator editing.

Active Region Calling Within Shared Accessible Peaks

Generation of Sliding Window Bins

We first merged all four ChrAcc peak sets (hg38 coordinates) into a single file with the UNIX *cat* function followed by BEDTools *merge* to generate a merged set of all peaks. Since ChrAcc peaks contain both active and silencing regulatory elements, it is important to divide peaks into smaller windows to best identify the element driving activity (Hansen & Hodges, 2022a). To do this, we tiled the merged peak set with sliding windows usingBEDTools *makewindows* and the -s 10 -w 50 parameters; bins smaller than 50 bp were removed. This generated 7.65 million bins for analysis.

Filtering Bins for Alignability and Shared Accessibility

To perform comparative analyses between human and macaque genomes, we required that all bins were mappable between hg38 and rheMac10 in a 1:1 orthologous fashion and with at least 90% alignability. To do this, we used liftOver with -minMatch=0.9 to convert our bins from hg38 coordinates to rheMac10 and bins that did not map from hg38 to rheMac10 were removed from the hg38 file. Furthermore, bins that changed size by more than +/- 2bp in the liftOver were excluded from the analysis. Altogether, this removed ~552,000 bins (~7.3%).

Because differentially accessible regions would be only assayed in one ATAC-STARR-seq plasmid library, they would confound differential activity measures when comparing the respective genomes. For this reason, we also required that our bins overlap shared ChrAcc accessible peaks by intersecting the alignability-filtered bins with the 29,531 shared ChrAcc peaks described above; we used BEDTools *intersect* with the -u option set. This resulted in 2,028,304 (26.5%) sliding window bins for further analysis.

Active Region Calling

We called active regions for each of the four experimental conditions using the 2,028,304 filtered sliding window bins as input. To control against sample-to-sample variability, we called the top 10,000 most significantly active regulatory regions in each condition. By comparing the same number of DNA regulatory elements across conditions, we assume that a similar number of regions are active in each of the four experiments. This is a more conservative assumption than comparing regions called with the same q-value threshold across experiments, which can be greatly influenced by data quality differences and may not accurately reflect biology in a comparative analysis. We compared the results of calling different active region thresholds including the top 5,000, 10,000, 25,000, and 50,000.

To call active regulatory regions, we first assigned reads to the filtered sliding window bins using the *featureCounts* function from the Subread package with the following parameters: -p -B -O --minOverlap 1 (Liao et al., 2014); for rheMac10 mapping reads, we used bins in rheMac10 coordinates (linked to hg38 coordinates by a unique bin ID). To avoid negative data interpretations, we next removed bins with a count of zero for any RNA or DNA replicate; between 8,775 and 70,819 bins were removed in each condition. We then quantified the activity of each bin by comparing RNA and DNA counts using DESeq2 (fitType="local") (Love et al., 2014). To obtain the top 10,000 most significantly active regions in each condition, we adjusted Benjamini-Hochberg adjusted p-value thresholds to yield active bins that when merged in genomic space resulted in about 10,000 active regions for each condition–padj thresholds ranged between 0.026 and 0.11. To ensure our active regions were robust regulatory elements, we required that each region be made up of at least 5 bins by using BEDTools merge with the -c option and a custom awk script. For the supplemental analysis investigating threshold effects on *cis* and *trans* divergent regions calls, we followed the same process of adjusted padj thresholds to yield the desired active region count and then performed the same methods as described above to identify *cis* and *trans* divergent regions. We used ChIPSeeker to annotate the active regions in each condition based on their distance to the nearest TSS (annotatePeak, *level = gene & tssRegion = -2000/+1000*). For the annotation plotting, we removed the *Downstream (<=300)* term from the legend to simplify, since we did not observe any assignments to that term.

Generation of ATAC-STARR-seq Activity bigWigs

We generated ATAC-STARR-seq activity signal files with the deepTools package; to streamline this, we created a custom python script, which is available on the ATAC-STARR-seq method GitHub (github link; *generate_ATAC-STARR_bigwig.py*). We compared the $\log_2$ ratio of cpm-normalized RNA and cpm-normalized files using the *bigwigCompare* function and the following parameters: --operation log2 --pseudocount 1 –skipZeroOverZero; the cpm-normalized bedGraph files for RNA and DNA were generated using the *bamCoverage* function and the following parameters: -bs 10 --normalizeUsing CPM. MH and MM activity signal files were converted from bigwig to bedGraph (with the bigWigToBedGraph function from UCSC), lifted over to hg38 coordinates from rheMac10 coordinates with Crossmap (Zhao et al., 2014), and then converted back to bigwig files using the bedGraphToBigWig function from UCSC. We generated bigwigs for individual replicates, as well as for merged replicate bam files.

Heatmaps of ATAC-STARR-seq Activity at Active and Inactive Bins

We first subsampled the inactive bins for each condition using the Unix *shuf* command (-n 150000) to reduce the number of regions plotted. ATAC-STARR-seq activity signal files for each replicate were plotted at their respective active and randomly subsampled inactive bins using the *computeMatrix* function (parameters: -a 500 -b 500 --referencePoint center -bs 25 --missingDataAsZero) and the *plotHeatmap* function (parameters: --sortRegions no --zMin -0.5 --zMax 0.5), both from deepTools.

Differential Activity Analysis

HH vs MM Activity Comparison

To identify conserved and species-specific active regions, we intersected the HH active regions with the MM active regions using BEDTools *intersect*. We called regions with at least a 50% reciprocal overlap as conserved active regions, whereas HH active regions that did not reciprocally overlap by at least 50% were classified as human-specific active regions and MM active regions that did not reciprocally overlap by at least 50% were classified as macaque-specific active regions. For all intersections, we used the following parameters: -f 0.5 -F 0.5 -e. This turns the 50% reciprocal into an "or" operation where either regions A&B are considered conserved active if either A or B overlaps the other by greater than 50%. This avoids mislabeling nested overlaps as differentially active where A could overlap B with 100% but B could be two times larger than A and therefore not overlap A by 50%. For the conserved active regions, we wrote the entire interval of the two overlapping regions using a combination of BEDTools *intersect* and *merge* in a custom script. We used the -v option in addition to the parameters listed above to write differentially active.

Identification of Cis Divergent Regions and Trans Divergent Regions

We determined if divergent active regions were a result of a change in the DNA sequence (*cis*) or a change in the cellular environment (*trans*) by intersecting species-specific active regions with the active region set from the relevant condition. For example, human-specific *cis* divergent regions were determined by intersecting the human-specific active regions with the MH active region set using BEDTools intersect. Human-specific active regions that did not reciprocally overlap by at least 50% were determined to be Human-specific *cis* divergent regions (parameters: -v -f 0.5 -F 0.5 -e). The other comparisons were performed in the same way as described above.

Identification of Cis & Trans Regions

To identify regions that were divergent in both *cis* & *trans*, we asked if the exact same region was contained in both the *cis* and *trans* divergent region sets using BEDTools *intersect* and the -f 1.0 -r parameters; we maintained species-specificity by only comparing human-specific *cis* with human-specific *trans* and macaque-specific *cis* with macaque-specific *trans.* Regions that were unique to the *cis* region set were classified as *cis only*, while regions that were unique to the *trans* region set were classified as *trans only*.

Observed vs. Expected Analysis of Active Region Overlaps

We calculated the expected overlap assuming random distribution in shared accessible chromatin for all differential activity comparisons. To do this, we first randomly shuffled the MM, HM, and MH active region sets within shared accessible chromatin with BEDTools *shuffle* (1000 iterations with the -noOverlapping parameter). This yielded 1000 sets of randomly positioned active region sets for MM, HM, and MH within the analytical space of shared accessible chromatin. For each of the 1000 shuffled region sets per condition, we determined the expected number overlaps by intersecting them with either the HH active, the human-specific active, or the macaque-specific active regions using BEDTools *intersect* in the same manner done for the observed value. We then compared the expected overlap distribution with the observed value and performed Grubb’s Test in R to test if the observed value was a statistical outlier.

Heatmaps Comparing ATAC-STARR-seq Activity Between Conditions

ATAC-STARR-seq activity signal files were plotted at the respective regions using the *computeMatrix* function (parameters: -a 1000 -b 1000 --referencePoint center -bs 10 --missingDataAsZero) and the *plotHeatmap* function (parameters: --sortRegions no --zMin -0.5 --zMax 0.5), both from deepTools.

Functional Characterization of Cis and Trans Divergent Regions

Annotation

We used ChIPSeeker to annotate *cis only*, *trans only*, *cis* & *trans*, and conserved active regions based on their distance to the nearest TSS (annotatePeak, *level = gene & tssRegion = -2000/+1000*). For the annotation plotting, we removed the *Downstream (<=300)* term from the legend to simplify, since we did not observe assignments to that term.

TF Motif Enrichment

We first generated background regions for each region set by shuffling the respective regions within shared accessible chromatin 10 times using bedtools *shuffle* and the -chrom -noOverlapping -maxTries 5000 parameters. We then performed motif enrichment using the *findMotiftsGenome.pl* script from the HOMER package using the respective background and the -size given and -mset vertebrates parameters. The top 15 motifs for each region set were selected for plotting using pheatmap and the following parameters: scale="row", cluster_cols = FALSE, cluster_rows = TRUE, cutree_rows = 7, cellheight = 15, cellwidth = 30, method = "ward.D2". Motifs within the same motif archetype (Vierstra et al., 2020) were collapsed so that only one motif of that archetype was displayed on the heatmap in the main figure.

Gene Ontology

We performed gene ontology on the putative target genes for *cis* only, *trans* only, *cis* & *trans*, and conserved active regions using GREAT (McLean et al., 2010) (http://great.stanford.edu/public/html/). We used the whole genome as background and assigned genes with the default *Basal plus extension* option. The top 10 terms were plotted in R.

Histone Modification Heatmaps

GM12878 ChIP-seq bigwig files for H3K27ac (ENCFF469WVA), H3K4me3 (ENCFF564KBE), and H3K4me1 (ENCFF280PUF) were downloaded from the ENCODE consortium(Moore et al., 2020) and plotted at conserved active, human-specific *cis only*, human-specific *trans only*, and human-specific *cis* & *trans* regions with deepTools. Specifically, we used the *computeMatrix* function, with the following parameters: -a 2000 -b 2000 --referencePoint center -bs 10 –missingDataAsZero and the *plotHeatmap* function with the following key parameters: --sortUsing mean –sortUsingSamples 1 (the H3K27ac file).

Distance to ChrAcc Peak Summits

We first extracted region centers in R using the following operation: center = ((End-Start)/2)+start; decimals were rounded up to integers. The ChrAcc peak summits are provided in the original narrowPeak file for GM12878 ChrAcc peaks, so we obtained peak summits for the shared accessible peaks by intersecting shared peaks with the human-active peak file. The distance between region center and peak summit was calculated using the bedtools *closest* function and the -D ref parameter. This distance was then plotted as a density plot with ggplot2 in R.

To generate the H3K27ac profile plot, we plotted the GM12878 H3K27ac bigwig from ENCODE at ChrAcc peak summits using deepTools with the *computeMatrix* function (parameters: -a 500 -b 500 --referencePoint center -bs 10 –missingDataAsZero) and the *plotProfile* function.

We repeated for the 17-way PhyloP bigwig after downloading from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP17way/hg38.phyloP17way.bw).

Generating expected background datasets from shared accessible, inactive regions

We identified all shared accessible peaks from any of the four (HH, HM, MH, MM) experiments. We then used BEDTools to subtract active, shared accessible peaks, leaving a set of shared accessible, but inactive peaks. Then, we shuffled active regions with BEDTools (-noOverlapping -maxTries 5000) in this shared accessible, inactive genomic background 10x to produce length-matched expectation datasets. We used these elements as our background to interpret evolutionary and genomic features of active and divergent elements.

TF Footprinting

Transcription factor footprinting was performed using the TOBIAS software package (Bentsen et al., 2020). For both the GM12878inGM12878 and LCL8664inLCL8664 samples, we used *ATACorrect* to generate Tn5-bias corrected cut count signal files from deduplicated bam files. We then used the corrected cut-counts files to calculate TF binding in the respective genomes using the *ScoreBigWig* function. We then paired all core non-redundant vertebrate JASPAR motifs (Fornes et al., 2020) with the GM12878 and LCL8664 TF binding profiles to call individual transcription factor footprints in the two genomes using the *BINDetect* function and the --bound-pvalue parameter set to 0.05 . Motifs with a footprint were classified as bound, while motifs without a footprint were classified as unbound. Aggregate plots were generated using the deepTools package. Tn5-corrected signal was measured at bound and unbound sites for each respective TF using the computeMatrix reference-point function with the following key parameters: -a 75 -b 75 --referencePoint center --missingDataAsZero -bs 1. The resulting matrix was plotted using the plotProfile function.

To determine differential footprinting at specific loci, we compared the TF motifs that footprinted in human and rhesus. We mapped the position of rhesus TF footprints in hg38 by lifting those footprint coordinates from rheMac10 using LiftOver software from UC Santa Cruz.

Trans only TF footprint enrichment vs. differential expression

We evaluated footprints for each TF for enrichment in human-specific and macaque-specific *trans* only regions compared to 10x length-matched expected regions. Enrichment scores were computed using Fisher's Exact Test with a BH adjusted p-value < 0.05. We intersected the enrichment score with the differential expression values of the specified TF. We removed footprints associated with TF multimers, for example the SMAD2-SMAD3-SMAD4 motif, so that only individual TFs, such as SMAD3, were assigned differential expression values. We also removed TFs that were not analyzed in the differential expression analysis, likely because they did not meet the 1:1 orthology requirement. Altogether, 386 TFs were retained for plotting. Scatterplots were made with ggplot2 and text was plotted for TFs with a footprint enrichment log2OR > 0, footprint enrichment padj < $1x10^{-10}$, differential expression log2FC > 0 (log2FC < 0 for macaque-specific), and a differential expression padj < $1x10^{-50}$ (padj < $1x10^{-20}$ for macaque-specific). For the TFs that met these criteria, which we defined as *putative trans regulators*, we intersected their footprints (BEDTools *intersect*: default parameters) with the respective trans only regions to determine the percentage with the given footprint. In a few cases we merged TF footprints, because some of the TFs shared the same motif archetype (Vierstra et al., 2020), for example IRF4, IRF7, and IRF8.

Evolutionary Characterization of Cis and Trans Divergent Regions

PhastCons Enrichment Analysis

We intersected active regions with 30-way MultiZ PhastCons elements—derived from an alignment of 27 primate species and three mammalian outgroup species (Lindblad-Toh et al., 2011; Siepel et al., 2005) (last downloaded September 22nd, 2021 from http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons30way/) using BEDTools with standard parameters. A region was considered conserved when overlapped >= 1 bp of a PhastCons element. For each category with activity differences between humans and rhesus macaques, we quantified PhastCons element enrichment in that category versus the matched 10x expectation sets using Fisher's Exact Test with a BH adjusted p-value < 0.05. Unless specified, in the evolutionary analyses, we combined human and macaque elements and evaluated their characteristics in the human genome.

Human Acceleration Enrichment Analysis

We estimated human acceleration from ATAC-STARR-seq bins using the phyloP function from the Phast tools suite (http://compgen.cshl.edu/phast/). Short term estimates of human acceleration and conservation (--mode CONACC) were calculated between the human and chimp branches against the 30-way neutral tree model (--g hg38.phastCons30way.mod) using the likelihood ratio test (--method LRT). For long term estimates of human acceleration, we first trimmed the model tree to remove any species on the human branch that emerged after the most recent common ancestor between humans and rhesus macaques, then used this trimmed neutral tree model to quantify acceleration and conservation (described above). Bins with a phyloP score cutoff < -1 were considered accelerated. We removed any bins from the acceleration analysis that overlapped human duplicated regions (hg38 SELF-CHAIN) with >= 1 bp overlap using BEDTools with standard parameters. To assign a single human acceleration value per divergently active region and matched-expectation, we assigned the bin with the minimum PhyloP score to entire region. We estimated human acceleration enrichment as the number of human accelerated regions (phylop < -1.0, corresponding to a p-value <0.05) in a divergently active group versus matched expected acceleration values. We assigned each region in the observed and expected dataset with the lowest phyloP bin value (i.e. the most accelerated value).

Repeatmasker Transposable Element Enrichment

We downloaded hg38 repeatmasker coordinates from the UCSC genome browser (last downloaded August 21st, 2021). Active regions and matched expectation sets were intersected with TE coordinates and active regions were assigned TE if a TE overlapped >=1bp of a region. To test for enrichment, we used Fisher's Exact Test with a BH adjusted p-value < 0.05 to compute the enrichment of TEs overlapping active elements versus matched expectation datasets. For family-specific analysis, we stratified by TE family overlap and quantified TE enrichment as the number of elements overlapping a TE family per activity category (e.g. *cis* only) and all other activity category datasets using Fisher's Exact Test with a BH adjusted p-value < 0.05.

TF footprint Enrichment for SINE/Alu Cis & Trans Regions

We evaluated GM12878 TF footprints for enrichment in *cis* & *trans* regions that overlapped SINE/Alu transposable elements compared to 10x expected regions. Enrichment scores were computed using Fisher's Exact Test with a BH adjusted p-value < 0.05.

Assigning Sequence Ages

The genome-wide hg38 100-way vertebrate multiz multiple species alignment was downloaded from the UCSC genome browser. Each syntenic block was assigned an age based on the most recent common ancestor (MRCA) of the species present in the alignment block in the

UCSC all species tree model. Regions and matched shuffles were intersected with syntenic blocks and the maximum age for each region was selected as the representative age. For most analyses, we focus on the MRCA-based age, but when a continuous estimate is needed, we use evolutionary distances from humans to the MRCA node in the fixed 100-way neutral species phylogenetic tree. Estimates of the divergence times of species pairs in millions of years ago (MYA) were downloaded from TimeTree (Hedges, Marin, Suleski, Paymer, & Kumar, 2015). Sequence age provides a lower-bound on the evolutionary age of the sequence block. Sequence ages could be estimated for 94% of the autosomal bp in the hg38 human genome.

Multiple Sequence Origin Enrichment Analysis

After assigning sequence ages to regions (above), we quantified how often regions overlapped multiple sequence ages (referred to as multi-origin sequences) with >=6 base pairs in length per age. We compared the number of multi-origin sequences in cis-, trans- and cis & trans categories with their length-matched expectation sets (see above section Generating genomic background - shared accessible, inactive expectation datasets) and computed enrichment using Fisher's Exact Test.

Human Variant Enrichment Analysis

eQTL Enrichment

We intersected each divergent activity category with eQTL from GTEx (version 8; last downloaded April 30th 2018) using BEDTools with standard parameters. To measure whether the observed number of eQTL variants was more than expected, we shuffled each divergent set of regulatory elements 1000x in a background set of length-matched shared accessible, inactive peaks and quantified the fold-changes as the number of observed eQTL variants divided by the median number of expected eQTL variants. We calculated the empirical p-values from the number of eQTL overlaps in the expected sets that were equal to or more extreme than the observed number of eQTL overlaps. We bootstrapped the 95% confidence intervals by estimating the distribution of fold-changes from the observed count with each of the 1000 expected overlaps.

UKBB GWAS Trait Enrichment

We selected a set of immune, inflammatory, and B cell related traits from the UKBB pan-GWAS. For each trait, we included only the tag-SNPs with genome-wide significance (p<5.5-e8) and LD-expanded those tag-SNPs to include variants in perfect LD (R2=1.0) in European populations from 1000 genomes (1000 genomes consortium). We removed any active regions that overlapped the HLA locus in hg38 (chr6:28898751-33807669), including 4 *cis* only elements, 1 *cis* & *trans*, 1 *trans* only, and 0 conserved active. We then intersected the accessible peaks containing divergently active regions with LD-expanded, significant GWAS SNPs using BEDTools with standard parameters. To measure whether the observed number of GWAS variants was more than expected, we shuffled each divergent set of regulatory elements 1000x in a background set of length-matched shared accessible, inactive regions and quantified the fold-changes as the number of observed GWAS variants divided by the median number of expected GWAS variants. We calculated the empirical p-values from the number of GWAS overlaps in the expected sets that were equal to or more extreme than the observed number of GWAS overlaps. We bootstrapped the 95% confidence intervals by estimating the distribution of fold-changes from the observed count with each of the 1000 expected overlaps.

Gene Expression Analysis

Data Collection

In addition to the RNA-seq experiments described above, we downloaded and analyzed FASTQ files from the following publications: Cain et al., 2011 - GSE24111 (SRR066745-7, SRR066751-3); Blake et al., 2020 - GSE112356 (SRR6900782-SRR6900812); Calderon et al., 2019 - GSE118165 (SRR11007061, 071, 082, 090, 092, 094, 096, 113, 121, 124, 126, 127, 137, 147, 156, 158, 160, 170, 183, 186, 188, 190; SRR7647654, 656, 658, 696, 698, 700, 731, 767, 768, 769, 807, 808), and the ENCODE GM12878 Wold (total RNA-seq: ENCFF248MER, ENCFF006YWA, ENCFF294LGZ, ENCFF995BLA) and Gingeras (polyA plus RNA-seq: ENCFF001REH - ENCFF001REK) GM12878 datasets. The FASTQ files from these datasets and our GM12878 and LCL8664 data were processed in the same way.

Fastq Processing of RNA-seq Data

Raw reads were trimmed and analyzed for quality with Trim Galore! using the --fastqc and --paired parameters. To avoid bias arising from duplicated genes, we restricted our analysis to 1:1 orthologous exons that we obtained from XSAnno (Zhu, Li, Sousa, & Sestan, 2014) (https://hbatlas.org/xsanno/files/Ensembl-v64-Human-Macaque). The hg19 file was converted to hg38 coordinates using liftOver. Because no rheMac2 to rheMac10 map chain file existed, we first converted rheMac2 coordinates to rheMac8 and then to rheMac10. We then mapped trimmed reads to the 1:1 orthologous exons in the respective genome using the STAR aligner(Dobin et al., 2013) (alignReads function); we built a STAR index for each genome for each illumina read length type (150nt, 50nt, 35nt, and 100nt) and applied it to the respective sample. We next counted reads in each 1:1 orthologous exon using the *featureCounts* function from subread(Liao et al., 2014); for our samples, we set the -s parameter to 1 because they were stranded RNA-seq datasets, while all others were set to 0 (unstranded). For paired datasets, we also specified the -p and -B options. We applied the -O option to all datasets.

Differential Expression Analysis

For all pairwise comparisons presented, we performed differential expression analysis with DESeq2 (fitType="local") and extracted results using the *lfcShrink* function and apeglm shrinkage algorithm, which shrinks the effect size of low count data (cite deseq and apeglm). Before comparing GM12878 and LCL8664, we removed sex chromosomes. We defined human-specific expressed genes as those with a log2FC > 2 and a padj < 0.001, while macaque-specific expressed genes had a log2FC < -2 and a padj < 0.001. We used ChIPSeeker and ClusterProfiler to perform Reactome pathway enrichment analysis using the differentially expressed gene sets (Yu et al., 2012); we plotted the top five to six categories in each case.

TPM normalization and Correlation Between Human and Macaque LCLs

For each of our GM12878 and LCL8664 replicates, we normalized read counts so they represented transcripts per million (TPM); we first calculated RPKM [10^9 * (reads mapped to transcript / (total reads * length of transcript))] and then converted to TPM [10^6 * (RPKM/(sum(RPKM)))]. We then calculated the mean TPM for each gene between the two replicates, added a pseudo count of 1, and $\log_{10}$ normalized the values. We then plotted the GM12878 and LCL8664 values on a 2D bin plot; both Pearson and Spearman’s correlation coefficients were calculated using the mean TPM values.

Principle Component Analysis

For each of the samples plotted in each PCA, we first extracted variance stabilizing transformed (VST) count values from the DESeq Dataset (dds) with the *vst* function (blind=TRUE) and then plotted principal components 1 and 2 using the *plotPCA* function (both functions from the DESeq2 package).

CHAPTER IV

DISCUSSION AND FUTURE DIRECTIONS

Discussion

In Chapter II of this dissertation, I described a new approach called ATAC-STARR-seq that simultaneously quantifies regulatory activity, chromatin accessibility, and TF occupancy in the human genome. With this approach, we identified 30,078 active regions and 21,125 silent regions in human lymphoblastoid cells that are contained within ~101,000 chromatin accessible regions; the number of regulatory regions we identify is substantially less than the number of accessible regions because most chromatin accessible DNA has neutral activity. We also identified TF bound sites within accessible chromatin for 746 different TF motif sequences and used them to stratify active and silent regions into DNA regulatory networks. Altogether, we showed that ATAC-STARR-seq can identify five different layers of transcriptional regulation from a single DNA fragment library: 1) chromatin accessibility, 2) TF binding, 3) active regulatory activity, 4) silent regulatory activity, and 5) DNA regulatory element networks. This new method substantially expands the capabilities of previous methods and allows us and others to address critical and new questions in transcriptional regulation across several disciplines, from developmental biology to evolutionary biology.

Massively parallel reporter assays like ATAC-STARR-seq are only one of several approaches to identify putative enhancers in the human genome. Each approach measures a slightly different aspect of DNA regulatory element biology and has its own set of limitations. Therefore, the field is best served when multiple orthogonal approaches are performed to identify a set of high confidence DNA regulatory elements in a given biological context (Gasperini et al., 2020). Compared to these other methods, there are three major advantages of ATAC-STARR-seq. First, the quantitative measure in ATAC-STARR-seq is direct; it is a functional quantification of the ability of a DNA sequence to drive transcription of a target gene. Second, the dense tiling of accessible regions allows for highly-resolved calls of regulatory activity such that we can identify ~100bp active and silent components of individual accessible regions—this is not possible for approaches like PRO-seq that use bidirectional transcription to call DNA regulatory elements. Third, while most assays only provide one measure, ATAC-STARR-seq provides five different layers of transcriptional regulatory information from one assay. The major caveat of ATAC-STARR-seq is that it is episomal, so it does not measure DNA sequences in their endogenous chromatin environments. Yet, by leveraging this caveat, we were able to identify the effect of cellular environment on DNA regulatory element activity in Chapter III of this dissertation. ATAC-STARR-seq also does not identify which genes are regulated by the elements it identifies, which is a significant challenge for most assays in the field. Altogether, ATAC-STARR-seq provides a substantial improvement to the MPRA arm of DNA regulatory identification strategies and should be used in combination with other approaches to obtain high-confident active and silent regions in future experiments.

ATAC-STARR-seq allowed investigation of global effects on regulatory element activity due to changes in *cis*, i.e. mutations to regulatory element DNA sequence, and changes in *trans, i.e,* changes in cellular environment. This investigation, described in Chapter III, aimed to answer a critical question in evolution: what was the primary mode of regulatory activity evolution between closely related primate species that only diverged 25 million years ago? Does evolution over this timeframe favor global changes on the cellular environment with pleiotropic effects (*trans)*, or does it favor precise changes to individual regulatory elements (*cis)*? By applying a comparative ATAC-STARR-seq framework to human and macaque lymphoblastoid cells, we observed roughly the same levels of *cis* and *trans* regulatory activity divergence between LCLs derived from each of these species, ~11,000 compared to ~10,500, respectively. This observation questions current dogma, which previously thought most regulatory element activity changes between closely related species were in *cis* (Irene Gallego Romero & Lea, 2022; Gordon & Ruvinsky, 2012; Mattioli et al., 2020; Whalen et al., 2023). Our results conclude that *trans*-regulatory divergence between human and non-human primates is much greater than previously thought.

In hindsight, the high frequency of *trans*-regulatory divergence is not surprising considering that alterations to the *trans*-regulatory environment can occur via many mechanisms. To name a few, mutations to regulatory elements that control TF expression, mutations to TF protein sequences that alter their function, mutations that alter post-transcriptional and post-translational processing of TFs, and alterations to signalling pathways that affect the phosphorylation state of a TF all can affect DNA regulatory element activity in *trans*. Comparatively, *cis* mutations are confined to the size of the DNA regulatory element, which is typically less than 1000bp. Simply put, there is a lot more opportunity to mutate *trans*-regulatory processes than those in *cis*. This principle is evident in a yeast study which identified that random mutations affect regulatory element activity in *trans* substantially more often than random mutations in *cis* (B. P. Metzger et al., 2016). However random mutations are not under selective pressures and since *trans* changes are pleiotropic and more likely to be deleterious overall, *cis* changes are preferentially selected during evolution (Signor & Nuzhdin, 2018). Indeed, a previous study between *Drosophila* species observed a greater proportion of *cis* divergence on gene expression as evolutionary distance increases (Coolon et al., 2014). Because non-selected random mutations more often affect gene expression in *trans*, our observation that *cis* and *trans* divergence occurs as the same frequency indicates there is an evolutionary preference for *cis,* but this preference is not so strong that *trans* divergence is completely lost.

We also observed that most regulatory regions diverge in both *cis* and *trans*, such that changes to either the DNA sequence or the cellular environment affect regulatory activity of a single DNA regulatory element. This observation indicates that these processes act in tandem with each other. There are three possible mechanisms to explain this phenomenon. First, this could be a coordinated redundancy mechanism to stabilize phenotypes. Such a mechanism would be under stabilizing selection and could operate on phenotypes that are sensitive to change. A second mechanism would be that after either a *cis* or *trans* change occur to inactivate a given regulatory element, its activity would no longer be selected. For example, if a *trans* change renders a DNA regulatory element inactive, mutation to the element can occur without affecting its function. A third mechanism could occur for regulatory elements that gain activity where a *cis* or *trans* change occurs but is not sufficient for DNA regulatory element activity. For example, a *cis* change could create a high affinity binding site for a TF that is not expressed. Likely it is a combination of all three, but future investigations are required to tease out these mechanisms.

It is unlikely that regions divergent in both *cis* and *trans* acquired these changes simultaneously, so one event must precede the other for a given regulatory element. Because *trans* divergence occurs more frequently with shorter evolutionary time scales (Hill et al., 2020; Signor & Nuzhdin, 2018), I would hypothesize that, for a majority of regulatory elements, *trans* changes occur first. Moreover, we identified over a thousand regions that were divergent only in *cis* or only in *trans*. It is intriguing to wonder if these regions represent an intermediate state and will ultimately acquire divergence in both *cis* and *trans*, given more evolutionary time. We cannot dissect this with our data, so future studies would be required to investigate these questions further.

Our results are important when considering MPRA design. For technical reasons, many studies identify enhancers in one tissue or cell type using an epigenetic annotation method and then assay them for activity as a MPRA in a different cell type (Johnson et al., 2018; Klein et al., 2018; Uebbing et al., 2021; Vockley et al., 2015; Weiss et al., 2021). For example, one study first identified candidate enhancers in liver tissues for a variety of primate species and tested them for activity in HepG2 cells. While HepG2 is a human liver cancer cell line, it may have a very different *trans*-regulatory environment than non-diseased liver cells. Moreover, this approach assumes that the *trans* regulatory environments between primate species are the same. Our failure to reproduce most activity measures across species highlights how such approaches may be misleading and confounded by effects from the cellular environment. In other words, the choice of cell type is critically important when performing MPRAs and interpreting their results.

Future Directions

Improvements to ATAC-STARR-seq

Current MPRA bioinformatic tools and pipelines did not account for unique aspects of ATAC-STARR-seq, so it was necessary to develop a custom bioinformatic workflow. In designing this workflow, I made a critical assumption that most regions are neutral—they do not affect the basal transcription rate of the plasmid. This strategy normalizes activity scores so that the average value is zero, meaning there are the same number of reporter RNAs as plasmid DNA—activity is $\log_2$(RNA/DNA). This assumption could be confounding if the true average is non-zero. For example, if the average is positive, regions we call as silencers may actually not inhibit the basal transcription rate. To remove this assumption altogether, I propose curating and implementing a negative control spike-in plasmid library made up of validated neutral activity test sequences. This technique is common among synthetically derived MPRA approaches, so these strategies could be adapted for ATAC-STARR-seq. The negative control library containing validated neutral sequences would be co-transfected with the ATAC-STARR-seq plasmid and normalized so that "zero" reflects the average activity of these neutral sequences.

In ATAC-STARR-seq, we had to decide whether to include or remove molecular duplicates from our analysis. Although I showed that including PCR duplicates was preferred over collapsing duplicates in the ATAC-STARR-seq workflow, it is potentially confounding that we cannot distinguish between PCR duplicates and biological duplicates (multiple transcripts off the same plasmid) in our assay. This limitation could be addressed by implementing a unique molecular identifier (UMI) to the system—such as the strategy employed by UMI-STARR-seq (Neumayr et al., 2019)—to collapse only the duplicates arising from PCR. To implement a UMI strategy in ATAC-STARR-seq, we would integrate a UMI into the reverse transcription reaction so that the UMI represents a single cDNA, which is a true biological duplicate that would not be collapsed. The UMI sequence would replace the i5 index so that paired-end sequencing would provide both an i7 index to tag the sample and deconvolute it from others, as well as a UMI sequence for each read. PCR duplicates would have the same UMI, so these could be collapsed and differentiated from biological duplicates which would have different UMI sequences. This could be done on a platform, such as HiSeq, that would allow us to deconvolute with just one index alone.

Future Applications of ATAC-STARR-seq

In addition to investigating *cis* and *trans* regulatory divergence between primates ATAC-STARR-seq presents the opportunity to investigate many unique questions. One of these would be to apply the method to the human embryonic stem cell to neural progenitor cell time course model that the Hodges Lab has already established. Another project in the lab, spearheaded by Lindsey Guerin, looks at the interplay of DNA methylation, chromatin accessibility, and gene expression on cellular differentiation during this time-course. Adding ATAC-STARR-seq to this would provide a regulatory activity layer that could be used to differentiate poised and active enhancers in this system and give a highly resolved view of the functional genomic mechanisms involved in driving early differentiation.

In Chapter III, I created an ATAC-STARR-seq plasmid library from one cell-type and tested it in another. We applied this unique aspect of MPRAs to investigate questions in evolutionary biology, but this concept could be applied elsewhere. *Trans* regulatory changes are probably very significant in cancer, and so it would be interesting to apply a similar logic to identify *trans* regulatory changes in cancer cell lines and the extent of their effects on regulatory element activity. This approach also has the potential to identify dysfunctional gene regulatory networks in diseases like cancer where neoplastic transformation can be driven by the dysfunction of a specific TF. One place to start would be to compare our current data in GM12878 cells with a B cell lymphoma cell line, such as SU-DHL-6. While GM12878 cells are immortalized and "cancer-like" their karyotype is normal and could act as the non-diseased state in a pilot experiment. A more elegant experiment would be to compare enhancer activity states of cancer and non-cancerous tissues from the same patient donor, but key limitations in cell count requirements for ATAC-STARR-seq would need to be solved prior to conducting this project.

There are many more possibilities for ATAC-STARR-seq. Ultimately it was designed to identify active enhancers genome-wide, so it could be used to identify active enhancers in a variety of different contexts or conditions. It would be valuable to have an ENCODE-like consortium-type resource of ATAC-STARR-seq data to identify active and silent regions in dozens of cell lines. Such a dataset would add a direct regulatory activity measure to the growing list of functional data available to researchers world-wide. Even on an individual experiment scale, ATAC-STARR-seq could be applied to important questions in human health such as understanding the effect of a given drug on global enhancer activity. Several drugs inhibit transcriptional regulation programs, so it would be beneficial to see where and how the active enhancer landscape changes to such drugs.

Deeper Investigation of Silencers

The identification of 21,000 silent regions led us to investigate whether they are real—*i.e.,* whether they are the result of technical artifacts or true biological activity. Overall, our investigation concluded that they most likely reflect true silent regions in the cell, but we cautioned other possible technical reasons could explain their presence. More needs to be done to tease out the biology of our putative silencer regions, and they present a very intriguing future direction of this research. One stark difference between active regions, silent regions, and neutral regions, is that silent regions are highly promoter enriched. This suggests that the majority of silencing activity occurs nearby the gene. We investigated some promoter-specific mechanisms, like transcription initiation from the 3’ UTR, as a way to explain this promoter bias but did not find anything of merit. Future investigations could explore the promoter bias further.

Implementation of the neutral control spike-in, as discussed above, would help support the validity of silencers because they provide a more controlled assessment of basal transcription rate. Another idea would be to validate our putative silencers with an orthogonal approach. One group recently developed an approach to identify silencers by leveraging the expression of Caspase 9 from a plasmid (Pang & Snyder, 2020). Caspase 9 induces cell death, so test sequences that inhibit Caspase 9 expression would linger in the cell population after many divisions because non-silencer cells would die. It would be illuminating to see if our silencers could be identified in this setting as well.

Notably, we did not consider silencers in the *cis* and *trans* investigation presented in Chapter III. It would be interesting to look at their activity differences in *cis* and *trans* and observe whether they have species specific activity and if differential activity occurs more predominantly in *cis* or in *trans*.

Future Directions to Address Limitations of *Cis/Trans* Study

In the discussion of chapter III, I highlighted some key limitations of our study. Future experiments could be performed to address these limitations. There are few non-human primate cell lines available for purchase and so we were limited to only one rhesus macaque cell line for this project. Furthermore, it is challenging to investigate many genotypes from one species with our comparative ATAC-STARR-seq workflow—the complexity of the analysis and experiment become too great. This is problematic for a few reasons. First, the effects we see may reflect differences between these two individuals specifically and not generalize to the entire species (Kelley & Gilad, 2020). Second, because these two cell lines were generated by separate immortalization processes, the latent viral load contained within them may explain some of the differences we observe. Moreover, they may represent different stages of B cell development and so differences may be due to cell stage differences, rather than species-level differences.

To address this in future studies, I propose using iPSC panels for human, chimpanzee, and macaque (I. Gallego Romero et al., 2015; H. Liu et al., 2008). These are freely available and multiple genotypes per species are contained within these panels. This would avoid any immortalization process or cell stage effects by comparing stem cells directly to each other. While many of the more obvious phenotypic differences between human and macaque involve differentiated cells, differences at the stem cell level could also be interesting and perhaps provide insight into the human-specific neoteny process, which is a developmental slowdown trait thought to explain many phenotypic differences between humans and non-human primates (Hirai, Imai, & Go, 2012).

This panel does not solve the feasibility limitation in applying a similar ATAC-STARR-seq workflow to investigate multiple genotypes per species. It is hard to envision how to navigate this aside from conducting a brute force type of approach where libraries A, B, and C from one species are transfected into cell types X, Y, and Z from the other species. The amount of pairwise combinations from this experiment ($\geq 3^3$) would be overwhelming, and so the most difficult hurdle would be developing an analytical scheme that can simplify the dataset.

Expanding our *Cis* and *Trans* Investigation to other Mammalian Species

In line with previous studies, we would expect to see a greater degree of *cis* divergence as evolutionary distance increases between the species being compared. To see if this is true, we could compare a more recent and at a more distant evolutionary relationship than human and macaque. For example, we could compare regulatory activity with similar comparative ATAC-

STARR-seq framework between human, chimpanzee, rhesus macaque, and mouse induced pluripotent stem cells (iPSCs), all of which are readily available. We would expect to see more *cis* divergence between mouse and human, and less between human and chimpanzee. We could also use this dataset to investigate the temporal order of *cis* and *trans* divergence by looking at whether a DNA regulatory element that is affected in *cis* and *trans* between human-macaque is only affected in *cis* or only in *trans* between human-chimpanzee.

Further Dissection of Individual Evolutionary-Relevant Loci

The ETS1 locus, described in Chapter III, provides an example of how a single substitution between human and macaque may have driven differential activity of ETS1 and therefore altered the cellular environment in *trans*. While we focused more on describing global trends, specific examples like this provide unique opportunities to investigate individual mutational events that had significant impacts on evolutionary outcomes. Future studies could use our comparative ATAC-STARR-seq dataset to find and characterize more examples like ETS1. I would recommend future studies use CRISPR to "rhesus-ize" these human enhancers to discover the effect of *cis* mutations in a highly controlled fashion.

Summary

In summary, I developed a new approach called ATAC-STARR-seq to identify enhancers in the human genome so that researchers could address new and important questions in gene regulation. I show how ATAC-STARR-seq can provide five different levels of gene regulatory information including the identification of active and silent regions. I then applied ATAC-STARR-seq to investigate the relative role of *cis* and *trans* regulatory changes on gene regulatory evolution between human and macaque. I discover a greater role for *trans* regulatory divergence than previously recognized and surprisingly find that most differentially active regulatory elements diverged in both *cis* and *trans*. These observations generate several hypotheses that can be investigated in future studies and add to our understanding of what it means to be human.

REFERENCES

Abugessaisa, I., Ramilowski, J. A., Lizio, M., Severin, J., Hasegawa, A., Harshbarger, J., . . . Kasukawa, T. (2021). FANTOM enters 20th year: expansion of transcriptomic atlases and functional annotation of non-coding RNAs. *Nucleic Acids Res, 49*(D1), D892-D898. doi:10.1093/nar/gkaa1054

Aebersold, R., Agar, J. N., Amster, I. J., Baker, M. S., Bertozzi, C. R., Boja, E. S., . . . Zhang, B. (2018). How many human proteoforms are there? *Nat Chem Biol, 14*(3), 206-214. doi:10.1038/nchembio.2576

Agoglia, R. M., Sun, D., Birey, F., Yoon, S. J., Miura, Y., Sabatini, K., . . . Fraser, H. B. (2021). Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature, 592*(7854), 421-427. doi:10.1038/s41586-021-03343-3

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., . . . Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature, 507*(7493), 455-461. doi:10.1038/nature12787

Andersson, R., & Sandelin, A. (2019). Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet*. doi:10.1038/s41576-019-0173-8

Andersson, R., Sandelin, A., & Danko, C. G. (2015). A unified architecture of transcriptional regulatory elements. *Trends Genet, 31*(8), 426-433. doi:10.1016/j.tig.2015.05.007

Arnold, C. D., Gerlach, D., Spies, D., Matts, J. A., Sytnikova, Y. A., Pagani, M., . . . Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet, 46*(7), 685-692. doi:10.1038/ng.3009

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science, 339*(6123), 1074-1077. doi:10.1126/science.1232542

Atlasi, Y., & Stunnenberg, H. G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. *Nat Rev Genet, 18*(11), 643-658. doi:10.1038/nrg.2017.57

Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell, 33*(3), 729-740. doi:10.1016/0092-8674(83)90015-6

Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C., & Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell, 23*(2), 276-288 e278. doi:10.1016/j.stem.2018.06.014

Barnett, K. R., Decato, B. E., Scott, T. J., Hansen, T. J., Chen, B., Attalla, J., . . . Hodges, E. (2020). ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. *Mol Cell, 77*(6), 1350-1364 e1356. doi:10.1016/j.molcel.2020.01.004

Barr, K. A. R., K. L.; Gilad, Y. (2022). Embryoid bodies facilitate comparative analysis of gene expression in humans and chimpanzees across dozens of cell types. *bioRxiv*(2022.07.20.500831). doi:https://doi.org/10.1101/2022.07.20.500831

Bauernfried, S., & Hornung, V. (2022). Human NLRP1: From the shadows to center stage. *J Exp Med, 219*(1). doi:10.1084/jem.20211405

Bauernfried, S., Scherr, M. J., Pichlmair, A., Duderstadt, K. E., & Hornung, V. (2021). Human NLRP1 is a sensor for double-stranded RNA. *Science, 371*(6528). doi:10.1126/science.abd0811

Benton, M. L., Talipineni, S. C., Kostka, D., & Capra, J. A. (2019). Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics, 20*(1), 511. doi:10.1186/s12864-019-5779-x

Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., . . . Looso, M. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun, 11*(1), 4267. doi:10.1038/s41467-020-18035-1

Bradner, J. E., Hnisz, D., & Young, R. A. (2017). Transcriptional Addiction in Cancer. *Cell, 168*(4), 629-643. doi:10.1016/j.cell.2016.12.013

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., . . . Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature, 478*(7369), 343-348. doi:10.1038/nature10532

Britten, R. J., & Davidson, E. H. (1969). Gene regulation for higher cells: a theory. *Science, 165*(3891), 349-357. doi:10.1126/science.165.3891.349

Britten, R. J., & Davidson, E. H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q Rev Biol, 46*(2), 111-138. doi:10.1086/406830

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods, 10*(12), 1213-1218. doi:10.1038/nmeth.2688

Cain, C. E., Blekhman, R., Marioni, J. C., & Gilad, Y. (2011). Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics, 187*(4), 1225-1234. doi:10.1534/genetics.110.126177

Calderon, D., Nguyen, M. L. T., Mezger, A., Kathiria, A., Muller, F., Nguyen, V., . . . Pritchard, J. K. (2019). Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat Genet, 51*(10), 1494-1505. doi:10.1038/s41588-019-0505-9

Cao, Z., Sun, X., Icli, B., Wara, A. K., & Feinberg, M. W. (2010). Role of Kruppel-like factors in leukocyte development, function, and disease. *Blood, 116*(22), 4404-4414. doi:10.1182/blood-2010-05-285353

Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L., & Pollard, K. S. (2013). Many human accelerated regions are developmental enhancers. *Philos Trans R Soc Lond B Biol Sci, 368*(1632), 20130025. doi:10.1098/rstb.2013.0025

Chaudhri, V. K., Dienger-Stambaugh, K., Wu, Z., Shrestha, M., & Singh, H. (2020). Charting the cis-regulome of activated B cells by coupling structural and functional genomics. *Nat Immunol, 21*(2), 210-220. doi:10.1038/s41590-019-0565-0

Chavarria-Smith, J., Mitchell, P. S., Ho, A. M., Daugherty, M. D., & Vance, R. E. (2016). Functional and Evolutionary Analyses Identify Proteolysis as a General Mechanism for NLRP1 Inflammasome Activation. *PLoS Pathog, 12*(12), e1006052. doi:10.1371/journal.ppat.1006052

Chen, A. F., Parks, B., Kathiria, A. S., Ober-Reynolds, B., Goronzy, J. J., & Greenleaf, W. J. (2022). NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat Methods, 19*(5), 547-553. doi:10.1038/s41592-022-01461-y

Cho, Y. G., Gordadze, A. V., Ling, P. D., & Wang, F. (1999). Evolution of two types of rhesus lymphocryptovirus similar to type 1 and type 2 Epstein-Barr virus. *J Virol, 73*(11), 9206-9212. doi:10.1128/JVI.73.11.9206-9212.1999

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science, 351*(6277), 1083-1087. doi:10.1126/science.aad5497

Chuong, E. B., Rumi, M. A., Soares, M. J., & Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet, 45*(3), 325-329. doi:10.1038/ng.2553

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M., & Zaret, K. S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell, 9*(2), 279-289. doi:10.1016/s1097-2765(02)00459-8

Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., . . . Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun, 9*(1), 781. doi:10.1038/s41467-018-03149-4

Consortium, G. T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science, 369*(6509), 1318-1330. doi:10.1126/science.aaz1776

Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R., & Wittkopp, P. J. (2014). Tempo and mode of regulatory evolution in Drosophila. *Genome Res, 24*(5), 797-808. doi:10.1101/gr.163014.113

Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., . . . Chang, H. Y. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet, 48*(10), 1193-1203. doi:10.1038/ng.3646

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., . . . Chang, H. Y. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods, 14*(10), 959-962. doi:10.1038/nmeth.4396

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., . . . Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A, 107*(50), 21931-21936. doi:10.1073/pnas.1016071107

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nat Methods, 10*(4), 325-327. doi:10.1038/nmeth.2375

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Doni Jayavelu, N., Jajodia, A., Mishra, A., & Hawkins, R. D. (2020). Candidate silencer elements for the human and mouse genomes. *Nat Commun, 11*(1), 1061. doi:10.1038/s41467-020-14853-5

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., . . . Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell, 66*(4), 568-576 e564. doi:10.1016/j.molcel.2017.04.018

Douillet, D., Sze, C. C., Ryan, C., Piunti, A., Shah, A. P., Ugarenko, M., . . . Shilatifard, A. (2020). Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. *Nat Genet, 52*(6), 615-625. doi:10.1038/s41588-020-0618-1

Duttke, S. H., Chang, M. W., Heinz, S., & Benner, C. (2019). Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res, 29*(11), 1836-1846. doi:10.1101/gr.253492.119

Edsall, L. E., Berrio, A., Majoros, W. H., Swain-Lenz, D., Morrow, S., Shibata, Y., . . . Allen, A. S. (2019). Evaluating Chromatin Accessibility Differences Across Multiple Primate Species Using a Joint Modeling Approach. *Genome Biol Evol, 11*(10), 3035-3053. doi:10.1093/gbe/evz218

Elbarbary, R. A., Lucas, B. A., & Maquat, L. E. (2016). Retrotransposons as regulators of gene expression. *Science, 351*(6274), aac7247. doi:10.1126/science.aac7247

Emerson, J. J., Hsieh, L. C., Sung, H. M., Wang, T. Y., Huang, C. J., Lu, H. H., . . . Li, W. H. (2010). Natural selection on cis and trans regulation in yeasts. *Genome Res, 20*(6), 826-836. doi:10.1101/gr.101576.109

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods, 9*(3), 215-216. doi:10.1038/nmeth.1906

Fenini, G., Karakaya, T., Hennig, P., Di Filippo, M., & Beer, H. D. (2020). The NLRP1 Inflammasome in Human Skin and Beyond. *Int J Mol Sci, 21*(13). doi:10.3390/ijms21134788

Fitzgerald, K. A., & Kagan, J. C. (2020). Toll-like Receptors and the Control of Immunity. *Cell, 180*(6), 1044-1066. doi:10.1016/j.cell.2020.02.041

Fong, S. L., & Capra, J. A. (2021). Modeling the Evolutionary Architectures of Transcribed Human Enhancer Sequences Reveals Distinct Origins, Functions, and Associations with Human Trait Variation. *Mol Biol Evol, 38*(9), 3681-3696. doi:10.1093/molbev/msab138

Fong, S. L., & Capra, J. A. (2022). Function and constraint in enhancer sequences with multiple evolutionary origins. *Genome Biol Evol*. doi:10.1093/gbe/evac159

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., . . . Mathelier, A. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res, 48*(D1), D87-D92. doi:10.1093/nar/gkz1001

Franchini, L. F., & Pollard, K. S. (2017). Human evolution: the non-coding revolution. *BMC Biol, 15*(1), 89. doi:10.1186/s12915-017-0428-9

Gallego Romero, I., & Lea, A. J. (2022). Leveraging massively parallel reporter assays for evolutionary questions. *arXiv*.

Gallego Romero, I., Pavlovic, B. J., Hernando-Herraez, I., Zhou, X., Ward, M. C., Banovich, N. E., . . . Gilad, Y. (2015). A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. *Elife, 4*, e07103. doi:10.7554/eLife.07103

Garcia-Perez, R., Esteller-Cucala, P., Mas, G., Lobon, I., Di Carlo, V., Riera, M., . . . Marques-Bonet, T. (2021). Epigenomic profiling of primate lymphoblastoid cell lines reveals the evolutionary patterns of epigenetic activities in gene regulatory architectures. *Nat Commun, 12*(1), 3116. doi:10.1038/s41467-021-23397-1

Garrett-Sinha, L. A. (2013). Review of Ets1 structure, function, and roles in immunity. *Cell Mol Life Sci, 70*(18), 3375-3390. doi:10.1007/s00018-012-1243-7

Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet, 21*(5), 292-310. doi:10.1038/s41576-019-0209-0

Glaser, L. V., Steiger, M., Fuchs, A., van Bommel, A., Einfeldt, E., Chung, H. R., . . . Meijsing, S. H. (2021). Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. *Nucleic Acids Res, 49*(21), 12178-12195. doi:10.1093/nar/gkab1100

Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., . . . Marioni, J. C. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Res, 22*(12), 2376-2384. doi:10.1101/gr.142281.112

Gordon, K. L., & Ruvinsky, I. (2012). Tempo and mode in evolution of transcriptional regulation. *PLoS Genet, 8*(1), e1002432. doi:10.1371/journal.pgen.1002432

Graze, R. M., McIntyre, L. M., Main, B. J., Wayne, M. L., & Nuzhdin, S. V. (2009). Regulatory divergence in Drosophila melanogaster and D. simulans, a genomewide analysis of allele-specific expression. *Genetics, 183*(2), 547-561, 541SI-521SI. doi:10.1534/genetics.109.105957

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol, 19*(10), 621-637. doi:10.1038/s41580-018-0028-8

Hansen, T. J., & Hodges, E. (2022a). ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome. *Genome Research, 32*, 1529-1541. doi:10.1101/gr.276766.122

Hansen, T. J., & Hodges, E. (2022b). Identifying transcription factor-bound activators and silencers in the chromatin accessible human genome using ATAC-STARR-seq (V2.1.0). *github.com/HodgesGenomicsLab/ATAC-STARR-seq*. doi:10.5281/zenodo.6640476

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol, 32*(4), 835-845. doi:10.1093/molbev/msv037

Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol, 16*(3), 144-154. doi:10.1038/nrm3949

Herz, H. M. (2016). Enhancer deregulation in cancer and other diseases. *Bioessays, 38*(10), 1003-1015. doi:10.1002/bies.201600106

Hill, M. S., Vande Zande, P., & Wittkopp, P. J. (2020). Molecular and evolutionary processes generating variation in gene expression. *Nat Rev Genet*. doi:10.1038/s41576-020-00304-w

Hirai, H., Imai, H., & Go, Y. (2012). *Post-genome biology of primates*. Tokyo ; New York: Springer.

Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., . . . Hannon, G. J. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell, 44*(1), 17-28. doi:10.1016/j.molcel.2011.08.026

Hubisz, M. J., & Pollard, K. S. (2014). Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. *Curr Opin Genet Dev, 29*, 15-21. doi:10.1016/j.gde.2014.07.005

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., . . . Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res, 27*(1), 38-52. doi:10.1101/gr.212092.116

International HapMap, C. (2003). The International HapMap Project. *Nature, 426*(6968), 789-796. doi:10.1038/nature02168

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., . . . D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res, 48*(D1), D498-D503. doi:10.1093/nar/gkz1031

Johnson, G. D., Barrera, A., McDowell, I. C., D'Ippolito, A. M., Majoros, W. H., Vockley, C. M., . . . Reddy, T. E. (2018). Human genome-wide measurement of drug-responsive regulatory activity. *Nat Commun, 9*(1), 5317. doi:10.1038/s41467-018-07607-x

Kelley, J. L., & Gilad, Y. (2020). Effective study design for comparative functional genomics. *Nat Rev Genet, 21*(7), 385-386. doi:10.1038/s41576-020-0242-z

Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., & Jones, P. A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res, 22*(12), 2497-2506. doi:10.1101/gr.143008.112

Kim, S., Yu, N. K., & Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med, 47*, e166. doi:10.1038/emm.2015.33

Kim, Y. S., Johnson, G. D., Seo, J., Barrera, A., Cowart, T. N., Majoros, W. H., . . . Reddy, T. E. (2021). Correcting signal biases and detecting regulatory elements in STARR-seq data. *Genome Res*. doi:10.1101/gr.269209.120

Kimble, J., & Hirsh, D. (1979). The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. *Dev Biol, 70*(2), 396-417. doi:10.1016/0012-1606(79)90035-6

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science, 188*(4184), 107-116. doi:10.1126/science.1090005

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., . . . Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun, 10*(1), 3583. doi:10.1038/s41467-019-11526-w

Klein, J. C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., . . . Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods, 17*(11), 1083-1091. doi:10.1038/s41592-020-0965-y

Klein, J. C., Keith, A., Agarwal, V., Durham, T., & Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. *Genome Biol, 19*(1), 99. doi:10.1186/s13059-018-1473-6

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet, 20*(4), 207-220. doi:10.1038/s41576-018-0089-8

Kreibich, E., Kleinendorst, R., Barzaghi, G., Kaspar, S., & Krebs, A. R. (2023). Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. *Mol Cell*. doi:10.1016/j.molcel.2023.01.017

Lamarre, S., Frasse, P., Zouine, M., Labourdette, D., Sainderichin, E., Hu, G., . . . Maza, E. (2018). Optimization of an RNA-Seq Differential Gene Expression Analysis Depending on Biological Replicate Number and Library Size. *Front Plant Sci, 9*, 108. doi:10.3389/fpls.2018.00108

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Larsson, J. (2021). eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. Retrieved from https://CRAN.R-project.org/package=eulerr

Ledford, H. (2008). Human genes are multitaskers. *Nature, 456*(7218), 9. doi:10.1038/news.2008.1199

Lee, C. M., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J. N., . . . Kent, W. J. (2020). UCSC Genome Browser enters 20th year. *Nucleic Acids Res, 48*(D1), D756-D761. doi:10.1093/nar/gkz1012

Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., . . . Gerstein, M. (2020). STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol, 21*(1), 298. doi:10.1186/s13059-020-02194-x

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, X. C., & Fay, J. C. (2017). Cis-Regulatory Divergence in Gene Expression between Two Thermally Divergent Yeast Species. *Genome Biol Evol, 9*(5), 1120-1129. doi:10.1093/gbe/evx072

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics, 30*(7), 923-930. doi:10.1093/bioinformatics/btt656

Lin, L., Shen, S., Jiang, P., Sato, S., Davidson, B. L., & Xing, Y. (2010). Evolution of alternative splicing in primate brain transcriptomes. *Hum Mol Genet, 19*(15), 2958-2973. doi:10.1093/hmg/ddq201

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., . . . Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature, 478*(7370), 476-482. doi:10.1038/nature10530

Liu, H., Zhu, F., Yong, J., Zhang, P., Hou, P., Li, H., . . . Deng, H. (2008). Generation of induced pluripotent stem cells from adult rhesus monkey fibroblasts. *Cell Stem Cell, 3*(6), 587-590. doi:10.1016/j.stem.2008.10.014

Liu, X., Li, Y. I., & Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell, 177*(4), 1022-1034 e1026. doi:10.1016/j.cell.2019.04.014

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell, 167*(5), 1170-1187. doi:10.1016/j.cell.2016.09.018

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Lynch, V. J., Nnamani, M. C., Kapusta, A., Brayer, K., Plaza, S. L., Mazur, E. C., . . . Wagner, G. P. (2015). Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep, 10*(4), 551-561. doi:10.1016/j.celrep.2014.12.052

Maricque, B. B., Dougherty, J. D., & Cohen, B. A. (2017). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res, 45*(4), e16. doi:10.1093/nar/gkw942

Mattioli, K., Oliveros, W., Gerhardinger, C., Andergassen, D., Maass, P. G., Rinn, J. L., & Mele, M. (2020). Cis and trans effects differentially contribute to the evolution of promoters and enhancers. *Genome Biol, 21*(1), 210. doi:10.1186/s13059-020-02110-3

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., . . . Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science, 337*(6099), 1190-1195. doi:10.1126/science.1222794

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., . . . Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol, 28*(5), 495-501. doi:10.1038/nbt.1630

McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., & Wittkopp, P. J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res, 20*(6), 816-825. doi:10.1101/gr.102491.109

Meiklejohn, C. D., Coolon, J. D., Hartl, D. L., & Wittkopp, P. J. (2014). The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Res, 24*(1), 84-95. doi:10.1101/gr.156414.113

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., . . . Mikkelsen, T. S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol, 30*(3), 271-277. doi:10.1038/nbt.2137

Metzger, B. P., Duveau, F., Yuan, D. C., Tryban, S., Yang, B., & Wittkopp, P. J. (2016). Contrasting Frequencies and Effects of cis- and trans-Regulatory Mutations Affecting Gene Expression. *Mol Biol Evol, 33*(5), 1131-1146. doi:10.1093/molbev/msw011

Metzger, B. P. H., Wittkopp, P. J., & Coolon, J. D. (2017). Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species. *Genome Biol Evol, 9*(4), 843-854. doi:10.1093/gbe/evx035

Mittleman, B. E., Pott, S., Warland, S., Barr, K., Cuevas, C., & Gilad, Y. (2021). Divergence in alternative polyadenylation contributes to gene regulatory differences between humans and chimpanzees. *Elife, 10*. doi:10.7554/eLife.62548

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., . . . Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature, 583*(7818), 699-710. doi:10.1038/s41586-020-2493-4

Morgan, M. A. J., & Shilatifard, A. (2020). Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. *Nat Genet, 52*(12), 1271-1281. doi:10.1038/s41588-020-00736-4

Mountjoy, E., Schmidt, E. M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., . . . Ghoussaini, M. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet, 53*(11), 1527-1533. doi:10.1038/s41588-021-00945-5

Muerdter, F., Boryn, L. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., . . . Stark, A. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat Methods, 15*(2), 141-149. doi:10.1038/nmeth.4534

Muhe, J., & Wang, F. (2015). Non-human Primate Lymphocryptoviruses: Past, Present, and Future. *Curr Top Microbiol Immunol, 391*, 385-405. doi:10.1007/978-3-319-22834-1_13

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., . . . Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics, 28*(14), 1919-1920. doi:10.1093/bioinformatics/bts277

Neumayr, C., Pagani, M., Stark, A., & Arnold, C. D. (2019). STARR-seq and UMI-STARR-seq: Assessing Enhancer Activities for Genome-Wide-, High-, and Low-Complexity Candidate Libraries. *Curr Protoc Mol Biol, 128*(1), e105. doi:10.1002/cpmb.105

Osada, N., Miyagi, R., & Takahashi, A. (2017). Cis- and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. *Genetics, 206*(4), 2139-2148. doi:10.1534/genetics.117.201459

Pang, B., & Snyder, M. P. (2020). Systematic identification of silencers in human cells. *Nat Genet, 52*(3), 254-263. doi:10.1038/s41588-020-0578-5

Pang, B., van Weerd, J. H., Hamoen, F. L., & Snyder, M. P. (2022). Identification of non-coding silencer elements and their regulation of gene expression. *Nat Rev Mol Cell Biol*. doi:10.1038/s41580-022-00549-9

Panigrahi, A., & O'Malley, B. W. (2021). Mechanisms of enhancer action: the known and the unknown. *Genome Biol, 22*(1), 108. doi:10.1186/s13059-021-02322-1

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., . . . Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol, 30*(3), 265-270. doi:10.1038/nbt.2136

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol, 27*(12), 1173-1175. doi:10.1038/nbt.1589

Peng, L., Li, E. M., & Xu, L. Y. (2020). From start to end: Phase separation and transcriptional regulation. *Biochim Biophys Acta Gene Regul Mech, 1863*(12), 194641. doi:10.1016/j.bbagrm.2020.194641

Phanstiel, D. H., Boyle, A. P., Araya, C. L., & Snyder, M. P. (2014). Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics, 30*(19), 2808-2810. doi:10.1093/bioinformatics/btu379

Picelli, S., Bjorklund, A. K., Reinius, B., Sagasser, S., Winberg, G., & Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res, 24*(12), 2033-2040. doi:10.1101/gr.177881.114

Planes, R., Pinilla, M., Santoni, K., Hessel, A., Passemar, C., Lay, K., . . . Meunier, E. (2022). Human NLRP1 is a sensor of pathogenic coronavirus 3CL proteases in lung epithelial cells. *Mol Cell, 82*(13), 2385-2400 e2389. doi:10.1016/j.molcel.2022.04.033

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res, 20*(1), 110-121. doi:10.1101/gr.097857.109

Ptashne, M. (1967). Specific binding of the lambda phage repressor to lambda DNA. *Nature, 214*(5085), 232-234. doi:10.1038/214232a0

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. *Nature, 322*(6081), 697-701. doi:10.1038/322697a0

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature, 470*(7333), 279-283. doi:10.1038/nature09692

Ramirez, F., Ryan, D. P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., . . . Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res, 44*(W1), W160-165. doi:10.1093/nar/gkw257

Rangan, S. R., Martin, L. N., Bozelka, B. E., Wang, N., & Gormus, B. J. (1986). Epstein-Barr virus-related herpesvirus from a rhesus monkey (Macaca mulatta) with malignant lymphoma. *Int J Cancer, 38*(3), 425-432. doi:10.1002/ijc.2910380319

Reilly, S. K., & Noonan, J. P. (2016). Evolution of Gene Regulation in Humans. *Annu Rev Genomics Hum Genet, 17*, 45-67. doi:10.1146/annurev-genom-090314-045935

Rickels, R., Herz, H. M., Sze, C. C., Cao, K., Morgan, M. A., Collings, C. K., . . . Shilatifard, A. (2017). Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet, 49*(11), 1647-1653. doi:10.1038/ng.3965

Rickels, R., & Shilatifard, A. (2018). Enhancer Logic and Mechanics in Development and Disease. *Trends Cell Biol, 28*(8), 608-630. doi:10.1016/j.tcb.2018.04.003

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi:10.1038/nature14248

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi:10.1038/nature14248

Sadowski, I., Ma, J., Triezenberg, S., & Ptashne, M. (1988). GAL4-VP16 is an unusually potent transcriptional activator. *Nature, 335*(6190), 563-564. doi:10.1038/335563a0

Sambrook, J., & Russell, D. W. (2006). Standard ethanol precipitation of DNA in microcentrifuge tubes. *CSH Protoc, 2006*(1). doi:10.1101/pdb.prot4456

Sandmann, L., & Ploss, A. (2013). Barriers of hepatitis C virus interspecies transmission. *Virology, 435*(1), 70-80. doi:10.1016/j.virol.2012.09.044

Santiago-Algarra, D., Dao, L. T. M., Pradel, L., Espana, A., & Spicuglia, S. (2017). Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res, 6*, 939. doi:10.12688/f1000research.11581.1

Schaffner, W. (2015). Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. *Biol Chem, 396*(4), 311-327. doi:10.1515/hsz-2014-0303

Schurch, N. J., Schofield, P., Gierlinski, M., Cole, C., Sherstnev, A., Singh, V., . . . Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA, 22*(6), 839-851. doi:10.1261/rna.053959.115

Shi, X., Ng, D. W., Zhang, C., Comai, L., Ye, W., & Chen, Z. J. (2012). Cis- and trans-regulatory divergence between progenitor species determines gene-expression novelty in Arabidopsis allopolyploids. *Nat Commun, 3*, 950. doi:10.1038/ncomms1954

Shibata, Y., Sheffield, N. C., Fedrigo, O., Babbitt, C. C., Wortham, M., Tewari, A. K., . . . Crawford, G. E. (2012). Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet, 8*(6), e1002789. doi:10.1371/journal.pgen.1002789

Sholtis, S. J., & Noonan, J. P. (2010). Gene regulation and the origins of human biological uniqueness. *Trends Genet, 26*(3), 110-118. doi:10.1016/j.tig.2009.12.009

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res, 15*(8), 1034-1050. doi:10.1101/gr.3715005

Signor, S. A., & Nuzhdin, S. V. (2018). The Evolution of Gene Expression in cis and trans. *Trends Genet, 34*(7), 532-544. doi:10.1016/j.tig.2018.03.007

Simonti, C. N., Pavlicev, M., & Capra, J. A. (2017). Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Mol Biol Evol, 34*(11), 2856-2869. doi:10.1093/molbev/msx219

Smith, E., & Shilatifard, A. (2014). Enhancer biology and enhanceropathies. *Nat Struct Mol Biol, 21*(3), 210-219. doi:10.1038/nsmb.2784

Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M., & Zaret, K. S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell, 161*(3), 555-568. doi:10.1016/j.cell.2015.03.017

Su, M., Han, D., Boyd-Kirkup, J., Yu, X., & Han, J. J. (2014). Evolution of Alu elements toward enhancers. *Cell Rep, 7*(2), 376-385. doi:10.1016/j.celrep.2014.03.011

Sulston, J. E., & Horvitz, H. R. (1977). Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev Biol, 56*(1), 110-156. doi:10.1016/0012-1606(77)90158-0

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol, 100*(1), 64-119. doi:10.1016/0012-1606(83)90201-4

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., . . . Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res, 24*(12), 1963-1976. doi:10.1101/gr.168872.113

Sundaram, V., & Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos Trans R Soc Lond B Biol Sci, 375*(1795), 20190347. doi:10.1098/rstb.2019.0347

Takahasi, K. R., Matsuo, T., & Takano-Shimizu-Kouno, T. (2011). Two types of cis-trans compensation in the evolution of transcriptional regulation. *Proc Natl Acad Sci U S A, 108*(37), 15276-15281. doi:10.1073/pnas.1105814108

The ENCODE Project Consortium, Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., . . . Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature, 583*(7818), 699-710. doi:10.1038/s41586-020-2493-4

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., . . . Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature, 489*(7414), 75-82. doi:10.1038/nature11232

Tirosh, I., Reikhav, S., Levy, A. A., & Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science, 324*(5927), 659-662. doi:10.1126/science.1169766

Tosato, G., & Cohen, J. I. (2007). Generation of Epstein-Barr Virus (EBV)-immortalized B cell lines. *Curr Protoc Immunol, Chapter 7*, Unit 7 22. doi:10.1002/0471142735.im0722s76

Triezenberg, S. J., LaMarco, K. L., & McKnight, S. L. (1988). Evidence of DNA: protein interactions that mediate HSV-1 immediate early gene activation by VP16. *Genes Dev, 2*(6), 730-742. doi:10.1101/gad.2.6.730

Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., . . . Brown, C. D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res, 27*(10), 1623-1633. doi:10.1101/gr.218149.116

Uebbing, S., Gockley, J., Reilly, S. K., Kocher, A. A., Geller, E., Gandotra, N., . . . Noonan, J. P. (2021). Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc Natl Acad Sci U S A, 118*(2). doi:10.1073/pnas.2007049118

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., . . . Ponten, F. (2015). Tissue-based map of the human proteome. *Science, 347*(6220), 1260419. doi:10.1126/science.1260419

Uhlen, M., Karlsson, M. J., Zhong, W., Tebani, A., Pou, C., Mikes, J., . . . Brodin, P. (2019). A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science, 366*(6472). doi:10.1126/science.aax9198

Vande Zande, P., Hill, M. S., & Wittkopp, P. J. (2022). Pleiotropic effects of trans-regulatory mutations on fitness and gene expression. *Science, 377*(6601), 105-109. doi:10.1126/science.abj7185

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., . . . Stamatoyannopoulos, J. A. (2020). Global reference mapping of human transcription factor footprints. *Nature, 583*(7818), 729-736. doi:10.1038/s41586-020-2528-x

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., . . . Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell, 160*(3), 554-566. doi:10.1016/j.cell.2015.01.006

Vockley, C. M., Guo, C., Majoros, W. H., Nodzenski, M., Scholtens, D. M., Hayes, M. G., . . . Reddy, T. E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res, 25*(8), 1206-1214. doi:10.1101/gr.190090.115

Vosa, U., Claringbould, A., Westra, H. J., Bonder, M. J., Deelen, P., Zeng, B., . . . Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet, 53*(9), 1300-1310. doi:10.1038/s41588-021-00913-z

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., . . . Soranzo, N. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell, 182*(5), 1214-1231 e1211. doi:10.1016/j.cell.2020.08.008

Wang, J., Dai, X., Berry, L. D., Cogan, J. D., Liu, Q., & Shyr, Y. (2019). HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res, 47*(D1), D106-D112. doi:10.1093/nar/gky864

Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., . . . Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun, 9*(1), 5380. doi:10.1038/s41467-018-07746-1

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., . . . Yue, F. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol, 19*(1), 151. doi:10.1186/s13059-018-1519-9

Weiss, C. V., Harshman, L., Inoue, F., Fraser, H. B., Petrov, D. A., Ahituv, N., & Gokhman, D. (2021). The cis-regulatory effects of modern human-specific variants. *Elife, 10*. doi:10.7554/eLife.63713

Whalen, S., Inoue, F., Ryu, H., Fair, T., Markenscoff-Papadimitriou, E., Keough, K., . . . Pollard, K. S. (2023). Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron*. doi:10.1016/j.neuron.2022.12.026

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Retrieved from https://ggplot2.tidyverse.org

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature, 430*(6995), 85-88. doi:10.1038/nature02698

Wittkopp, P. J., Haerum, B. K., & Clark, A. G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet, 40*(3), 346-350. doi:10.1038/ng.77

Wolf, B. K., Zhao, Y., McCray, A., Hawk, W. H., Deary, L. T., Sugiarto, N. W., . . . Wang, X. (2023). Cooperation of chromatin remodeling SWI/SNF complex and pioneer factor AP-1 shapes 3D enhancer landscapes. *Nat Struct Mol Biol, 30*(1), 10-21. doi:10.1038/s41594-022-00880-x

Wu, D. Y., Kalpana, G. V., Goff, S. P., & Schubach, W. H. (1996). Epstein-Barr virus nuclear protein 2 (EBNA2) binds to a component of the human SNF-SWI complex, hSNF5/Ini1. *J Virol, 70*(9), 6020-6028. doi:10.1128/JVI.70.9.6020-6028.1996

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol, 21*(1), 22. doi:10.1186/s13059-020-1929-3

Yao, X., Lu, Z., Feng, Z., Gao, L., Zhou, X., Li, M., . . . Liu, J. (2022). Comparison of chromatin accessibility landscapes during early development of prefrontal cortex between rhesus macaque and human. *Nat Commun, 13*(1), 3883. doi:10.1038/s41467-022-31403-3

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS, 16*(5), 284-287. doi:10.1089/omi.2011.0118

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics, 31*(14), 2382-2383. doi:10.1093/bioinformatics/btv145

Zhang, T., Zhang, Z., Dong, Q., Xiong, J., & Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol, 21*(1), 45. doi:10.1186/s13059-020-01957-w

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J. P., & Wang, L. (2014). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics, 30*(7), 1006-1007. doi:10.1093/bioinformatics/btt730

Zhou, X., Cain, C. E., Myrthil, M., Lewellen, N., Michelini, K., Davenport, E. R., . . . Gilad, Y. (2014). Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol, 15*(12), 547. doi:10.1186/s13059-014-0547-3

Zhu, Y., Li, M., Sousa, A. M., & Sestan, N. (2014). XSAnno: a framework for building ortholog models in cross-species transcriptome comparisons. *BMC Genomics, 15*, 343. doi:10.1186/1471-2164-15-343