# CIS-REGULATORY ADAPTATIONS IN THE PRIMATE IMMUNE SYSTEM

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lauren Ashley Choate

May 2020

# CIS-REGULATORY ADAPTATIONS IN THE PRIMATE IMMUNE SYSTEM

Lauren Ashley Choate, Ph. D.

Cornell University 2020

Divergent patterns in gene expression play an essential role in the development of new traits between species. Accordingly, evolutionary changes in the cis-regulatory elements that cause gene expression differences occur more quickly than changes in protein-coding genes. Individual cis-regulatory elements undergo turnover and are frequently under selection between primates, and mutations in cis-regulatory elements can fine-tune a response to the environment by adjusting quantitative traits and their tissue specificity. The goal of this work is to further understand the tuning of gene regulation between primate species and within populations and what functional consequences there may be.

Transcription factors are a primary determinant of regulatory interactions and therefore changes in transcription factor binding sites often contribute to changes in gene regulation. In order to determine how transcription factor binding patterns are evolutionary constrained in different tissues among humans, we developed a machine learning algorithm, dTOX, to predict transcription factor occupancy patterns based on DNase-I-seq, and used it to create an atlas of transcription factor binding across over a hundred human tissues. We found evidence of different evolutionary rates in the binding sites occupied by transcription factors among tissues, specifically between embryonic

and adult tissues. Our results show that evolution in transcription factor binding mirrors the tissue-driven model of protein-coding gene evolution.

By closely examining examples of cis-regulatory changes that contribute to a measurable difference in a trait in primate species, we can further understand the mechanisms of evolutionary change. We use the differential expression and regulation of the anthrax toxin receptor, *ANTXR2*, as a case study to understand how natural selection, the environment, and the regulatory landscape have contributed to a functional difference in the immune system of primates. We found evidence of a change in the cis-regulatory landscape of *ANTXR2*, signatures of selection between human populations, and population-specific polymorphisms that may contribute to differences in sensitivity to anthrax disease. Our results demonstrate how a regulatory change that was likely influenced by historical host-pathogen interactions has had lasting effects on immunity today both between species and within human populations.

# BIOGRAPHICAL SKETCH

Lauren was born in St. Louis, Missouri in 1991. She was raised in the town of Farmington, Missouri. For college, she attended Truman State University in Kirksville, Missouri. At Truman State, Lauren worked under Dr. Brent Buckner and studied the evolution of maize genome structure for four years. In 2014, she graduated summa cum laude with a Bachelor of Science in Biology. After developing a love for research as an undergraduate, Lauren moved to Cornell University to pursue a Ph.D. in Genetics, Genomics, and Development. She joined Dr. Charles Danko's lab to study gene regulation. In the Danko lab, Lauren focused on the role of gene regulation in the evolution of primate immune systems. She divided her time between doing experiments to understand the differences in regulation between immune cells in humans, chimpanzees, rhesus macaques, and baboons and developing computational algorithms to make predictions about gene regulation based on the data she generated.

I would like to dedicate my thesis to my family for their unwavering support.

# ACKNOWLEDGMENTS

First, I would like to thank my PI, Dr. Charles Danko for his continued support since the first day I started in his lab. I could not have chosen a better lab to do my graduate work in. Charles' never-ending positivity and scientific curiosity have shaped me as a scientist. I would like to thank Dr. Andrew Clark and Dr. Elia Tait Wojno for challenging me to think in an inter-disciplinary manner as members of my thesis committee. I would like to thank each member of the Danko lab for making my time at Cornell a terrific experience, even during rough days in the lab. In particular, I would like to thank my office mates, Shao-Pei Chou, Paul Munn, Dr. Zhong Wang, and Dr. Tinyi Chu for all of their help throughout the years and for making going into lab each day enjoyable. I thank Ed Rice, Gilad Barshad, and Pierce McMahon for all of their help with experiments throughout my time in graduate school. Finally, I would like to thank my family for all of their support. My parents, Karen and Bill Choate, and my grandfather, Earl Freitag have helped foster my love for science and always believed in me. I would like to thank Eric Fuemmeler for love and support and for being there every step of the way.

This work was supported through National Institutes of Health R01 grants R01HG009309-01 and 1R01HG010346-01 to Dr. Charles Danko, NIH F31 fellowship 1F31AI140050-01A1 to Lauren Choate, NIH training grant 5T32GM007617 to Genetics, Genomics, and Development, the Cornell University Center for Vertebrate Genomics scholar grant, and the Baker Institute of Animal Health.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# LIST OF ABBREVIATIONS AND SYMBOLS

| | |
|---|---|
| ANTXR2 | Anthrax toxin receptor gene 2 |
| CEU | European (from Utah) HapMap population |
| ChIP-seq | Chromatin immunoprecipitation sequencing |
| CHB | Han Chinese (from Beijing) HapMap population |
| CRE | Cis-regulatory element |
| CRISPR | Clustered regularly interspaced short palindromic repeats |
| DNase-I-seq | DNase-I hypersensitive sequencing |
| dTOX | Discriminative transcription factor occupancy extraction |
| eRNA | Enhancer RNA |
| E[A] | Expected number of adaptive substitutions |
| E[W] | Expected number of weakly deleterious polymorphic sites |
| EF | Edema factor |
| ELISA | Enzyme-linked immunosorbent assay |
| ENCODE | Encyclopedia of DNA Elements |
| Fst | Fixation index |
| GWAS | Genome-wide association study |
| HapMap | Haplotype map of the human genome |
| Hi-C | Chromosome conformation capture sequencing |
| INDEL | Insertion or deletion |
| INSIGHT | Program to measure the influence of natural selection |
| JPT | Japanese (from Tokyo) HapMap population |

| | |
|---|---|
| lincRNA | long intergenic non-coding RNA |
| LF | lethal factor |
| PBMC | Peripheral blood mononuclear cell |
| PRO-seq | Precision run-on and sequencing |
| PA | Protective antigen |
| Rho | Fraction of nucleotides under selection |
| RNA-seq | RNA sequencing |
| RTFBSDB | Transcription factor binding site identification |
| SNP | Single nucleotide polymorphism |
| SVM | Support vector machine |
| TF | Transcription factor |
| TAD | Topologically associated domain |
| TFBS | Transcription factor binding site |
| TIR | Transcription initiation site |
| TSS | Transcription start site |
| TU | Transcriptional unit |
| UMAP | Uniform manifold approximation and projection |
| VCF | Variant call format |
| YRI | Yoruban (from Nigeria) HapMap population |

# CHAPTER 1

# INTRODUCTION

A human genome contains approximately three billion bases of DNA, which contain all the information needed for its development and survival. The central dogma of molecular biology (Crick, 1970) describes how DNA is transcribed into RNA which is translated into proteins that enact all necessary functions in the cell. The concepts in the central dogma revolutionized the study of biology and genetics, but they only scratch the surface of the complex regulatory steps involved in turning the vast amount of information contained in DNA into functional products that can be used in a cell. Each step during the conversion of DNA into protein has its own forms of regulation, including (but certainly not limited to) what and when DNA is transcribed and which parts of the genome determine this (Aymoz et al., 2018; Lee & Young, 2013), the tuning of RNA stability and structure (Baker & Coller, 2006; Frye et al., 2018; Schaefer et al., 2017), and the modification of proteins that affect their turnover and ability to interact with other proteins (Duan & Walther, 2015; Johnson, 2009; Murn & Shi, 2017). The work in this thesis is centered on some of the first regulatory steps that are essential for controlling the process of transcription.

Multicellular organisms have the same genome in every cell of their body, yet these cells perform unique functions. The early regulatory steps in transcription control these distinct functionalities. The genome can be subdivided into different categories of functional elements that play unique roles in the process of transcribing DNA into RNA.

The broadest of these categories is coding versus non-coding DNA. Coding DNA contains discrete units called genes that encode the information needed to create RNAs. Non-coding DNA plays more of a regulatory role, containing elements that can that control the expression timing and levels of protein-coding genes. Non-coding functional elements play a key role in the process of gene regulation. In fact, ninety percent of disease-causing single nucleotide polymorphisms (SNP) are found in non-coding regions (Manolio et al., 2009). The process of gene expression takes the information contained in the DNA of protein-coding genes and converts it to functional products using the information contained in non-coding elements. The regulation of gene expression, or gene regulation, controls which genes are turned on or off in a given cell, how highly or lowly they are expressed, and the timing of their expression (Lelli et al., 2012).

## *KEY COMPONENTS OF EARLY GENE REGULATION*

In order for genes to be regulated, there are many essential components that need to be brought together to work as a complex functional unit. The DNA that encodes protein-coding genes needs to be accessible so that it can be regulated by proteins, the DNA needs to be positioned in such a way that non-coding elements are in close proximity with protein-coding genes, and transcriptional machinery needs to be recruited.

DNA is packaged into chromatin which prevents DNA damage, regulates gene expression, and helps segregate chromosomes during cell division (Kornberg, 1977;

Ruiz-Velasco & Zaugg, 2017; Widom, 1998). In order to be packaged into the chromatin structure, DNA is first wrapped around proteins called histones forming nucleosome units (Cutter & Hayes, 2015; Zhang et al., 2011). The nucleosome units are then wrapped into a more compact form, known as heterochromatin, which is tightly packed, making the DNA inaccessible to transcriptional machinery (Allshire & Madhani, 2018; Zhang et al., 2011). The level of compactness of chromatin, either when it is fully packed as heterochromatin or when it is more loosely packed as euchromatin, determines the ability of a region of the genome to be transcribed (Babu & Verma, 1987). Chemical modifications to histones give instructions about the transcriptional activity of region, which allows for chromatin binding proteins to regulate the level of compactness (Clapier & Cairns, 2009; Taverna et al., 2007). For example, methylation of histone 3's lysine residue 36 is often found in areas of the genome with gene bodies that are being transcribed (Kouzarides, 2007). The regulation of chromatin compactness allows for genes to be accessible and is an essential early step in gene expression.

At its most basic level, gene expression is dependent on the binding of proteins called transcription factors, which control many steps involved in transcription including the recruitment of transcriptional machinery (Spitz & Furlong, 2012). Transcription factors bind to two classes of non-coding DNA elements: those close to the transcription start and those that are more distally located in the genome. These non-coding elements are known as regulatory elements and contain conserved sequences, known as motifs, that can be recognized by transcription factors. Transcription factor binding to these distal regulatory elements starts a cascade of factor recruitment that results in transcription

(Lelli et al., 2012). Distal regulatory elements vary based on their function to either increase gene expression levels (enhancers), decrease gene expression levels (silencers), or buffer expression levels depending on environment cues (insulators) (Ong & Corces, 2011). Enhancers are capable of activating transcription independent of their distance from their target gene or orientation (Long et al., 2016). In order for enhancers to control the regulation of a gene that is not physically close in the nucleus, the chromatin must loop to bring them into contact.

The three-dimensional architecture of chromatin is organized based on transcriptional activity level, with localized regions of active euchromatin at the center of the nucleus and heterochromatin remaining tightly coiled at the periphery (Lieberman-Aiden et al., 2009). At a smaller scale, regions of the genome are organized into topologically associated domains (TADs) where there is an enrichment of DNA interactions within the domain compared to neighboring domains (Dixon et al., 2012; Nora et al., 2012). Genes and their corresponding enhancers are often found within the same TAD so that chromatin can easily loop to connect them (Robson et al., 2019). Exactly how enhancers are brought into contact with genes still remains unclear. The loop-extrusion model of TAD formation, in which a ring-shaped protein, cohesin, progressively extrudes chromatin loops until reaching boundary proteins at the edges of TADs, is one possibility for a scanning mechanism of connection between the two loci (Brackley et al., 2017; Nuebler et al., 2018). Another possibility is a model of phase separation where the disordered regions of associated proteins (such as transcription factors at enhancers and transcriptional machinery at promoters) form many weak interactions, promoting

the assembly of condensed physical structures where transcription can occur (Banani et al., 2017; Hnisz et al., 2017).

Once the transcription factors bound at enhancers are brought into close enough proximity with their target genes, they begin a cascade of recruitment. Sequence-specific transcription factors recruit the general transcription factors, which bind to the core promoter element within proximal regulatory elements near the transcription start site (Sainsbury et al., 2015; Woychik & Hampsey, 2002). From here, the general transcription factors and coactivator proteins recruit RNA Polymerase II (Pol II) to form the preinitiation complex that directs Pol II to the transcription start site to begin transcribing the gene (Kadonaga, 2012; Rhee & Pugh, 2012). The process of transcription begins with promoter clearance, where Pol II moves from the promoter to the transcription start site (Luse, 2013). After transcribing a small number of nucleotides, Pol II pauses as an additional step of regulation and to ensure that repressive chromatin is not formed (Adelman & Lis, 2012; Gilchrist et al., 2010; Kwak et al., 2013). Pol II then escapes from pausing and moves into productive elongation throughout the gene body and is eventually terminated after transcription is complete (Jonkers & Lis, 2015; Kwak & Lis, 2013).

## *THE EVOLUTION OF GENE REGULATION*

The small effect sizes of mutations on the early steps of gene regulation can be amplified during the RNA life cycle to create divergent phenotypes between species. In fact, there

is a growing body of evidence that suggests that differences in the interaction of regulatory elements may play a key role in hybrid dysfunction, resulting in speciation (Mack & Nachman, 2017). Understanding these inter-species regulatory changes provides insight into how gene regulation is tuned over evolutionary time.

Multiple studies in various species show that expression differences between species are driven more frequently by mutations within cis-regulatory elements such as promoters and enhancers than at regions of protein-coding DNA (Ong & Corces, 2011; Pennacchio et al., 2013). Transcription factor binding sites show higher frequencies of mutations than coding sequences (Arbiza et al., 2013) and evolutionary changes in TF binding sites often contribute to changes in gene regulation (Kilpinen et al., 2013). Adaptive substitutions in the binding sites of transcription factors (Arbiza et al., 2013; McLean et al., 2011; Prabhakar et al., 2009; Rockman et al., 2005) and turnover of binding sites (Ballester et al., 2014; Bradley et al., 2010; Doniger & Fay, 2007; Schmidt et al., 2010; Zheng et al., 2010) play a role in the rate-limiting steps of transcriptional activation (Fuda et al., 2009). TF binding patterns have also been shown to vary greatly between primates and within humans, suggesting that changes occur rapidly during the course of evolution (Kasowski et al., 2010). Despite the high turnover rate of individual transcription factor motifs within cis-regulatory elements, their effect on gene expression differences is lower than expected (Cusanovich et al., 2014). This suggests that while transcription factor binding differences play a role in the evolution of gene regulation, other key steps in regulation also contribute to differences across species.

Changes in chromatin conformation and architecture over evolutionary time contribute to the changes in gene regulation between species. The insulator protein CTCF is essential for maintaining boundaries between topological domains (TAD), along with cohesin (Dixon et al., 2012). The boundaries of TADs and the associated CTCF binding site locations are relatively constrained across mammalian species (Schmidt et al., 2012). There are conflicting reports about the conservation of higher-order TADs between species. Some data suggests that large scale domains often remain intact as modules between species and that differences in chromatin organization between species often occur within TADs in the form of local sequence divergence (Vietri Rudan et al., 2015). Other data suggests the opposite, that lower order pairwise contacts are conserved between species (Eres et al., 2019). Whether the rearrangement of chromatin contacts occurs at large scale TADs or within TADs, it is clear that the shuffling of chromatin contacts over evolutionary time has consequences for gene expression levels. The chromatin contacts of promoters with differentially expressed genes are reorganized between species (Eres et al., 2019). Additionally, the evolutionary rate of changes in enhancer and promoters is correlated with the number of chromatin contacts within a locus (Danko et al., 2018).

There is a high rate of turnover at the level of individual cis-regulatory elements between species (Villar et al., 2015), yet gene expression remains relatively stable between species (Prescott et al., 2015). This occurs because of redundancy and compensation between groups of cis-regulatory elements that work together to control gene expression levels. Developmental genes are frequently regulated by ensembles of enhancers,

known as shadow enhancers, that have functional redundancy, likely to protect the required stable expression level from mutations in individual enhancers (Cannavò et al., 2016; Hong et al., 2008). In general, the higher the number of enhancers that control a gene, the more likely a gene is involved in disease pathogenesis (X. Wang & Goldstein, 2020). Genes that are regulated by a large number of enhancers have more stable levels of expression over evolutionary time, despite changes at the individual enhancers that regulate them between species (Danko et al., 2018). The converse pattern is also true with enhancers that control expression of the same gene being under less evolutionary constraint, as long as target gene expression remains the same (Danko et al., 2018).

The effects of evolutionary changes at various steps of gene regulation can be measured by comparing gene expression conservation between species. Comparative studies in primates show that the evolution of gene regulation plays a foundational role in the development of new traits and differentiating species in the primate lineage (Siepel & Arbiza, 2014). By analyzing gene expression levels between species, we can learn about the different forces of selection acting on the levels of gene expression in primates. Comparative studies of gene expression show that the expression level of genes is highly conserved among primate species (Perry et al., 2012; Somel et al., 2009). The variation in gene expression levels within a primate species explains most of the inter-species variation in gene expression, which is consistent with stabilizing selection (Romero et al., 2012). However, expression levels of all genes are not under stabilizing selection. Lineage-specific changes in gene expression are present in a significant proportion of genes (Enard et al., 2002; Gilad et al., 2006). Those genes with species-specific

expression levels show evidence of positive selection. Additionally, the conservation of expression level is also dependent on which tissue genes are being expressed in. For example, comparisons of humans with chimpanzees demonstrate that expression levels in the brain are much more stable than expression in the testes (Khaitovich et al., 2006).

### *RESEARCH QUESTIONS*

We can further understand the role of gene regulation in creating differences between species and individuals by learning about how different transcription factor binding patterns are evolutionary constrained in different tissues among humans. In chapter two, I will describe a computational tool I developed to predict transcription factor binding using functional genomic data such DNase-I-sequencing (DNase-I-seq). DNase-I-seq is more readily available than experimental data like Chromatin Immunoprecipitation sequencing (ChIP-seq) data, which is capable of directly measuring transcription factor binding. By being able to predict transcription factor binding using DNase-I-seq data, we can discover binding patterns in a diverse set of tissues and use this transcription factor atlas as a way to learn about the natural selection pressures that constrain these binding patterns. I show that the sequences of motifs bound by transcription factors are under different evolutionary constraints in different tissues. In particular, I find that occupied transcription factor binding sites in the immune system are evolving more rapidly than all other tissues. Thus, learning more about how evolution shapes transcription factor binding, a key step in the gene regulatory landscape, leads to new insights into how the evolution of gene regulation differs among tissues.

The work in chapter two demonstrates that the immune system is rapidly evolving at the level of transcription factor binding. This rapid evolution is due to the immune system's role of protecting an organism from the ever-evolving landscape of pathogens it encounters. Over evolutionary time, pathogens have remained a major selective pressure for animals (Fumagalli et al., 2011). In fact, immunity related genes are frequently the most overrepresented category of genes in genome-wide screens for positive selection in humans and other animals (Barreiro & Quintana-Murci, 2009; Kosiol et al., 2008). In chapter three, I will describe my work to characterize evolutionary differences in another important component of gene regulation, cis-regulatory elements, of human and non-human primate immune systems.

Mutations in cis-regulatory elements are a major driver of the evolution of new phenotypes. By closely examining examples of cis-regulatory changes that contribute to a measurable difference in a trait between and within primate species, we can further understand the mechanisms of evolutionary change. In chapter three, I use the differential expression and regulation of the anthrax toxin receptor, *ANTXR2*, as a case study of the evolution of gene regulation in response to dynamic host-pathogen interactions in primates. I hypothesize that increased exposure to the pathogen that causes anthrax disease in humans due to the advent of agriculture selected for the decreased *ANTXR2* expression present in humans compared to non-human primates. The goal of this work is to understand how natural selection, the environment, and the regulatory landscape have contributed to a functional difference in the immune system response to anthrax disease between species and within human populations. I find

evidence of a change in the cis-regulatory landscape of *ANTXR2* between humans and non-human primates, along with population-specific polymorphisms that may contribute to differences in sensitivity to anthrax disease. Together, my work highlights what evolutionary forces are shaping the immune system, how genes are regulated, and an example of the functional implications of these changes for humans today.

## *REFERENCES*


Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nature Reviews. Genetics, 13(10), 720–731. https://doi.org/10.1038/nrg3293

Allshire, R. C., & Madhani, H. D. (2018). Ten principles of heterochromatin formation and function. Nature Reviews Molecular Cell Biology, 19(4), 229–244. https://doi.org/10.1038/nrm.2017.119

Arbiza, L., Gronau, I., Aksoy, B. A., Hubisz, M. J., Gulko, B., Keinan, A., & Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. Nature Genetics, 45(7), 723–729. https://doi.org/10.1038/ng.2658

Aymoz, D., Solé, C., Pierre, J., Schmitt, M., de Nadal, E., Posas, F., & Pelet, S. (2018). Timing of gene expression in a cell-fate decision system. Molecular Systems Biology, 14(4), 1–13. https://doi.org/10.15252/msb.20178024

Babu, A., & Verma, R. S. (1987). Chromosome Structure: Euchromatin and Heterochromatin. International Review of Cytology, 108, 1–60.

Baker, K. E., & Coller, J. (2006). The many routes to regulating mRNA translation. Genome Biology, 7(12). https://doi.org/10.1186/gb-2006-7-12-332

Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A. J., Funnell, A. P. W., Goncalves, A., Kutter, C., Lukk, M., Menon, S., McLaren, W. M., Stefflova, K., Watt, S., Weirauch, M. T., Crossley, M., Marioni, J. C., … Wilson, M. D. (2014). Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. ELife, 3(October2014), 1–29. https://doi.org/10.7554/eLife.02626

Banani, S. F., Lee, H. O., Hyman, A. A., & Rosen, M. K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology, 18(5), 285–298. https://doi.org/10.1038/nrm.2017.7

Brackley, C. A., Johnson, J., Michieletto, D., Morozov, A. N., Nicodemi, M., Cook, P. R., & Marenduzzo, D. (2017). Nonequilibrium Chromosome Looping via Molecular Slip Links. Physical Review Letters, 119(13), 95–103. https://doi.org/10.1103/PhysRevLett.119.138101

Bradley, R. K., Li, X. Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., Tonkin, L. A., Biggin, M. D., & Eisen, M. B. (2010). Binding site turnover produces

pervasive quantitative changes in transcription factor binding between closely related drosophila species. PLoS Biology, 8(3). https://doi.org/10.1371/journal.pbio.1000343

Cannavò, E., Khoueiry, P., Garfield, D. A., Geeleher, P., Zichner, T., Gustafson, E. H., Ciglar, L., Korbel, J. O., & Furlong, E. E. M. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Current Biology, 26(1), 38–51. https://doi.org/10.1016/j.cub.2015.11.034

Clapier, C. R., & Cairns, B. R. (2009). The Biology of Chromatin Remodeling Complexes. Annual Review of Biochemistry, 78(1), 273–304. https://doi.org/10.1146/annurev.biochem.77.062706.153223

Crick, F. (1970). Central dogma of molecular biology. Nature, 227(5258), 561–563. https://doi.org/10.1038/227561a0

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., & Gilad, Y. (2014). The Functional Consequences of Variation in Transcription Factor Binding. PLoS Genetics, 10(3). https://doi.org/10.1371/journal.pgen.1004226

Cutter, A. R., & Hayes, J. J. (2015). A brief review of nucleosome structure. FEBS Letters, 589(20), 2914–2922. https://doi.org/10.1016/j.febslet.2015.05.016

Danko, C. G., Choate, L. A., Marks, B. A., Rice, E. J., Wang, Z., Chu, T., Martins, A. L., Dukler, N., Coonrod, S. A., Tait Wojno, E. D., Lis, J. T., Kraus, W. L., & Siepel, A. (2018). Dynamic evolution of regulatory element ensembles in primate CD4+T cells. Nature Ecology and Evolution. https://doi.org/10.1038/s41559-017-0447-5

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398), 376–380. https://doi.org/10.1038/nature11082

Doniger, S. W., & Fay, J. C. (2007). Frequent gain and loss of functional transcription factor binding sites. PLoS Computational Biology, 3(5), 0932–0942. https://doi.org/10.1371/journal.pcbi.0030099

Duan, G., & Walther, D. (2015). The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. PLoS Computational Biology, 11(2), 1–23. https://doi.org/10.1371/journal.pcbi.1004049

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R., Doxiadis, G. M., Bontrop, R. E., & Pääbo, S. (2002). Intra- and interspecific variation in primate gene expression patterns. Science, 296(5566), 340–343. https://doi.org/10.1126/science.1068996

Eres, I. E., Luo, K., Hsiao, C. J., Blake, L. E., & Gilad, Y. (2019). Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates. In PLoS Genetics (Vol. 15, Issue 7). https://doi.org/10.1371/journal.pgen.1008278

Frye, M., T. Haranda, B., Behm, M., & He, C. (2018). Expression During Development. Science, 361(September), 1346–1349.

Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature, 461(7261), 186–192. https://doi.org/10.1038/nature08449

Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P., & White, K. P. (2006). Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature, 440(7081), 242–245. https://doi.org/10.1038/nature04559

Gilchrist, D. A., Dos Santos, G., Fargo, D. C., Xie, B., Gao, Y., Li, L., & Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell, 143(4), 540–551. https://doi.org/10.1016/j.cell.2010.10.004

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., & Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. Cell, 169(1), 13–23. https://doi.org/10.1016/j.cell.2017.02.007

Hong, J. W., Hendrix, D. A., & Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. Science, 321(5894), 1314. https://doi.org/10.1126/science.1160631

Johnson, L. N. (2009). The regulation of protein phosphorylation. Biochemical Society Transactions, 37(4), 627–641. https://doi.org/10.1042/BST0370627

Jonkers, I., & Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nature Reviews Molecular Cell Biology, 16(3), 167–177. https://doi.org/10.1038/nrm3953

Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. Wiley Interdisciplinary Reviews: Developmental Biology, 1(1), 40–51. https://doi.org/10.1002/wdev.21

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M.-Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., & Snyder, M. (2010). Variation in transcription factor binding among humans. Science (New York, N.Y.), 328(5975), 232–235. https://doi.org/10.1126/science.1183621

Khaitovich, P., Enard, W., Lachmann, M., & Pääbo, S. (2006). Evolution of primate gene expression. Nature Reviews Genetics, 7(9), 693–702. https://doi.org/10.1038/nrg1940

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., … Dermitzakis, E. T. (2013). Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. Science, 342, 744–747. https://doi.org/10.1126/science.1242463

Kornberg, R. (1977). Structure of chromatin. Annual Review of Biochemistry, 46, 931–954. https://doi.org/10.1038/newbio229101a0

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell, 128(4), 693–705. https://doi.org/10.1016/j.cell.2007.02.005

Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. Science, 339(6122), 950–953. https://doi.org/10.1126/science.1229386

Kwak, H., & Lis, J. T. (2013). Control of Transcriptional Elongation. Annual Review of Genetics, 47(1), 483–508. https://doi.org/10.1146/annurev-genet-110711-155440

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. Cell, 152(6), 1237–1251. https://doi.org/10.1016/j.cell.2013.02.014

Lelli, K. M., Slattery, M., & Mann, R. S. (2012). Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. Annual Review of Genetics, 46(1), 43–68. https://doi.org/10.1146/annurev-genet-110711-155437

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, N.Y.), 326(5950), 289–293. https://doi.org/10.1126/science.1181369

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

Luse, D. S. (2013). Promoter clearance by RNA polymerase II. Biochimica et

Biophysica Acta - Gene Regulatory Mechanisms, 1829(1), 63–68. https://doi.org/10.1016/j.bbagrm.2012.08.010

Mack, K. L., & Nachman, M. W. (2017). Gene Regulation and Speciation. Trends in Genetics, 33(1), 68–80. https://doi.org/10.1016/j.tig.2016.11.003

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. In Nature (Vol. 461, Issue 7265, pp. 747–753). Nature Publishing Group. https://doi.org/10.1038/nature08494

McLean, C. Y., Reno, P. L., Pollen, A. A., Bassan, A. I., Capellini, T. D., Guenther, C., Indjeian, V. B., Lim, X., Menke, D. B., Schaar, B. T., Wenger, A. M., Bejerano, G., & Kingsley, D. M. (2011). Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature, 471(7337), 216–219. https://doi.org/10.1038/nature09774

Murn, J., & Shi, Y. (2017). The winding path of protein methylation research: Milestones and new frontiers. Nature Reviews Molecular Cell Biology, 18(8), 517–527. https://doi.org/10.1038/nrm.2017.35

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., Van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature, 485(7398), 381–385. https://doi.org/10.1038/nature11049

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proceedings of the National Academy of Sciences of the United States of America, 115(29), E6697–E6706. https://doi.org/10.1073/pnas.1717730115

Ong, C. T., & Corces, V. G. (2011). Enhancer function: New insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics, 12(4), 283–293. https://doi.org/10.1038/nrg2957

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. Nature Reviews Genetics, 14(4), 288–295. https://doi.org/10.1038/nrg3458

Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., Michelini, K., Zehr, S., Yoder, A. D., Stephens, M., Pritchard, J. K., & Gilad, Y. (2012). Comparative RNA sequencing reveals substantial genetic variation in endangered

primates. Genome Research, 22(4), 602–610. https://doi.org/10.1101/gr.130468.111

Prabhakar, S., Visel, A., Akiyama, J. A., Shoukry, M., Keith, D., Holt, A., Plajzer-frick, I., Morrison, H., Fitzpatrick, D. R., Pennacchio, L. A., Rubin, E. M., Noonan, J. P., Division, G., & Berkeley, L. (2009). Human-specific gain of function in a developmental enhancer. 321(5894), 1346–1350. https://doi.org/10.1126/science.1159974.Human-specific

Prescott, S. L., Srinivasan, R., Marchetto, M. C., Gage, F. H., Swigut, T., Selleri, L., Gage, F. H., Swigut, T., & Wysocka, J. (2015). Enhancer Divergence and cis - Regulatory Evolution in the Human and Chimp Neural Crest Article Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest. 68–83. https://doi.org/10.1016/j.cell.2015.08.036

Rhee, H. S., & Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature, 483(7389), 295–301. https://doi.org/10.1038/nature10799

Robson, M. I., Ringel, A. R., & Mundlos, S. (2019). Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. Molecular Cell, 74(6), 1110–1122. https://doi.org/10.1016/j.molcel.2019.05.032

Rockman, M. V., Hahn, M. W., Soranzo, N., Zimprich, F., Goldstein, D. B., & Wray, G. A. (2005). Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biology, 3(12), 1–12. https://doi.org/10.1371/journal.pbio.0030387

Romero, I. G., Ruvinsky, I., & Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics, 13(7), 505–516. https://doi.org/10.1038/nrg3229

Ruiz-Velasco, M., & Zaugg, J. B. (2017). Structure meets function: How chromatin organisation conveys functionality. Current Opinion in Systems Biology, 1(i), 129–136. https://doi.org/10.1016/j.coisb.2017.01.003

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nature Reviews Molecular Cell Biology, 16(3), 129–143. https://doi.org/10.1038/nrm3952

Schaefer, M., Kapoor, U., & Jantsch, M. F. (2017). Understanding RNA modifications: The promises and technological bottlenecks of the “epitranscriptome.” Open Biology, 7(5). https://doi.org/10.1098/rsob.170077

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonalves, Â., Kutter, C.,

Brown, G. D., Marshall, A., Flicek, P., & Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell, 148(1–2), 335–348. https://doi.org/10.1016/j.cell.2011.11.058

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., & Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. 1036(May), 1036–1040. https://doi.org/10.1126/science.1186176

Siepel, A., & Arbiza, L. (2014). Cis-regulatory elements and human evolution. Current Opinion in Genetics and Development, 29, 81–89. https://doi.org/10.1101/005652

Somel, M., Franz, H., Yan, Z., Lorenc, A., Guo, S., Giger, T., Kelso, J., Nickel, B., Dannemann, M., Bahn, S., Webster, M. J., Weickert, C. S., Lachmann, M., Pääbo, S., & Khaitovich, P. (2009). Transcriptional neoteny in the human brain. Proceedings of the National Academy of Sciences of the United States of America, 106(14), 5743–5748. https://doi.org/10.1073/pnas.0900544106

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. Nature Reviews Genetics, 13(9), 613–626. https://doi.org/10.1038/nrg3207

Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D., & Patel, D. J. (2007). How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. Nature Structural and Molecular Biology, 14(11), 1025–1040. https://doi.org/10.1038/nsmb1338

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Reports, 10(8), 1297–1309. https://doi.org/10.1016/j.celrep.2015.02.004

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M. A., Bertelsen, M. F., Murchison, E. P., Flicek, P., & Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. Cell, 160(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Wang, X., & Goldstein, D. B. (2020). Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. American Journal of Human Genetics, 106(2), 215–233. https://doi.org/10.1016/j.ajhg.2020.01.012

Widom, J. (1998). Structure, Dynamics, and Function of Chromatin in Vitro. Annual

Review of Biophysics and Biomolecular Structure, 27, 285–327.

Woychik, N. A., & Hampsey, M. (2002). The RNA polymerase II machinery: Structure illuminates function. Cell, 108(4), 453–463. https://doi.org/10.1016/S0092-8674(02)00646-3

Zhang, Z., Wippo, C. J., Wal, M., Ward, E., Philipp, K., & Pugh, F. (2011). A Packing Mechanism for Nucleosome Organization Reconstituted Across A Eukaryotic Genome. Science, 332(May), 977–980. https://doi.org/10.1126/science.1200508

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M., & Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. Nature, 464(7292), 1187–1191. https://doi.org/10.1038/nature08934

# CHAPTER 2

# SIGNALS OF NATURAL SELECTION ON IMPUTED TRANSCRIPTION FACTOR BINDING SITES IN HUMAN TISSUES


## *ABSTRACT*

**Summary:** Predicting transcription factor binding remains challenging due to high false positive rates, cell type specific differences in DNA recognition, and experimental bias. We developed a motif-based discriminative method, dTOX (discriminative Transcription factor Occupancy eXtraction), to predict transcription factor binding using a single data type-- DNase-I-seq. Our method is based on predicting if a given motif is bound by any TF with a motif at a given genomic position, which we term 'transcription factor occupancy'. We used dTOX to predict transcription factor binding across ENCODE cell types and among clusters of similar motifs. Based on the predicted bound TFBS, we used INSIGHT to detect signatures of recent natural selection across tissues. We found evidence of purifying selection in motifs bound in the brain, positive selection in motifs bound in blood cells, and an increase in selection of embryonic versus adult cells among all tissues. Evolution of occupied transcription factor motifs in regulatory elements follows similar patterns to the evolution of tissue-specific protein-coding genes.

**Availability and implementation:** dTOX is freely available on GitHub (https://github.com/Danko-Lab/dTOX).

## *INTRODUCTION*

Transcription factors (TF) coordinate complex transcriptional regulatory programs by binding to DNA sequence motifs and coordinating steps during the RNA Polymerase II (Pol II) transcription cycle (Fuda et al., 2009). Detecting transcription factor binding using experimental techniques is difficult, labor intensive, and depends on the availability and validation of high-quality antibodies. Developing accurate computational methods to predict TF binding sites has therefore become a major goal for the genomics community.

DNase-I hypersensitivity can contain patterns that are characteristic of particular transcription factors (Galas & Schmitz, 1978; Song et al., 2011). However, many transcription factors are not associated with characteristic footprints (Guertin et al., 2012; He et al., 2014). Additionally, most transcription factors bind only a small fraction of DNA sequences matching their motif (Guertin & Lis, 2010). Transcription factor families tend to bind similar DNA sequence motifs. There is some specific information that can be used to distinguish between particular proteins (Bulyk, 2006; Samee et al., 2019). However, motifs alone are often sub optimized and features such as interactions with different cofactors can vary the binding affinity of a transcription factor in different cell types (Farley et al., 2015; Luna-Zurita et al., 2016). Thus, there is a risk of overfitting models to one cell type. These characteristics of transcription factors make transcription factor binding site prediction a challenging computational problem.

Existing computational approaches largely work by enumerating the set of TF binding sites in the genome using a known motif for a TF of interest, and then predicting whether each motif occurrence is bound on the basis of chromatin accessibility data. Tools such as CENTIPEDE (Pique-Regi et al., 2011) and PIQ (Sherwood et al., 2014) have been introduced to predict TFBSs with reasonably high accuracy, allowing the prediction of virtually any TF with a known motif in any cell type with chromatin accessibility data. Although both of these tools have a high specificity, there is such a large excess of unbound motifs in mammalian genomes that the majority of predictions are false positives. The high false positive rate has limited the widespread deployment of these tools.

More recently, newer methods have focused on using more accurate discriminative machine learning models to predict the location of all ChIP-seq peaks, without conditioning on a specific motif of interest. Virtual ChIP-seq (Karimzadeh & Hoffman, 2018), an elegant example of this alternative strategy, predicts the location of motifs based on ChIP-seq training data with remarkably high accuracy. These tools have the advantage of being able to predict the location of ChIP-seq supported binding sites at loci which do not contain the motif of interest. However, there are two important limitations of this alternative strategy. First, the extent to which ChIP-seq signals without a canonical TF binding motif reflect biologically relevant events, or systematic artifacts due to crosslinking or antibody cross reactivity, remains debated (Steube et al., 2017; Wreczycka et al., 2017). Second, training a model for each TF in a manner that is independent of a motif requires having collected ChIP-seq data in at least one cell type

for model training, limiting the use of this strategy to ~200 of the estimated 1,800 TFs for which a validated antibody exists.

Here we describe dTOX (**d**iscriminative **T**ranscription **f**actor **O**ccupancy e**X**traction), a discriminative support vector machine (SVM) classifier that predicts whether motifs identified in a genome of interest are bound by a transcription factor on the basis of nuclease accessibility (DNase-I-seq). We introduce the concept of transcription factor occupancy, where we predict if a given motif is bound by any TF with a motif at the position. By predicting motif occupancy, we can avoid the usual high false positive signal due to cofactors or motif similarity between factors. To improve the classification accuracy over CENTIPEDE and PIQ, dTOX used a single large training dataset under the assumption that many TFs share a common pattern in functional marks, consisting of a nucleosome depleted core region flanked by divergent transcription initiation (Core et al., 2014; Scruggs et al., 2015). By using a single model trained on a dataset of many motifs, the training of the model can be done on millions of motif examples and is generalizable to any TF with motif data. We use this model to identify patterns of selection in TF binding sites between varying cell types and tissues to further understand the evolutionary role of transcription factors.

## *RESULTS*

### A single model predicts transcription factor motif occupancy

Whereas previous methods predict whether a motif occurrence is occupied by a specific transcription factor, dTOX predicts whether a motif occurrence is occupied by any member of a transcription factor family. dTOX was motivated by reports that DNA sequence and factor-general assays, like DNase-I-seq, provide weak information about which member of a transcription factor protein family is bound. For example, MYC and MAX are transcription factors in the bHLHz family that bind to an E-box consensus motif (Amati et al., 1992; Blackwood & Eisenman, 1991; Kretzner et al., 1992). Using ChIP-seq data in K562 cells, we found that MYC and MAX show similar binding patterns at genomic locations with an annotated motif for both MYC and MAX or an annotated motif for only one of the factors (**Supplementary Figure 2.1**). Likewise, MYC and MAX occupancy affected the pattern of DNase-I-seq accessibility near E-box occurrences in similar ways. To generalize beyond MYC and MAX, we trained a discriminative support vector machine (SVM) classifier that accurately predicted the occupancy of 34 transcription factors using DNase-I-seq data from a holdout cell type (see **Methods**). The factor-general SVM predicted motif occupancy as well as models trained to recognize individual transcription factors, indicating that different transcription factors often share similar patterns of DNase-I-seq cut sites.

Motivated by examples like MYC and MAX, we designed dTOX to classify candidate binding sites in a reference genome as bound or unbound by any member of a

transcription factor protein family which can recognize that motif (**Figure 2.1A**). dTOX takes as input the genomic coordinates of occurrences similar to CisBP binding motifs in the RTFBSDB database (Wang et al., 2016; Weirauch et al., 2014). dTOX classifies each motif occurrence as bound or unbound in a cell or tissue sample using DNase-I-seq, which is factor-general. To accelerate computation, dTOX excludes candidate motif occurrences from prediction for two reasons: (i) motifs that do not meet a signal cutoff threshold and therefore would not be predicted as bound by the classifier, and (ii) those which are classified as unbound using a random forest tuned to achieve nearly perfect sensitivity (see **Methods**). These candidate motif occurrences are classified as unbound. dTOX classifies remaining motif occurrences using a SVM implemented on a graphical processing unit (GPU) to speed up computation time.

**dTOX performs well on a holdout set**

We trained dTOX using ChIP-seq data profiling 50 sequence-specific transcription factors in K562 and GM12878 (**Supplementary Table 2.1**). We used motif occurrences that intersect a ChIP-seq peak call of a transcription factor that recognizes that motif and has a minimum signal threshold as the set of true positive training examples. Negative examples were defined as motifs that were outside of a ChIP-seq peak for all transcription factors reported to recognize that motif. We trained dTOX using an unbalanced dataset of 2 million motif positions, 30% of which were bound to a transcription factor (true positives). All arbitrary parameters used during dTOX training, for instance the shape and size of feature vectors and the proportion of bound and

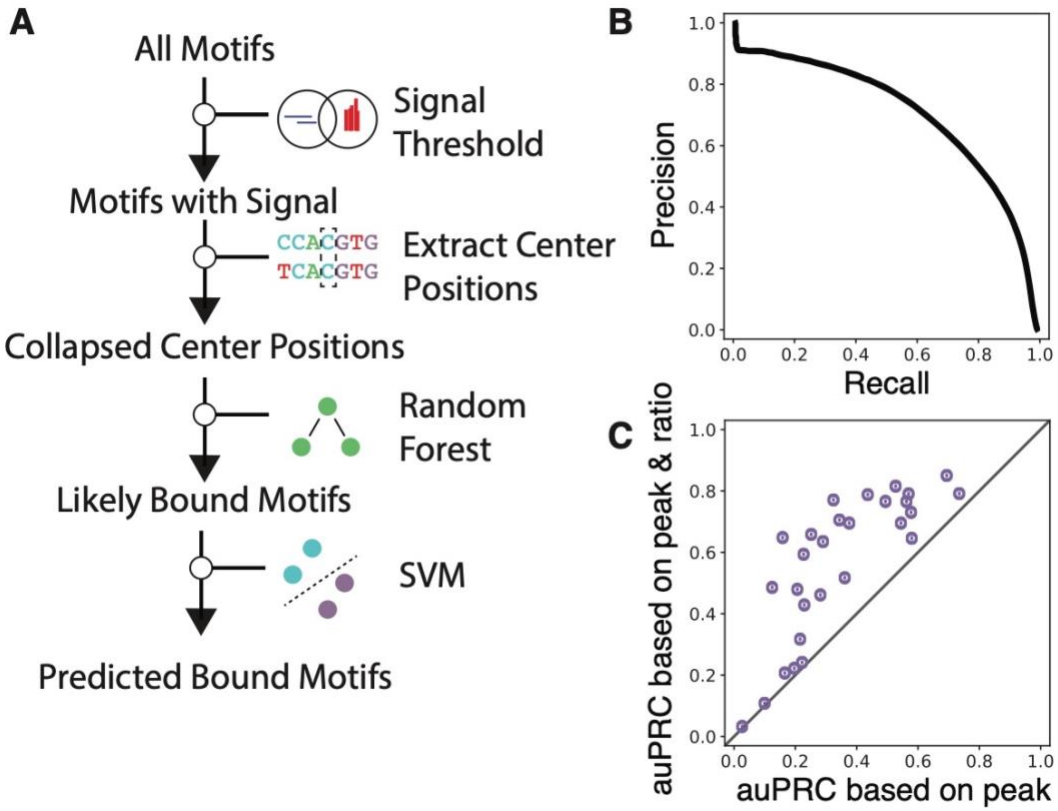



**Figure 2.1. dTOX prediction and performance. A)** Workflow of dTOX motif prediction reduces the total number of motifs that are predicted with the SVM which reduces total computation time. **B)** Precision-recall curve for all transcription factors in the HeLa holdout set based on a true positive set based on ChIP-seq peak calls and a signal to input ratio threshold. auPRC=0.78. **C)** Comparison of auPRC between true positive set based solely on peak calls for ChIP-seq and true positive set based on ChIP-seq peak calls and a signal to input ratio threshold for individual motifs in HeLa.

unbound motifs in the training dataset, were optimized empirically to maximize performance on holdout sites in K562 and GM12878 (see **Methods**).

We evaluated the accuracy of dTOX binding predictions on 34 ChIP-seq datasets in HeLa, a complete holdout cell type. As motif occupancy is highly unbalanced, with larger numbers of negative examples, we used the area under the precision recall curve (auPRC) as a performance metric. Using the HeLa holdout set, we found that the average auPRC for models based on all motifs was 0.78 (**Figure 2.1B; Supplementary Table 2.2**). Performance improved when we defined true bound samples using a combination of ChIP-seq peak and a minimum signal to input ratio similar to how we defined bound samples for the training set (**Figure 2.1C**; see **Methods**). For individual TFs, models trained on the union of all training TFs performed equally to models trained on only their own ChIP-seq and motif data, demonstrating the utility of our one model approach (**Supplementary Figure 2.2**). Notably, 8/34 HeLa TFs we predicted on were not included in the training set. The auPRC for TFs included in the training set compared to the auPRC for TFs not included in the training set was not statistically different (0.79 vs. 0.77). Thus, dTOX gives a good estimate of TF binding and performs better than existing motif-based models.

**Selection on transcription factor binding sites across tissues**

Motifs are not discrete units and there is often redundancy between motifs with similar binding partners and members of large motif families (Kuntz et al., 2012). We defined motif family clusters in the human genome by merging TFBS within composite DNase-

I hypersensitive sites (from >200 reference cell types) and clustering based on location in the genome. Using this method, we found 447 clusters of TF families (**Figure 2.2A**) with clusters ranging from 1 to 77 individual motifs. By combining predictions of TF binding using our occupancy method with clustering results of TF motif families, we can make new observations about the patterns of TF family binding across cell types.

We assessed TF binding across cell types using ENCODE DNase-I-seq data from diverse cell types that were consolidated based on correlation between similar datasets and UMAP clustering (**Figure 2.2B**). We predicted TF binding on 118 DNase-I-seq datasets on tissues including lung, kidney, blood, brain, dermis, stomach, large intestine, small intestine, spinal cord, adrenal gland, and thymus.

We then used this large dataset of TF binding predictions to not only understand the differences in TF binding across tissues, but also to understand the evolutionary constraints in place on conserved and tissue-specific TF binding sites. An analysis of natural selection on TF binding sites bound in a limited number of cancer cell types in the human genome showed that deleterious mutations in regulatory regions occur at higher frequencies than in coding regions (Arbiza et al., 2013). To extend this analysis to motifs that are bound and look at patterns of selection across tissues, we ran INSIGHT (Gronau et al., 2013) to detect signatures of recent natural selection in our predicted bound TF binding sites based on the ENCODE DNase-I-seq data. INSIGHT detects three signatures of selection: the fraction of nucleotides under selection (Rho), the

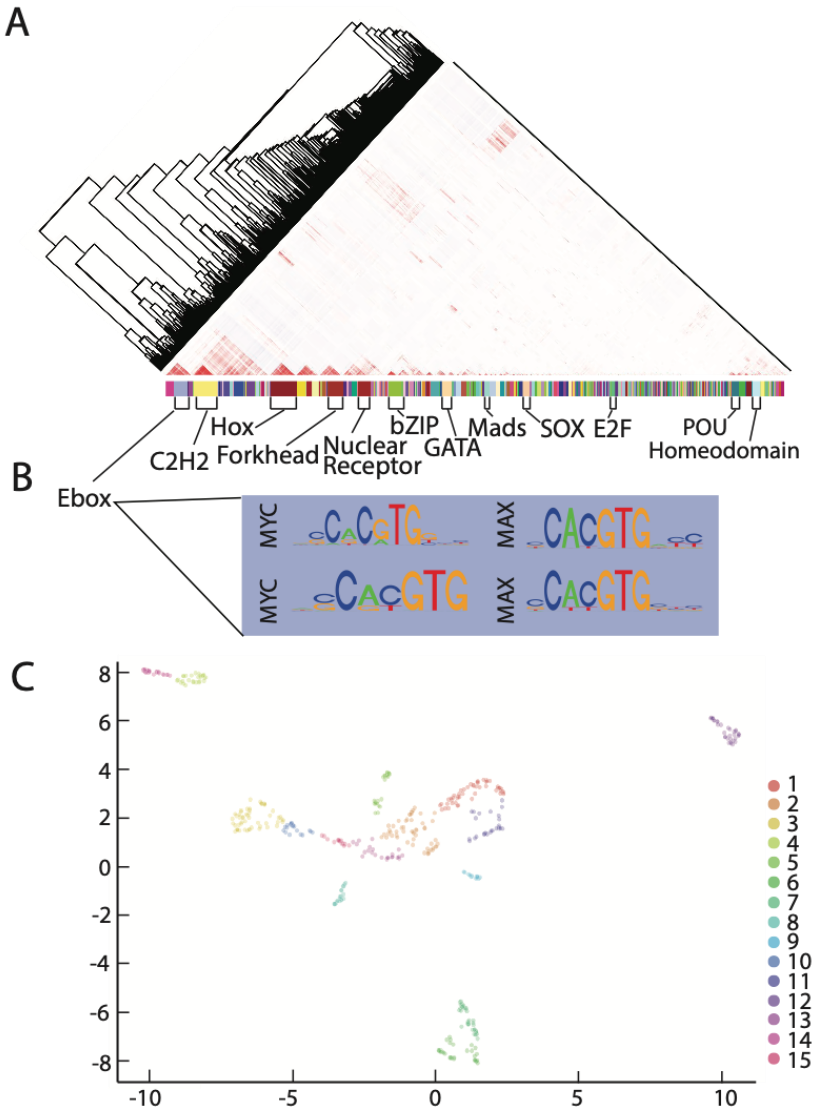



**Figure 2.2. Clustering of motifs and ENCODE DNase-I-seq data. A)** Clustering of 1777 PWMs results in 447 clusters. Clusters represent different enrichments of functional classes of transcription factors. **B)** Motifs for MYC and MAX are members of cluster 2 (Ebox). **C)** UMAP clustering ENCODE DNase-I-seq data with 15 clusters.

expected number of adaptive substitutions per kilobase (E[A]), and the expected number of polymorphic sites subject to weak negative selection per kilobase (E[W]).

We evaluated predicted TF occupancy across all motif clusters for each tissue (**Figure 2.3A-C**) and as a matrix of each individual cluster for each tissue (**Figure 2.3D-F**). We find that DNase-I-seq samples from brain tissues have a significantly higher fraction of nucleotides under selection with an average of value of 0.21, compared to an average Rho value of 0.16 for all other tissues ($p<0.0001$) (**Figure 2.3A**). Many samples had little to no evidence of adaptive substitutions (**Figure 2.3B**). However, blood samples had a significantly higher expected number of adaptive substitutions than other tissues (0.05/kB vs. 0.007/kB, $p<0.05$). All tissues had a greater number of polymorphic sites under weak negative selection per kilobase compared to expected adaptive substitutions (**Figure 2.3C**). Samples from the stomach had the greatest number of expected deleterious polymorphic sites per kilobase (1.09 vs. 0.74 for all other tissues, $p<0.01$).

Some motif clusters appear to have high levels of selection regardless of tissue. To further understand what these motif clusters are and how they may be functioning within the context of evolution, we analyzed the average Rho, E[A], E[W] for each cluster among all tissues (**Figure 2.4A-C**). We see several outlier motif clusters with high average measures of selection, regardless of tissue (**Figure 2.4D-4F**). Members of the highest average motif clusters for Rho and E[W] (**Figure 2.4D,2.4F**) are generally GC rich, which suggests that they may be enriched near transcription start sites of protein-coding genes and this placement in the genome may be driving the signal of selection.

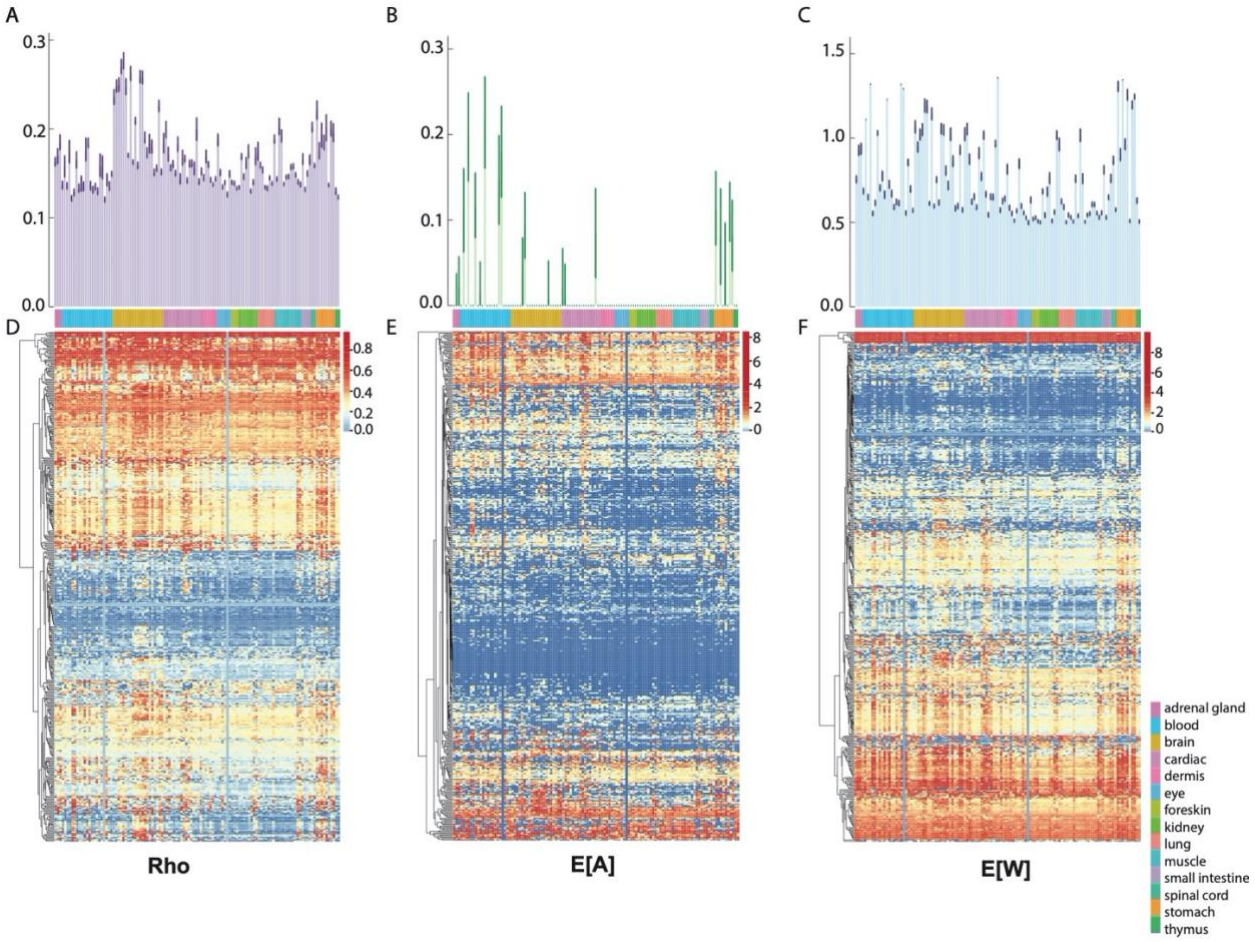



**Figure 2.3. Natural selection on predicted bound motifs. A)** Rho (fraction of nucleotides under selection) for all bound transcription factors in each tissue. **B)** E[A] (number of adaptive substitutions/kB) for all bound transcription factors in each tissue. **C)** E[W] (number of deleterious substitutions/kB) for all bound transcription factors in each tissue. **D)** Rho for each motif cluster in each tissue. **E)** E[A] for each motif cluster in each tissue. **F)** E[W] for each motif cluster in each tissue.

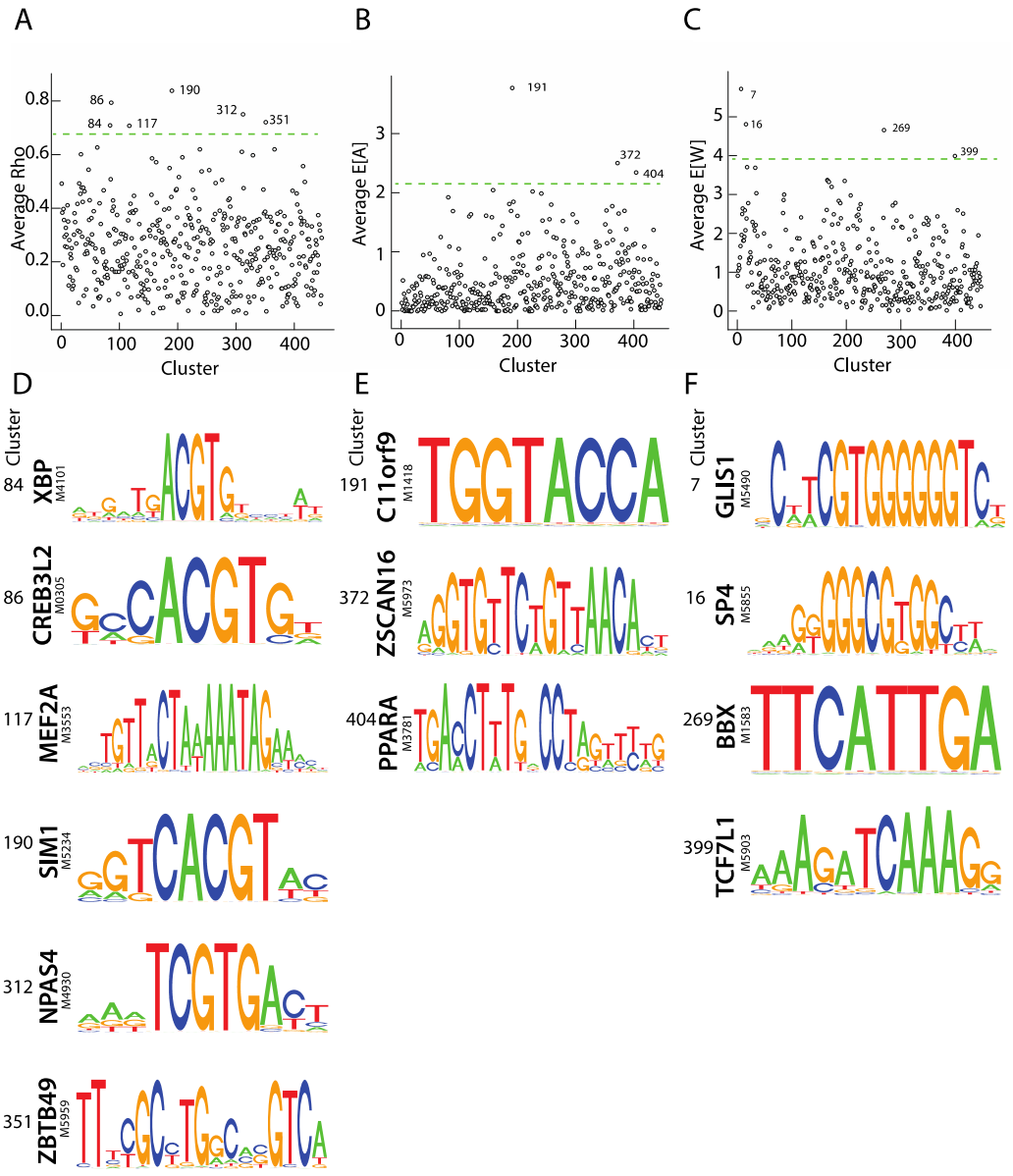



**Figure 2.4. Average selection for each motif cluster. A)** Average Rho across all tissues for each motif cluster. **B)** Average E[A] across all tissues for each motif cluster. **C)** Average E[W] across all tissues for each motif cluster. **D)** Motif PWMs for motif clusters that had the highest average Rho across all tissues. **E)** Motif PWMs for motif clusters that had the highest average E[A] across all tissues. **F)** Motif PWMs for motif clusters that had the highest average E[W] across all tissues.

We see localized bands of higher values for Rho and E[W] among specific tissues in the majority of clusters (**Figure 2.3D-F**). Upon investigating these tissues, we see a trend of increased selection in embryonic tissues. For each measure of selection, we compared values in the combined set of all bound sites for tissues labelled as embryonic and adult in all tissues (**Figure 2.5A-C**). We see a significant increase in Rho for embryonic tissues compared to adult tissues (p=0.001942). When this analysis is narrowed to a per tissue basis, we see that this pattern is especially evident across brain tissues for Rho (average Rho=0.96 for embryonic tissues and average Rho=0.76 for adult tissues, p= 0.0003543) (**Figure 2.5D**) and a similar pattern in E[W] (average E[W]=0.24 for embryonic and average E[W]=0.18 for adult tissues, p=0.03324). Thus, embryonic tissues have a higher fraction of nucleotides under selection than adult tissues and in some tissues, this difference is more evident along with a change in the expected number of polymorphic sites under weak negative selection.

## *DISCUSSION*

Here we present a new method, dTOX, that reformulates the problem of transcription factor occupancy prediction. Previous methods predict whether a motif is occupied by a certain sequence-specific transcription factor. We reformulated this problem, recognizing that DNA sequence and factor-general assays provide only weak information that can distinguish between transcription factors which recognize similar DNA binding motifs. Instead, dTOX predicts whether each motif occurrence is occupied by any transcription factor that is known to recognize that motif. Additionally,

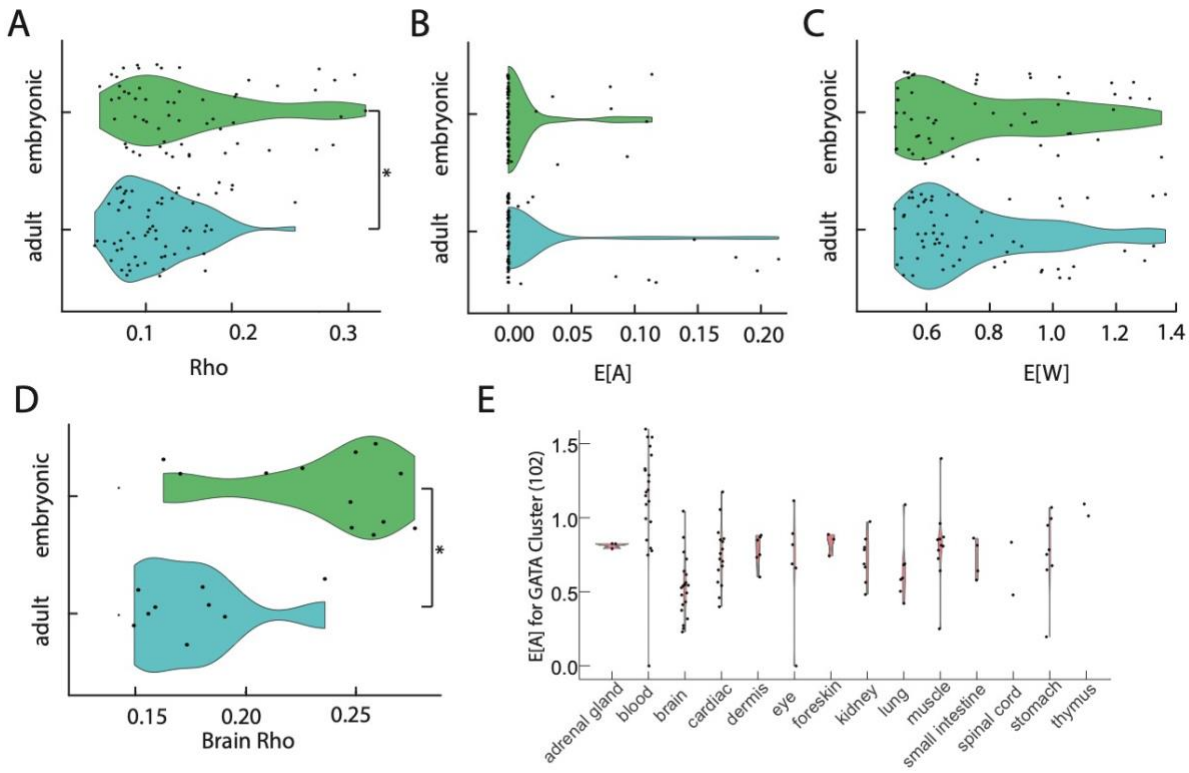



**Figure 2.5. Selection is dependent on tissue. A)** Rho across all clusters for embryonic vs. adult tissues. Embryonic tissues have a significantly higher Rho than adult tissues (p=0.001942) **B)** E[A] across all clusters for embryonic vs. adult tissues. **C)** E[W] across all clusters for embryonic vs. adult tissues. **D)** Brain embryonic vs. adult tissues show a significant difference in Rho (p=0.03324). **E)** E[A] values for the GATA cluster in each tissue. Blood E[A] for GATA is significantly higher than all other tissues (p=4.235e-05).

dTOX uses an accurate discriminative support vector machine (SVM) that we trained to recognize the general pattern in DNase-I-seq associated with motif occupancy. We show that this pattern effectively generalizes across most transcription factors with available ChIP-seq data. Using this model, we identified high-confidence occupied motifs for 447 transcription factor families in 118 human tissues.

We see a positive correlation between Rho and E[W], suggesting that weak negative selection is driving the majority of selection at bound transcription factor motifs. In motif clusters with universally high signatures of selection we see patterns in the types of motifs included. For clusters with high average Rho and E[W] values, we find motifs that are GC rich. This suggests that they may be localized near promoters of protein-coding genes and be under selection due to their functional role in transcription. The role of these motifs in transcriptional regulation is supported by the presence of the highly conserved motif for SP4, a transcriptional activator that is a member of the SP family of general transcription factors, which activates the transcription of many genes that contain CG-rich Sp-binding sites (Suske, 1999).

We saw a relatively high signal for the expected number of adaptive substitutions in samples of blood cells, despite most tissues showing no evidence of expected adaptive substitutions. This signature of positive selection in blood cells is likely influenced by the essential role of blood cells in immunity. In genome-wide scans for positive selection, immune genes are consistently enriched (Raj et al., 2013; Voight et al., 2006) so it is likely that the genomic positions where transcription factors are binding are under

similar selective pressures and favor the fixation of new derived alleles. These increased levels of positive selection in blood cells are apparent when we analyze motif clusters for transcription factors that play essential roles in blood cells. The GATA family of transcription factors are involved in hematopoiesis and adaptive and innate immunity (Tindemans et al., 2014; Tremblay et al., 2018). When analyzing the GATA cluster of motifs we see that they have significantly higher E[A] than all other cell types (1.13 vs. 0.71, p-value = 4.235e-05) (**Figure 2.5E**). This suggests that the positive selection signal differences our analysis picks up in transcription factor binding may be driven by adaption to differences in immune challenges among humans.

We see the highest fraction of nucleotides are under selection (Rho) in motifs that are bound in the brain with certain brain tissues also showing among the highest levels of natural selection across all tissues. This is consistent with previous reports of purifying selection and slow rates of evolution for genes that are highly expressed in the brain (Duret & Mouchiroud, 2000; Tuller et al., 2008). We also see relatively high levels of weak negative selection on polymorphic sites in some brain tissues, suggesting evolutionary constraint on essential transcription factor binding sites. In the future, we would like to look more closely at the functional roles of the constrained bound motifs in the brain to see what role they play in the highly complex biochemical networks in the brain that contribute to slower evolutionary rates (Kuma et al., 1995).

We dissected rates of evolution at different life stages for all tissues. When all tissues are analyzed together, we see a significant increase in Rho values for embryonic tissues

compared to adult tissues. The high level of conservation of embryonic body plans likely contributes to the high level of selection in embryonic tissues compared to adult tissues. The high E[W] values for stomach tissues may be driven by embryonic tissues, since 6 out 8 stomach tissues are embryonic in origin. A limitation of our analysis is the imbalance of embryonic and adult stages for the majority of tissues, making it challenging to fully understand the differences in regulation between the stages. This trend is most evident in brain tissues, with other tissues that contain both embryonic and adult stages not having significant differences (cardiac and kidney). We found that embryonic brain tissues have higher fractions of nucleotides under selection (Rho) and higher amounts of negative selection (E[W]) than adult brain tissues. This additional relaxed evolutionary constraint on motif occupancy in adult brain tissue may play a role in the neural plasticity in adult brains. It has been shown that transcription factors can regulate key steps of neural plasticity (Engelmann & Haenold, 2016; Veyrac et al., 2014) so it is possible that it is evolutionarily advantageous to have lower levels of purifying selection to allow for increased plasticity in adults.

Our analyses show that transcription factor motifs that are bound in different tissues are evolving at different rates. We characterize clusters of motifs that are under high levels of selection which may be connected to their sequence characteristics. We show an enrichment of high levels of selection at motifs that are bound in embryonic tissues compared to their adult counterparts. We confirm previously shown examples of different rates of evolution in different tissues by showing that motifs bound in the brain evolve at slower rates than motifs bound in the immune system likely because of their

differing functional roles (Kuma et al., 1995). Our work studying differences in evolution of occupied transcription factor binding sites mirrors the 'tissue-driven' hypothesis of the evolution of protein-coding genes (Park & Choi, 2010).

## *METHODS*

### SVM TRAINING

**Motif selection:**

We identified motif occurrences for each TF with ENCODE data in either GM12878 or K562 using the RTFBSDB database and narrowed the number of training motifs by only selecting motifs inside bins that have reads within 400bp on either the plus or minus strand in the training dataset. Motifs were Intersected with ChIP-seq narrow peaks called by ENCODE to give them a rough label of bound or unbound. These categories were further refined by using raw ChIP-seq and input data for each TF. Only motifs within a ChIP-seq peak and with a raw read ratio over input from a 100bp bin surrounding the motif that was >2 were included for training as bound. Motifs defined outside a ChIP-seq peak and with a ratio of <1 were included for training as unbound. These restrictions were applied to the training data to ensure a clear distinction between bound and unbound sites, but were not applied to the holdout testing set.

**SVM feature vectors:**

Training sets were composed of the restricted motifs from K562 and GM12878, with an adjusted ratio of bound to unbound reads of 30:70. For each site included in model

training, a feature vector of the genomic data was extracted using logistic scaling to remove differences in feature magnitudes.

Feature vector: 501 features (500 read count features + motif score)

Read count features

| Size of window | Number of windows |
|---|---|
| 1,4,10, 25, 50,500 | 10,50,10, 10, 30,15 |

To further refine the training set for each model, Precision Run-On and Sequencing (PRO-seq) data was used to remove possible outlier training data. In addition to the training feature vectors, PRO-seq data was collected for 100bp surrounding each site. Bound sites within the DNase-I-seq model that were at the bottom 10% for PRO-seq reads were removed from the training set. This allowed for a more selective training set that was more likely to be, thus avoiding adding any unbound samples to the bound training set.

**SVM training:**

The dTOX model was built using the gtSVM package for e1071 implementation on a GPU, with the following parameters: gamma=0.05, tolerance=0.001, biased=F.

**Random forest filtering:**

In order to reduce the computational time to the limitation of the GPU queue, we designed a filter trained by a Random Forest algorithm to reduce the number of loci predicted on with the SVM through elimination of the majority of negative samples. The filter uses few features and takes a short amount of time to do this first filtering step prediction. To train the filtering model, motifs were extracted from RTFBSDB and run through dTOX to get bound and unbound calls. A random forest was built based on these binding calls with the following features: 8 read count windows (1kb, 5kb, 25kb, 50kb around motif on + and - strand), 3 motif score summaries (minimum, maximum, and mean of motif score in 25bp window around motif), and 1 motif score. To filter motifs before dTOX prediction, motifs were extracted from RTFBSDB and run through the random forest model. Motifs predicted as bound were then run through dTOX to get final binding calls.

RUNNING dTOX ON ENCODE DNase-I-seq DATASETS

**Clustering of motifs in the human genome:**

DHS sites (merged from 216 reference cell types defined by the min150 data from (Chu et al., 2018) were scanned for motifs. This accounted for 1/5 regions of hg19. We selected the TFBS with the top 1000 scores for each motif. TFBS were merged and each merged region was extended by +-30bp (which corresponds to the length of the longest PWM). This resulted in ~900,000 regions. These 900,000 regions were re-scanned with the ～1800 PWM and defined with a score of either 0 if the max score in a region is

below 7, or 1 if otherwise. Pairwise correlation coefficients were calculated between each PWM. 1-correlation coefficient was used as the distance for clustering with the ward.D2 method.

**Combining ENCODE DNase-I-seq data:**

All human tissue DNase-I-seq data from ENCODE that passed ENCODE benchmarks for quality control was downloaded. Datasets were initially partitioned by cell/organ type (lung, kidney, blood, brain, dermis, stomach, large intestine, small intestine, spinal cord, adrenal gland, and thymus). Individual cell types within the larger categories were manually curated. For datasets with the same manually curated categories of cell type, the Spearman correlation coefficients were calculated between each individual tissue type. Individual datasets with a correlation coefficient of >=0.85 and the same manually curated annotation were combined, resulting in a total of 118 DNase-I-seq datasets for downstream analysis.

**UMAP clustering of DNase-I-seq data:**

Combined DNase-I-seq datasets were clustered using the AgglomerativeClustering from sklearn.cluster to make clusters for coloring the UMAP. The distance threshold was set to 0.9 to generate clusters with correlations > 0.9. A PCA was run on gene body counts and the top 50 PCs were selected as input for the Scikit-learn UMAP package.

**INSIGHT on dTOX output:**

INSIGHT was run on all motifs that had a dTOX prediction of >0.5 for each DNase-I-seq dataset with the default settings. For individual clusters of transcription factor motifs, the binding predictions for each cluster member were extracted from the dTOX output and combined. Predictions were filtered to remove duplicate bound sites between members.

## *REFERENCES*


Amati, B., Dalton, S., Brooks, M. W., Littlewood, T. D., Evan, G. I., & Land, H. (1992). Transcriptional activation by the human c-Myc oncoprotein in yeast requires interaction with Max. Nature, 359(6394), 423–426.

Arbiza, L., Gronau, I., Aksoy, B. A., Hubisz, M. J., Gulko, B., Keinan, A., & Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. Nature Genetics, 45(7), 723–729.

Blackwood, E. M., & Eisenman, R. N. (1991). Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science, 251(4998), 1211–1217.

Bulyk, M. L. (2006). Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays. Methods in Enzymology, 410, 279–299.

Chu, T., Rice, E. J., Booth, G. T., Salamanca, H. H., Wang, Z., Core, L. J., Longo, S. L., Corona, R. J., Chin, L. S., Lis, J. T., Kwak, H., & Danko, C. G. (2018). Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. Nature Genetics, 50(11), 1553–1564.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics, 46(12), 1311–1320.

Duret, L., & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Molecular Biology and Evolution, 17(1), 68–74.

Engelmann, C., & Haenold, R. (2016). Transcriptional Control of Synaptic Plasticity by Transcription Factor NF-κB. Neural Plasticity, 2016, 7027949.

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

Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature, 461(7261), 186–192.

Galas, D. J., & Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Research. https://academic.oup.com/nar/article-abstract/5/9/3157/2380868

Gronau, I., Arbiza, L., Mohammed, J., & Siepel, A. (2013). Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Molecular Biology and Evolution, 30(5), 1159–1171.

Guertin, M. J., & Lis, J. T. (2010). Chromatin landscape dictates HSF binding to target DNA elements. PLoS Genetics, 6(9), e1001114.

Guertin, M. J., Martins, A. L., Siepel, A., & Lis, J. T. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. PLoS Genetics, 8(3), e1002610.

He, H. H., Meyer, C. A., Hu, S. S., Chen, M.-W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., & Brown, M. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nature Methods, 11(1), 73–78.

Karimzadeh, M., & Hoffman, M. M. (2018). Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome. In bioRxiv (p. 168419). https://doi.org/10.1101/168419

Kretzner, L., Blackwood, E. M., & Eisenman, R. N. (1992). Myc and Max proteins possess distinct transcriptional activities. Nature, 359(6394), 426–429.

Kuma, K., Iwabe, N., & Miyata, T. (1995). Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. Molecular Biology and Evolution, 12(1), 123–130.

Kuntz, S., Williams, B. A., Sternberg, P. W., Wold, B. J. (2012). Transcription facotr redundacy and tissue-specific regulation: Evidence from functional and physical network connectivity. Genome Research, 22(10), 1907-1919.

Luna-Zurita, L., Stirnimann, C. U., Glatt, S., Kaynak, B. L., Thomas, S., Baudin, F., Samee, M. A. H., He, D., Small, E. M., Mileikovsky, M., Nagy, A., Holloway, A. K., Pollard, K. S., Müller, C. W., & Bruneau, B. G. (2016). Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. Cell, 164(5), 999–1014.

Park, S. G., & Choi, S. S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evolutionary Biology, 10, 241.

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research, 21(3), 447–455.

Raj, T., Kuchroo, M., Replogle, J. M., Raychaudhuri, S., Stranger, B. E., & De Jager, P. L. (2013). Common risk alleles for inflammatory diseases are targets of recent positive selection. American Journal of Human Genetics, 92(4), 517–529.

Samee, M. A. H., Bruneau, B. G., & Pollard, K. S. (2019). A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs. Cell Systems, 8(1), 27–42.e6.

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., & Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. Molecular Cell, 58(6), 1101–1112.

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T., & Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature Biotechnology, 32(2), 171–178.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. A., Simon, J. M., Vales, T., Wang, T., Winter, D., … Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Research, 21(10), 1757–1767.

Steube, A., Schenk, T., Tretyakov, A., & Saluz, H. P. (2017). High-intensity UV laser ChIP-seq for the study of protein-DNA interactions in living cells. Nature Communications, 8(1), 1303.

Suske, G. (1999). The Sp-family of transcription factors. Gene, 238(2), 291–300.

Tindemans, I., Serafini, N., Di Santo, J. P., & Hendriks, R. W. (2014). GATA-3 function in innate and adaptive immunity. Immunity, 41(2), 191–206.

Tremblay, M., Sanchez-Ferras, O., & Bouchard, M. (2018). GATA transcription factors in development and disease. Development , 145(20). https://doi.org/10.1242/dev.164384

Tuller, T., Kupiec, M., & Ruppin, E. (2008). Evolutionary rate and gene expression across different brain regions. Genome Biology, 9(9), R142.

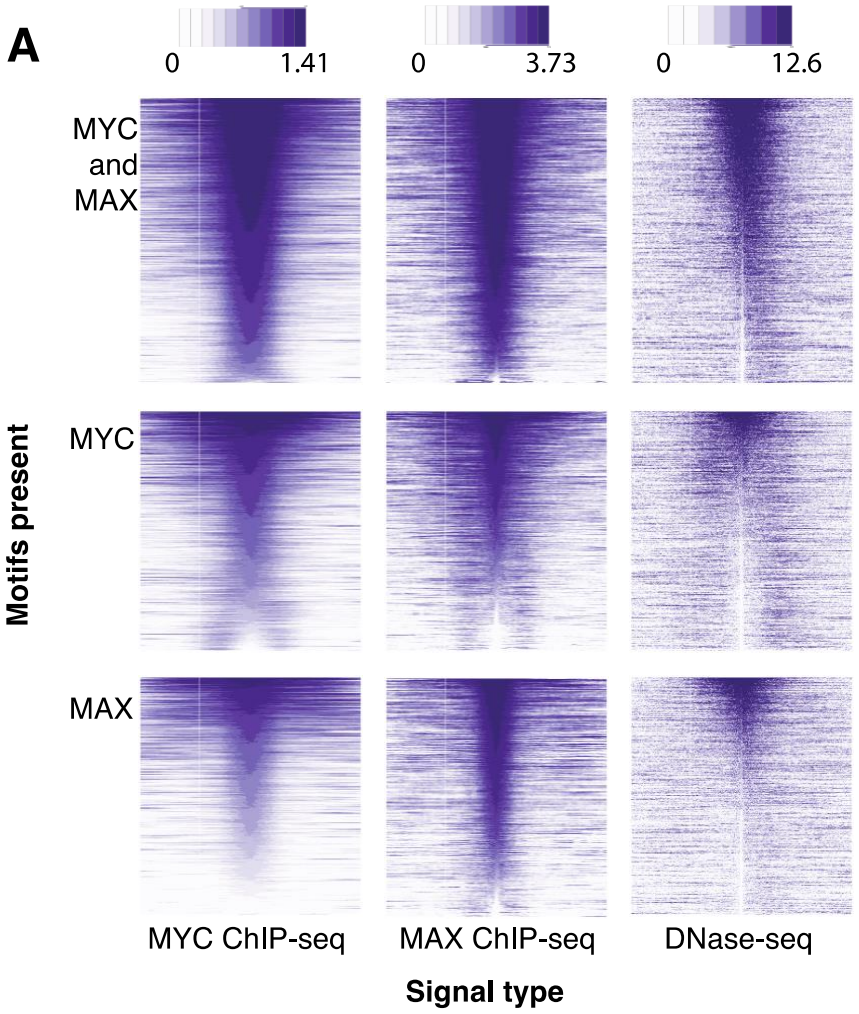
Veyrac, A., Besnard, A., Caboche, J., Davis, S., & Laroche, S. (2014). The transcription factor Zif268/Egr1, brain plasticity, and memory. Progress in Molecular Biology and Translational Science, 122, 89–129.

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. PLoS Biology, 4(3), e72.

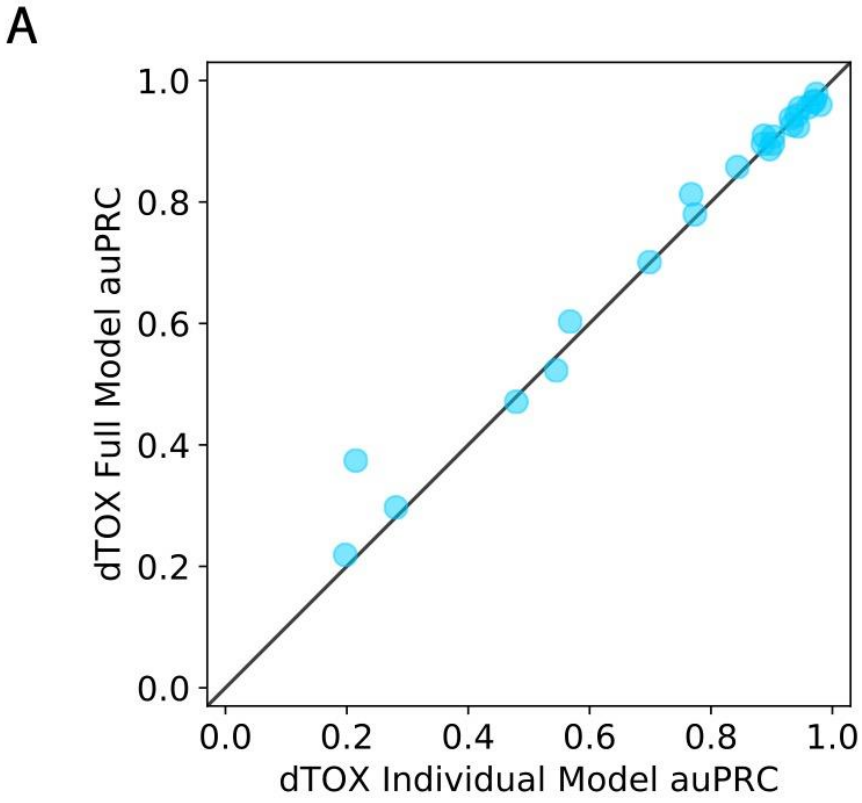
Wang, Z., Martins, A. L., & Danko, C. G. (2016). RTFBSDB: an integrated framework for transcription factor binding site analysis. Bioinformatics . https://doi.org/10.1093/bioinformatics/btw338

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., … Hughes, T. R. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell, 158(6), 1431–1443.

Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., & Akalin, A. (2017). HOT or not: Examining the basis of high-occupancy target regions. In bioRxiv (p. 107680). https://doi.org/10.1101/107680




**Supplementary Figure 2.1. Comparison of MYC and MAX binding. A)** Comparison of MYC and MAX ChIP-seq and DNase-I-seq for regions that have both MYC and MAX ChIP-seq peaks, only MYC peaks, or only MAX peaks.




**Supplementary Figure 2.2. Individual vs. full model performance. A)** A general model based on all transcription factors performs as well as individual models built for each transcription factor.

**Supplemental Table 1.** Transcription factors that the dTOX SVM was trained on.

| **K562** | **GM12878** |
|---|---|
| ARID3A | BATF |
| ATF1 | BCL11A |
| ATF3 | BHLHE40 |
| BACH1 | BRCA1 |
| BHLHE40 | CEBPB |
| CEBPB | CTCF |
| CTCF | E2F4 |
| CTCFL | EBF1 |
| E2F6 | EGR1 |
| EGR1 | EGR1 |
| ELF1 | ELF1 |
| ELK1 | ELK1 |
| ETS1 | EP300 |
| FOS | ETS1 |
| FOSL1 | FOS |
| GABPA | FOXM1 |
| GATA1 | GABPA |
| GATA2 | IKZF1 |
| JUN | IRF4 |
| JUNB | JUND |
| JUND | MAX |
| MAFF | MAZ |
| MAFK | MEF2A |
| MAX | MEF2C |
| MAZ | MXI1 |
| MEF2A | NFATC1 |
| MXI1 | NFE2 |
| MYC | NFIC |
| NFE2 | NFYA |
| NFYA | NFYB |
| NFYB | NR2C2 |
| NR2C2 | NRF1 |
| NR2F2 | PAX5 |
| NRF1 | PBX3 |
| REST | POU2F2 |

| | |
|---|---|
| RFX5 | REST |
| SP1 | RFX5 |
| SPI1 | RUNX3 |
| SRF | RXRA |
| STAT5A | SIX5 |
| TAL1 | SP1 |
| TBP | SPI1 |
| TEAD4 | SRF |
| THAP1 | STAT1 |
| USF1 | STAT3 |
| USF2 | STAT5A |
| YY1 | TBP |
| ZBTB33 | TCF12 |
| ZBTB7A | USF1 |
| ZNF143 | USF2 |
| ZNF263 | YY1 |
| ZNF274 | ZBTB33 |
| | ZEB1 |
| | ZNF143 |

**Supplemental Table 2.** Area under the Precision-Recall curve of transcription factors in the HeLa complete holdout set.

| | Dnase-I-seq |
|---|---|
| BRCA1 | 0.9377293 |
| CEBPB | 0.9076302 |
| CTCF | 0.9408423 |
| E2F1 | 0.55667436 |
| E2F4 | 0.52261895 |
| E2F6 | 0.47087342 |
| ELK1 | 0.9426953 |
| ELK4 | 0.9437501 |
| FOS | 0.96009252 |
| GABPA | 0.60311646 |
| IRF3 | 0.20841823 |
| JUN | 0.96463354 |
| JUND | 0.97801263 |
| MAFK | 0.70095223 |
| MAX | 0.96614459 |
| MAZ | 0.95602217 |
| MXI1 | 0.92438095 |
| MYC | 0.89565138 |
| NFYA | 0.89548851 |
| NFYB | 0.85740809 |
| NR2C2 | 0.29694781 |
| NRF1 | 0.81273542 |

| | |
|---|---|
| PRDM1 | 0.73673496 |
| REST | 0.37418515 |
| RFX5 | 0.92706971 |
| SMARCC1 | 0.83944721 |
| SMARCC2 | 0.8576676 |
| STAT3 | 0.88707339 |
| TBP | 0.77961091 |
| TCF7L2 | 0.89422163 |
| USF2 | 0.95452187 |
| ZKSCAN1 | 0.9285918 |
| ZNF143 | 0.90882246 |
| ZNF274 | 0.21882296 |

# CHAPTER 3

# ADAPTIVE CHANGES IN ANTHRAX TOXIN RECEPTOR EXPRESSION IN HUMANS


## *ABSTRACT*

The advent of animal husbandry and hunting increased human exposure to zoonotic pathogens, placing new sources of selective pressure on the ancestors of modern humans. Here we report that adaptive immune cells in humans have an 8-fold decrease in the expression of the anthrax toxin receptor 2 (*ANTXR2*) compared with chimpanzee and rhesus macaque. Using CRISPRa to recover the ancestral expression of *ANTXR2* in human cells revealed that expression differences affected cellular sensitivity to recombinant anthrax toxins. We identified cis-regulatory elements (CRE) correlated with *ANTXR2* expression by integrating genomic data profiling enhancer RNA transcription and chromatin architecture in multiple primate species, with follow-up validation using luciferase assays. At least one candidate DNA sequence change was polymorphic in human populations and had an allele frequency that was geographically correlated with anthrax endemic areas. Our results document how changes in human ecology and exposure to zoonotic pathogens may have contributed to regulatory evolution.

## *INTRODUCTION*

Pathogens are a primary source of morbidity in humans and other animals (Karlsson et al., 2014). Host genetics play a central role in pathogen susceptibility, and as a result, infectious diseases were a major source of selective pressure during the evolution of modern humans (Fumagalli et al., 2011; J. B. S. Haldane, 2006; J. B. S. Haldane, 1949). During the past 60 years, numerous case-studies have revealed evidence of positive selection on alleles that provide resistance to infectious disease in human populations in endemic areas of the globe. These include resistance alleles for *P. falciparum* (Malaria) (Allison, 1954), *T. brucei* (African trypanosomiasis) (Genovese et al., 2010; Ko et al., 2013), and *Arenaviridae* viruses (Lassa fever) (Sabeti et al., 2007). More recently, global signatures of selection on human "host" genes, as well as patterns of introgression between humans and Neanderthals, have supported a major role for infectious disease in selection during the evolution of modern humans (Enard & Petrov, 2018, 2020; Fumagalli et al., 2011; Kosiol et al., 2008). Genes with a primary function in immunity are consistently overrepresented in genome-wide screens of positive selection in humans and other animals (Kosiol et al., 2008; Shultz & Sackton, 2019). Several well-characterized examples show how genetic changes affecting immune cells have led to phenotypic differences in host immunity (Chrousos et al., 1982; Denny et al., 2000; Nguyen et al., 2006; Reynolds et al., 1999; Scammell et al., 2001).

Changes in ecology and behavior as the ancestors of humans began hunting, developed animal husbandry, and agriculture has long been speculated to affect the pathogens with which humans interact (Gonzalez et al., 2000; Johnson et al., 1993; Wolfe et al., 2007). One pathogen that began affecting human ancestors after the development of hunting is *B. anthracis*, the bacterium that causes anthrax disease (Kamal et al., 2011; Spencer, 2003). *B. anthracis* primarily afflicts domestic grazing species during the course of its natural life-cycle, and has driven selective pressures in cattle, sheep, and other ruminant species (Lv et al., 2014; F. Zhao et al., 2015). However, anthrax disease has also been a source of mortality in humans (Kamal et al., 2011; Spencer, 2003). Anthrax disease was speculated to be the causative agent behind plagues in antiquity based in part on the description of characteristic black boils on the host (Kamal et al., 2011). More recently, inhaled anthrax from hides was the causative agent behind Woolsorter's disease, primarily affecting textile workers during the 18th and 19th century (Eurich, 1926; Laforce, 1978). *B. anthracis* remained endemic across the world until the first half of the 20th century, causing an estimated 20 to 100 thousand annual human cases (Kamal et al., 2011). The prevalence of anthrax disease has decreased markedly in the latter half of the 20th century due to more effective mitigation strategies, although recent outbreaks have been documented in humans (Meselson et al., 1994; Mwenye et al., 1996).

Here we show that expression of the anthrax toxin receptor 2 (*ANTXR2*), the membrane receptor granting anthrax toxins access to host cells, decreased in human CD4+ T cells compared to non-human primates. Using CRISPR activation (CRISPRa) to recover the

ancestral expression of *ANTXR2* in human cells revealed that expression differences affect cellular sensitivity to recombinant anthrax toxins. By integrating genomic data profiling enhancer activity and chromatin architecture with reporter assays, we identified and validated cis-regulatory elements (CREs) correlated with changes in *ANTXR2* expression. Much of the variation in *ANTXR2* expression was present in all modern humans. However, at least one candidate DNA sequence change was polymorphic in human populations and had an allele frequency which correlated with the historical distribution of *B. anthracis* worldwide. Finally, we identified a candidate selective sweep in the gene desert upstream of *ANTXR2* which may affect its expression in multiple tissues. Collectively, our results document a change in *ANTXR2* expression in humans, potentially revealing an example in which changes in primate ecology during the advent of hunting led to new interactions with zoonotic pathogens.

## ***RESULTS***

To identify candidate transcriptional changes that influenced the human immune system, we used recently published maps of RNA polymerase collected using PRO-seq in primary CD4+ T cells from human, chimpanzee, and rhesus macaque (Danko et al., 2018). We previously demonstrated that PRO-seq data in each species interrogated a relatively pure population of T-helper type 1 (Th1), Th2, Th17, T-regulatory and T-follicular helper CD4+ cells (Danko et al., 2018). Cells were grown on standard medium under non-stressed conditions and were harvested in stationary phase. We used this

PRO-seq dataset to identify 552 candidate genes (of 9081 GENCODE annotations expressed in at least one primate) whose transcription changed in humans compared to both non-human primates (**Figure 3.1A**; DESeq2 FDR-corrected $p < 0.05$, (Love et al., 2014)). Candidate transcriptional changes included *SIGLEC5* (**Figure 3.1A**), which was reported previously as the causal gene responsible for changes in the sensitivity of human CD4+ T cells to activation (Nguyen et al., 2006).

We found an enrichment of differentially expressed genes encoding extracellular surface proteins. Of the 552 differentially transcribed genes in humans, 182 encoded genes annotated as an internal component of the plasma membrane or related gene ontology terms, which represents up to a 3-fold enrichment (**Figure 3.1A-B**). Similar results were obtained using newly collected RNA-seq data from one human and one rhesus macaque (**Supplementary Figure 3.1**). Regulatory changes in genes encoding transmembrane receptors may be influenced by interactions between T cells and pathogens. Indeed, several genes that were differentially transcribed encoded putative pathogen receptors, including the alanyl aminopeptidase (*ANPEP*; required for viral entry of coronavirus (Forni et al., 2017) 229E) and anthrax toxin receptor 2 (*ANTXR2*) (receptor for *B. anthracis* toxins (Banks et al., 2005)) (**Figure 3.1A; Supplementary Figure 3.2**). These findings are consistent with a hypothesis that some changes in transmembrane receptor expression may be influenced by interactions between CD4+ T cells and pathogens.

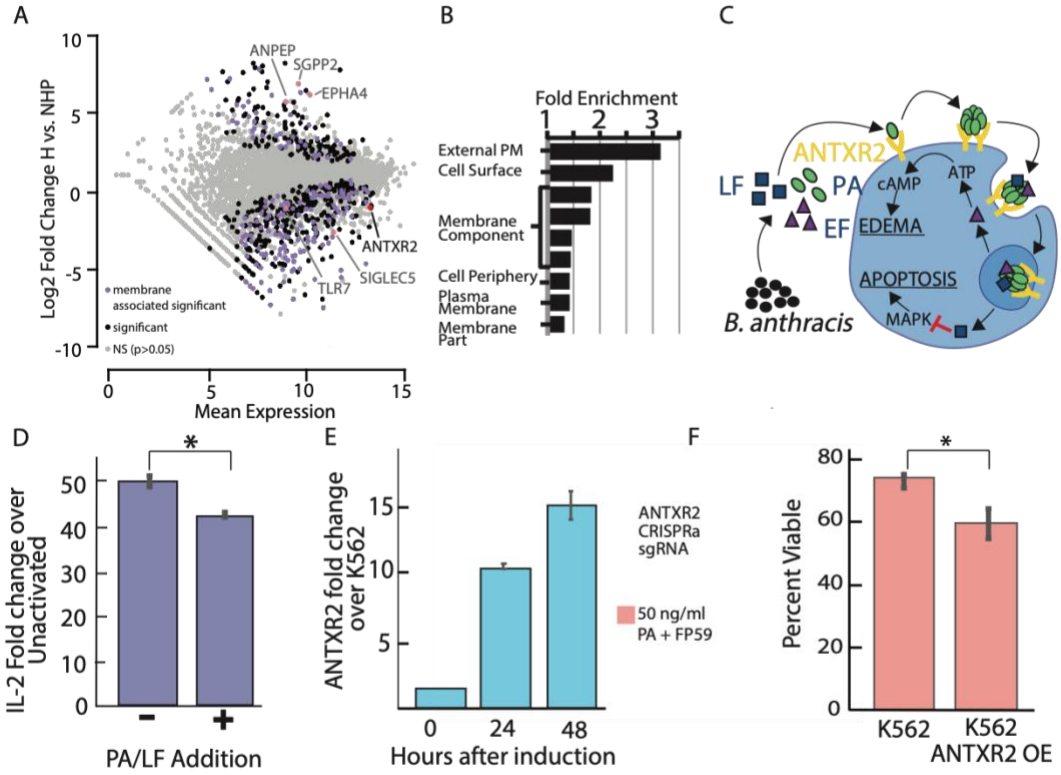



**Figure 3.1. *ANTXR2* is expressed at lower levels in human CD4+ T-cells compared to non-human primates. A)** Differentially transcribed protein-coding genes between humans and non-human primates in CD4+ T cells. The GO term 'integral component of the membrane' is enriched in differentially transcribed genes. **B)** Differentially transcribed genes between human and non-human primates are enriched for GO terms associated with the periphery of the cell. **C)** *B. anthracis* toxins lethal factor (LF) and edema factor (EF) cause apoptosis and edema in target host cells. **D)** CD4+ T cells produce less IL-2 upon activation with PMA and ionomycin when treated with anthrax toxins protective antigen (PA) and LF. **E)** CRISPRa induction in K562 cells results in increased *ANTXR2* expression. **F)** CRISPRa K562 cells that overexpress ANTXR2 are less viable after an anthrax toxin challenge as measured by alamar blue.

We selected the human-specific down-regulation of *ANTXR2* for follow-up study. *ANTXR2* was 8-fold downregulated in humans compared to non-human primates at the level of both transcription and mRNA (**Figure 3.1A; Supplementary Figure 3.1**). The *ANTXR2* gene encodes a transmembrane receptor which aids in the cellular entry of three toxins secreted by the *B. anthracis* bacterium (Moayeri & Leppla, 2004): protective antigen (PA), lethal factor (LF), and edema factor (EF). Toxins enter the cytoplasm of host cells by binding *ANTXR2*, which is ubiquitously expressed on mammalian cells (Scobie et al., 2003; Sun & Jacquez, 2016) (**Figure 3.1C**). In most cell types, anthrax toxins cause cell death by activating apoptosis and necrosis pathways, leading to anthrax disease (Moayeri & Leppla, 2009) (**Figure 3.1C**). Anthrax toxins do not cause apoptosis in CD4+ T cells but are reported to affect T cell activation (Comer et al., 2005; Paccani et al., 2005), which may limit the efficacy of downstream adaptive immune response to the *B. anthracis* bacterium. T cells initiate adaptive immune responses to foreign pathogens by secreting signaling proteins, including IL-2 (Luckheeram et al., 2012). We used an ELISA to determine that recombinant anthrax toxins PA and LF lead to decreased IL-2 secretion from stimulated human CD4+ T cells (**Figure 3.1D**). This result suggests that anthrax toxins can affect T cells by dampening activation of host immunity.

We asked whether *ANTXR2* expression changes of the magnitude observed between primate species can affect sensitivity to anthrax toxins. We overexpressed *ANTXR2* in a human myelogenous leukemia cell line (K562) using CRISPR activation (CRISPRa), in which catalytically dead CAS9 is fused to a VP64 transcriptional activator (Perez-

Pinera et al., 2013). CRISPRa using a single guide RNA targeting the *ANTXR2* promoter increased mRNA levels by 10-fold relative to an empty vector control 24 hours after transfection, a similar change to the differences observed between humans and non-human primates (**Figure 3.1E**). Next, we treated *ANTXR2* overexpressing and empty vector control cells with recombinant anthrax toxins PA and FP59, an LF analog reported to induce apoptosis in hematopoietic cells (Lui et al., 2001), and measured cell viability using alamar blue (**Figure 3.1F**). *ANTXR2* overexpressing cells had a significantly lower viability following treatment with recombinant anthrax toxins (**Figure 3.1F**). Thus, our experimentally induced increase in human *ANTXR2* expression was of sufficient magnitude to decrease cell viability in the presence of anthrax toxins.

Genetic changes that affect *ANTXR2* transcription are likely to reside in cis-regulatory elements (CREs). We identified candidate CREs near *ANTXR2* in CD4+ T cells of each primate species using dREG to analyze PRO-seq data (Danko et al., 2015; Wang et al., 2018). We identified numerous CREs near *ANTXR2*: several in introns of *ANTXR2,* a broad promoter with numerous proximal regulatory sites, a proximal CRE within ~5kb of the *ANTXR2* promoter, three CREs situated within a gene desert that separated *ANTXR2* and *PRDM8*, and CREs within the *PRDM8* transcription unit (**Figure 3.2A**). Notably, a comparison with DNase-I hypersensitivity and H3K27ac ChIP-seq data in humans did not identify additional candidate CREs worth considering (**Supplementary Figure 3.3**). We observed rapid turnover in the activity of CREs within the first intron, which were often species-specific. By contrast, CREs in the gene desert or near *PRDM8*

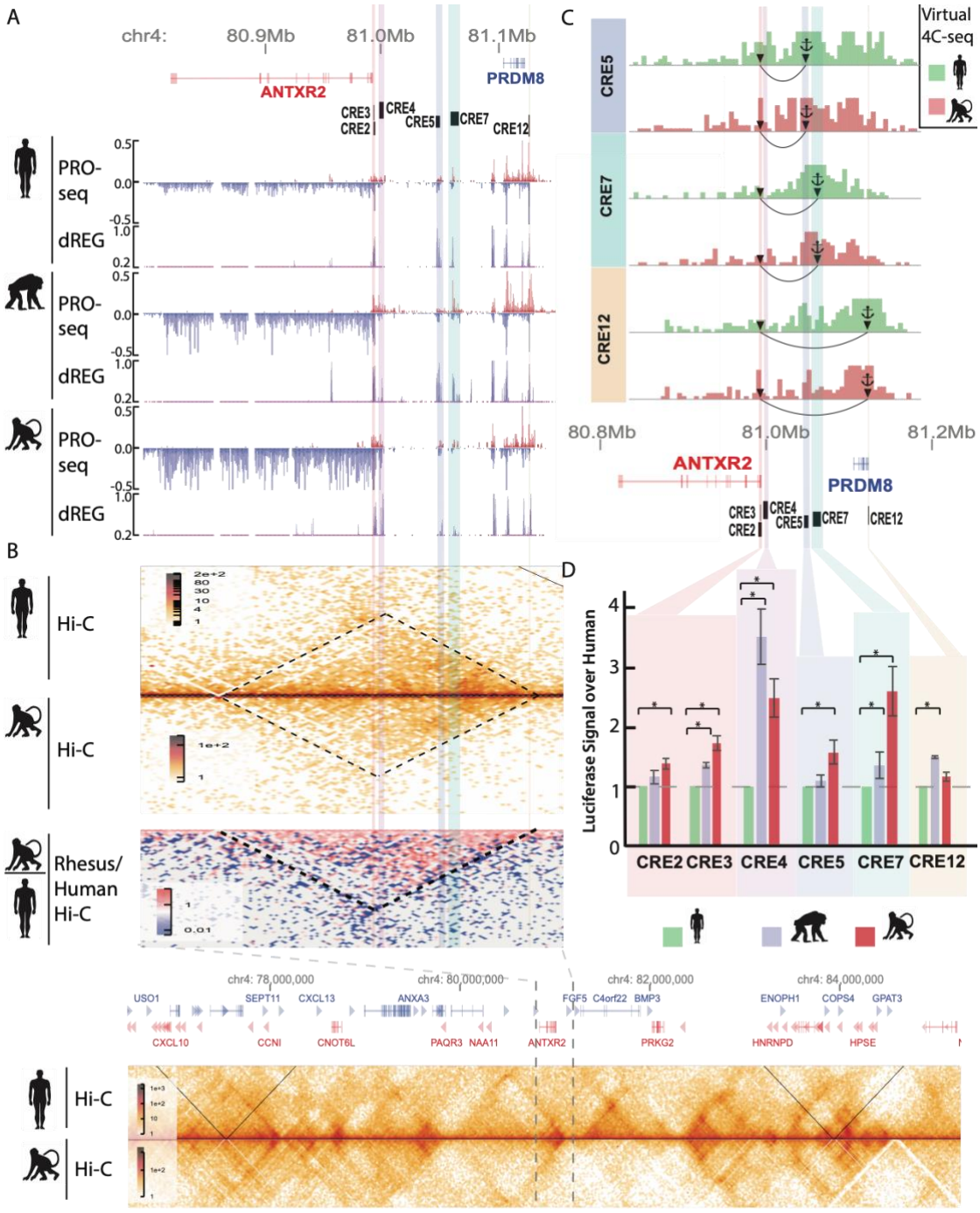



**Figure 3.2. Changes in *ANTXR2* cis-regulatory element activity. A)** Genome browser shot of CD4+ T cell PRO-seq in human, chimpanzee, and rhesus macaque and regulatory elements and regulatory elements predicted by dREG. **B)** Hi-C in CD4+ T cells from human and rhesus macaque. *ANTXR2* is located in the same topological domain (TAD) as the upstream gene *PRDM8*. Rhesus macaque Hi-C signal divided by

human Hi-C shows an increase in contacts within *ANTXR2*'s TAD. **C)** Virtual 4C-seq plots for distal regulatory elements tested in the luciferase assay show contacts to the *ANTXR2* promoter. **D)** Luciferase assay performed in Jurkat cells to test the activity of regulatory elements in human, chimpanzee, and rhesus macaque shows an increase in activity for either chimpanzee or rhesus macaque compared to humans for CRE2, CRE3, CRE4, CRE5, CRE7, and CRE12.

were often conserved between both non-human primate species. Since the majority of CREs change activity between chimpanzee and rhesus macaque (Danko et al., 2018), the conservation of gene desert CREs may indicate that they have a biological function.

To determine whether CREs interacted with the *ANTXR2* promoter we performed *in situ* Hi-C in CD4+ T cells from human and rhesus macaque. Hi-C data revealed that *ANTXR2*, *PRDM8*, and all candidate CREs were found within the same topological associated domain (TAD), a structure reported to insulate the effects of distal enhancers from affecting expression (Symmons et al., 2014) (**Figure 3.2B**). Nearly all regions within the *ANTXR2* TAD had a higher contact frequency in rhesus macaque compared with humans, potentially reflecting species-specific differences in *ANTXR2* transcription (**Figure 3.2B**). Virtual 4C-seq analysis showed that several of the distal candidate CREs had a focal increase in contact frequency with the *ANTXR2* promoter (**Figure 3.2C; Supplementary Figure 3.4**). Additionally, we identified an *ANTXR2* expression quantitative trait locus (eQTL) that overlapped CREs in the gene desert and near *PRDM8* using RNA-seq data from human CD4+ T cells (Schmiedel et al., 2018) (**Supplementary Figure 3.5**). Taken together, these findings indicate that the presence of evolutionarily conserved CREs throughout the upstream gene desert can affect *ANTXR2* expression in CD4+ T cells.

We used a luciferase assay to measure the activity of 12 CREs in the gene desert and near the *ANTXR2* promoter using DNA sequences from human, chimpanzee, and rhesus macaque. We used human Jurkat CD4+ T cells as a model trans-environment, which

recapitulates the pattern of transcription in the *ANTXR2* locus observed in primary T cells (**Supplementary Figure 3.6**). Six of the 12 cis-regulatory elements showed higher luciferase activity in chimpanzee or rhesus macaque than in human, whereas only one was higher in human (**Figure 3.2D; Supplementary Figure 3.7**). The one CRE with a higher activity in human was one element inside of the complex *ANTXR2* promoter, which overall decreased activity in human (**Supplementary Figure 3.8**). Although most changes were modest in magnitude, several were 2-3-fold lower in humans than in one or both primates. These findings may be consistent with reports that multiple causal DNA sequence differences, some with a modest magnitude, contribute to expression changes between or within species (Cusanovich et al., 2018; Kalay & Wittkopp, 2010).

Of the six candidate CREs with changes in activity, CRE4 stood out for its larger magnitude of effect (2-3 fold) and consistently higher luciferase activity in both chimpanzee and rhesus macaque (**Figure 3.2D**). Transcriptional activity is controlled by DNA sequences that lie within a stereotypical chromatin and transcriptional architecture that can be identified by divergently oriented transcription initiation and pause sites, which provide a higher resolution for regulatory sequences that control the activity of a CRE (Andersson et al., 2014; Core et al., 2014; Scruggs et al., 2015; Tippens et al., 2019). The genomic region cloned for CRE4 consisted of two separate divergent initiation sites, both of which had substantially higher quantities of Pol II loading in chimpanzee and rhesus macaque than in humans (**Figure 3.3A**). Both of the divergent initiation sites corresponded to conserved non-coding regions identified by

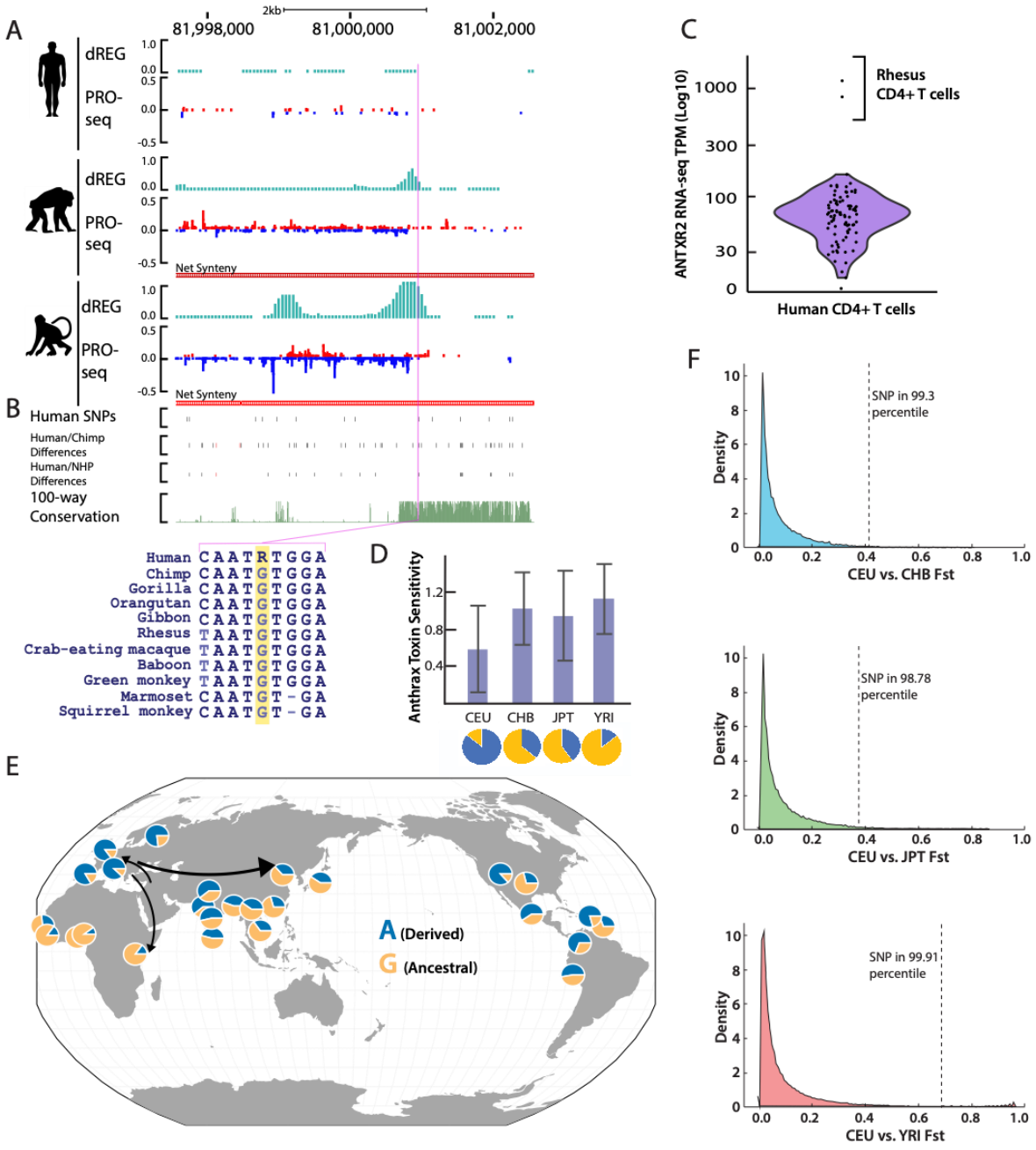



**Figure 3.3. Differences in allele frequencies of upstream regulatory element. A)** PRO-seq and dREG signal at CRE4. rs41407844 falls within a dREG peak shared by chimpanzee and rhesus macaque and the ancestral allele G is conserved within the primate lineage. **B)** Conservation of CRE4 between humans and non-human primates. Tracks show human-specific SNPs, human/chimpanzee differences (SNPs in black, INDELS in red), human/non-human primate differences (SNPs in black, INDELS in red), and 100-way conservation. **C)** Comparison of *ANTXR2* RNA expression levels

among a large set of humans (n=85) to rhesus macaque. Human variation in expression does not overlap with rhesus macaque expression. **D)** Martchenko et al. showed that CEU lymphoblastoid cell lines have the lowest sensitivity to anthrax toxins. **E)** rs41407844 allele frequencies align with the predicted expansion patterns of *B. anthracis*. **F)** rs41407844 is above the 98% percentile for Fst comparisons between CEU and CHB, JPT, and YRI HapMap populations.

a 100-way PhyloP (Pollard et al., 2010) comparison (**Figure 3.3B**). We found 16 single nucleotide differences within the entire 5kb region that were shared in both rhesus macaque and chimpanzee, but in which the human reference sequence diverged. Only one of these human divergences was found inside divergent initiation sites in a conserved non-coding sequence in CD4+ T cells (**Figure 3.3C**): a human divergence in which the human reference allele, *A*, differed from the orthologous base in all sequenced primate species, *G*.

We asked whether lower *ANTXR2* expression was fixed in humans. *B. anthracis* is speculated to have evolved in southern Africa (Keim et al., 1997), potentially providing opportunities for exposure throughout the evolution of hominids. Analysis of 85 human CD4+ T cell RNA-seq datasets (Schmiedel et al., 2018) showed that none of the sampled humans had *ANTXR2* expression levels close to non-human primates (**Figure 3.3C**). Nevertheless, *ANTXR2* expression varied by nearly one order of magnitude within humans. Previous reports have noted differences in anthrax toxin sensitivity among humans (Martchenko et al., 2012), and indeed the candidate single nucleotide difference in CRE4 was polymorphic in humans, corresponding to rs41407844. The derived allele of rs41407844 appears to confer reduced sensitivity to anthrax, as its allele frequency is correlated with reported anthrax toxin sensitivity (Martchenko et al., 2012): high frequency in Europeans (~0.85), intermediate in East Asians (~0.38), and low frequency in African populations (~0.17; **Figure 3.3D**). Although *B. anthracis* is believed to have evolved in southern Africa (Keim et al., 1997), population genetic data suggest one strain recently radiated across the globe in a manner that correlates with rs41407844

allele frequency: beginning in Europe (Schmid & Kaufmann, 2002) and spreading through trade during the past several thousand years (Sternbach, 2003; Van Ert et al., 2007) (**Figure 3.3E**). The fixation index (Fst) of rs41407844 was substantially higher between Europeans and other human populations than >98% of SNPs across the genome (**Figure 3.3F, Figure 3.4A**). Collectively these data support a model in which multiple genetic changes each arising over different time-scales during the divergence of modern humans from other primates have collectively contributed to reductions in *ANTXR2* expression.

We asked whether variants that reduce in *ANTXR2* expression were under positive selection in the ancestors of modern humans. We computed the composite-likelihood-ratio (CLR) of a selective sweep across chromosome 4 in four human populations analyzed by the 1000 Genomes Project (Europeans [CEU], East Asians [CHB and JPT], and Africans [YRI]) using SweepFinder2 (DeGiorgio et al., 2016). We found a candidate selective sweep in the gene desert upstream of the *ANTXR2* locus (**Figure 3.4B**). The selective sweep had a higher CLR in European (CEU) than 98% of other loci on chromosome 4 (**Supplementary Figure 3.9**). Moreover, the CLR was substantially higher in Europeans compared to 1000 Genomes populations representative of East Asian (CHB, JPT) or African (YRI) ancestry (**Figure 3.4B**). Thus, we conclude that a selective sweep affected the *ANTXR2* gene desert in populations with a higher historical anthrax exposure.

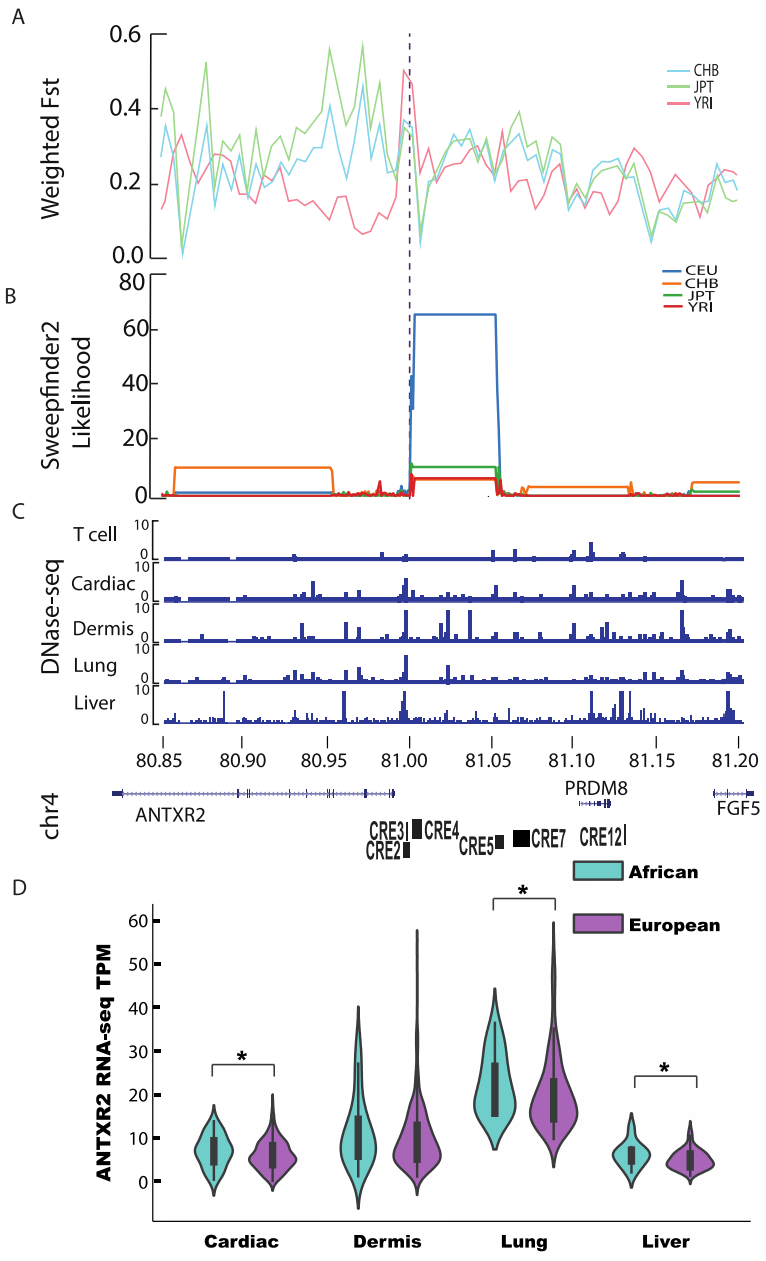



**Figure 3.4. Population genetics of the *ANTXR2* locus. A)** Weighted Fst comparisons between CEU and CHB, JPT, and YRI HapMap populations at the *ANTXR2* locus show a peak around *ANTXR2*. **B)** Sweepfinder2 shows a predicted selective sweep in the CEU population upstream of *ANTXR2*. **C)** ENCODE DNase-I-seq data for T cells, cardiac cells, dermis, lung, and liver show differential regulatory landscapes within the predicted selective sweep in CEU. **D)** GTEX data for cardiac, dermis, lung, and liver show increased *ANTXR2* expression in African American individuals compared to European individuals.

The candidate sweep was adjacent to, but did not overlap, the candidate causal SNP in CD4+ T cells, rs41407844. We therefore also considered whether there may be additional genetic variation in humans which affects *ANTXR2* expression in other tissues. Anthrax disease is a systemic disorder affecting tissues and organs in different manners. The main effects of anthrax disease are apoptosis (by anthrax lethal toxin), edema (by anthrax edema toxin), and a suppressed immune response (Liu et al., 2014; Moayeri & Leppla, 2004). The main targets of anthrax lethal toxin are cardiomyocytes and smooth muscle cells, whereas the main targets of anthrax edema toxin are the dermis, hepatocytes, and lungs (Liu et al., 2013; Lowe & Glomski, 2012). Anthrax disease progresses with attacks on multiple organ systems, but often the cause of host death is by the induction of apoptosis in the heart (Brojatsch et al., 2014; Golden et al., 2009). We therefore examined human DNase-I hypersensitivity data in 42 tissues from the Epigenome Roadmap project. We found substantial numbers of DNase-I hypersensitive sites overlapping the interval of high CLR values in heart, lung, and dermis (**Figure 3.4C, Supplementary Figure 3.10**). To determine whether *ANTXR2* expression varies among populations, we identified the population in GTEx RNA-seq data into European, East Asian, and African ancestry. Europeans consistently had lower expression on average than other populations (**Figure 3.4D**).

## *DISCUSSION*

During its natural life cycle, *B. anthracis* primarily affects domestic livestock species which ingest spores from the contaminated soil while grazing in endemic areas.

However, *B. anthracis* spores can be passed to a secondary host through either inhalation, ingestion, or cutaneous contact (Kamal et al., 2011). Thus, the prevalence of anthrax disease in humans is likely to have increased after humans began hunting or cultivating grazing herbivores in endemic areas. Historical accounts point to numerous cases where anthrax disease affected human populations, for instance as one of the ten plagues in Egypt, which affected both humans and their livestock (Schwartz, 2009), and Woolsorter's disease in Europe, which affected textile workers in close contact with hides from infected animals (Eurich, 1926; Laforce, 1978). Anthrax toxins enter cells primarily by binding to *ANTXR2*, which has around 1000-fold higher affinity for PA than the orthologous receptor *ANTXR1 (Young & Collier, 2007)*. The anthrax toxins cripple both the innate and adaptive immune system by causing apoptosis and/or rendering immune responses such as activation and cytokine production inoperative, such as in CD4+ T cells (Baldari et al., 2006). After evading the immune system, systemic anthrax disease causes failure of critical organs, including the heart, liver, and lung (Liu et al., 2014), leading to host death.

Here we present data demonstrating that the toxin receptor by which anthrax toxins gain entry to host cells, *ANTXR2*, decreased in humans relative to non-human primates. We identified CREs which may harbor the causal change, and validated that several have changes in activity in humans. In particular, CRE4 was a candidate causal change that stood out to us because it had decreased luciferase activity in human compared with both non-human primate species, exhibited changes in Pol II loading in humans, and was situated close to the *ANTXR2* transcription start site where most functional enhancers lie (Gasperini et al., 2019). Intriguingly, one of the candidate SNPs

between divergent TREs is still polymorphic in humans, and has an allele frequency that correlates with historical patterns of *B. anthracis* pathogens estimated from population genetic analyses of the *B. anthracis* bacterium (Kollek, 2004; Van Ert et al., 2007).

Differences in *ANTXR2* expression between human and rhesus macaque are only partially explained by variation within humans. In part, this may be explained by ascertainment bias in the currently available RNA-seq data, which is heavily biased for individuals of European or East Asian ancestry. Indeed, *ANTXR2* expression is higher in blood cells isolated from hunter gatherers than in nearby agricultural populations within Africa, suggesting that we are undersampling human variation (Harrison et al., 2019). Alternatively, changes in *ANTXR2* expression between species may have been driven by either drift or by selection on other roles of *ANTXR2* within the host. *ANTXR2* has been extensively studied in the context of anthrax disease, and the primary function of the receptor is not well understood. *ANTXR2* is a receptor for extracellular matrix proteins, such as collagen type IV and laminin, as part of the assembly of the basement membrane matrix, homeostasis of the extracellular matrix, and angiogenesis (Bell et al., 2001; Reeves et al., 2012; Ye et al., 2014). Either selection on these host-specific functions, or drift, may in part explain changes between humans and non-human primate species.

Haldane famously postulated that certain human genetic diseases are caused by selection to avoid pathogens. Haldane's hypothesis was centered around sickle-cell disease being correlated with areas endemic for malaria (J. B. S. Haldane, 1949), and it was later determined that the same SNP which confers malaria resistance causes sickle-

cell anemia (Allison, 1954). In examples of positive selection due to both malaria and African trypanosomiasis, the selected allele has a downstream consequence in disease (Allison, 1954; Genovese et al., 2010). The *ANTXR2* intron is associated with a GWAS SNP for autoimmunity (Farh et al., 2015) and for ankylosing spondylitis risk (Ou, 2015). In addition, SNPs within *ANTXR2* introns and exons are casual for hyaline fibromatosis syndrome (Bürgi et al., 2017; Nakai et al., 1986). *ANTXR2* may therefore be another example that supports notions originally formulated by Haldane that selection by pathogens can increase the allele frequency of disease-associated genetic changes in endemic areas.

## *METHODS*

### *EXPERIMENTAL METHODS*

#### Isolation of CD4+ T cells from humans and non-human primates

All human and animal experiments were done in compliance with Cornell University IRB and IACUC guidelines. We obtained peripheral blood samples (60–80 mL) from healthy adult male humans, chimpanzees, and rhesus macaques. Informed consent was obtained from all human subjects. To account for within-species variation in gene transcription we used three individuals to represent each primate species. Blood was collected into purple top EDTA tubes. Human samples were maintained overnight at 4C to mimic shipping non-human primate blood samples. Blood was mixed 50:50 with phosphate buffered saline (PBS). Peripheral blood mononuclear cells (PBMCs) were isolated by centrifugation (750× g) of 35 mL of blood:PBS over 15 mL Ficoll-Paque for

30 minutes at 20C. Cells were washed three times in ice cold PBS. CD4+ T cells were isolated using CD4 microbeads (Miltenyi Biotech, 130-045-101 [human and chimp], 130-091-102 [rhesus macaque]). Up to 10^8 PBMCs were resuspended in binding buffer (PBS with 0.5% BSA and 2mM EDTA). Cells were bound to CD4 microbeads (20uL of microbeads/107 cells) for 15 minutes at 4C in the dark. Cells were washed with 1–2 mL of PBS/BSA solution, resuspended in 500uL of binding buffer, and passed over a MACS LS column (Miltenyi Biotech, 130-042-401) on a neodymium magnet. The MACS LS column was washed three times with 2mL PBS/BSA solution, before being eluted off the neodymium magnet. Cells were counted in a hemocytometer.

**Luciferase assays**

Genomic DNA was isolated from human, chimp, and rhesus macaque PBMCs depleted for CD4+ cells using a Quick-DNA Miniprep Plus Kit (#D4068S; Zymo research) following the manufacturer's instructions. Putative enhancer regions were amplified from the genomic DNA, restriction digested with KpnI and MluI, and cloned into the pGL3-promoter vector (Promega). The same orthologous regions were amplified from all three species with identical primers where possible or species-specific primers covering orthologous DNA in diverged regions. The media (RPMI-1640) was changed in Jurkat cells one day before transfection. RPMI-1640 with 20% FBS was equilibrated in plates in a 37C incubator prior to transfection. On the day of transfection, Jurkat cells were centrifuged at 100xg for 10 minutes, washed with PBS, and centfrigured again. After centrifugation, Jurkat cells (2 million per reaction) were resuspended in 100 ul room temperature Mirus electroporation solution. Vectors were co-transfected with

pRL-SV40 Renilla (Promega) in a 20:1 ratio (2ug pGL3 to 100ng pRL-SV40). Electroporation was done in a Lonza Nucleofector 2b device using program X-001. Immediately after electroporation, 1 ml equilibrated media was added to the cuvette and then the cell mixture was added to 6 well plates and incubated at 37C. 18 hours post-transfection, luminescence was measured in triplicate using the Dual-Luciferase® Reporter Assay System (Promega).

**CRISPRa in K562**

Generation of dCas9-KRAB K562 line—

Lentivirus was made using lipofectamine 3000 from Invitrogen. Phoenix Hek cells (grown in DMEM with 10% FBS and antibiotics) were seeded in a 6-well plate at 400,000 cells/plate. Cells were grown until ~90% confluent. 1ug of pHAGE_EF1a_dCas9-KRAB plasmid from addgene (#50919) plasmid was transfected. 24 hours later 3ml/well of virus was mixed with 10ug/ml polybrene and incubated for 5 minutes at room temperature. This mix was added to 300,000 K562 cells and centrifuged for 40 minutes at 800g at 32C. 12–24 hours later the virus was removed and fresh media was added. 24–48 hours later the cells were selected with 150ug/ml Hygromycin B for 2 weeks. The K562 dCas9-KRAB stable cell lines was grown and maintained in Hygromycin B.

sgRNA cloning—

Primers for sgRNAs were designed using ChopChop (http://chopchop.cbu.uib.no/) for the *ANTXR2* gene and scrambled controls. Primers were located -400 to -50 bp away

from the TSS for CRISPRa. A G was added at the 5' end of primers for use with a U6 promoter, along with restriction sites for cloning. Forward and reverse sgRNAs were synthesized separately by IDT and annealed. T4 Polynucleotide Kinase (NEB) was used to phosphorylate the forward and reverse sgRNA during the annealing. 10× T4 DNA Ligase Buffer, which contains 1mM ATP, was incubated for 30 minutes at 37°C and then at 95C for 5 minutes, decreasing by 5°C every 1 minute until 25°C. Oligos were diluted 1:200 using Molecular grade water. sgRNAs were inserted into the pLenti SpBsmBI sgRNA Hygro plasmid from addgene (#62205) by following the authors protocol (26501517). The plasmid was linearized using BsmBI digestion (NEB) and purified using gel extraction (QIAquick Gel Extraction Kit). The purified linear plasmid was then dephosphorylated using Alkaline Phosphatase Calf Intestinal (CIP) (NEB) to ensure the linear plasmid did not ligate with itself. A second gel extraction was used as before to purify the linearized plasmid. The purified dephosphorylated linear plasmid and phosphorylated annealed oligos were ligated together using the Quick Ligation Kit (NEB). The ligated product was transformed into One Shot Stbl3 Chemically Competent E. coli (ThermoFisher Scientific). 100ul of the transformed bacteria were plated on Ampicillin (200ug/ml) plates. Single colonies were picked, sequenced, and the plasmid was isolated using endo free midi-preps from Omega.

Transfection of sgRNA plasmid—

The day prior to transfection, the media (RPMI-1640 with 10% media) was changed in K562 cells and cells were diluted to a concentration of 1 million cells/mL. On the day of transfection, K562 cells were centrifuged at 100xg for 10 minutes, washed with PBS,

and centrifuged again. RPMI-1640 with 10% FBS was equilibrated in plates in a 37C incubator prior to transfection. After centrifugation, K562 cells (1 million per reaction) were resuspended in 100 ul room temperature Mirus electroporation solution. 2ug of the sgRNA plasmid was added to each reaction. Electroporation was done in a Lonza Nucleofector 2b device using the program for K562 cells. Immediately after electroporation, 1 ml equilibrated media was added to the cuvette and then the cell mixture was added to 6 well plates and incubated at 37C. 12 hours after transfection, 2ug/ml doxycycline was added to cells to activate dCas9-KRAB expression. To confirm overexpression of *ANTXR2*, 4 hours after the addition of doxycycline a portion of the cells were collected for RNA extraction using Trizol. cDNA was generated from RNA samples using the Thermo Fisher High Capacity RNA-to-cDNA kit and qPCR was performed using SsoAdvanced Universal SYBR Green master mix with primers to assay *ANTXR2* expression.

**Toxin viability assays**

K562 cells transfected with CRISPRa plasmids were confirmed to have overexpression of *ANTXR2*. After confirmation and within the 12 hour half-life window of the activating doxycycline, 50,000 K562 cells were added to wells of 96-well plates in 100 ul RPMI-1640, supplemented with 10% RPMI. Anthrax toxin PA (List Biological Laboratories 171D) was added to wells at a concentration of 1ug/ml (or vehicle control). FP59, a recombinant anthrax lethal factor fused to the Pseudomonas Exotoxin A Catalytic Domain (Kerafast ENH013), which is capable of killing blood cells, was added at a concentration of 50ng/ml (or vehicle control). 20 hours after the addition of

PA and FP59, 10ul of Alamar Blue (Thermo Fisher #DAL1025) was added to each well. Four hours after the addition of Alamar Blue, fluorescence was measured using a plate reader with an excitation of 570 and emission of 610.

**Activation of CD4+ T cells**

CD4+ T cells were isolated using the above procedure for two human subjects. After equilibration in RPMI-1640 with 10% FBS, cells were stimulated with 25ng/mL PMA and 1mM Ionomycin (P/I or $\pi$) or vehicle control (2.5uL EtOH and 1.66uL DMSO in 10mL of culture media). Thirty minutes after activation or addition of the vehicle control, cells were treated with 2.5ug/ml Recombinant PA (List Biological Laboratories 171D) and 500ng/ml Recombinant LF (List Biological Laboratories 172A) or vehicle control. 24 hours later, media from non-activated, non-activated with toxin treatment, activated, and activated with toxin treatment wells was collected for ELISA.

**IL-2 ELISA on CD4+ T cells**

ELISA was done using the R&D Human IL-2 DuoSet Kit (Catalog #DY202).

Plate preparation—16 hours prior to the ELISA experiment, the capture antibody was added to 96-well plates at a concentration of 0.5ug/ul in 100ul of PBS. Plates were sealed and left at room temperature. The following day, the diluted capture antibody was aspirated and washed with 400ul Wash Buffer three times. 300ul of Block Buffer was added to each well and incubated for at least one hour. After incubation, plates were washed with 400ul Wash Buffer.

ELISA assay—100ul of RPMI from the samples or standards diluted in the Reagent Diluent were added to the prepared 96-well plate. The plate was covered and left to incubate for 2 hours at room temperature. 100ul of the Detection Antibody diluted at 1:60 with the Reagent Diluent was added to each well. The plate was washed with Wash Buffer. 100ul of Streptavidin HRP diluted at 1:40 in the Reagent Diluent was added to each well. The plate was covered, protected from light, and incubated for 20 minutes at room temperature. The plate was washed and 100ul of the Substrate Solution was added to each well, the plate was covered, protected from light, and incubated for 20 minutes at room temperature. 50ul of Stop Solution was added to each well and mixed. Fluorescence was measured on a plate reader at 450nm and wavelength corrected at 540nm.

**Hi-C library preparation**

Cell preparation—

CD4+ T cells were isolated according to the above procedure. After >1 hour of equilibration in RPMI-1640 supplemented with 10% FBS, cells were centrifuged at 300xg, washed with PBS, and centrifuged again. Cells were resuspended in a mixture of 1% paraformaldehyde in 1x PBS. Cells were incubated at room temperature for 10 min on a rocker. Paraformaldehyde was quenched by the addition of 2.5M Glycine to a Cf=0.2M. Cells were incubated for room temperature for 5 minutes on a rocker. Cells were centrifuged at 4C, washed in cold PBS, centrifuged, and PBS was aspirated. Pellets were flash frozen using dry ice and stored at -80C prior to library preparation.

Hi-C—

The protocol detailed in (Rao et al., 2014) was followed with the following adjustments. After the addition of lysis buffer, cells were incubated on ice for 30 min. MboI (NEB #R0147) was used for restriction digestion. Following DNA purification, unligated biotin was removed using a mixture of 0.5uL of 10mM dATP, 0.5uL dGTP, 20uL 3000U/ml T4 DNA polymerase (NEB #M02030) for each sample. Samples were incubated for 4 hours at room temperature and then T4 DNA polymerase was inactivated at 72C for 20min. Shearing was done using a Bioruptor sonicator using the LOW setting 30S ON/ 90S OFF for 2 cycles of 10 minutes. Libraries were prepared with the NEBNext Ultra II Library Preparation Kit (NEB #E7103). Samples were sequenced on a combination of Illumina's NovaSeq 6000 and HiSeq 4000 at Novogene.

*DATA ANALYSIS*

**Hi-C analysis and visualization**

Individual Hi-C samples were mapped to hg19 (for human samples) or rheMac8 (for rhesus macaque samples) using Juicer (Durand et al., 2016). Replicates for each species were combined using Juicer's mega function. The combined rhesus macaque aligned dataset was lifted over to hg19 using Crossmap (H. Zhao et al., 2014).

**Mapping orthologs between species**

Cross-species comparison of genomic coordinates and genes was based on the methods using in (Danko et al., 2018). Briefly, all datasets for chimpanzee and rhesus macaque were converted to the human assembly (hg19) using CrossMap (H. Zhao et al., 2014).

Reciprocal-best (rbest) nets were used to convert genomic coordinates between genome assemblies using (Kent et al., 2003).

**PRO-seq and RNA-seq differential expression**

We mapped PRO-seq reads using standard informatics tools. Our PRO-seq mapping pipeline begins by removing reads that fail Illumina quality filters and trimming adapters using cutadapt with a 10% error rate. Reads were mapped with BWA (Li & Durbin, 2010) to the appropriate reference genome (either hg19, panTro4, or rheMac3) and a single copy of the Pol I ribosomal RNA transcription unit (GenBank ID# U13369.1). Mapped reads were converted to bigWig format for analysis using BedTools (Quinlan & Hall, 2010) and the bedGraphToBigWig program in the Kent Source software package (Kuhn et al., 2013). The location of the RNA polymerase active site was represented by the single base, the 3′ end of the nascent RNA, which is the position on the 5′ end of each sequenced read.

**DICE RNA-seq and eQTL analysis**

RNA-seq from 85 human CD4+ naive T cell samples from DICE (Schmiedel et al., 2018) was downloaded under dbGap protocol 23187. RNA-seq data was mapped using Salmon (Patro et al., 2017) and NCBI RefSeq genes to hg19. Rhesus macaque RNA-seq data was mapped to hg19 using the same parameters. *ANTXR2* transcripts per million were compared between samples for transcript NM_001145794.1. We retrieved eQTLs for *ANTXR2* from the DICE online database.

**Sweepfinder2 CLR scan**

Human genome data was taken from phase3 of the 1,000 Genomes Project (1000 Genomes Project Consortium et al., 2015). VCFs were subsetted based on their population group. The b-value maps that are used to compute the effect of background selection on the human genome were taken from (McVicker et al., 2009). Recombination maps from the deCODE database (Kong et al., 2010) were used. Sweepfinder2 (DeGiorgio et al., 2016) was used to calculate the composite-likelihood-ratio for CEU, JPT, CHB, and YRI on chromosome 4.

**Fst analysis**

All Fst values were computed using VCFtools (Danecek et al., 2011) using the --weir-fst command with a window size of 5kb and a step size of 5kb. In addition to the Fst calculations for bins of 5kb, Fst was calculated for each SNP of chromosome 4. Fst was calculated for all pairwise comparisons of CEU, YRI, JPT and CHB.

## *REFERENCES*


1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malareal infection. *British Medical Journal*, *1*(4857), 290–294.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461.

Baldari, C. T., Tonello, F., Paccani, S. R., & Montecucco, C. (2006). Anthrax toxins: A paradigm of bacterial immune suppression. *Trends in Immunology*, *27*(9), 434–440.

Banks, D. J., Barnajian, M., Maldonado-Arocho, F. J., Sanchez, A. M., & Bradley, K. A. (2005). Anthrax toxin receptor 2 mediates Bacillus anthracis killing of macrophages following spore challenge. *Cellular Microbiology*, *7*(8), 1173–1185.

Bell, S. E., Mavila, A., Salazar, R., Bayless, K. J., Kanagala, S., Maxwell, S. A., & Davis, G. E. (2001). Differential gene expression during capillary morphogenesis in 3D collagen matrices: regulated expression of genes involved in basement membrane matrix assembly, cell cycle progression, cellular differentiation and G-protein signaling. *Journal of Cell Science*, *114*(Pt 15), 2755–2773.

Brojatsch, J., Casadevall, A., & Goldman, D. L. (2014). Molecular determinants for a cardiovascular collapse in anthrax. *Frontiers in Bioscience* , *6*, 139–147.

Bürgi, J., Kunz, B., Abrami, L., Deuquet, J., Piersigilli, A., Scholl-Bürgi, S., Lausch, E., Unger, S., Superti-Furga, A., Bonaldo, P., & van der Goot, F. G. (2017). CMG2/ANTXR2 regulates extracellular collagen VI which accumulates in hyaline fibromatosis syndrome. *Nature Communications*, *8*, 15861.

Chrousos, G. P., Renquist, D., Brandon, D., Eil, C., Pugeat, M., Vigersky, R., Cutler, G. B., Jr, Loriaux, D. L., & Lipsett, M. B. (1982). Glucocorticoid hormone resistance during primate evolution: receptor-mediated mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(6), 2036–2040.

Comer, J. E., Chopra, A. K., Peterson, J. W., & König, R. (2005). Direct inhibition of T-lymphocyte activation by anthrax toxins in vivo. *Infection and Immunity*, *73*(12), 8275–8281.

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian

promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320.

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H. A., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J., & Furlong, E. E. M. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, *555*(7697), 538–542.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics* , *27*(15), 2156–2158.

Danko, C. G., Choate, L. A., Marks, B. A., Rice, E. J., Wang, Z., Chu, T., Martins, A. L., Dukler, N., Coonrod, S. A., Tait Wojno, E. D., Lis, J. T., Kraus, W. L., & Siepel, A. (2018). Dynamic evolution of regulatory element ensembles in primate CD4+ T cells. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-017-0447-5

Danko, C. G., Hyland, S. L., Core, L. J., Martins, A. L., Waters, C. T., Lee, H. W., Cheung, V. G., Kraus, W. L., Lis, J. T., & Siepel, A. (2015). Identification of active transcriptional regulatory elements from GRO-seq data. *Nature Methods*, *12*(5), 433–438.

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics* , *32*(12), 1895–1897.

Denny, W. B., Valentine, D. L., Reynolds, P. D., Smith, D. F., & Scammell, J. G. (2000). Squirrel monkey immunophilin FKBP51 is a potent inhibitor of glucocorticoid receptor binding. *Endocrinology*, *141*(11), 4107–4113.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, *3*(1), 95–98.

Enard, D., & Petrov, D. A. (2018). Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell*, *175*(2), 360–371.e13.

Enard, D., & Petrov, D. A. (2020). Ancient RNA virus epidemics through the lens of recent adaptation in human genomes. In *bioRxiv* (p. 2020.03.18.997346). https://doi.org/10.1101/2020.03.18.997346

Eurich, F. W. (1926). THE HISTORY OF ANTHRAX IN THE WOOL INDUSTRY OF BRADFORD, AND OF ITS CONTROL. *The Lancet*, *207*(5341), 107–110.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., … Bernstein, B. E. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, *518*(7539), 337–343.

Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular Evolution of Human Coronavirus Genomes. *Trends in Microbiology*, *25*(1), 35–48.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L., & Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genetics*, *7*(11), e1002355.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., & Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, *176*(1-2), 377–390.e19.

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., Bernhardy, A. J., Hicks, P. J., Nelson, G. W., Vanhollebeke, B., Winkler, C. A., Kopp, J. B., Pays, E., & Pollak, M. R. (2010). Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. In *Science* (Vol. 329, Issue 5993, pp. 841–845). https://doi.org/10.1126/science.1193032

Golden, H. B., Watson, L. E., Lal, H., Verma, S. K., Foster, D. M., Kuo, S.-R., Sharma, A., Frankel, A., & Dostal, D. E. (2009). Anthrax toxin: pathologic effects on the cardiovascular system. *Frontiers in Bioscience , 14,* 2335–2357.

Gonzalez, J. P., Nakoune, E., Slenczka, W., Vidal, P., & Morvan, J. M. (2000). Ebola and Marburg virus antibody prevalence in selected populations of the Central African Republic. *Microbes and Infection / Institut Pasteur*, *2*(1), 39–44.

Haldane, J. B. S. (1949). The rate of mutation of human genes. *Hereditas*, *35*(S1), 267–273.

Haldane, J. B. S. (2006). Disease and Evolution. In K. R. Dronamraju & P. Arese (Eds.), *Malaria: Genetic and Evolutionary Aspects* (pp. 175–187). Springer US.

Harrison, G. F., Sanz, J., Boulais, J., Mina, M. J., Grenier, J.-C., Leng, Y., Dumaine, A., Yotova, V., Bergey, C. M., Nsobya, S. L., Elledge, S. J., Schurr, E., Quintana-Murci, L., Perry, G. H., & Barreiro, L. B. (2019). Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. *Nature Ecology & Evolution*, *3*(8), 1253–1264.

Johnson, E. D., Gonzalez, J. P., & Georges, A. (1993). Filovirus activity among selected ethnic groups inhabiting the tropical forest of equatorial Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *87*(5), 536–538.

Kalay, G., & Wittkopp, P. J. (2010). Nomadic enhancers: tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species. *PLoS Genetics*, *6*(11), e1001222.

Kamal, S. M., Rashid, A. K. M. M., Bakar, M. A., & Ahad, M. A. (2011). Anthrax: an update. *Asian Pacific Journal of Tropical Biomedicine*, *1*(6), 496–501.

Karlsson, E. K., Kwiatkowski, D. P., & Sabeti, P. C. (2014). Natural selection and infectious disease in human populations. *Nature Reviews. Genetics*, *15*(6), 379–393.

Keim, P., Kalif, A., Schupp, J., Hill, K., Travis, S. E., Richmond, K., Adair, D. M., Hugh-Jones, M., Kuske, C. R., & Jackson, P. (1997). Molecular evolution and diversity in Bacillus anthracis as detected by amplified fragment length polymorphism markers. *Journal of Bacteriology*, *179*(3), 818–824.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., & Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11484–11489.

Kollek, D. (2004). "The history of anthrax", by Sternbach G [Review of *"The history of anthrax", by Sternbach G*]. *The Journal of Emergency Medicine*, *26*(3), 354; author reply 354.

Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G. B., Jonasdottir, A., Gylfason, A., Kristinsson, K. T., Gudjonsson, S. A., Frigge, M. L., Helgason, A., Thorsteinsdottir, U., & Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, *467*(7319), 1099–1103.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., & Siepel, A. (2008). Patterns of positive selection in six Mammalian genomes. *PLoS Genetics*, *4*(8), e1000144.

Ko, W.-Y., Rajan, P., Gomez, F., Scheinfeldt, L., An, P., Winkler, C. A., Froment, A., Nyambo, T. B., Omar, S. A., Wambebe, C., Ranciaro, A., Hirbo, J. B., & Tishkoff, S. A. (2013). Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. *American Journal of Human Genetics*, *93*(1), 54–66.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161.

Laforce, F. M. (1978). Woolsorters' disease in England. *Bulletin of the New York Academy of Medicine*, *54*(10), 956–963.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* , *26*(5), 589–595.

Liu, S., Bugge, T. H., Leppla, S. H. (2001). Targeting of tumor cells by cell surface urokinase plasminogen activator-dependent anthrax toxin. *Journal of Biological Chemistry*, 276(21), 17976-17984.

Liu, S., Moayeri, M., & Leppla, S. H. (2014). Anthrax lethal and edema toxins in anthrax pathogenesis. *Trends in Microbiology*, *22*(6), 317–325.

Liu, S., Zhang, Y., Moayeri, M., Liu, J., Crown, D., Fattah, R. J., Wein, A. N., Yu, Z.-X., Finkel, T., & Leppla, S. H. (2013). Key tissue targets responsible for anthrax-toxin-induced lethality. *Nature*, *501*(7465), 63–68.

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

Lowe, D. E., & Glomski, I. J. (2012). Cellular and physiological effects of anthrax exotoxin and its relevance to disease. *Frontiers in Cellular and Infection Microbiology*, *2*, 76.

Luckheeram, R. V., Zhou, R., Verma, A. D., & Xia, B. (2012). CD4+T cells: differentiation and functions. *Clinical & Developmental Immunology*, *2012*, 925135.

Lv, F.-H., Agha, S., Kantanen, J., Colli, L., Stucki, S., Kijas, J. W., Joost, S., Li, M.-H., & Ajmone Marsan, P. (2014). Adaptations to climate-mediated selective pressures in sheep. *Molecular Biology and Evolution*, *31*(12), 3324–3343.

Martchenko, M., Candille, S. I., Tang, H., & Cohen, S. N. (2012). Human genetic variation altering anthrax toxin sensitivity. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(8), 2972–2977.

McVicker, G., Gordon, D., Davis, C., & Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, *5*(5), e1000471.

Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A., & Yampolskaya, O. (1994). The Sverdlovsk anthrax outbreak of 1979. *Science*, *266*(5188), 1202–1208.

Moayeri, M., & Leppla, S. H. (2004). The roles of anthrax toxin in pathogenesis. *Current Opinion in Microbiology*, *7*(1), 19–24.

Moayeri, M., & Leppla, S. H. (2009). Cellular and systemic effects of anthrax lethal toxin and edema toxin. *Molecular Aspects of Medicine*, *30*(6), 439–455.

Mwenye, K. S., Siziya, S., & Peterson, D. (1996). Factors associated with human anthrax outbreak in the Chikupo and Ngandu villages of Murewa district in Mashonaland East Province, Zimbabwe. *The Central African Journal of Medicine*, *42*(11), 312–315.

Nakai, A., Nagasaka, A., Ohyama, T., Aono, T., Iwase, K., Hasegawa, H., Hayami, S., Hidaka, H., Tanaka, T., & Niinomi, M. (1986). High activity of cyclic 3',5'-nucleotide phosphodiesterase in sera of patient with phaeochromocytoma. *Clinical Endocrinology*, *24*(4), 409–414.

Nguyen, D. H., Hurtado-Ziola, N., Gagneux, P., & Varki, A. (2006). Loss of Siglec expression on T lymphocytes during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(20), 7765–7770.

Ou, Y. (2015). Anthrax toxin receptor 2 gene (ANTXR2) rs4333130 is associated with ankylosing spondylitis. *International Journal of Clinical and Experimental Medicine*, *8*(5), 7679–7683.

Paccani, S. R., Tonello, F., Ghittoni, R., Natale, M., Muraro, L., D'Elios, M. M., Tang, W.-J., Montecucco, C., & Baldari, C. T. (2005). Anthrax toxins suppress T lymphocyte activation by disrupting antigen receptor signaling. *The Journal of Experimental Medicine*, *201*(3), 325–331.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Perez-Pinera, P., Kocak, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., Thakore, P. I., Glass, K. A., Ousterout, D. G., Leong, K. W., Guilak, F., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature Methods*, *10*(10), 973–976.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121.

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

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

Reeves, C. V., Wang, X., Charles-Horvath, P. C., Vink, J. Y., Borisenko, V. Y., Young, J. A. T., & Kitajewski, J. K. (2012). Anthrax toxin receptor 2 functions in ECM homeostasis of the murine reproductive tract and promotes MMP activity. *PloS One*, *7*(4), e34862.

Reynolds, P. D., Ruan, Y., Smith, D. F., & Scammell, J. G. (1999). Glucocorticoid Resistance in the Squirrel Monkey Is Associated with Overexpression of the Immunophilin FKBP51. *The Journal of Clinical Endocrinology and Metabolism*, *84*(2), 663–669.

Sabeti, P. C., The International HapMap Consortium, Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. In *Nature* (Vol. 449, Issue 7164, pp. 913–918). https://doi.org/10.1038/nature06250

Scammell, J. G., Denny, W. B., Valentine, D. L., & Smith, D. F. (2001). Overexpression of the FK506-binding immunophilin FKBP51 is the common cause of glucocorticoid resistance in three New World primates. *General and Comparative Endocrinology*, *124*(2), 152–165.

Schmid, G., & Kaufmann, A. (2002). Anthrax in Europe: its epidemiology, clinical characteristics, and role in bioterrorism. *Clinical Microbiology and Infection: The Official*

*Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *8*(8), 479–488.

Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J. A., McVicker, G., Seumois, G., Rao, A., Kronenberg, M., Peters, B., & Vijayanand, P. (2018). Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*, *175*(6), 1701–1715.e16.

Schwartz, M. (2009). Dr. Jekyll and Mr. Hyde: a short history of anthrax. *Molecular Aspects of Medicine*, *30*(6), 347–355.

Scobie, H. M., Rainey, G. J. A., Bradley, K. A., & Young, J. A. T. (2003). Human capillary morphogenesis protein 2 functions as an anthrax toxin receptor. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(9), 5170–5174.

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., & Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell*, *58*(6), 1101–1112.

Shultz, A. J., & Sackton, T. B. (2019). Immune genes are hotspots of shared positive selection across birds and mammals. *eLife*, *8*. https://doi.org/10.7554/eLife.41815

Spencer, R. C. (2003). Bacillus anthracis. *Journal of Clinical Pathology*, *56*(3), 182–187.

Sternbach, G. (2003). The history of anthrax. *The Journal of Emergency Medicine*, *24*(4), 463–467.

Sun, J., & Jacquez, P. (2016). Roles of Anthrax Toxin Receptor 2 in Anthrax Toxin Membrane Insertion and Pore Formation. *Toxins*, *8*(2), 34.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., & Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, *24*(3), 390–400.

Tippens, N. D., Liang, J., Leung, K. Y., Ozer, A., Booth, J. G., Lis, J., & Yu, H. (2019). Transcription imparts architecture, function, and logic to enhancer units. In *bioRxiv* (p. 818849). https://doi.org/10.1101/818849

Van Ert, M. N., Easterday, W. R., Huynh, L. Y., Okinaka, R. T., Hugh-Jones, M. E., Ravel, J., Zanecki, S. R., Pearson, T., Simonson, T. S., U'Ren, J. M., Kachur, S. M., Leadem-Dougherty, R. R., Rhoton, S. D., Zinser, G., Farlow, J., Coker, P. R., Smith, K. L., Wang, B., Kenefic, L. J., … Keim, P. (2007). Global genetic population structure of Bacillus anthracis. *PloS One*, *2*(5), e461.

Wang, Z., Chu, T., Choate, L. A., & Danko, C. G. (2018). Identification of regulatory elements from nascent transcription using dREG. *Genome Research*, *29*(2), 293–303.

Wolfe, N. D., Dunavan, C. P., & Diamond, J. (2007). Origins of major human infectious diseases. *Nature*, *447*(7142), 279–283.

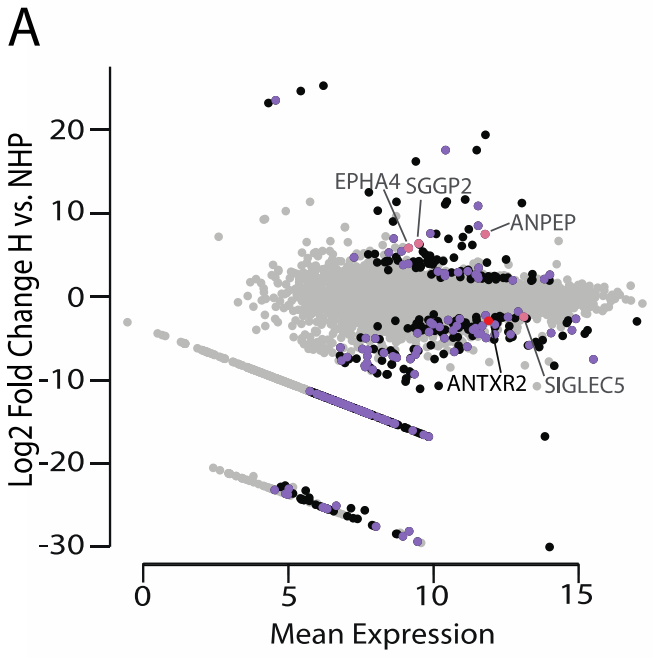
Ye, L., Sun, P.-H., Sanders, A. J., Martin, T. A., Lane, J., Mason, M. D., & Jiang, W. G. (2014). Therapeutic potential of capillary morphogenesis gene 2 extracellular vWA domain in tumour-related angiogenesis. *International Journal of Oncology*, *45*(4), 1565–1573.

Young, J. A. T., & Collier, R. J. (2007). Anthrax toxin: receptor binding, internalization, pore formation, and translocation. *Annual Review of Biochemistry*, *76*, 243–265.

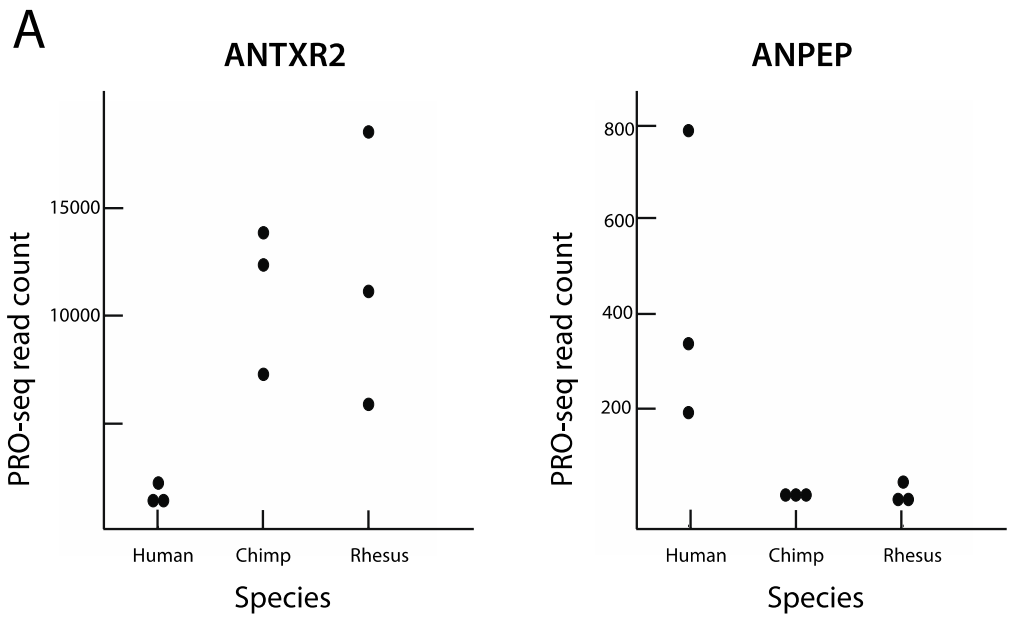
Zhao, F., McParland, S., Kearney, F., Du, L., & Berry, D. P. (2015). Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genetics, Selection, Evolution: GSE*, *47*, 49.

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.

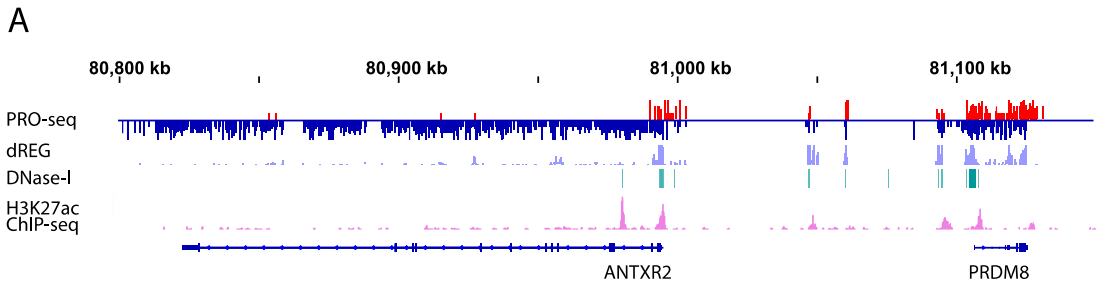
***SUPPLEMENTARY FIGURES***




**Supplementary Figure 3.1. Differential RNA expression between humans and non-human primates. A)** RNA-seq comparison of protein-coding genes between humans and non-human primates in CD4+ T cells. The GO term 'integral component of the membrane' is enriched in differentially expressed genes.

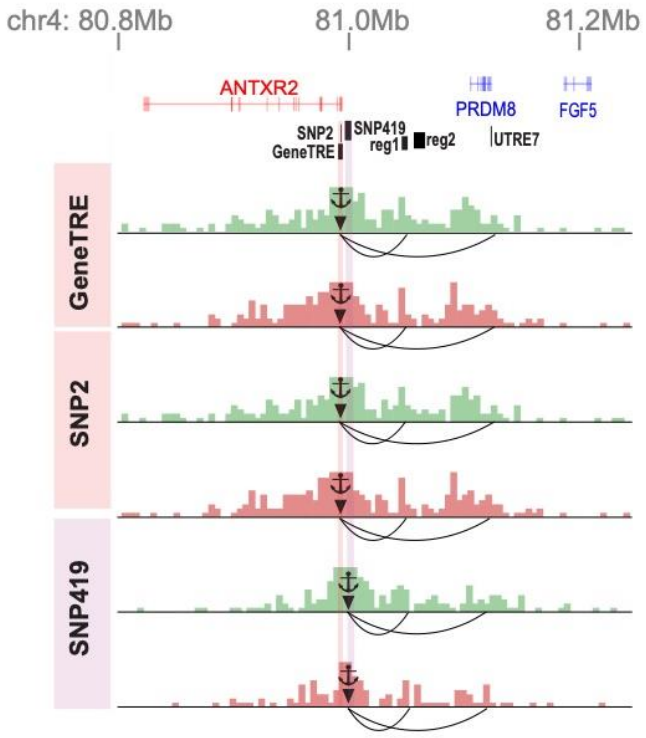



**Supplementary Figure 3.2. PRO-seq expression for *ANTXR2* and *ANPEP*. A)** PRO-seq of human, chimpanzee, and rhesus macaque for *ANTXR2* and *ANPEP*.

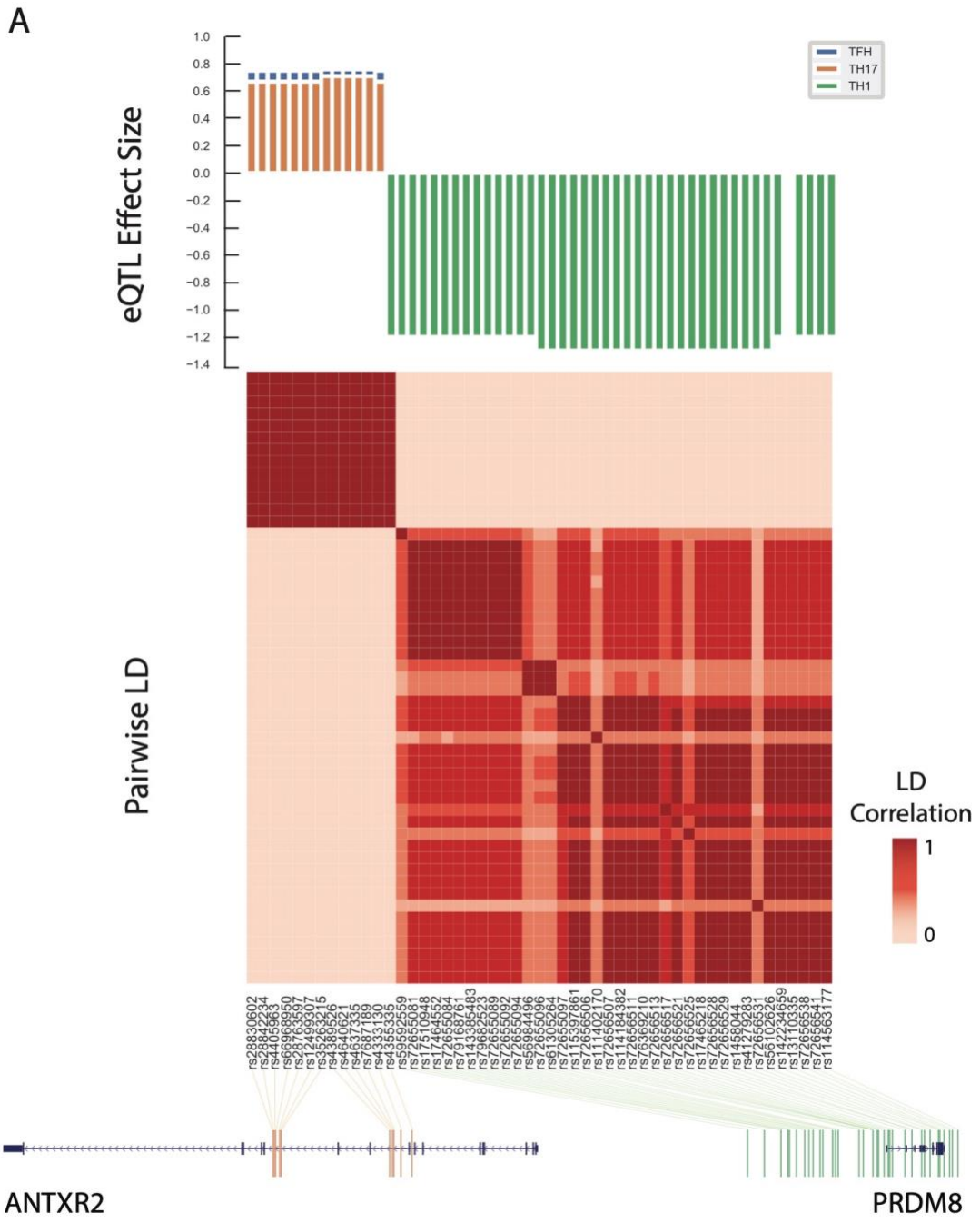



**Supplementary Figure 3.3. H3K27ac and DNase-I-seq at *ANTXR2*. A)** PRO-seq, dREG signal, DNase-I-seq peaks, and H3K27ac ChIP-seq at the *ANTXR2* locus.

A

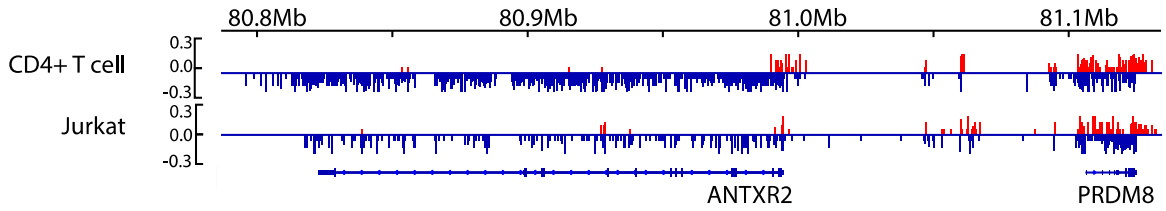



**Supplementary Figure 3.4. Virtual 4C-seq of proximal CREs. A)** Virtual 4C-seq plots of CRE2, CRE3, and CRE4 show contact with the upstream CREs that had significantly higher activity in non-human primates.

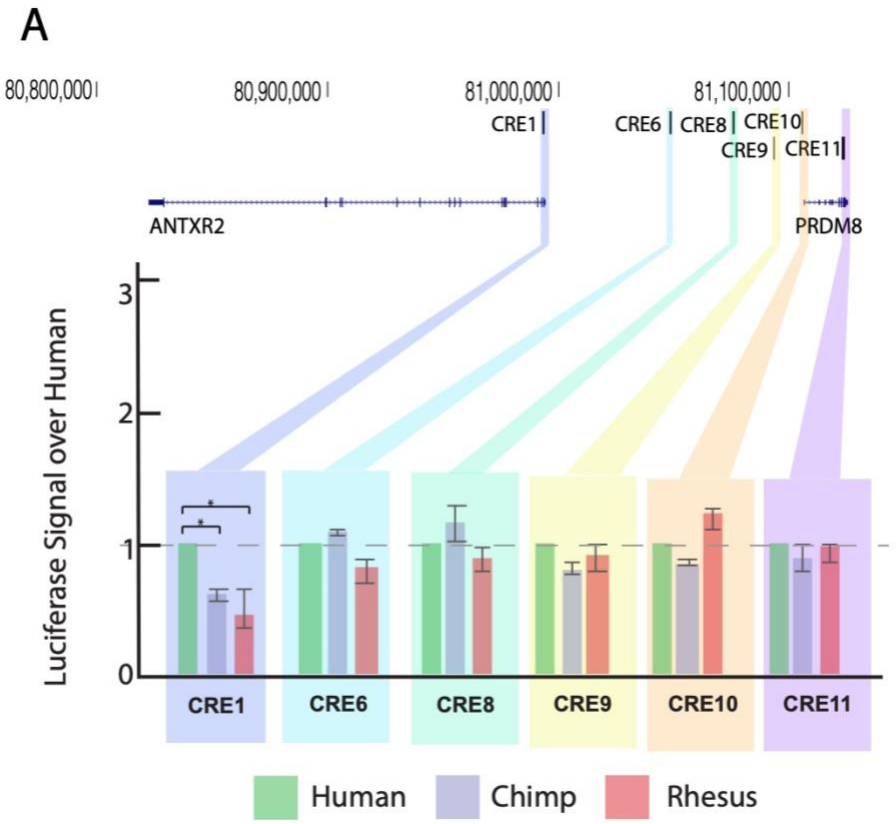



**Supplementary Figure 3.5. *ANTXR2* eQTLs in humans. A)** Expression eQTLs for *ANTXR2* fall into two main regions: within the gene and upstream of *ANTXR2* around *PRDM8*. Genic eQTLs are in linkage disequilibrium (LD) and have a positive effect on *ANTXR2* expression. Upstream eQTLs fall within two blocks of LD and have a negative effect on *ANTXR2* expression.

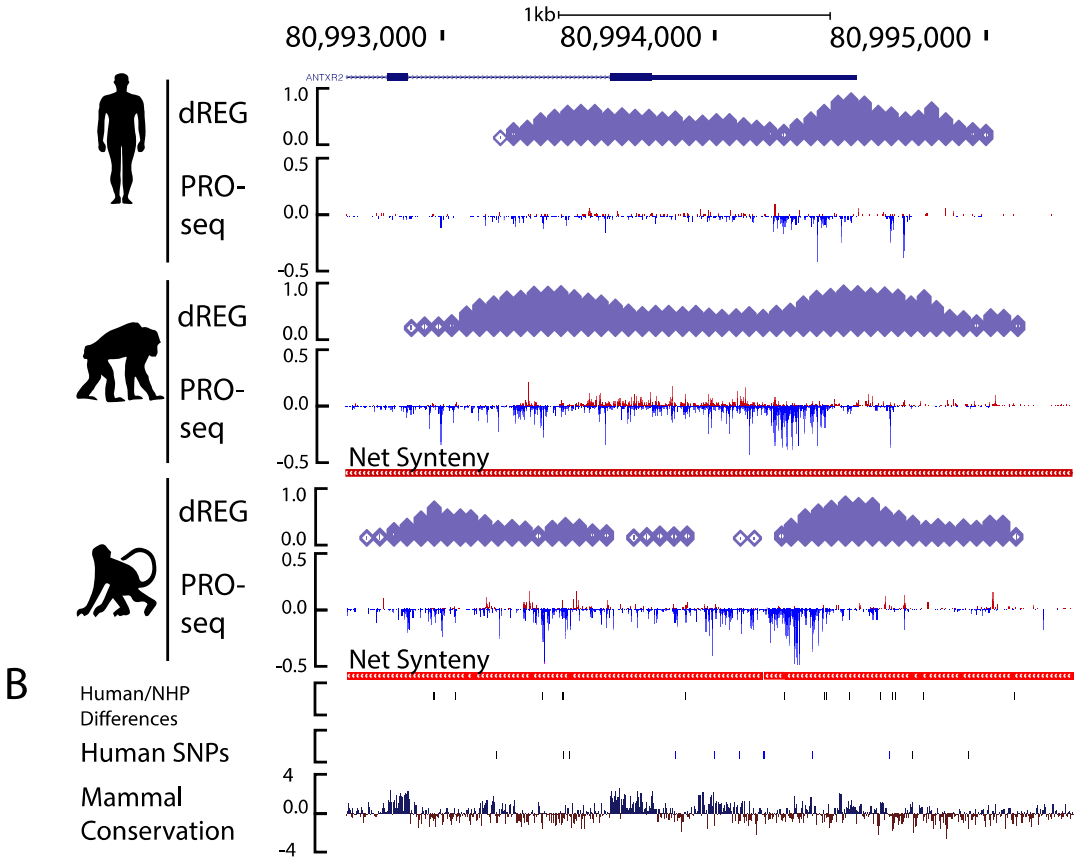
A



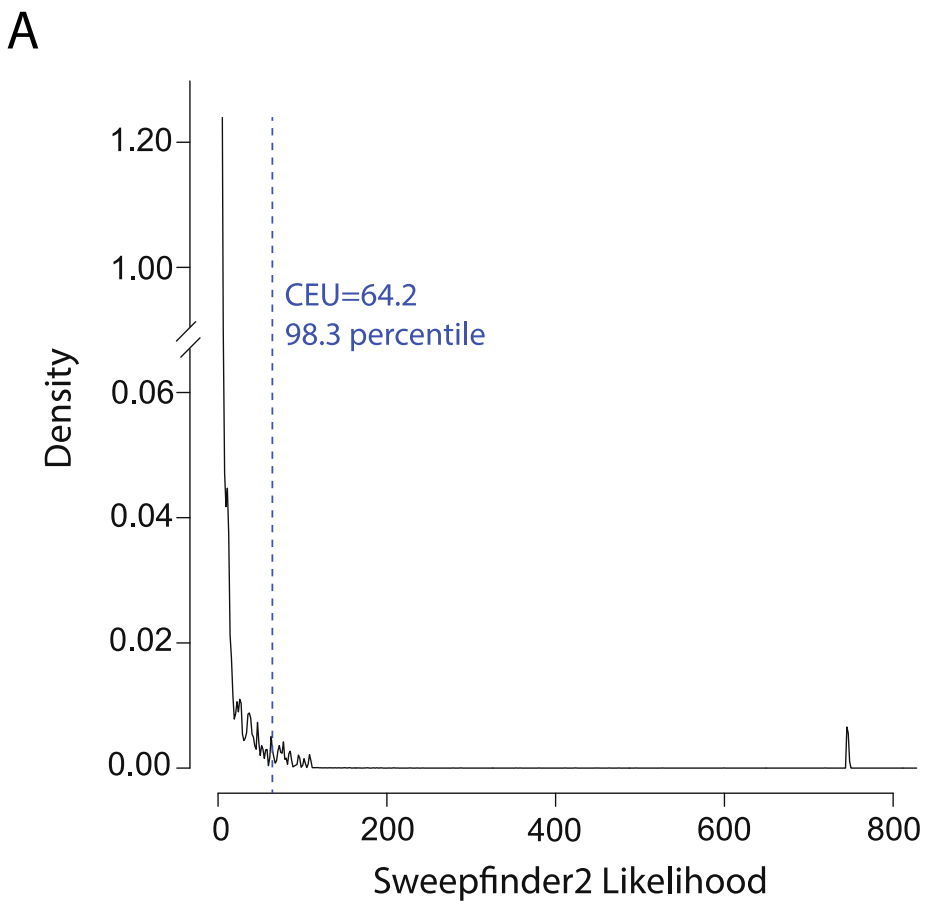

**Supplementary Figure 3.6. Jurkat vs. CD4 PRO-seq. A)** PRO-seq from Jurkat and human CD4+ T cells shows a similar regulatory landscape around *ANTXR2*.



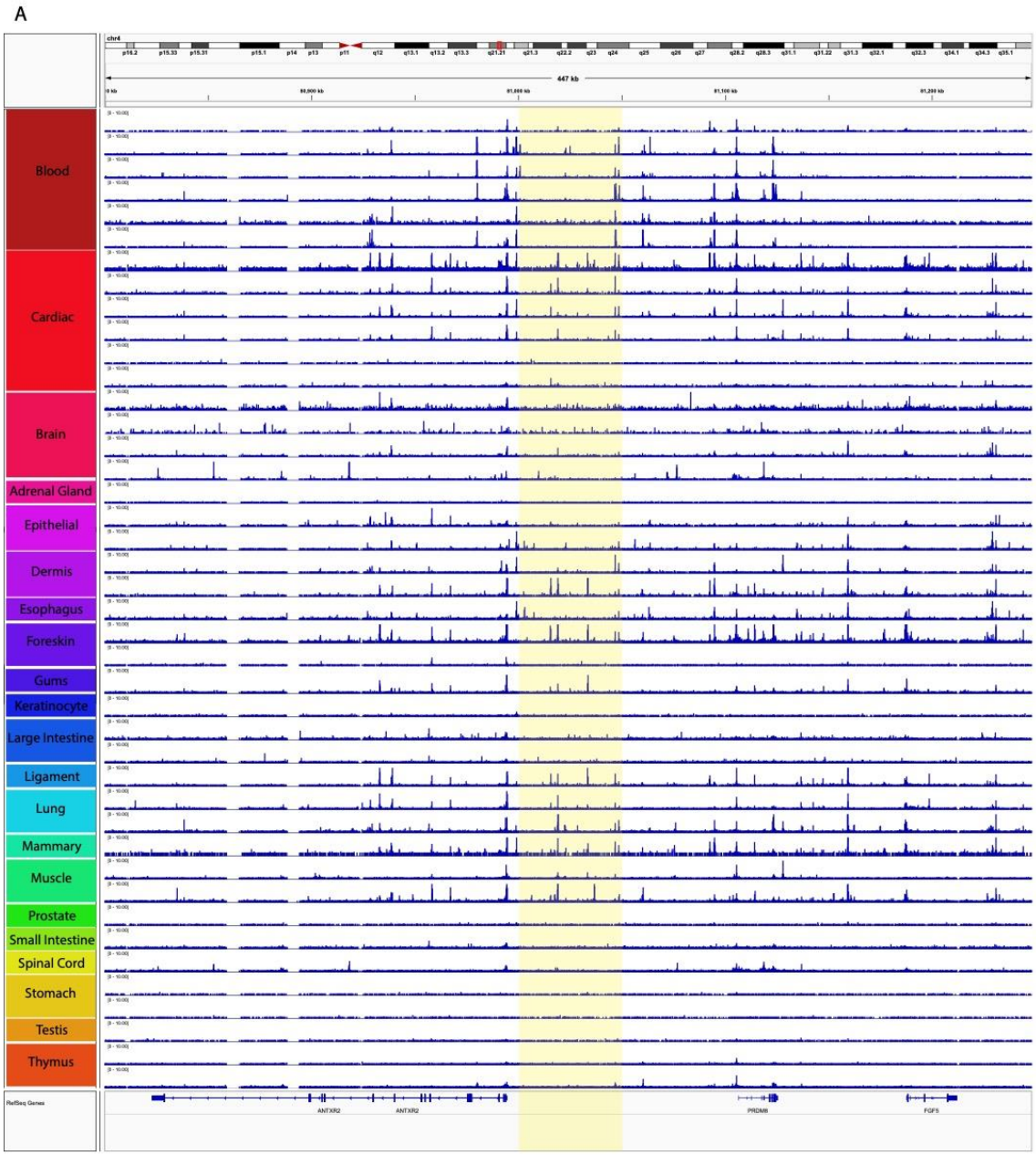

**Supplementary Figure 3.7. Luciferase data of other CREs. A)** Luciferase assay performed in Jurkat cells to test the activity of regulatory elements in human, chimpanzee, and rhesus macaque shows a decrease in activity for CRE1 in humans compared to chimpanzee and rhesus macaque and no change for other CREs.




**Supplementary Figure 3.8. Complex promoter of *ANTXR2*. A)** PRO-seq and dREG signal from human, chimpanzee, and rhesus macaque at the *ANTXR2* promoter. **B)** Human-specific changes, human common SNPs, and PhyloP conservation of the *ANTXR2* promoter.




**Supplementary Figure 3.9. CLR percentile in CEU. A)** The predicted selective sweep upstream of *ANTXR2* falls within the 98th percentile for all Sweepfinder2 likelihoods on chromosome 4 for the CEU population.




**Supplementary Figure 3.10. DNase-I-seq profiles across diverse ENCODE tissues at ANTXR2. A)** Patterns in DNase-I-seq data reveal differential regulatory landscapes between tissues at the ANTXR2 locus. The selective sweep predicted by Sweepfinder2 is colored in yellow.

# CHAPTER 4

# DISCUSSION AND FUTURE DIRECTIONS

## *SUMMARY*

Gene regulation is a complex process of interwoven steps that result in the precise control of the timing, localization, and levels of gene expression (Lelli et al., 2012). The packaging of chromatin is modified to allow for accessibility of the gene to be regulated (Cutter & Hayes, 2015), chromatin looping brings distal enhancer elements into close proximity with the promoter region of target genes (Robson et al., 2019), and transcription factors bind to enhancers and recruit general transcription factors and RNA Polymerase II (Sainsbury et al., 2015), which transcribes the gene. Variation in each of these steps is a target for natural selection, creating differences in regulatory landscapes between species. This results in phenotypic differences between individuals and species. The aim of my graduate work was to further characterize the effects of natural selection on different levels of gene regulation and to understand the functional implications of these changes over evolutionary time between primate species and within human populations.

The focus of my first project was on the selective pressures of occupied transcription factor binding sites across tissues in humans. Transcription factors are the primary determinant of chromatin accessibility and regulatory interactions and therefore changes in transcription factor binding sites often contribute to changes in gene regulation

(Kilpinen et al., 2013). By studying the variation underlying the differing localizations of transcription factor binding among tissues, we sought to learn more about which tissues have the greatest selective constraints on transcription factor binding and families of transcription factors that are under selection. In order to study these binding differences among a large collection of tissues, we needed a method to determine transcription factor binding patterns for a diverse set of transcription factors. We developed a machine learning algorithm, dTOX, to predict transcription binding patterns based on DNase-I-seq data, which is available for many cell types. Our machine learning strategy was based the concept of motif occupancy where we detect if a motif location in the genome is bound by any factor, which is not only biologically relevant due to transcription factor ensembles and co-factors often being present at motifs but also provides us with a lower false positive rate for predictions. We predicted binding patterns for 447 clusters of transcription factor motifs in 118 human tissues and then computed measures of selection on the predictions. We found an enrichment of high levels of selection at bound motif clusters in embryonic tissues compared to adult tissues. This trend is especially evident in the brain, which has significantly higher levels of negative selection in embryonic tissues. We detected different rates of evolution in tissues and specifically found that motifs bound in the brain evolve at slower rates than motifs bound in the immune system. Our analysis of selection on bound transcription factor motifs resembles the 'tissue-driven' hypothesis of protein-coding gene evolution (Park & Choi, 2010).

The results of my first project demonstrated that the regulatory landscape of the immune system is rapidly evolving compared to other tissues. Based on this observation, the motivation for the second project was to understand the causative changes in cis-regulatory elements that contribute to the rapid evolution of the immune system and differential transcription of humans and non-human primates. The adaptive immune system is the primary interface between an organism and the environment. As such, the cells of the adaptive immune system, such as CD4+ T cells, are continually under evolutionary pressures to recognize and respond to new pathogens (Abi-Rached et al., 2011; Worobey et al., 2007). Accordingly, we found an enrichment of differential transcription for membrane receptor genes in our dataset of PRO-seq from the CD4+ T cells of humans, chimpanzees, and rhesus macaques, demonstrating the importance of differing environmental pressures on shaping immune functionality between species. We used the change in transcription of one of the membrane receptors, *ANTXR2*, between humans and non-human primates as a case study. *ANTXR2*, anthrax toxin receptor 2, binds toxins produced by *B. anthracis* and plays a central role in anthrax disease pathogenesis (Tournier et al., 2009). Our aim was to understand the functional consequences of the reduced levels of *ANTXR2* in humans in the context of anthrax disease and to identify the cis-regulatory element changes responsible for the reduced expression levels specific to humans. We found that increased *ANTXR2* expression is causal for reduced viability of cells in response to anthrax toxin treatment. There are six cis-regulatory elements that show decreased ability to drive expression in a reporter assay in humans compared to non-human primates and that have chromatin contacts with the promoter of *ANTXR2*. Thus, they may drive the reduced *ANTXR2* expression

in humans. We find evidence of selection at the *ANTXR2* locus in Europeans, which have a lower sensitivity to anthrax disease compared to Africans and East Asians (Martchenko et al., 2012). There is also evidence for a selective sweep in Europeans upstream of *ANTXR2* which overlaps with DNase-I-seq peaks from tissues that are the main targets for anthrax toxins in anthrax pathogenesis. This case study highlights an example of a difference in gene regulation that was likely influenced by historical host-pathogen interactions which has had lasting effects on immunity today both between species and within human populations.

## *IMPLICATIONS*

One of the main conclusions from my work studying natural selection in transcription factor binding sites is the difference in the fraction of nucleotides under selection in embryonic compared to adult tissues. On a fundamental level, this makes sense because body plans are highly conserved and embryogenesis must be tightly regulated (Lafond & Vaillancourt, 2009). The effects of fluctuations in gene regulation of embryos are amplified throughout development making turnover of transcription factor binding sites disadvantageous. Enhancer pleiotropy, where one enhancer controls the expression of multiple genes, may contribute to this difference because it is common during development and associated with purifying selection (Fish et al., 2017; Preger-Ben Noon et al., 2018). This trend is most evident in the brain where embryonic tissues show significantly higher amounts of nucleotides under selection and weakly deleterious substitutions, which are indicative of purifying selection. Purifying selection may have

many causes, including pleiotropy of binding sites that are used by many tissues or binding sites that are highly specific to one tissue and carry out an essential function, both of which need to remain highly conserved. The brain follows the second option with very specialized regulatory elements compared to the rest of the body (Carullo & Day, 2019). It was already known that protein coding genes in the brain evolve at a relatively slow rate (H. Y. Wang et al., 2007), but our work has provided additional insight into the evolution of non-coding elements during brain development.

It has been suggested that the transition to an agrarian lifestyle contributed to strong signatures of positive selection in immune genes in agricultural cultures (Dounias & Froment, 2006; Suzuki & Nei, 2002). Based on our work studying the gene regulatory changes in the anthrax toxin receptor, it is likely that historical exposure to *B. anthracis* led to selective pressures that may have affected the immune system in the ancestors of modern humans and contributed to selective pressures for a decrease in *ANTXR2* transcription. Differences in *ANTXR2* expression levels and sensitivity to anthrax disease between human populations support this notion, with Europeans having the least sensitivity to an anthrax toxin challenge (Martchenko et al., 2012). In fact, data from a study measuring RNA expression in agricultural and hunter-gatherer populations shows significantly higher expression of *ANTXR2* in hunger gatherers (Harrison et al., 2019). Based on the genetic diversity in *B. anthracis* strains, anthrax most likely originated either in Africa or Europe (Pearson et al., 2004). Early divergences between strains date the origin time of *B. anthracis* prior to 12,000-25,000 years ago (Van Ert et al., 2007), within the same timeframe as the domestication of livestock (McTavish et al., 2013).

Approximately 3,500-6,500 years ago, *B. anthracis* underwent a radiation and expanded to Asia from Europe, which coincides with the start of long-distance trade of farming goods (Van Ert et al., 2007). The introduction of anthrax to Asia occurred several thousand years after the origin of *B. anthracis* in Europe and northern Africa, suggesting that human populations in Asia had less time to adapt to the pathogen. Thus, it is possible that a shift to agrarian lifestyle in Europe, which coincided with the radiation of *B. anthracis*, caused the population difference we see today, similar to the well-characterized example of lactase persistence (Bersaglieri et al., 2004).

We can apply what we have learned from the anthrax case study to further understand the results of the scan for natural selection of transcription factor binding motifs across tissues. We found that motifs bound in the brain evolve at slower rates than motifs bound in the immune system, a phenomenon that is well-studied in protein-coding genes (Kuma et al., 1995). In particular, immune cells represent some of the only cells with evidence of adaptive substitutions. An organism's immune system must be able to adapt to new pathogen pressures in the environment in order to survive. We see evidence of this in our PRO-seq data of CD4+ T cells, which are key component of the adaptive immune system and subsequently show many differences in transcription between species, particularly at membrane receptors. There is a correlation between pathogen richness and genetic variation at immune genes (Prugnolle et al., 2005). Furthermore, genome-wide scans for positive selection consistently report an enrichment in genes involved in immunity (Kimura et al., 2007; Mukherjee et al., 2009; Raj et al., 2013; Shultz & Sackton, 2019). The vast majority of work looking at the evolution of immune

system has been done at the level of protein-coding genes, but we are beginning to gain insight into how regulatory element evolution influences immunity. Between species comparisons reveal extensive turnover of the enhancers responsible for certain immune responses (Jubb et al., 2016). Our lab has previously shown that lineage-specific regulatory elements are evolving under positive selection and that there is a rapid turnover of cis-regulatory elements between primate species in CD4+ T cells (Danko et al., 2018). The results of both projects give another clue about the rapid turnover of transcription factor binding sites in certain immune cell types.

### *FUTURE DIRECTIONS*

There are many analyses that can be done to unite the two major themes of this thesis. We can apply dTOX to the CD4+ T cell PRO-seq dataset to learn more about the role of transcription factor binding in the immune systems of primates and with a more specific focus of the regulation of *ANTXR2*. We have preliminary dTOX models to predict transcription binding based on PRO-seq data. The use of these models will not only allow for an analysis of differential binding in our CD4+ T cell dataset, but also allow for the correlation between binding and expression levels of target genes. From the combined PRO-seq and dTOX datasets, we can start to learn more about how different stages of regulation coordinate gene expression across the primate lineage. We can look at transcription factor binding at different classes of genes in our PRO-seq dataset to see if the enrichment of membrane receptor genes is also present at the level of transcription factor binding. At a broader level, we can use this dataset to further

understand the rapid rate of evolution at regulatory elements in the immune system across primates. We can use the combination of dTOX and INSIGHT to understand if the high rate of evolution for immune cells seen in our dTOX analysis is similar between primate species.

The work described in this thesis focused on evolutionary changes at the levels of transcription factor binding genome-wide and cis-regulatory elements for one case study gene. My lab is interested in evolutionary changes at all levels of gene regulation and we recently started working on a project to understand enhancer cooperativity across evolutionary time. Genes are often regulated by "ensembles" comprised of multiple distal enhancer elements (Spitz, 2016). It is still unclear how enhancers within these ensembles work collectively to encode the expression of target genes. Our goal is to understand how changes in the activity of an enhancer within an ensemble change across primate species and the effects of these changes on the conservation of gene expression levels. Two competing models have been proposed: one in which individual enhancers work in an additive fashion, and another in which enhancers encode redundant or partially overlapping functions that can "fill-in" for one another (Osterwalder et al., 2018; Shin et al., 2016). So far, we have collected a combination of Hi-C, PRO-seq, and RNA-seq data in CD4+ T cells of humans and non-human primates. With this data, we plan to identify these enhancer ensembles (using Hi-C) and measure the effects of differences in these ensembles on gene expression (using PRO-seq and RNA-seq). Preliminary analysis of this data points to a model in which enhancers largely work in an additive fashion in ensembles but there can still be enhancer turnover

at long evolutionary time scales by the accumulation of changes with a small effect on organism fitness.

There are many ways in which evolution can act on the regulation of gene expression. In this thesis, I described how natural selection has shaped both transcription factor binding patterns and cis-regulatory elements. Since gene regulation was described nearly sixty years ago (Jacob & Monod, 1961), much has been learned about the evolutionary mechanisms that shape gene regulation. However, additional work is required to understand how gene regulation shapes the differences between species and individuals.

## *REFERENCES*


Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F. A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S. G. E., Maiers, M., Guethlein, L. A., Tavoularis, S., … Parham, P. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. Science (New York, N.Y.), 334(6052), 89–94. https://doi.org/10.1126/science.1209202

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., & Hirschhorn, J. N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. American Journal of Human Genetics, 74(6), 1111–1120. https://doi.org/10.1086/421051

Carullo, N. V. N., & Day, J. J. (2019). Genomic enhancers in brain health and disease. Genes, 10(1). https://doi.org/10.3390/genes10010043

Cutter, A. R., & Hayes, J. J. (2015). A brief review of nucleosome structure. FEBS Letters, 589(20), 2914–2922. https://doi.org/10.1016/j.febslet.2015.05.016

Danko, C. G., Choate, L. A., Marks, B. A., Rice, E. J., Wang, Z., Chu, T., Martins, A. L., Dukler, N., Coonrod, S. A., Tait Wojno, E. D., Lis, J. T., Kraus, W. L., & Siepel, A. (2018). Natural Selection has Shaped Coding and Non-coding Transcription in Primate CD4+ T-cells. Nature Ecology and Evolution, Accepted for Publication.

Dounias, E., & Froment, A. (2006). When forest-based hunter-gatherers become sedentary: Consequences for diet and health. Unasylva, 57(224), 26–33.

Fish, A., Chen, L., & Capra, J. A. (2017). Gene regulatory enhancers with evolutionarily conserved activity aremore pleiotropic than those with species-specific activity. Genome Biology and Evolution, 9(10), 2615–2625. https://doi.org/10.1093/gbe/evx194

Harrison, G. F., Sanz, J., Boulais, J., Mina, M. J., Grenier, J. C., Leng, Y., Dumaine, A., Yotova, V., Bergey, C. M., Nsobya, S. L., Elledge, S. J., Schurr, E., Quintana-Murci, L., Perry, G. H., & Barreiro, L. B. (2019). Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. Nature Ecology and Evolution, 3(8), 1253–1264. https://doi.org/10.1038/s41559-019-0947-6

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology, 3(3), 318–356. https://doi.org/10.1016/S0022-2836(61)80072-7

Jubb, A. W., Young, R. S., Hume, D. A., & Bickmore, W. A. (2016). Enhancer Turnover Is Associated with a Divergent Transcriptional Response to Glucocorticoid in Mouse and Human Macrophages. The Journal of Immunology, 196(2), 813–822. https://doi.org/10.4049/jimmunol.1502009

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., … Dermitzakis, E. T. (2013). Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. Science, 342, 744–747. https://doi.org/10.1126/science.1242463

Kimura, R., Fujimoto, A., Tokunaga, K., & Ohashi, J. (2007). A Practical Genome Scan for Population-Specific Strong Selective Sweeps That Have Reached Fixation. PLoS ONE, 2(3), e286. https://doi.org/10.1371/journal.pone.0000286

Kuma, K., Iwabe, N., & Miyata, T. (1995). Functional Constraints against Variations on Molecules from the Tissue Level: Slowly Evolving Brain-Specific Genes Demonstrated by Protein Kinase and Immunoglobulin Supergene Families. Molecular Biology and Evolution, 12(1), 123–130.

Lafond, J., & Vaillancourt, C. (2009). Human Embryogenesis: Methods and Protocols. In Embryogenesis. https://doi.org/10.5772/36871

Lelli, K. M., Slattery, M., & Mann, R. S. (2012). Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. Annual Review of Genetics, 46(1), 43–68. https://doi.org/10.1146/annurev-genet-110711-155437

Martchenko, M., Candille, S. I., Tang, H., & Cohen, S. N. (2012). Human genetic variation altering anthrax toxin sensitivity. Proceedings of the National Academy of Sciences, 109(8), 2972–2977. https://doi.org/10.1073/pnas.1121006109

McTavish, E. J., Decker, J. E., Schnabel, R. D., Taylor, J. F., & Hillis, D. M. (2013). New World cattle show ancestry from multiple independent domestication events. Proceedings of the National Academy of Sciences of the United States of America, 110(15), E1398-406. https://doi.org/10.1073/pnas.1303367110

Mukherjee, S., Sarkar-Roy, N., Wagener, D. K., & Majumder, P. P. (2009). Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proceedings of the National Academy of Sciences of the United States of America, 106(17), 7073–7078. https://doi.org/10.1073/pnas.0811357106

Osterwalder, M., Barozzi, I., Tissiéres, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A., & Pennacchio, L. A. (2018). Enhancer redundancy provides phenotypic robustness in mammalian development. Nature, 554(7691), 239–243. https://doi.org/10.1038/nature25461

Park, S. G., & Choi, S. S. (2010). Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evolutionary Biology, 10(1). https://doi.org/10.1186/1471-2148-10-241

Pearson, T., Busch, J. D., Ravel, J., Read, T. D., Rhoton, S. D., U'Ren, J. M., Simonson, T. S., Kachur, S. M., Leadem, R. R., Cardon, M. L., Van Ert, M. N., Huynh, L. Y., Fraser, C. M., & Keim, P. (2004). Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proceedings of the National Academy of Sciences of the United States of America, 101(37), 13536–13541. https://doi.org/10.1073/pnas.0403844101

Preger-Ben Noon, E., Sabarís, G., Ortiz, D. M., Sager, J., Liebowitz, A., Stern, D. L., & Frankel, N. (2018). Comprehensive Analysis of a cis-Regulatory Region Reveals Pleiotropy in Enhancer Function. Cell Reports, 22(11), 3021–3031. https://doi.org/10.1016/j.celrep.2018.02.073

Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V., & Balloux, F. (2005). Pathogen-Driven Selection and Worldwide HLA Class I Diversity. Current Biology, 15(11), 1022–1027. https://doi.org/10.1016/j.cub.2005.04.050

Raj, T., Kuchroo, M., Replogle, J. M., Raychaudhuri, S., Stranger, B. E., & De Jager, P. L. (2013). Common Risk Alleles for Inflammatory Diseases Are Targets of Recent Positive Selection. The American Journal of Human Genetics, 92(4), 517–529. https://doi.org/10.1016/j.ajhg.2013.03.001

Robson, M. I., Ringel, A. R., & Mundlos, S. (2019). Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D. Molecular Cell, 74(6), 1110–1122. https://doi.org/10.1016/j.molcel.2019.05.032

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nature Reviews Molecular Cell Biology, 16(3), 129–143. https://doi.org/10.1038/nrm3952

Shin, H. Y., Willi, M., Yoo, K. H., Zeng, X., Wang, C., Metser, G., & Hennighausen, L. (2016). Hierarchy within the mammary STAT5-driven Wap super-enhancer. Nature Genetics, 48(8), 904–911. https://doi.org/10.1038/ng.3606

Shultz, A. J., & Sackton, T. B. (2019). Immune genes are hotspots of shared positive selection across birds and mammals. ELife, 8, 1–33. https://doi.org/10.7554/eLife.41815

Spitz, F. (2016). Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. Seminars in Cell and Developmental Biology, 57, 57–67. https://doi.org/10.1016/j.semcdb.2016.06.017

Suzuki, Y., & Nei, M. (2002). Origin and evolution of influenza virus hemagglutinin genes. Molecular Biology and Evolution, 19(4), 501–509. https://doi.org/10.1093/oxfordjournals.molbev.a004105

Tournier, J. N., Rossi Paccani, S., Quesnel-Hellmann, A., & Baldari, C. T. (2009). Anthrax toxins: A weapon to systematically dismantle the host immune defenses. Molecular Aspects of Medicine, 30(6), 456–466. https://doi.org/10.1016/j.mam.2009.06.002

Van Ert, M. N., Easterday, W. R., Huynh, L. Y., Okinaka, R. T., Hugh-Jones, M. E., Ravel, J., Zanecki, S. R., Pearson, T., Simonson, T. S., U'Ren, J. M., Kachur, S. M., Leadem-Dougherty, R. R., Rhoton, S. D., Zinser, G., Farlow, J., Coker, P. R., Smith, K. L., Wang, B., Kenefic, L. J., … Keim, P. (2007). Global Genetic Population Structure of Bacillus anthracis. PLoS ONE, 2(5), e461. https://doi.org/10.1371/journal.pone.0000461

Wang, H. Y., Chien, H. C., Osada, N., Hashimoto, K., Sugano, S., Gojobori, T., Chou, C. K., Tsai, S. F., Wu, C. I., & Shen, C. K. J. (2007). Rate of evolution in brain-expressed genes in humans and other primates. PLoS Biology, 5(2), 0335–0342. https://doi.org/10.1371/journal.pbio.0050013

Worobey, M., Bjork, A., & Wertheim, J. O. (2007). Point, Counterpoint: The Evolution of Pathogenic Viruses and their Human Hosts. Annual Review of Ecology, Evolution, and Systematics, 38, 515–540. http://dx.doi.org/10.1146/annurev.ecolsys.38.091206.095722