Functional characterization of regulatory changes in sequence and genome structure underlying modern human evolution

by
Lana Harshman

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

in

Genetics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Elphege Nora
B6D98EABE1914FD...

Elphege Nora
Chair

DocuSigned by:
Nadav Ahituv
DocuSigned by:
F4...

Nadav Ahituv

DocuSigned by:
Yin Shen
AD18806FAF2442B...

Yin Shen

Committee Members

DEDICATION

I dedicate my thesis work to my Dad, Gary Harshman, who first inspired me to pursue a career in science. He was always there to remind me that all my hard work and effort would be worth it in the end. I wish he were here to see that effort pay off. I love you, Dad.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Dr. Nadav Ahituv. Nadav has been such a supportive mentor, in terms of my scientific projects and my personal well-being. He always encouraged me to pursue projects that interested me and never pressured me to work on anything that did not. He allowed me to approach my PhD studies on my terms and always respected my boundaries with work. I have learned so much about science and about being a mentor from him and will forever be grateful that I had him as an advisor through such an important phase of my life. I hope to have advisors like Nadav in my future, and eventually, be an advisor like him.

The Ahituv lab has undergone a major expansion since I joined in 2018, and all members have helped me reach this milestone in one way or another. I specifically want to thank Dr. Fumitaka Inoue and Dr. Aki Ushiki, who were two senior postdocs when I joined the lab and who set the bar for what great science is. Thank you to Dr. Serena Tamura, the only graduate student in the lab when I joined, for not only setting an example of how to be a successful graduate student, but for also for being my friend, confidante, and biggest cheerleader throughout graduate school. I also want to thank Wei Gordon, who has become a great friend and huge supporter of mine; I will miss our walks and lunches together. Thank you to Dianne Laboy-Cintrón, my desk neighbor and fellow Husky (Go Dawgs). I'll miss being constantly distracted by one another and our mutual respect for the noise cancelling headphones. Thank you to Rachael Bradley for keeping the lab running and being the honorary graduate student that fills out our "Lab Gals" group. A huge thank you to Ryder Easterlin for the help on my ongoing projects, but also for

being my Gen-Z friend who never misses an opportunity to remind me how old I am. And a very special thank you to Dr. Hai Nguyen. Hai joined the lab at a crucial point in my training and gave me the help and confidence I needed to finish out the last stretch of my PhD and I will miss our morning coffees together. Thank you to all members of the Ahituv lab, current and present.

My thesis committee has been a major source of support throughout the pursuit of my PhD. Thank you to Dr. Yin Shen, who not only served on my thesis committee, but was also the chair of my qualifying exam committee and one of my rotation advisors. She has done so much to push my project forward and to challenge me as a scientist. I cannot thank my thesis committee chair, Dr. Elphège Nora, enough for his help throughout my preparation for my qualifying exam, putting together my thesis work, and especially for helping me get through the last six months of my PhD. Our meetings together were instrumental in putting my project in perspective and for framing my entire PhD in a way that felt attainable.

I would be remiss if I did not thank the first scientists I even worked with at the University of Washington, when I thought working in a lab might just be a fun thing to do. To Dr. Evan Eichler, who gave me my first opportunity to work in a lab when I had absolutely zero experience. I first met Dr. Eichler while taking one of his Genome Sciences courses and was hooked. I am so grateful that he saw my enthusiasm as potential and gave me a chance to grow my skills as a scientist. Also, a huge thank you to Dr. Megan Dennis, who was a postdoc in the Eichler lab when I joined. She taught me so much about science and gave me one of the best examples of how to be a woman and mother working in the field. Years after working together,

she helped me apply to graduate school and has continued to be an outstanding mentor and friend through the years.

Scientists succeed through collaborations with our peers, but we are nothing without the friends and family outside of that circle that support us in our endeavors. To this point, it's of the utmost importance that I thank all those people in my life that lifted me up and got me to this point in my career. My friends and family have been my rock throughout this process, and it means the world to have such supportive people in my life. First, I want to thank the people who helped me through the toughest time in my life; the beginning of graduate school which coincided with my father's passing. I would not have made it back to UCSF and I would not be finishing my degree if they hadn't stepped up and helped me and my siblings when we needed them the most.

A special thank you to my brother, Andrew, and sister, Natalie. I feel so lucky to have grown up with these two and it is very special how we have been able to remain close through everything. I am and always will be grateful for the level of effort they put in to understanding my research just so that they can accurately tell others about it. Thank you for always pumping me up! I also want to thank my parents. My dad, Gary Harshman, instilled a desire for knowledge in me that persists today; I would not have made it this far without his encouragement. Thank you to my mother, Nancy Hanks, for showing me how to be a woman working in a male dominated field. I will forever carry with me the lessons she taught me.

Although both my parents have passed, I am forever grateful for their love and for making me the person I am today.

Most importantly, I want to thank my husband, Tom, for never hindering my desire to better myself and experience life. He moved to San Francisco with me without even questioning where we would build a life together and there is no other person I would rather go through this life with. He has always been supportive of my pursuits, and I absolutely would not have made it through the last five years without him. Now on to the next adventure!

CONTRIBUTIONS

Chapter 2 of this dissertation is a peer reviewed and published paper co-first authored by Lana Harshman (see full citation below). Chapter 3 is an ongoing project with work contributed from Ryder Easterlin, Dr. Philip Kleinert, Dr. Martin Kircher, and Dr. Nadav Ahituv. Dr. Alex Pollen contributed samples for this project, namely the human, chimpanzee, and orangutan cells from brain organoids. Dr. Lucia Carbone also contributed unpublished TAD boundary calls that were crucial for our analysis.

Citation for Chapter 2:

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

Functional characterization of regulatory changes in sequence and genome structure underlying modern human evolution

Lana Mae Harshman

ABSTRACT

The coding regions of the human and chimpanzee genomes are 99% identical, which implies changes to the noncoding genome are likely responsible for the divergent phenotypes between these species. A large body of research has focused on noncoding regions in an attempt to decode their function and implications on phenotypes and disease. In this work, I examined noncoding regulatory sequences unique to modern humans and identified candidate regulatory sequences that may have contributed to modern human speciation. First, I tested the regulatory potential of all fixed or nearly fixed single nucleotide changes in the modern human lineage as compared to Neanderthal and Denisovan using Massively Parallel Reporter Assays (MPRA). We found that a subset of these sequences are potentially active regulatory elements, and an additional subset were differentially active between the archaic and modern sequence versions. The differentially active sequences were linked to genes involved in brain development, vocal tract function, and other phenotypes. Additionally, we annotated changes in CTCF binding sites between humans, great apes, and archaic humans. CTCF is an architectural protein that helps facilitate genome looping and enhancer-promoter interactions. We identified CTCF binding sites that were uniquely gained or lost in the human genome, as compared to great apes (human specific), and separately compared to Neanderthal and Denisovan (recent human specific). We identified 2,230 human specific gained CTCF sites and 24 human specific

lost CTCF sites, as compared to great apes. As compared to Neanderthal and Denisovan, we found 24 recent human specific gained sites. We observed an enrichment of human gained sites at TAD boundaries, but no enrichment for the human lost CTCF sites. Additionally, we found that human specific gained CTCF sites were enriched near genes involved in cognitive function and recent human specific CTCF sites were enriched near genes related to chondrocyte differentiation. Finally, I created a stable human iPSC line in which I deleted one human specific gained CTCF site. I differentiated these cells into neurons for further characterization but found that this deletion had no effect on gene expression in the region. My work provides a list of single nucleotide changes and CTCF sites that are interesting targets for future research in determining the effects of noncoding sequence changes on modern human evolution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: Introduction

Understanding where we come from has been a focal point of human fascination for thousands of years with myths about human origins appearing in nearly every ancient civilization. Scientifically, humans have been compared to apes and other non-human primates since the 1800s, primarily due to the obvious similarities in phenotypes. Even contemporary research still capitalizes on the commonalities between humans and other primates. Our ability to sequence genomes has greatly furthered our understanding of human evolution and provides vital tools for expanding our knowledge. Genomic comparisons between humans and non-human primates reveal notable differences that may contribute to shaping human specific traits. Although using closely related species in our study of modern human evolution has been very informative, there remains a large distance, evolutionarily, between humans and our closest living relative: the chimpanzee. The chimpanzee-human split is estimated to have occurred between 9.3 and 6.5 million years ago[1], and a timeframe of this magnitude leaves a large gap in our understanding of human specificity. Ideally, more closely related species with more recent common ancestors would be used to compare with the modern human genome. Although resources for these species exist, these species are extinct, which greatly hinders our ability to study them.

1.1 Extinct archaic humans: Neanderthal and Denisovan

Probably the most well-known extinct species from the homo lineage is the Neanderthal. The first Neanderthal skeleton was discovered in 1856 in the Neander Valley in Germany[2]. Initially it was believed to be human remains, but later became key in the debate about modern human

evolution. Since this initial discovery, many other Neanderthal remains have been found throughout Europe. Denisovans are a lesser-known extinct archaic human species whose first skeletal remains, a single finger bone, were only discovered in 2008 in a cave in Siberia[3]. This discovery yielded some important information, including the realization that this was a novel, previous unidentified extinct hominin. This classification was only possible through novel techniques in ancient DNA sequencing[4].

Estimating the exact split between any species is difficult, but the difficulty of parsing out the split between Neanderthal, Denisovan, and modern humans is compounded by the fact that admixture occurred between all three species. Today, most non-sub-Saharan African human populations have about 2% Neanderthal DNA and some modern human populations have as much as 3.5% Denisovan DNA[3,5]. Admixture is estimated to have been occurring as recently as 45 thousand years ago (kya)[6]. Despite this admixture, it is estimated that modern humans diverged from archaic humans around 550 kya, and Denisovans and Neanderthals diverged from each other about 450 kya.

Since numerous Neanderthal skeletal remains have been found, researchers have been able to identify key skeletal differences between Neanderthals and modern humans. It is generally believed that Neanderthals were shorter than modern humans, with a wider rib cage and are often described as “barrel chested”[7]. Skull phenotypes unique to Neanderthal include a lower forehead and more protruding face. Unique skull morphologies have allowed for the identification of likely differences in outer brain structures between Neanderthals and modern

humans[8,9]. Vocal tract differences have also been inferred between the two species[10]. Comparing the modern human skeletal structure to Denisovan is much harder given that very few remains have been found. However, through reconstructed DNA methylation maps, a rendition of the full Denisovan skeleton has been put forth showing some key differences, notably in skull shape and facial structures[11]. **Figure 1.1**, from Gokhman, et al. 2019, shows predicted skeletal differences between all three of these homo species.

Genomic sequences from these extinct archaic species became available in the early 2010s[4,5] and has greatly increased the resources available to study them. Some examples of their application have been mentioned previously. As new technologies improve, improved genomic sequences, assemblies, and variant annotations will become available for these species and will no doubt become an even more integral part of studying modern human evolution.

1.2 Gene regulation in modern human evolution: progress and challenges

Nearly half a decade ago, it was found that approximately 99% of coding regions in the human and chimpanzee genomes were the same[12]. Research focusing on the roughly 1.5% differences in coding sequences found a subset of genes that showed evidence of purifying selection in humans; notably genes involved in the olfactory system, skeletal development, and neurogenesis, among others[13]. Additionally, it has been shown that some human specific duplicated genes have gained novel function in humans, which provides other examples of adaptation stemming from changes to the coding regions of the genome[14,15].

Although some human specific traits appear to be the result of coding changes, these changes alone cannot account for all human specific traits. Most sequence differences between human and chimpanzee exist in noncoding regions of the genome. Noncoding regions account for nearly 99% of the entire genome and contain gene regulatory elements. Regulatory elements help dictate the location, timing, and level of expression of each gene. Changes in gene expression are predicted to be the driving force behind phenotypic differences between human and chimpanzees, hence why a lot of effort has gone into decoding the noncoding genome.

Understanding gene regulation is crucial to understanding resulting phenotypes, however genomic sequences do not provide enough information to completely parse this out. Data about abundance of RNA molecules, epigenetic modifications, and chromatin interactions are needed to fully understand a given gene regulatory network. Despite having genomic sequences for Neanderthal and Denisovan, not having live cells greatly hinders our ability to study gene regulation in the context of these species, further complicating the study of modern human specific gene regulation.

Comparing noncoding genomic sequences can provide glimmers of information about human specific gene regulation. In the case of human accelerated regions (HARs), sequences that experienced rapid changes in the human lineage, expedited evolution can imply function. To this point, it was found that HARs tend to overlap promoters and enhancers important for human development[16]. Additionally, the importance of HARs is implicated through their observed association with human diseases, such as schizophrenia[17] and autism[18]. Interrogating

human specific deleted sequences (hDELs) has also identified sequence changes of interest. It was found that the majority of hDELs are near genes envovled in hormone pathways and neural function[19], implying an important role in human specific gene regulation.

Despite lack of direct ancient human epigenetic information, reconstructed DNA methylation maps[20], whose applications have been mentioned previously, provide a small glimpse into one aspect of gene regulation in extinct archaic species. However, other than methylation maps, there is little other epigenetic data for these species.

1.3 Massively parallel reporter assay as a tool to study modern human gene regulation

Massively parallel reporter assay (MPRA) provides the ability to test the gene regulatory potential of thousands of sequences in one experiment. Candidate regulatory sequences (CRSs) are cloned upstream of a minimal promoter and reporter gene, along with a unique barcode, which allows thousands of sequences to be pooled together. Libraries of cloned CRSs are packaged into lentivirus and used to infect a cell type of interest. If a given sequence is capable of driving gene expression, the unique barcode will be transcribed and can be measured via RNA sequencing. DNA is also sequenced in order to measure of how often a given sequence is integrated into the genome. A ratio of RNA to DNA gives the relative expression of each CRS in the library. Details of this protocol and subsequent analysis are explained in Gordon, et al.[21] and a schematic of the protocol is shown in **Figure 1.2**.

MPRA is an apt tool for studying modern human evolution for a few reasons. First, MPRA uses synthetically created DNA sequences. This allows for testing sequences directly from archaic extinct species because the underlying sequence is all that is required, not live cells. Additionally, MPRA is high throughput. Because of the lack of epigenetic information for archaic human species, it is very difficult to determine which sequences are acting as gene regulatory elements in their native genomic context. This makes it very difficult to rank CRSs on likely functionality in these species. Being able to screen many CRSs in one experiment, therefore, becomes crucial for this line of research.

A downside of MPRA is its inability to test a given CRS in its endogenous genomic context; the CRSs randomly integrate into the genome at many different loci. Therefore, this method only tests the capability of a sequences to drive expression and does not actually prove that a given sequence drives expression in its native locus.

1.4 Long-range gene regulation

It is tempting to envision a straightforward method of gene regulation in which an enhancer physically interacts with a nearby promoter causing its activation and subsequent gene expression. But as is customary in biology, things are not so simple. It is known that enhancers can have target promoters very far away, sometimes up to thousands of kilobases away, and may not even target the closest gene[22]. The interactions between enhancers and their far away target promoters are traditionally believed to be facilitated by interactions between proteins bound at each site. This interaction creates a loop structure that brings the enhancer into close

proximity with a promoter[22]. However, this classic enhancer-promoter looping model doesn't account for recent data that found an increasing, rather than decreasing, distance between enhancer-promoter contacts at the transcriptionally active *Shh* locus via fluorescence in situ hybridization (FISH) and chromosome conformation capture carbon copy (5C)[23]. In another recent study, live imaging revealed a similar pattern of no correlation between enhancer-promoter proximity and expression at the *Sox2* locus[24]. An intriguing model has been suggested to account for these types of observations. It proposes that enhancers imprint their target promoters and this imprinting slowly decays over time, suggesting that enhancers may not need to contact promoters consistently in order to cause gene expression[25].

Additional important elements of long-range gene regulation are insulators. Insulators are regulatory elements that function to create or maintain chromatin domain barriers. This is achieved in one of two ways. The first is preventing the spread of heterochromatin into euchromatin, thus maintaining inactive and active chromatin domains, respectively. The second is by blocking enhancer-promoter interactions. Insulators can do this by creating a physical barrier in the form of a genome loop. Enhancer-blocking elements can interact with each other, tethering the DNA together and preventing interaction between a promoter and enhancer located on either side of the newly formed genome loop[26]. It is becoming increasingly evident that genome domains and insulators play a key role in gene regulation.

1.5 3D genome structures and domains

The organization of chromatin within a nucleus can be broken down into subcategories. Within the nucleus, chromosomes tend to occupy certain regions, which are called chromosomal territories. Chromosomes are further sub-divided into transcriptionally active (A) and inactive (B) compartments. Within these compartments, topologically associated domains (TADs) have been identified. TADs are genomic regions, typically hundreds of kilobases in size, that are characterized by preferential self-interaction rather than interactions with outside regions. The regions flanking these domains are referred to as TAD boundaries and are highly conserved across cell types and species[27,28]. Such a high level of conservation can be indicative of function and indeed there are examples of TAD structures being important for gene expression and downstream phenotypes that will be discussed more thoroughly in a later section. Important progress has been made in terms of determining how these domains are created and maintained. An early study found that binding sites for the transcription factor CCCTC-binding factor (CTCF) are enriched at TAD boundaries, suggesting a role for CTCF in TAD boundary formation[27]. Additional experiments using a CTCF degron system solidified CTCF as an important factor in genome folding, and ultimately vital in instructing TAD formation and maintenance, by showing that TADs are not properly formed when CTCF is depleted[29]. However, not all TAD boundaries are demarcated by CTCF sites and how these CTCF-independent domains are formed is not as clearly defined.

1.6 3D genome structure and human evolution

It is difficult to determine what effect 3D genome structures have had on human evolution but comparing domains across closely related species can provide some insight. Using the highly rearranged gibbon genome as a framework, researchers compared TAD boundaries across species and found that breaks of synteny between species tend to co-localize with TAD boundaries, which could be indicative of evolutionary pressure against the disruption of TAD structures[30]. Examining human specific chromatin domains and structures can also yield important information. One study found that human specific genome loops tend to be enriched for HARs and human specific structural variants, suggesting novel genome loop formation as a mechanism by which modern human specific gene regulatory patterns could have arisen[31]. Additionally, this study found human specific genome loops were enriched for enhancer-enhancer interactions, indicative of multi-enhancer regulatory networks, that are active in the developing brain. Including Neanderthal and Denisovan data into analysis of human specific 3D genome structures is not straightforward since there is no way to directly observe chromatin domains or loops in these extinct species. Researchers have, however, computationally inferred 3D genome structure in these extinct species and found a set of regions with predicted diverged 3D organization near genes involved in cognitive function, among others phenotypes[32]. These studies together imply an important role of 3D genome structures in modern human adaptation, but much remains to be explored.

1.7 CCCTC-binding factor

CTCF is a highly conserved architectural protein[33,34] that binds DNA via its central 11 zinc finger domains[35] in an orientation dependent manner[36]. The exact number of potential and occupied CTCF binding sites in a genome can vary drastically based on species, cell type, and assay used. A study in mouse cells found over 100,000 potential CTCF binding sites, the majority of which were bound in multiple cell types[37]. In human, about 55,000 bound CTCF sites have been observed, around half of which were detected in the majority of the 19 cell types tested, and 72% were in at least two of the cell types[38].

CTCF is believed to be involved in creating genome loops via the loop extrusion model (**Figure 1.3**). In this model, a bound CTCF site is thought to stall the circular protein complex, cohesin, as it translocates across the DNA. Cohesin continues to extrude one end of the DNA until it encounters another bound and convergently oriented CTCF site[39,40]. CTCF also has an important function as an insulator, mentioned previously in that it is important for TAD boundary formation and maintenance. In the proposed model, the looping caused by the clustered CTCF sites at TAD boundaries is thought to prevent interactions between TADs, thus providing an insulating function[39]. Additionally, CTCF competes with the protein WAPL, which can unload the cohesin complex from DNA, therefore helping to maintain genome loops once they are formed[41,42].

1.8 CTCF and gene regulation

CTCF seems to play a role in nearly all aspects of gene regulation. The function as a gene up-regulator was first recognized over twenty years ago when researchers observed a co-localization of CTCF with the amyloid β-protein precursor promoter[43]. In addition to its indirect role in gene down-regulation via its insulator function, there is early evidence suggesting CTCF can act as a direct repressor, like for the oncogene *c-myc*[44].

Altering CTCF sites can have a range of effects on gene expression, 3D genome interactions, and even downstream phenotypes. An example of an extreme phenotype caused by CTCF perturbation is within the *SHH* locus. The deletion of three CTCF sites leads to truncated limbs in humans, a disease called acheiropodia, caused by the inability of the *SHH* limb enhancer to interact with its promoter[45]. Molecular changes have also been detected when CTCF sites are perturbed. For example, altered chromatin domains were observed within the *Hox* cluster when CTCF binding sites were deleted[46]. Additionally, while interrogating the *Sox9-Kcnj2*, researchers found that inverting a TAD boundary, including its CTCF sites, caused novel CTCF-mediated loops to form which ultimately altered gene expression[47]. These examples demonstrate that CTCF, at least in some instances, can drive formation of genome domains, facilitate gene expression, and ultimately influence phenotypes.

Although CTCF perturbations can change gene expression, changes to individual CTCF sites generally do not have an effect by themselves. As mentioned previously, inversions at the *Sox9-Kcnj2* locus caused changes in gene expression, but the researchers did not see a change in

gene expression when deleting individual CTCF sites[47]. Additionally, the *Shh* locus and subsequent gene expression was found to be highly robust to CTCF perturbations[48]. Likewise, the locus spanning the genes *WNT6*/*IHH*/*EPHA4*/*PAX3* only showed aberrant gene expression and altered chromosome contacts when entire TAD structures were perturbed and not when individual CTCF sites were altered[49].

There are many more studies in addition to the ones mentioned that demonstrate how CTCF perturbations can and cannot change genome structure, gene expression, and phenotypes. Combined, these studies demonstrate that CTCF function is highly dependent on the genomic environment. Although we have some idea of what we can expect given a particular CTCF perturbation, we still don't fully understand the exact rules governing CTCF function and its role in gene expression.

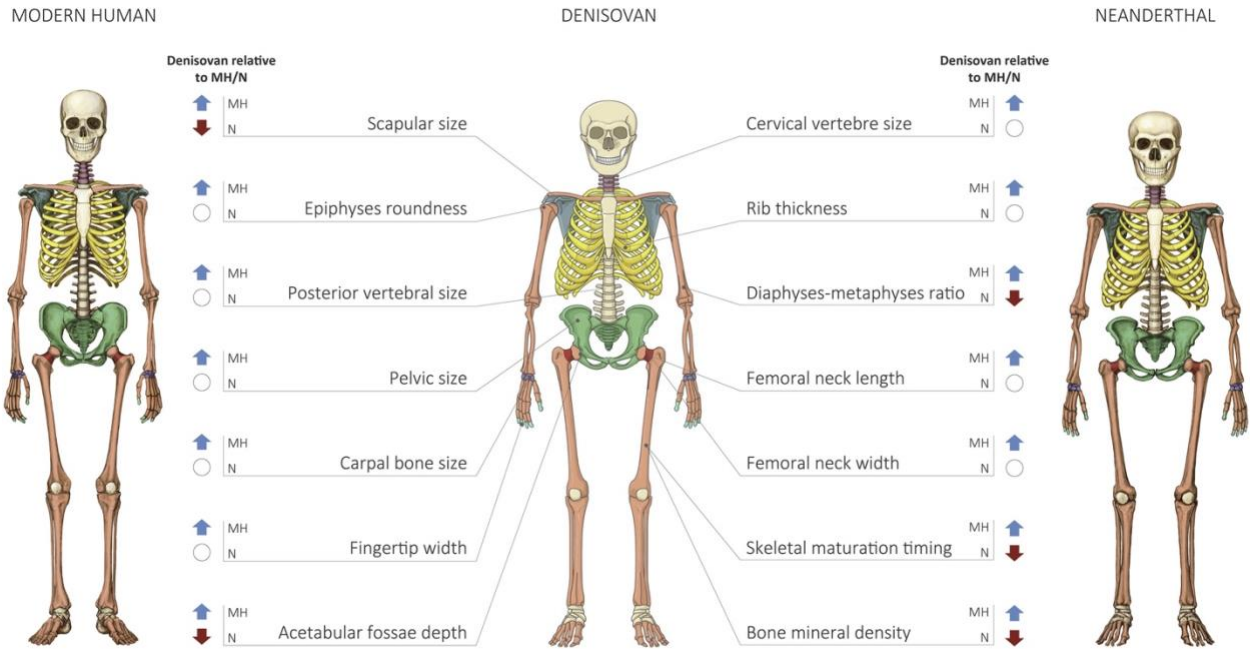

Figure 1.1: Reconstructed Denisovan skeletal structure compared to modern human and Neanderthal

Figure taken from Gokhman, et al. 2019[11]. The colors (green, yellow, red, and orange) on the Denisovan skeleton represent the reconstructed skeletal traits inferred by differentially methylated regions. The equivalent regions in the other skeletons are marked with the same color. The skeletal structures that could not be reconstructed are depicted in a general way. The blue and red arrows show the direction of predicted change in the Denisovan as compared to Neanderthals (N) and modern humans (MH). No predicted change is represented with an empty circle.

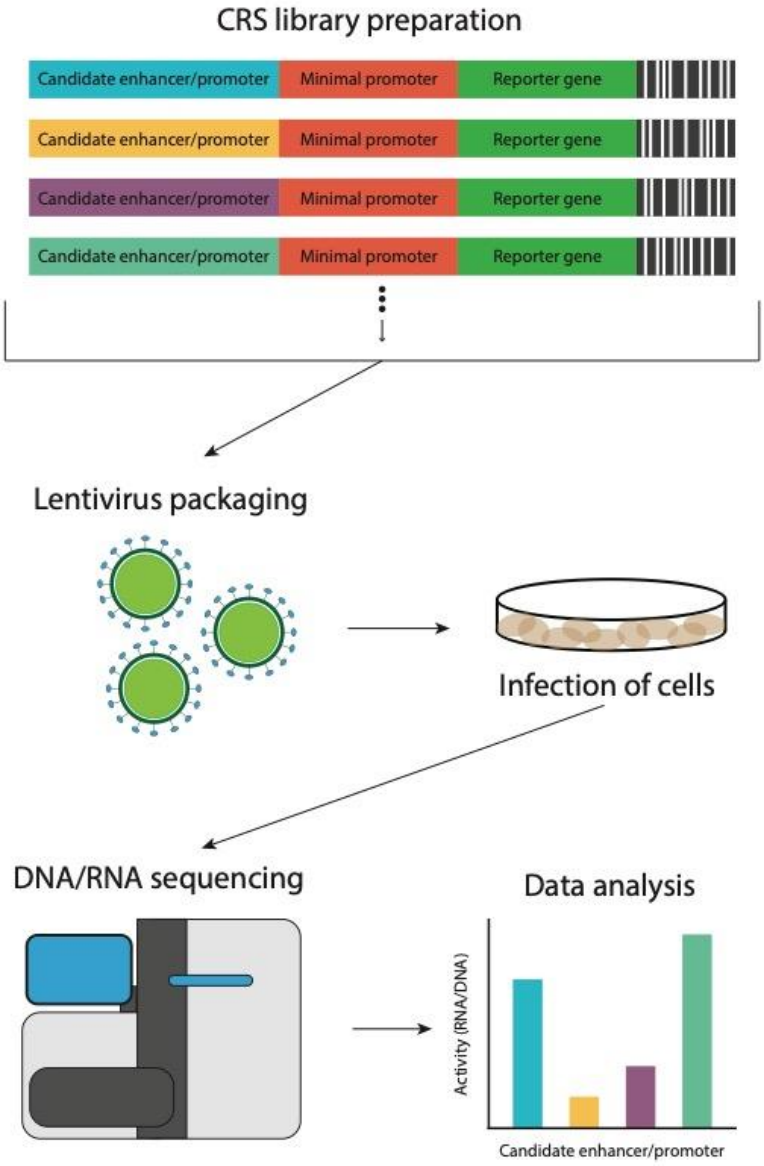

Figure 1.2: Massively Parallel Reporter Assay (MPRA) schematic
A schematic showing the general workflow for an MPRA experiment. The first step is cloning the candidate regulatory sequences (CRSs) into a vector to create a CRS library. The library is packaged into lentivirus, used to infect cells, then the DNA and RNA are sequenced. Expression level is determined by taking the ratio of RNA to DNA for each CRS.

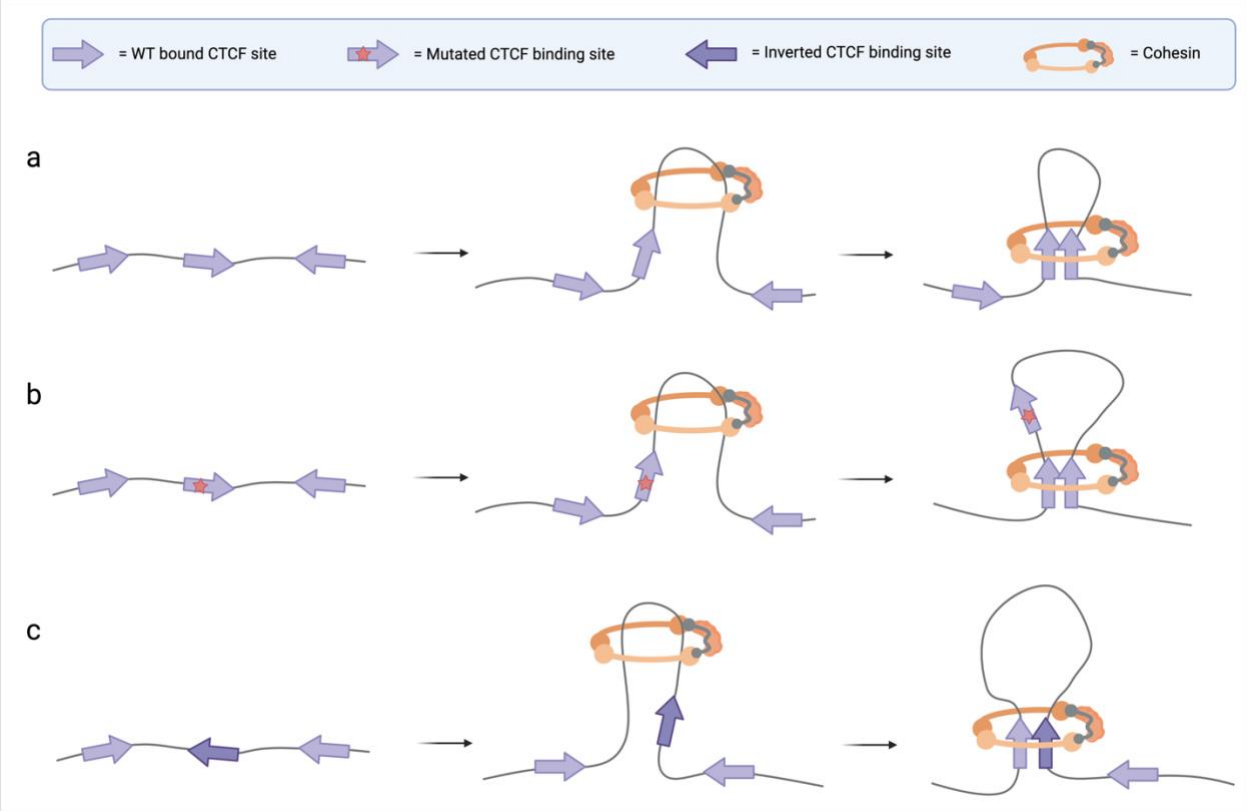

Figure 1.3: CTCF looping mechanics

A simplified schematic depicting how CTCF is generally thought to interact with other CTCF sites to form genome loops. (a) How wildtype bound CTCF sites (blue arrows) are thought to facilitate genome loops. Namely, two convergently oriented and bound CTCF sites stall cohesin (purple circle), creating the loop. (b) How a mutated CTCF site (blue arrow with a red star) which is mutated in a way that prevents CTCF from binding at that site, will not stall cohesin and will cause an alternate, in this case larger, genome loop to form. (c) The consequence of a CTCF site being inverted (cyan arrow). This causes the anchor for one side of the genome loop to change, again in this case causing a larger genome loop to form.

1.9 REFERENCES

1. Almécija, S. *et al.* Fossil apes and human evolution. *Science* **372**, (2021).
2. Schmitz, R. W. *et al.* The Neandertal type site revisited: interdisciplinary investigations of skeletal remains from the Neander Valley, Germany. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 13342–13347 (2002).
3. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
4. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
5. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
6. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. Origins of modern human ancestry. *Nature* **590**, 229–237 (2021).
7. Sawyer, G. J. & Maley, B. Neanderthal reconstructed. *Anat. Rec. B New Anat.* **283**, 23–31 (2005).
8. Neubauer, S., Hublin, J.-J. & Gunz, P. The evolution of modern human brain shape. *Sci Adv* **4**, eaao5961 (2018).
9. Kochiyama, T. *et al.* Reconstructing the Neanderthal brain using computational anatomy. *Sci. Rep.* **8**, 6296 (2018).
10. Gokhman, D. *et al.* Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun.* **11**, 1189 (2020).
11. Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps. *Cell*

179, 180-192.e10 (2019).

12. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
13. Clark, A. G. *et al.* Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**, 1960–1963 (2003).
14. Charrier, C. *et al.* Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**, 923–935 (2012).
15. Florio, M. *et al.* Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science* **347**, 1465–1470 (2015).
16. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130025 (2013).
17. Xu, K., Schadt, E. E., Pollard, K. S., Roussos, P. & Dudley, J. T. Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. *Mol. Biol. Evol.* **32**, 1148–1160 (2015).
18. Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**, 341-354.e12 (2016).
19. McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011).
20. Gokhman, D. *et al.* Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science* **344**, 523–527 (2014).
21. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional

characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).

22. Vernimmen, D. & Bickmore, W. A. The Hierarchy of Transcriptional Activation: From Enhancer to Promoter. *Trends Genet.* **31**, 696–708 (2015).
23. Benabdallah, N. S. *et al.* Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol. Cell* **76**, 473-484.e7 (2019).
24. Alexander, J. M. *et al.* Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *Elife* **8**, (2019).
25. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *Nature* **604**, 571–577 (2022).
26. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **7**, 703–713 (2006).
27. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
28. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
29. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).
30. Lazar, N. H. *et al.* Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.* **28**, 983–997 (2018).
31. Luo, X. *et al.* 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell* **184**, 723-740.e21 (2021).
32. McArthur, E. *et al.* Reconstructing the 3D genome organization of Neanderthals reveals

that chromatin folding shaped phenotypic and sequence divergence. *bioRxiv* 2022.02.07.479462 (2022) doi:10.1101/2022.02.07.479462.

33. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17**, 520–527 (2001).
34. Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. & Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17507–17512 (2012).
35. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
36. Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol. Cell* **66**, 711-720.e3 (2017).
37. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488**, 116–120 (2012).
38. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
39. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
40. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6456-65 (2015).
41. Nora, E. P. *et al.* Molecular basis of CTCF binding polarity in genome folding. *Nat. Commun.* **11**, 5612 (2020).

42. Li, Y. *et al.* The structural basis for cohesin–CTCF-anchored loops. *Nature* **578**, 472–476 (2020).

43. Vostrov, A. A. & Quitschke, W. W. The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J. Biol. Chem.* **272**, 33353–33359 (1997).

44. Filippova, G. N. *et al.* An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.* **16**, 2802–2813 (1996).

45. Ushiki, A. *et al.* Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia. *Nat. Commun.* **12**, 2282 (2021).

46. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015).

47. Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).

48. Paliou, C. *et al.* Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12390–12399 (2019).

49. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

CHAPTER 2: The *cis*-regulatory effects of modern human-specific variants

Carly V. Weiss[1,*], Lana Harshman[2,3,*], Fumitaka Inoue[2,3,4], Hunter B. Fraser[1], Dmitri A. Petrov[1,†], Nadav Ahituv[2,3,†], David Gokhman[1,†]

Affiliations:

[1] Department of Biology, Stanford University, Stanford, CA 94305, USA

[2] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA, 94158, USA.

[3] Institute for Human Genetics, University of California San Francisco, San Francisco, CA, 94158, USA.

[4] Present address: Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, 606-8501, Japan

[*] Equal contributors

[†] Corresponding authors

2.1 SUMMARY

The Neanderthal and Denisovan genomes enabled the discovery of sequences that differ between modern and archaic humans, the majority of which are noncoding. However, our understanding of the regulatory consequences of these differences remains limited, in part due to the decay of regulatory marks in ancient samples. Here, we used a massively parallel reporter assay in embryonic stem cells, neural progenitor cells and bone osteoblasts to investigate the regulatory effects of the 14,042 single-nucleotide modern human-specific variants. Overall, 1,791 (13%) of sequences containing these variants showed active regulatory activity, and 407 (23%) of these drove differential expression between human groups. Differentially active sequences were associated with divergent transcription factor binding motifs, and with genes enriched for vocal tract and brain anatomy and function. This work provides insight into the regulatory function of variants that emerged along the modern human lineage and the recent evolution of human gene expression.

2.2 BACKGROUND

The fossil record allows us to directly compare skeletons between modern humans and their closest extinct relatives, the Neanderthal and the Denisovan. From this we can make inferences not only about skeletal differences, but also about other systems, such as the brain. These approaches have uncovered a myriad of traits that distinguish modern from archaic humans. For example, our face is flat with smaller jaws, our development is slower, our pelvises are narrower, our limbs tend to be slenderer, and our brain differs in its substructure proportions[1–3] (especially the cerebellum[4]). Despite our considerable base of knowledge of how modern humans differ from archaic humans at the phenotypic level, we know very little about the genetic changes that have given rise to these phenotypic differences.

The Neanderthal and the Denisovan genomes provide a unique insight into the genetic underpinnings of recent human phenotypic evolution. The vast majority of genetic changes that separate modern and archaic humans are found outside protein-coding regions, and some of these likely affect gene expression[5]. Such regulatory changes may have a sizeable impact on human evolution, as alterations in gene regulation are thought to underlie most of the phenotypic differences between closely related groups[6–9]. Indeed, there is mounting evidence that many of the noncoding variants that emerged in modern humans have altered gene expression in *cis*, shaped phenotypes, and have been under selection[5,10–18]. Fixed variants, in particular, could potentially underlie phenotypes specific to modern humans, and some of these variants might have been driven to fixation by positive selection.

Unfortunately, our ability to infer the regulatory function of noncoding variants is currently limited[19]. In archaic humans, incomplete information on gene regulation is further exacerbated by the lack of RNA molecules and epigenetic marks in these degraded samples[5]. We have previously used patterns of cytosine degradation in ancient samples to reconstruct whole-genome archaic DNA methylation maps[12,20,21]. However, despite various approaches to extract regulatory information from ancient genomes[5,13,21–26], our understanding of gene regulation in archaic humans remains minimal, with most archaic regulatory information being currently inaccessible[5]. Additionally, whereas expression quantitative locus (eQTL) mapping can be used to identify variants that drive differential expression between individuals, it can only be applied to loci that are variable within the present-day human population. Therefore, fixed noncoding variants are of particular interest in the study of human evolution, but are also particularly difficult to characterize.

Massively parallel reporter assays (MPRAs) provide the ability to interrogate the regulatory effects of thousands of variants *en masse*[27]. By cloning a candidate regulatory sequence upstream to a short transcribable sequence-based barcode, thousands of sequences and variants can be tested for regulatory activity in parallel. Thus, MPRA is an effective high-throughput tool to identify variants underlying divergent regulation, especially in organisms where experimental options are limited[28–31]. Here, we conducted a lentivirus-based MPRA (lentiMPRA[32]) on the 14,042 fixed or nearly fixed single-nucleotide variants that emerged along the modern human lineage. We generated a library of both the derived (modern human) and ancestral (archaic human and ape) sequences of each locus and expressed them in three

human cell types: embryonic stem cells (ESCs), neural progenitor cells (NPCs), and primary fetal osteoblasts. By comparing the transcriptional activities of each pair of sequences, we generated a comprehensive catalog providing a map of sequences capable of promoting expression, and those that alter gene expression. We found that 1,791 (13%) of the sequence pairs promote expression and that 407 (23%) of these active sequences drive differential expression between the modern and archaic alleles. These differentially active sequences are associated with differential transcription factor binding affinity and are enriched for genes that affect the vocal tract and brain. This work provides a genome-wide catalog of the *cis*-regulatory effects of genetic variants unique to modern humans, allowing us to systematically interrogate recent human gene regulatory evolution.

2.3 RESULTS

2.3.1 LentiMPRA design and validation

To define a set of variants that likely emerged and reached fixation or near fixation along the modern human lineage, we took all the single-nucleotide variants where modern humans differ from archaic humans and great apes (based on three Neanderthal genomes[33–35], one Denisovan genome[36], and 114 chimpanzee, bonobo, and gorilla genomes[37]). We excluded any polymorphic sites within modern humans (in either the 1000 Genomes Project[38] or in dbSNP[39]), or within archaic humans and great apes[33–37] (see Methods). The resulting set of 14,042 variants comprises those changes that likely emerged and reached fixation or near fixation along the modern human lineage (**Supplementary File 1a-c**). The vast majority of these variants are intergenic (**Figure S2.1a**). By definition, this list does not include variants that introgressed from archaic humans into modern humans and spread to detectable frequencies. We refer to the derived version of each sequence as the *modern human sequence* and the ancestral version as the *archaic human sequence*.

We synthesized a library composed of 200 base pair (bp) sequences (due to oligonucleotide synthesis length limitations) per each of the 14,042 variants (one sequence for the modern human allele and one for the archaic human allele, **Figure 2.1, Supplementary File 1a-c**). Each sequence contained at its center either the modern or archaic human variant. 13,680 out of 14,042 sequence pairs (90%) had a single variant separating the human groups. For the 1,362 sequence pairs containing additional variants within the 200 bp window, we used either the modern-only or archaic-only variants throughout the sequence. We amplified this library of

sequences, each along with a minimal promoter and barcode. We then inserted these constructs into the lentiMPRA vector, so that the barcode, which is the readout of activity, is located within the 5'UTR of the reporter gene and is transcribed if the assayed sequence is an active regulatory element[32]. We associated each sequence with multiple barcodes to achieve a high number of independent replicates of expression per sequence, thereby reducing potential site-of-integration effects. 97% of sequences had at least 10 barcodes associated with them, with a median of 96 barcodes per sequence (**Figure S2.2a**). Furthermore, we used a chromosomally integrating construct rather than an episomal construct due to the improved technical reproducibility and correlation of results from chromosomally integrating constructs with functional genomic signals like transcription factor ChIP-seq and histone acetylation marks[43]. To further reduce lentivirus site-of-integration effects, this vector contained antirepressors on either side and was integrated in multiple independent sites, with each sequence marked by multiple barcodes. (see Discussion for additional lentiMPRA limitations). Importantly, despite the caveat of interrogating sequences outside of their endogenous context, MPRAs were shown to generally replicate the endogenous activity of sequences[43–45].

The brain and skeleton have been the focus of evolutionary studies due to their extensive phenotypic divergence among human lineages[3]. Therefore, we chose human cells related to each of these central systems: NPCs and primary fetal osteoblasts. In addition, we used ESCs (line H1, from which the NPCs were derived) to gain insight into early stages of development. Finally, the abundance of previously published regulatory maps for these three cell types[20,41,42] also enables the investigation of the dynamics of evolutionary divergence at different

regulatory levels. While these cell types represent diverse systems, further studies are needed in order to characterize the activity of these sequences in other cell types.

We used the library of 14,042 pairs of archaic and modern human sequences, together with positive and negative control sequences, to infect each cell type. As positive controls for ESCs and NPCs, we added a set of 199 sequences with known regulatory capacity from previous MPRAs (**Supplementary File 1d**). To our knowledge, there have not been any MPRAs conducted in osteoblasts, so we searched the literature for putative regulatory regions in osteoblasts and other bone cell types and used these as putative positive controls (**Supplementary File 1d**, see Methods.). As negative controls, in all cell types, we randomly chose 100 sequences from the library and scrambled the order of their bases, creating a set of GC-content matching sequences that had not been previously established to drive expression (**Supplementary File 1e**).

We performed three replicates of library infection in each cell type and quantified barcode abundance for each sequence in RNA and DNA (**Figure 2.1**). To assess the reproducibility of our lentiMPRA results, we calculated the RNA/DNA ratio (a measure of expression normalized to the number of integrated DNA molecules) for each sequence and compared it across the three replicates per cell type. We saw a strong correlation of RNA/DNA ratios between replicates for all cell types (Pearson's r = 0.76 – 0.96, $P < 10^{-100}$, **Figure S2.2b**), with the lower correlation scores being in ESC, likely due to our use of lower multiplicity of infection (MOI) in these cells due to their increased sensitivity to lentivirus infection. High barcode and read coverage in MPRA generally provides increased power to detect differences in allelic expression[32,45]. Thus,

to determine how variability depended on our barcode counts, we downsampled the number of barcodes per sequence and calculated the RNA/DNA ratio at each step for each of the three replicates. In agreement with previous studies[43], we found that the number of barcodes used in this study is well within the plateau, suggesting that the number of barcodes is not a limiting factor in our experiment (**Figure 2.2c**). Finally, we assessed the distribution of RNA/DNA ratios across our scrambled sequences and positive controls. The mean RNA/DNA ratio of the scrambled sequences was lower than that of the positive control sequences in ESCs and NPCs ($P = 2.7\text{x}10^{-8}$ for ESCs and $P = 1.8\text{x}10^{-6}$ for NPCs, *t*-test, see Methods, **Figure S2.2d**), but not in osteoblasts ($P = 0.25$). This is unlikely due to a problem with the osteoblasts, as the osteoblast-related controls show similar expression in all three cell types. Moreover, ESC and NPC positive controls are active in osteoblasts ($P = 1.1\text{x}10^{-3}$). The correlation between replicates was also similar between osteoblasts and the other two cell types (**Figure S2.2b**). Thus, the lack of activity of the osteoblast putative positive controls is likely because, in contrast to the ESC and NPC confirmed positive controls, the osteoblast putative positive controls were not previously tested in an MPRA, and some of these putative enhancers were identified in mouse and were not validated in human. Overall, these results suggest that the lentiMPRA was technically reproducible and adequately powered to detect expression.

2.3.2 Characterization of active regulatory sequences

We first examined which of the assayed sequences are able to drive expression. To do so, we utilized MPRAnalyze[40], which uses a model for each of the RNA and DNA counts, estimates transcription rate and then identifies sequences driving significant expression. We also added

an additional stringency filter whereby a sequence is only considered expressed if it had an RNA/DNA ratio significantly higher than that of the scrambled sequences (FDR < 0.05). We found that in ESCs, 8% (1,183) of sequence pairs drove expression in at least one of the alleles, 6% (814) in osteoblasts, and 4% (602) in NPCs (FDR < 0.05, **Supplementary File 1a-c**, **Figure S2.2d**, see Methods). Hereinafter, we refer to these sequences as *active* sequences. Overall, 13% (1,791) of archaic and modern human sequence pairs were active in at least one cell type, 4% (586) in at least two cell types, and 2% (222) in all three cell types (overlap of 75-fold higher than expected, $P < 10^{-100}$, Super Exact test[46], **Figure 2.2a**).

Some of these sequences may show activity in the lentiMPRA experiment, but not in their endogenous genomic context. To test whether activity in our lentiMPRA reflects true biological function, we investigated whether our active sequences had expected regulatory characteristics in the modern human genome. Active regulatory sequences in the genome tend to bear active chromatin marks. Therefore, we examined whether active sequences in lentiMPRA tend to be enriched for markers of active chromatin in their endogenous context. We first tested overlap with five histone modification marks and one histone variant associated with active chromatin (H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H2A.Z), as well as with two histone modification marks associated with repressed chromatin (H3K9me3 and H3K27me3, see Methods)[42]. We found that on average, active sequences were 1.6-2.7-fold more likely than inactive sequences to have active chromatin marks, depending on cell type. Also, these sequences tended to show relatively fewer repressive marks compared to active marks (**Figure 2.2b-d, Supplementary File 2**). These trends get stronger when looking at more highly active

sequences. For example, while only 18% of inactive sequences in ESCs overlap H3K4me2 peaks, 70% of active sequences with an RNA/DNA ratio ≥ 3 in ESCs overlap H3K4me2 peaks (*FDR* = $4.4x10^{-16}$, Fisher's exact test, **Figure 2.2b-d, Supplementary File 2**). To further test the functional characteristics of active sequences, we analyzed chromHMM annotation[41,42], which uses chromatin signatures to subdivide the genome into functional regions. 2,163 of the 14,042 sequences (15%) overlapped promoter or enhancer chromHMM annotations in at least one of the three cell types. Additional 2,658 sequences (19%) overlapped such marks in other cell types not included in this study. Compared to inactive sequences, we found that active sequences are enriched for promoter and enhancer marks (FDR < 0.05 in each of the cell types for overlap with *Active TSS* and *Enhancers*, **Figure 2.2e, Figure S2.1**, **Supplementary File 1f, Supplementary File 2)**. We also found that compared to inactive sequences, active sequences are 6-32% closer to GTEx[47] eQTLs, depending on cell type (FDR < 0.05, *t*-test). Active sequences are also 1.2-1.3x closer to transcription start sites (TSSs), with 32-39% of them located within 10 kb of a TSS, depending on cell type (FDR < 0.05, *t*-test, **Supplementary File 2**).

Active genomic regions often show reduced DNA methylation levels compared to inactive regions[48]. To further test if the activity we detected in the lentiMPRA reflects true biological function, we tested whether the active sequences in the lentiMPRA tend to be hypomethylated in their endogenous genomic context. To do so, we used our previously published modern and archaic human DNA methylation maps[12,20,21]. Because the DNA methylation maps originate from skeletal samples, we compared them to the osteoblast lentiMPRA data. We found that active sequences are significantly hypomethylated compared to inactive sequences ($P = 5.5x10^{-}$

[13], *t*-test, **Figure 2.2f**) and that their activity level (RNA/DNA ratio) is negatively correlated with methylation levels ($6.0x10^{-9}$, Pearson's r = -0.24).

Finally, compared to inactive sequences, active sequences show slightly higher sequence conservation in primates, indicating a potential functional role (PhyloP, -0.05 on average for inactive, -0.04 for active, FDR = $1.1x10^{-3}$, *t*-test) with more highly active sequences showing higher conservation levels (e.g., 0.24 for active sequences with RNA/DNA ratio ≥ 4, **Figure S2.3a**, **Supplementary File 2**). In summary, we found that sequences that are capable of driving expression tend to overlap active chromatin marks, are depleted of repressive chromatin marks, closer to TSSs and eQTLs, and have higher sequence conservation, giving us confidence that the MPRA provides us with biologically meaningful results.

2.3.3 Differentially active sequences between modern and archaic humans

We next set out to identify modern and archaic human sequences driving differential expression. We used MPRAnalyze[40] to compare expression driven by the modern and archaic sequences. Out of the active sequence pairs in each cell type, 110 (9%) in ESCs drive significantly differential expression between modern and archaic humans, 243 (30%) in osteoblasts, and 153 (25%) in NPCs (FDR ≤ 0.05, see Methods, **Figure 2.3a-c, Figure S2.2**, see Discussion for cell-type differences). We refer to these sequence pairs hereinafter as *differentially active* sequences. Overall, we see significant overlap between cell types in differentially active sequences: 407 sequences (23% of active sequences) were differentially active in at least one cell type, 89 (5%) in at least two cell types, and 10 (0.6%) in all three cell

types (8-fold higher than expected compared to active sequences, $P = 5x10^{-7}$, Super Exact test[46], **Figure 2.3d**).

As expected from such closely related organisms, and similar to other MPRAs that compared nucleotide variants (see Discussion), including one that compared human and chimp sequences[30], most sequences drove modest magnitudes of expression difference; of the 407 differentially active sequences, the median fold-change was 1.2x, and only five sequences had a fold-change greater than 2x (**Figure 2.3a-c**). We refer to differentially active sequences where modern human expression is higher/lower than archaic human expression as up/downregulating sequences, respectively. In ESCs and NPCs, sequences were equally likely to be up- or downregulating (51% and 52% of differentially active sequences were downregulating, $P = 0.92$ and $P = 0.63$, respectively, Binomial test), while in osteoblasts downregulation was observed slightly more often (59%, $P = 6.9x10^{-3}$). Finally, we examined the 89 sequence pairs that were differentially active in two cell types and the 10 sequence pairs that were differentially active in all three cell types, and tested how often the direction of differential activity in one cell type matched the direction in the other cell types. We found a strong agreement in the direction of differential activity across cell types (87 out of 89 of sequence pairs that are differentially active in two cell types, $P = 6.5x10^{-24}$, and 10 out of 10 for three cell types, $P = 9.5x10^{-7}$, Binomial test). We also observed a high correlation between the magnitudes of differential activity (Pearson's $r = 0.82$, $P = 1.6x10^{-27}$). That differentially active sequences from one cell type are predictive of differential activity in other cell types, even of cell types as disparate as those used here, suggests that these sequences are likely to be differentially active in other cell types not assayed in this lentiMPRA.

To further test the replicability of these results, we examined the relationship between pairs of overlapping differentially active sequences (i.e., variants that are < 200bp apart and thus appear in more than one sequence, three overlapping pairs in ESCs, five in osteoblasts, and two in NPCs). We found that the direction of expression change is identical in all pairs of overlapping sequences (P = 2.0x10^{-3}, binomial test), and that the magnitude of their expression change is highly correlated (Pearson's r = 0.95, 2.4x10^{-5}, **Figure S2.3b**). To validate these results with an orthogonal method, we tested four differentially active sequences from each cell type in a luciferase reporter assay and found that the direction and magnitude of differential expression tended to replicate the lentiMPRA results (9 out of 12 sequences, Pearson's r = 0.67, P = 3.7x10^{-4}, **Figure S2.3c, Supplementary File 1g**). These results suggest that the lentiMPRA was both technically reproducible across cell types and assays and also indicative of true biological signal.

Finally, we examined the endogenous genomic locations of differentially active sequences, focusing on promoters and enhancers. Between 33-45% of these sequences are within 10 kb of a TSS (depending on cell type, **Supplementary File 1h**). Analyzing chromHMM[41,42], we found that between 20-25% of the differentially active sequences are within putative promoter or enhancer regions (**Supplementary File 1f**). To test if differentially active sequences are enriched within regulatory elements, we compared the proportion overlapping chromHMM promoters and enhancers in differentially active sequences to that proportion in the other active sequences. We found that differentially active sequences are over-represented within putative enhancer regions in NPCs (2.2-fold, FDR = 0.03, Fisher's exact test, **Figure S2.1c,d**). These results

support a model of rapid enhancer evolution in modern humans, as previously reported for other mammals[50] (see Discussion).

2.3.4 Molecular mechanisms underlying differential activity

Next, we sought to understand what regulatory mechanisms might be associated with differential activity. Changes in expression are often linked to changes in regulatory marks. For example, increased DNA methylation tends to be associated with reduced activity[48]. We therefore tested methylation levels in each pair of sequences and examined if the human group with the lower sequence activity tends to show higher methylation levels. Here too, because the DNA methylation maps originate from bone samples[12,20,21], we compared them to the osteoblast lentiMPRA data. We found that upregulating sequences indeed have a slight but significant tendency to be hypomethylated in modern compared to archaic humans, and that downregulating sequences tend to be hypermethylated in modern compared to archaic humans (on average -2% methylation in upregulating sequences, and +1% methylation in downregulating sequences, in the modern compared to the archaic genomes, $P = 0.028$, paired *t*-test, **Figure S2.4a**). This trend is slightly more pronounced when looking at the most differentially regulating sequences. For example, the top ten most downregulating sequences show on average +8% methylation in modern compared to archaic humans, whereas the top ten most upregulating sequences show -7% methylation in modern compared to archaic humans. We also examined promoter regions (5 kb upstream to 1 kb downstream of a TSS), where the association between methylation and reduced activity is known to be stronger compared to the rest of the genome[48]. Indeed, we found that upregulating promoter

sequences have +5% methylation on average in the modern compared to the archaic genomes, while downregulating promoter sequences have -8% methylation (P = 0.034, paired t-test; **Figure S2.4b**). This trend is more pronounced in CpG-poor promoters, where the link between methylation and expression is known to be stronger[51–53] (-15% methylation in upregulating sequences, and +15% methylation in downregulating promoter sequences in modern compared to archaic humans; $P = 6 \times 10^{-3}$, paired t-test; **Figure S2.4c**).

We conjectured that some of the differential activity in these loci might have been driven by alterations in transcription factor (TF) binding. To investigate this, we compared predicted TF binding affinity to the modern and archaic sequences using FIMO[54]. We found that: (1) compared to other active sequences, the difference in predicted binding between the modern and archaic human alleles tends to be larger for differentially active sequences (combined across cell types: 4.3x, P = 0.02, t-test, **Figure S2.4d**); (2) the directionality of differential expression tends to match the directionality of differential binding, i.e., upregulating sequences tend to have stronger predicted binding for the modern human sequence, whereas downregulating sequences tend to have stronger predicted binding for the archaic sequence ($P = 3.7 \times 10^{-6}$ for ESCs, $P = 1.7 \times 10^{-6}$ for osteoblasts, and $P = 1.3 \times 10^{-5}$ for NPCs, binomial test, **Figure 2.3e,** see Methods); and (3) the magnitude of expression difference is correlated with the magnitude of predicted binding difference (Pearson's r = 0.43 and $P = 1.2 \times 10^{-3}$ for ESCs, Pearson's r = 0.23 and P = 0.02 for osteoblasts, and Pearson's r = 0.35 and $P = 2.4 \times 10^{-3}$ for NPCs, **Figure S2.5a-c** and **Supplementary File 3**). These results support the notion that alterations in

TF binding played a role in shaping some of the expression differences between modern and archaic humans.

To identify the TFs that primarily drove these observations, we investigated which motif changes are most predictive of expression changes. For each TF and the sequences it is predicted to differentially bind, we examined the correlation between binding and expression fold-change (either positive or negative). We found that changes to the motifs of 14 TFs were predictive of expression changes (**Figure S2.5d, Supplementary File 3b**). All of these TFs had a positive correlation between changes in their predicted binding affinity and changes in expression of their bound sequences, reflective of their known capability to promote transcription[55–63]. Of note, the use of a minimal promoter with basal activity in the MPRA design means that transcriptional repression is less likely to be detected, and therefore, further investigation is required in order to identify potential repressive activity in these sequences (see Discussion).

Next, we sought to explore if some motif changes are particularly over-represented within differentially active sequences, suggestive of a more central role in shaping modern human regulatory evolution. To control for sequence composition biases, we used active sequences as a background to search for motif enrichment within differentially active sequences. We found that ZNF281, an inhibitor of neuronal differentiation[64], is significantly enriched: out of 153 differentially active sequences in NPCs, 14 are predicted to be bound by ZNF281 (4.6-fold, FDR = 0.04, **Supplementary File 3c**). Notably, ZNF281 is also one of the TFs whose predicted

differential binding is most tightly linked with differential expression (**Figure S2.5d,e**). Overall, these data support a model whereby variants in ZNF281 motifs might have modulated ZNF281 binding in NPCs, thereby contributing to neural expression differences between modern and archaic humans.

2.3.5 Potential phenotypic consequences of differential expression

In an attempt to assess the functional effects of the differential transcriptional activity we detected, we first sought to link each sequence to the gene(s) it might regulate in its endogenous genomic location. While most regulatory sequences are known to affect their closest gene[66,67], some exert their function through interactions with more distal genes, often reflected in chromatin conformation capture assays, such as Hi-C interactions[68], or eQTL associations[68,69]. To predict the genes linked to each sequence we combined data from four sources: (1) proximity to transcription start sites; (2) proximity to eQTLs[47]; (3) proximity to putative enhancers[70]; and (4) spatial interaction with promoters using Hi-C data[69] (see Methods). Using these data, we generated for each cell type a list of genes potentially regulated by each sequence. Overall, 1,341 out of the 1,791 active sequences (75%) were linked to at least one putative target gene (**Supplementary File 1h**).

To study the potential functional effects of differentially active sequences, we analyzed functions associated with their linked genes. To control for confounders such as cell type-specific regulation, gene length, and GC content, we compared differentially active sequences to other active sequences (instead of the genomic background), which minimizes inherent

biases in the active sequences. First, we tested Gene Ontology terms and found an enrichment of the following terms within downregulating sequences: *vesicle-mediated transport* (6.6-fold, FDR = $1.9x10^{-3}$, in osteoblasts), *regulation of apoptotic process* (6.0-fold, FDR = $1.9x10^{-3}$, in ESCs), *protein ubiquitination* (4.7-fold, FDR = $1.9x10^{-3}$, in ESCs), multicellular organism development (3.3-fold, FDR = 0.01, in ESCs), and *protein transport* (3.3-fold, FDR = 0.02, in osteoblasts, **Figure S2.5f**, **Supplementary File 4a**). No enriched terms were found within upregulating sequences. To obtain a more detailed picture of phenotypic function, we ran Gene ORGANizer, a tool that uses monogenic disorders to link genes to the organs they affect[71]. We analyzed the genes linked to differentially active sequences and found that for genes linked to sequences driving upregulation, the most enriched body parts belong to the vocal tract, i.e., the vocal cords (5.0-fold, FDR = $1.3x10^{-3}$), voice box (larynx, 3.8-fold, FDR = $4.8x10^{-3}$), and pharynx (3.3-fold, FDR = $9.5x10^{-3}$, all within ESCs, **Figure 2.4a**). Interestingly, we have previously reported that the most extensive DNA methylation changes in modern compared to archaic humans arose in genes affecting the vocal cords and voice box[12]. Conversely, within sequences driving downregulation, the enriched body part is the cerebellum (3.0-fold, FDR = $9.2x10^{-3}$, in NPCs, **Figure 2.4a, Supplementary File 4b**). This is in line with previous reports of cerebellar anatomy differences between modern humans and Neanderthals[1–3], including results suggesting that the biggest differences in brain anatomy are in the cerebellum[4]. These data also provide leads into the functional divergence of organs, like the voice box, that are not preserved in the fossil record.

Next, we delved into individual phenotypes associated with the differentially active sequences. To this end, we used the Human Phenotype Ontology (HPO) database[72], a curated database of genes and the phenotypes they underlie in monogenic disorders. HPO covers a broad range of phenotypes related to anatomy, physiology, and behavior. We found that enriched phenotypes were involved in speech, heart morphology testicular descent, and kidney function (FDR < 0.05, **Figure 2.4b, Supplementary File 4b**). These results reveal body parts and phenotypes that were particularly associated with gene expression changes between modern and archaic humans, and could be new candidates for phenotypes under selection.

2.3.6 Downregulation of *SATB2* potentially underlies brain and skeletal differences

This catalog of *cis*-regulatory changes allows us to explore specific sequence changes that potentially underlie divergent phenotypes observed from fossils. To use the most robust data from lentiMPRA, we examined the ten sequences that are differentially active across all three cell types, and looked at their linked genes. To investigate the phenotypes that are potentially linked to these genes, we looked for those genes whose phenotypes can be compared to the fossil record (i.e., skeletal phenotypes). The only gene that fit these criteria was *SATB2*, a regulator of brain and skeletal phenotypes[73]. First, we analyzed its linked variant (C to T transition), which is at a position that is relatively conserved in vertebrates (GRCh38: 199,469,203 on chromosome 2, PhyloP score = 0.996). This position is found within a CpG island between two alternative TSSs of *SATB2* (**Figure 2.4c**). It is also found in the first intron of *SATB2-AS1*, an antisense lncRNA which upregulates SATB2 protein levels[74]. To determine if this position lies within a regulatory region, we investigated it for chromatin marks in modern

humans. We found that it overlaps a DNase I-hypersensitive site[75] and shows many peaks of active histone modification marks in all three cell types (**Figure 2.4c, Supplementary File 1f**). Indeed, this sequence drives high expression in all three cell types (fourth, eighth, and 19th percentile among active sequences, in ESCs, osteoblasts, and NPCs, respectively, FDR < 10^{-5} across all). Although this sequence shows hallmarks of activity in modern humans, compared to the archaic sequence the modern human sequence is downregulating in all three cell types (-9% in ESCs, FDR = $6.8x10^{-4}$, -27% in osteoblasts, FDR = $2.2x10^{-42}$, and -12% in NPCs, FDR = $1.1x10^{-7}$, **Figure 2.4d**). These results suggest that the ancestral version of this sequence possibly promoted even higher expression in archaic humans.

SATB2 encodes a transcription factor expressed in developing bone and brain. Its activity promotes bone formation, jaw patterning, cortical upper layer neuron specification, and tumorigenesis[73]. Genome-wide association studies (GWAS) show that common variants near and within *SATB2* are mainly associated with brain and bone phenotypes, such as reaction time, anxiety, mathematical abilities, schizophrenia, autism, bone density, and facial morphology[76,77]. Heterozygous loss-of-function mutations in *SATB2* result in the *SATB2*-associated syndrome, which primarily affects neurological and craniofacial phenotypes. This includes speech delay, behavioral anomalies (e.g., jovial personality, aggressive outbursts, and hyperactivity), autistic tendencies, small jaws, dental abnormalities, and morphological changes to the palate[78]. Additionally, reduced functional levels of SATB2 due to heterozygous loss-of-function have been shown to be the cause of these phenotypes in both human[73,78–80] and mouse[81–83]. Because these phenotypes are driven by changes to functional SATB2 levels[73], we conjectured that the

differential expression of *SATB2* predicted from lentiMPRA might be linked to divergent modern human phenotypes. Thus, we examined whether the phenotypes *SATB2* affects are divergent between archaic and modern humans (e.g., if modern human jaw size is different than the jaw size of archaic humans). We focused on phenotypes available for examination from the fossil record, primarily skeletal differences between modern humans and Neanderthals. From HPO, we generated a list of 17 phenotypes known to be affected by *SATB2* and found that 88% (15) of them are divergent between these human groups (**Supplementary File 5**). These include the length of the skull, size of the jaws, and length of the dental arch. Next, based on *SATB*2 downregulation in modern humans predicted from lentiMPRA, we examined whether the direction of a phenotypic change between patients and healthy individuals matches the direction of phenotypic change between modern and archaic humans. For example, given that *SATB2*-associated syndrome patients have smaller jaws, we tested if modern human jaws are smaller compared to archaic humans. If *SATB2* expression is not in fact related to phenotypic divergence, there is a 50% likelihood for a given phenotype to match the fossil record. Yet, we observed a match in direction in 80% of the phenotypes (12 out of 15, **Supplementary File 5**). This includes smaller jaws, flatter face, and higher forehead in modern compared to archaic humans. Overall, the observed number of phenotypes that are both divergent and match in their direction of change is 2.3-fold higher than expected by chance ($P = 1.3x10^{-4}$, hypergeometric test, **Supplementary File 5,** see Methods). Together, these data support a model whereby the C→T substitution in the putative promoter of *SATB2*, which emerged and reached fixation in modern humans, possibly reduced the expression of *SATB2* and possibly

affected brain and craniofacial phenotypes. However, further evidence is required to elucidate the potential role of this variant in modern human evolution.

2.4 DISCUSSION

Identifying noncoding sequence changes underlying human traits is one of the biggest challenges in genetics. This is particularly difficult in ancient samples, where regulatory information is scarce[5,21]. Here, we use an MPRA-based framework to study how sequence changes shaped human gene regulation. By comparing modern to archaic sequences, we investigated the regulatory potential of each of the 14,042 single-nucleotide variants that emerged and reached fixation or near fixation in modern humans. We found an association between divergent TF motifs and the sequences driving expression changes, suggesting that changes to TF binding might have played a central role in shaping divergent modern human expression. Our results also suggest that genes affecting the vocal tract and cerebellum might have been particularly affected by these expression changes, which is in line with previous comparisons based on the fossil record[1–4] and DNA methylation[12]. More importantly, these results provide candidate sequence changes underlying these evolutionary trends.

LentiMPRA is designed for linking DNA sequence changes to expression changes *en masse*. Notably, it has limitations that could influence our results, mainly by potentially generating false negatives. First, our lentiMPRA library inserts were limited to ~200bp in length, due to oligonucleotide synthesis technical restrictions, which may be insufficient to detect the activity of longer enhancer sequences[43]. Second, some minimally active sequences may not be expressed at a high enough level to pass our limit of detection. At the same time, some minimally active sequences may not be biologically significant. Third, some sequences may regulate expression post-transcriptionally, which lentiMPRA is not designed to detect. Fourth,

since test sequences are randomly integrated into the genome, sequences that are dependent on their endogenous genomic environments (e.g., on nearby TF binding sites) might show reduced activity when inserted in new locations, while others might show activity that they otherwise would not have. Our design partially addresses this through the use of antirepressors and multiple independent integrations, which are intended to dilute location-specific effects. Additionally, all biases are expected to similarly affect the modern and archaic human versions of each sequence[43]. Fifth, transcriptional repression is less likely to be detected due to the low basal activity of the minimal promoter used. Sixth, the level of sequence activity may depend on more than one variant (including non-fixed variants, which we have not tested here). In the cases of non-fixed variants, the extent of differential activity could vary between individuals. At the same time, in the 10% of sequences that include more than one fixed variant, it is generally impossible to determine which of the variants drives the differential activity (with the exception of cases with more than two variants where the tiled sequences include a different combination of these variants).

Finally, differences in the *trans* environment of a cell could have an effect on the ability of a sequence to exert its *cis*-regulatory effect, resulting in cell-type-specific cis-regulatory effects, as we observed in our data. The *trans* environment of the same cell type might also differ between two organisms. However, the majority of the *cis*-regulatory changes we observed would be expected to be present in archaic human cells as well, considering that such conservation has been observed between substantially more divergent organisms (e.g., human-chimpanzee[30] and human-mouse[84]). In other words, while *trans*-regulatory changes play a key

role in species divergence, the *trans* environments of the same cell type in two closely related organisms tend to affect *cis*-regulation similarly. Despite these caveats, MPRAs have been repeatedly shown to be able to replicate the activity of sequences in their endogenous context[43–45].

Importantly, when genomes from additional modern human individuals are sequenced and new variants mapped, it might become clear that some of the variants we analyzed have not reached fixation. However, regardless of whether they are completely fixed or not, these variants represent derived substitutions that likely emerged in modern humans and spread to considerable frequency. Further investigation is required to determine when they emerged, how rapidly they spread, and whether their effect was neutral or adaptive.

As expected, we observed differences in activity and differential activity between cell types[28,45,84]. Although some of this variation is likely biological (i.e., cell type-specific gene regulation), it is difficult to determine what proportion of it is due to biological versus technical factors (e.g., differences in lentivirus preparation, infection rate, or cell growth, see Methods). Importantly, these differences are expected to result in false negatives rather than false positives. In other words, some of the sequences that appear as active or differentially active in one cell type might actually be active or differentially active in additional cell types (including cell types that were not tested in this study). Thus, we largely refrained from comparisons between cell types and the overlap observed in **Figure 2.2a and Figure 2.3a** should not be used to define such similarities. Rather, these diagrams should be used to examine the replicability of

our results. Despite these caveats and limitations, lentiMPRA is a powerful high-throughput tool to characterize the regulatory activity of derived variants, and indeed has become a common assay to study the capability of sequences to promote expression[19].

With this method, we found that 1,791 (13%) of the 14,042 sequence pairs can promote expression in at least one of the three cell types tested, and that 405 (23%) of these active sequences show differential activity between modern and archaic humans (average fold-change: 1.24x, standard deviation: 0.18, **Figure 2.2, Supplementary File 1a-c**). Interpreting these results in light of previous MPRAs is challenging, not only because of key differences in statistical power and experimental design (e.g., sequence length), but also because of differing variant selection processes for each MPRA. With the exception of highly repetitive regions, which were removed from our library for technical reasons, the sequences we selected included all known modern human-derived fixed or nearly fixed variants (see Methods). Conversely, previous reporter assays and MPRAs on human intra- or inter-species variation used biased sets of variants by selecting sequences with putative regulatory function (e.g., eQTLs[28], TF binding sites[16], ChIP-seq peaks[29], or TSSs[84]) and/or regions showing particularly rapid evolution (e.g., human accelerated regions[30,31,85,86]). In line with the fact that our data was not pre-filtered for putative regulatory regions, the proportion of active sequences we observed tends to be slightly lower than these previous studies. However, the magnitude of differential activity, as well as the fraction of differentially active sequences out of the active sequences was similar to previous studies[16,28–31,84–86]. At the same time, we were capable of measuring regulatory activity in regions that would otherwise be excluded by filtering for a

specific set of marks. Thus, future MPRAs on unfiltered sets of variants will enable the comparison of the patterns we observed to patterns within modern humans, between more deeply divergent clades, and of non-fixed modern-archaic differences.

Our results also suggest that differentially active sequences are over-represented within putative enhancers in NPCs (**Figure S2.1c-d, Supplementary File 2**). Enhancers have been suggested to be an ideal substrate for evolution because of their tissue-specificity and temporal modularity[87]. Indeed, previous studies of introgression between archaic and modern humans suggested that enhancers are some of the most divergent regions between modern and archaic humans[11,25,88]. In line with the enrichment we observed in NPCs, brain-related putative enhancers show particularly low introgression, perhaps suggesting that the modern human sequences in these regions were adaptive[25,88]. To fully characterize the underlying mechanisms of differential activity in enhancers, it is important to disentangle the various factors and confounders that might contribute to this enrichment. There are several alternative explanations for the enrichment we observe, namely that variants within enhancers could be more likely to alter expression compared to other active sequences, or they could be particularly detectable in lentiMPRA. This could be tested using saturation mutagenesis MPRAs[45] to compare the effect of random mutations in enhancer and non-enhancer modern human-derived active sequences.

Our results suggest that differentially active sequences are not randomly distributed across the genome, but rather tend to be linked to genes affecting particular body parts and phenotypes.

The most pronounced enrichment was in the vocal tract, i.e., the vocal cords, larynx, and pharynx. This was evident in the Gene ORGANizer analysis, where these organs are over-represented by up to 5-fold, as well as in the HPO phenotype analysis, where some of the most enriched phenotypes are *nasal speech*, *palate development*, *nasal passage opening*, and *laryngeal stiffness* (**Figure 2.4b, Supplementary File 4c**). Overall, 53 of the 407 differentially active sequences were linked to genes which are known to affect one or more vocal tract phenotypes. Previous reports have also suggested that the vocal tract went through particularly extensive regulatory changes between modern and archaic humans[12], as well as between humans and chimpanzees[65,89]. Intriguingly, the anatomy of the vocal tract differs between humans and chimpanzees, and has been suggested to affect human phonetic range[90]. Comparing the anatomy of archaic and modern human larynges is currently impossible because the soft tissues of the larynx rapidly decay postmortem. However, together with these previous reports[12,65,89], our results enable the study of vocal tract evolution from a genetic point of view and suggest that genes influencing the modern human vocal tract have possibly gone through regulatory changes that are not shared by archaic humans.

We also identified an enrichment of brain-related phenotypes, particularly those affecting the size of the cerebellum (**Figure 2.4, Supplementary File 4b,c**). The cerebellum is involved in motor control and perception, as well as more complex functions such as cognitive processing, emotional regulation, language, and working memory[91]. Interestingly, the cerebellum has been described as the most morphologically divergent brain region between modern and archaic humans[1,4]. Evidence of divergent brain and cerebellar evolution can also be found at the

regulatory level. Studies of Neanderthal alleles introduced into modern humans through introgression provide a clue as to the functional effects of divergent loci between archaic and modern humans. These works have shown that many of the introgressed sequences were likely negatively selected, with the strongest effect in regulatory regions[11,25], particularly in brain enhancers[88]. Studies of introgressed sequences have also shown that the cerebellum is one of the regions with the most divergent expression between Neanderthal and modern human alleles[10]. Together with our results, these data collectively suggest that sequences separating archaic and modern humans are particularly linked to functions of the brain, and especially the cerebellum.

Functional information on archaic human genomes is particularly challenging to obtain because of the postmortem decay of RNA and epigenetic marks in ancient samples. MPRA not only provides a new avenue to identify differential regulation in archaic samples, but also reveals the sequence changes underlying these differences. Here, we present a catalog providing regulatory insight into the sequence changes that separate modern from archaic humans. This resource will hopefully help assign functional context to various signatures of sequence divergence, such as selective sweeps and introgression deserts, and facilitate the study of modern human evolution through the lens of gene regulation.

2.5 METHODS

Code and data availability

Code is available for download on Github: https://github.com/weiss19/AH-v-MH. Data was deposited in GEO under accession number: GSE152404.

Selection of fixed, derived variants and design of DNA oligonucleotides

We selected the variants for our lentiMPRA in the following manner. As a basis, we used the list of 321,820 modern human-derived single nucleotide changes reported to differ between modern humans and the Altai Neanderthal genome[33]. We then filtered this list to include only positions where the Vindija Neanderthal[34] and Denisovan sequences[36] both match the Altai Neanderthal variant, and are also not polymorphic in any of the four ape species examined (61 *Pan troglodytes,* 10 *Pan paniscus,* 15 *Gorilla beringei*, and 28 *Gorilla gorilla*)[37]. Next, we excluded loci which had any observed variation within modern humans in dbSNP, as annotated by Prüfer et al.[33] or in the 1000 Genomes project (phase 3)[38]. Finally, for technical limitations in downstream synthesis and cloning, we excluded variants at which the surrounding 200 base pairs (bp) had >25% repetitive elements as defined by RepeatMasker[92]. The resulting list contained 14,297 sequences and was used to design the initial set of DNA fragments. Upon completion of the lentiMPRA, another high-coverage Neanderthal genome (the Chagyrskaya Neanderthal) was published[35], and we subsequently also filtered out loci at which the Chagyrskaya Neanderthal genome did not match the ancestral sequence, bringing the final list of analyzed loci to 14,042 (28,082 archaic and modern sequences, **Supplementary File 1a-c**).

We designed DNA fragments (oligonucleotides, hereinafter oligos) centered on each variant, including the 99 bp upstream and 100 bp downstream of each variant (200 bp total). For each variant we designed two fragments, one with the ancestral (archaic human and ape) sequence and one with the derived (modern human) sequence. For cases where two or more variants would be included in the same oligo, we used either derived-only (modern human) or ancestral-only (archaic human and ape) variants throughout the oligo. The average variants per oligo out of the 14,042 oligos was 1.1, with 12,680 containing one variant, 1,259 containing two, 96 containing three and seven containing four. We also included 100 negative control fragments, created by randomly picking 100 of the designed DNA fragments and scrambling their sequence (**Supplementary File 1e**). Lastly, we incorporated 299 positive control fragments[30,85,101,93–100] (i.e., expected to drive expression; **Supplementary File 1d**). As the library was infected into three cell types (see later), we designed positive controls for each of the cell types. For human embryonic stem cells (ESCs) and human neural progenitor cells (NPCs), we used sequences which were previously shown to drive expression in MPRA in each of these cell types (**Supplementary File 1d**). For fetal osteoblast cells (Hobs), we used putative and confirmed enhancers from mouse and human (**Supplementary File 1d**). 15 bp adapter sequences for downstream cloning were added to the 5' (5'-AGGACCGGATCAACT) and 3' (5'-CATTGCGTGAACCGA) ends of each fragment, bringing the total length of each fragment to 230 bp. We synthesized each fragment as an oligonucleotide through Agilent Technologies, twice independently to minimize synthesis errors (**Supplementary File 1i**).

Production of the plasmid lentiMPRA library and barcode association sequencing

The plasmid lentiMPRA library was generated as described in Gordon et al. [32]. In brief, the two independently synthesized Agilent Technology oligo pools were amplified separately via a 5-cycle PCR using a different pairs of primers for each pool (forward primers, 5BC-AG-f01.1 and 5BC-AG-f01.2; reverse primers, 5BC-AG-r01.1 and 5BC-AG-r01.2; **Supplementary File 1i**), adding a minimal promoter (mP) downstream of the test sequence. A second round of 5-cycle PCR was performed with the same primers for both pools (5BC-AG-f02 and 5BC-AG-r02; **Supplementary File 1i**) to add a 15-bp random barcode downstream of the mP. The two pools were then combined at a 1:1 ratio and cloned into a doubled digested (AgeI/SbfI) pLS-SceI vector (Addgene, 137725) with NEBuilder HiFi Master Mix (NEB). The resulting plasmid lentiMPRA library was electroporated into 10-beta competent cells (NEB) using a Gemini X2 electroporation system (BTX) [2kv, 25uF, 200Ω] and allowed to grow up overnight on twelve 15cm 100 mg/mL carbenicillin LB agar plates. Colonies were pooled and midiprepped (Qiagen). We collected approximately 6 million colonies, such that ~200 barcodes were associated with each oligo on average. To determine the sequences of the random barcodes and which oligos they were associated with, we first amplified a fragment containing the oligo, mP and barcode from each plasmid in the lentiMPRA library using primers that contain Illumina flow cell adapters (P7-pLSmp-ass-gfp and P5-pLSmP-ass-i#, **Supplementary File 1i**). We sequenced these amplified sequences with a NextSeq 150PE kit using custom primers (R1, pLSmP-ass-seq-R1; R2 (index read), pLSmP-ass-seq-ind1; R3, pLSmP-ass-seq-R2, **Supplementary File 1i**) to obtain approximately 150M total reads. We later did a second round of barcode association sequencing of these fragments to obtain approximately 76M additional reads, for a combined

total of 225,592,667 reads. To associate barcodes with oligos, we first mapped read pairs (R1 and R3) to the original list of 28,993 oligos using bowtie2 (--very-sensitive)[102]. Next, we filtered out pairs of reads that (1) did not map to the same oligo, (2) did not have at least one of the reads in the pair with a mapping quality of ≥ 6, or (3) did not have the "proper pair" SAM designation. We linked each pair of reads with the read covering its barcode (R2) and saved only those barcode reads having at least a quality score of 30 across all 15 bases in the R2 read. We removed any barcodes associated with more than a single unique oligo (i.e., "promiscuous" barcodes), as well as any barcodes where we did not see evidence of its oligo association at least three times. We then created a list of barcode-oligo associations – this final list comprised 3,495,698 unique barcodes spanning 28,678 oligos (98.9% of the original list of 14,297 variant sequence pairs, 100 negative sequences and 299 positive control sequences), which we refer to as the barcode-oligo association list.

Cell culture and differentiation

Human fetal osteoblasts were purchased from Cell Applications Inc. (406K-05f, tested negative for mycoplasma) and were maintained in osteoblast Growth Medium (Cell Applications Inc.). For passaging, cells were washed with 1x PBS, dissociated with Trypsin/EDTA (Cell Applications Inc.), and plated at approximately 5,000 cells/cm^2. H1-ESCs (embryonic stem cells, ESCs, WiCell WA-01, RRID:CVCL_9771, identity authenticated via STR profiling, and tested negative for mycoplasma) were cultured on Matrigel (Corning) in mTeSR1 media (STEMCELL Technologies) and medium was changed daily. For passaging, cells were dissociated using StemPro Accutase (Thermo Fisher Scientific), washed and re-plated on Matrigel-coated dishes at a dilution of 1:5

to 1:10 in mTeSR1 media supplemented with 10 µM Y-27632 (Selleck Chemicals). ESCs were differentiated into neural progenitor cells (NPCs) by dual-Smad inhibition as previously described (Chambers et al., 2009; Inoue et al., 2019). Briefly, ESCs were cultured in mTeSR1 media until the cells became 80% confluent and then the media was replaced with neural differentiation media consisting of: KnockOut DMEM (Life Technologies) supplemented with KnockOut Serum Replacement (Life Technologies), 2 mM L-glutamine, 1x MEM-NEAA (Life Technologies), 1x beta-mercaptoethanol (Life Technologies), 200 ng/mL Recombinant mouse Noggin (R&D systems), and 10 µM SB431542 (EMD Millipore). On day 4 of differentiation, the neural differentiation media was gradually replaced by N2 media [DMEM/F12 (Thermo Fisher Scientific) supplemented with N2 (Thermo Fisher Scientific)] every 2 days (3:1 ratio on day 6, 1:1 on day 8 and 1:3 on day 10) while maintaining 200 ng/mL Noggin and 10 µM SB431542. On day 12, cells were dissociated into single-cell using TrypLE Express (Thermo Fisher Scientific) and cultured in N2B27 media [1:1 mixture of N2 media and Neurobasal media (Thermo Fisher Scientific) with B27 (Thermo Fisher Scientific)] supplemented with 20 ng/mL bFGF (R&D systems) and 20 ng/mL EGF (Millipore sigma)] on Matrigel-coated dish. NPCs were maintained in N2B27 with bFGF and EGF for a month and used for the following experiments at passage 15.

NPCs were validated through RT-qPCR at passage 1 (after one week of culturing in N2B27 media supplemented with bFGF and EGF) and at passage 10. RT-qPCR primers were designed for neural marker genes: *SOX1/2*, *NES* (*NESTIN*), *MAP2*; glial marker genes: *GFAP*, *OLIG2*; mesoderm marker genes: *T*(*BRA*), *GSC*; and endoderm marker genes: *SOX17*, *FOXA2* (**Supplementary File 1j**). Expression of each marker was compared to *HPRT* expression

(**Supplemental fig. 5h**). Additionally, validation via RNA-seq at passage 1 was performed. Results can be found in Figure 7A and 7D of Inoue, et al.[93] (data in GEO under accession number: GSE115046).

Cell line infection with lentiMPRA library, RNA- and DNA-seq and read processing

Lentivirus was produced and packaged with the plasmid lentiMPRA library in twelve 15cm dishes of HEK293T cells using the Lenti-Pac HIV expression packaging kit, following the manufacturer's protocol (GeneCopoeia). Additional lentivirus was produced as needed in batches of ten 15cm dishes. Lentivirus containing the lentiMPRA library (referred to hereafter as lentivirus) was filtered through a 0.45μm PES filter system (Thermo Scientific) and concentrated with Lenti-X concentrator (Takara Bio). Titration reactions using varying amounts of lentivirus were conducted on each cell type to determine the best volume to add, based on an optimal number of viral particles per cell, as described in Gordon et al.[32]. Lentiviral infection, DNA/RNA extraction, and barcode sequencing were all performed as described in Gordon et al.[32]. Briefly, each replicate consisted of approximately 9.6 million cells each of ESC and osteoblast, and 20 million cells of NPC. ESC and osteoblast cells were seeded into four 10cm dishes per replicate (with approximately 2.4 million cells in each dish), while NPCs were seeded into five 10cm dishes per replicate (with approximately 4 million cells per dish). Additional cells were used for NPCs due to decreased efficiency of DNA/RNA extraction in NPCs). Three replicates were performed per cell type. Cells were infected with the lentiMPRA library at a MOI of 50 for NPCs and osteoblasts, and a MOI of 10 for ESCs. We used a lower MOI for ESC because the cells are very sensitive to infection and a MOI higher than 10 would result in cell

death. For ESC and osteoblasts, cell media was changed to include 8ug/mL polybrene before the addition of the lentiMPRA library to increase infection efficiency. The media was replaced with growth media without polybrene approximately 24 hours after infection. Infected cells were grown for three days before combining the plates of each replicate for extraction of RNA and DNA via the Qiagen AllPrep mini kit (Qiagen). We subsequently purified mRNA from the RNA using the Oligotex mRNA prep kit (Qiagen) and synthesized cDNA from the resulting mRNA with SuperScript II RT (Invitrogen), using a primer containing a unique molecular identifier (UMI) (P7-pLSmp-ass16UMI-gfp, **Supplementary File 1i**). DNA fragments were amplified from both the isolated DNA and generated cDNA, keeping each replicate and DNA type separate, with 3-cycle PCR using primers that include adapters necessary for sequencing (P7-pLSmp-ass16UMI-gfp and P5-pLSmP-5bc-i#, **Supplementary File 1i**). These primers also contained a sample index for demultiplexing and a UMI for consolidating replicate molecules (see later). A second round of PCR was performed to amplify the library for sequencing using primers targeting the adapters (P5, P7, **Supplementary File 1i**). The fragments were purified and further sequenced with six runs of NextSeq 15PE with 10-cycle dual index reads, using custom primers (R1, pLSmP-ass-seq-ind1; R2 (read for UMI), pLSmP-UMI-seq; R3, pLSmP-bc-seq; R4 (read for sample index), pLSmP-5bc-seq-R2, **Supplementary File 1i**). Later, an additional two runs of 15PE of only the ESC samples were performed due to lower lentivirus infection efficiency in this cell type. Each samples' R1 and R3 reads (containing the barcode) were mapped with bowtie2 [102] (--very-sensitive) to the barcode-oligo association list. Next, we applied several quality filters on the resulting alignments. We first filtered out read pairs that didn't map as proper pairs, and then ensured the mapped sequence completely matched the known barcode sequence by

requiring that both R1 and R3 reads have CIGAR stings = 15M, MD flags = 15 and a mapping quality of at least 20. Next, we consolidated read abundance per barcode by selecting only reads with unique UMIs, the result being abundance counts for each barcode, across each replicate library of each cell type for both RNA and DNA.

Data was deposited in GEO under accession number: GSE152404.

Measurement of expression and differential expression

We used the R package MPRAnalyze[40] (version 1.3.1, https://github.com/YosefLab/MPRAnalyze) to analyze lentiMPRA data. To determine which oligos were capable of promoting expression, we modeled replicate information into both the RNA and DNA models of MPRAnalyze's quantification framework (rnaDesign = ~ replicate and dnaDesign = ~ replicate) and extracted alpha, the transcription rate, for each oligo. MPRAnalyze used the expression of our 100 scrambled oligos as a baseline against which to measure the level of expression of each tested oligo. We corrected the mean absolute deviation (MAD) score-based *P*-values from MPRAnalyze for multiple testing across tested oligos, including positive controls and excluding scrambled sequences, using the Benjamini-Hochberg method, thus generating an MAD score-based expression false discovery rate (FDR) for each oligo. For each variant and for each cell type, we looked at both the archaic and modern sequence oligos and assigned an oligo as potentially capable of driving expression if it had an FDR ≤ 0.05 in at least one sequence, and at least 10 barcodes in both sequences (**Supplementary File 1a-c**). This left 2,097 sequences in ESCs, 1,059 in osteoblasts, and 664 in NPCs. Next, we applied a second test for activity, to account for potential overestimation of active sequences in ESCs due to the

lower lentiviral infection efficiency in these cells. We aggregated UMI-normalized read abundances across all barcodes of each oligo, across all replicates in a given cell type, and calculated a simple ratio of expression as RNA abundance normalized to DNA abundance (RNA/DNA ratio). Next, similarly to Kwasnieski et al.[103], we determined an RNA/DNA ratio threshold per cell type. This was done by first removing scrambled sequences that show RNA/DNA ratios >2 standard deviations away from the average RNA/DNA ratio of all of the scrambled sequences, as these likely represent oligos that are, by chance, capable of driving some expression. This left 95 scrambled sequences in ESCs, 94 in osteoblasts and 97 in NPCs. Then, we used the distribution of RNA/DNA ratios of the remaining scrambled sequences to assign an FDR for each of the non-scrambled oligos. FDR was calculated as the fraction of scrambled sequences that showed an RNA/DNA ratio as high or higher than each non-scrambled oligo. Only oligos that passed both tests described above (FDR ≤ 0.05 in each test) were considered as "active" (i.e., capable of driving expression). This resulted in 1,183 sequences in ESCs, 814 in osteoblasts and 602 in NPCs.

To measure differential expression between archaic and modern sequences, we used MPRAnalyze's comparative framework. In essence, this tool uses a barcode's RNA reads as an indicator of expression level and normalizes this to the DNA reads as a measure of the number of genomic insertions of that barcode (i.e., the number of fragments from which RNA can be transcribed). MPRAnalyze uses information across all the barcodes for both alleles of a given sequence, as well as information across all replicates. For the terms of the model, we included replicate information in the RNA, DNA and reduced (null) models, allele information in the RNA

and DNA models, and barcode information only in the DNA model (rnaDesign = ~ replicate + allele, dnaDesign = ~ replicate + barcode + allele, reducedDesign = ~ replicate). We extracted *P*-values and the differential expression estimate (fold-change of the modern relative to archaic sequence). Then, we corrected the *P*-values of the set of active oligos (see above) for multiple testing with the Benjamini-Hochberg method to generate an FDR for each sequence. We set a cutoff of FDR ≤ 0.05 to call a sequence capable of driving differential expression. From this we generated, for each cell type, a list of sequences with differential expression between the archaic and modern alleles (**Supplementary File 1a-c**).

We tested agreement between replicates by examining how many differentially active sequences show disagreement between the three replicates in the direction of their differential activity. We found that our dataset shows high between-replicate agreement, with the majority of sequences showing the same directionality across all three replicates (ESCs: 76%, osteoblasts: 78%, NPCs: 86%, compared to 25% expected by chance, $P < 10^{-16}$ for all three cell types, one-tailed Binomial test, **Supplementary File 1k**). Importantly, the log2(fold-change) of the disagreeing replicate tends to cross the 0 line only marginally: the median log2(fold-change) of the disagreeing replicate is 0.05 compared to 0.3 in the agreeing replicates. We also tested activity levels and found no evidence of lower activity in sequences with disagreement (P = 0.27, one-tailed *t*-test). However, their absolute log2(fold-change) tends to be slightly lower (0.25 vs. 0.32, $P = 6x10^{-5}$, one-tailed *t*-test).

Luciferase validation assays

Each assayed oligo was synthesized by Twist Biosciences and cloned into the pLS-mP-Luc vector (Addgene 106253) upstream of the luciferase gene. Lentivirus was generated independently for each vector using techniques as described for MPRA (see above), with the omission of the filtering and concentration step, which was replaced with the collection of the entirety of the cell culture media for use in subsequent infections. In addition, pLS-SV40-mP-Rluc (Addgene 106292), to adjust for infection efficiency, was added at a 1:3 ratio to the assayed vector for a total of 4ug for lentivirus production. We infected each cell type individually with each viral prep. The amount of lentivirus added was based on titrations in which varying amounts of a subset of viral preps were added to each cell type and cell death was observed 3 days post infection; the virus volume that produced between 30-50% death was used for subsequent experiments. Approximately 20,000 cells were plated in 96-well plates and grown for 24-48 hours (~70% confluent) before the addition of lentivirus. For osteoblasts and ESCs, 8ug/mL polybrene was added to the culture media at the same time as the addition of the lentivirus. The media was changed 24 hours after infection and cells were grown for an additional 48 hours. The cells were then washed with PBS and lysed. Firefly and renilla luciferase expression were measured using the Dual-Luciferase Reporter Assay System (Promega) on the GloMax plate reader (Promega). Each oligo was tested using two biological replicates on different days and each biological replicate consisted of three technical replicates. Activity of a given oligo was calculated by normalizing the firefly luciferase activity to the renilla luciferase. We then calculated the $\log_2$ fold change (LFC) between the modern and archaic alleles as $\log_2$(modern / archaic). A full list of oligos tested and their LFC can be found in **Supplementary File 1a-c**.

We found that the mean difference in fold-change between replicates was threefold lower for the differentially active vs other active sequences (0.22 vs 0.60), and that the variance of these differences was ninefold lower for differentially active sequences compared to other active sequences (0.09 vs 0.83, **Supplementary File 1k**), suggesting that differentially active sequences reflect a true biological signal.

Predicting target genes

To connect the surrounding locus of each variant to genes it potentially regulates, we combined four data sources. For each locus, we generated four types of gene lists, based on four largely complementary approaches: (1) overlap with known expression quantitative trait loci (eQTLs); (2) spatial interaction with promoters; (3) proximity to putative enhancers; and (4) proximity to a transcription start site (TSS, **Supplementary File 1h**). Each data source was obtained and incorporated into each type of list as described below:

1) Proximity to known eQTLs

eQTLs are genetic variants between individuals shown to be associated with expression differences. We reasoned that the target genes of the sequence surrounding a variant are potentially similar to the target genes of nearby eQTLs. We downloaded eQTLs and their associated genes from GTEx[47] (www.gtexportal.org, v8 on August 26, 2019) and overlapped the locations of each eQTL with our list of sequences. We linked the target genes of any eQTLs within +/-1 kb to each variant. We used all tissue types reported by GTEx, for each cell type in the lentiMPRA. 9,503 out of the 14,042 loci were found within +/- 1 kb of an eQTL, with 83,777 eQTLS overall overlapping them.

2) Spatial interaction with a promoter via Hi-C data

High-throughput chromosome conformation capture (Hi-C) techniques map spatial interactions between segments of DNA. We reasoned that if a variant is found within or near a region that was shown to interact physically with a promoter, that variant could be in a region involved in regulating that promoter. We downloaded promoter capture Hi-C data from Jung et al.[69], containing a list of all the significant interactions between promoters and other segments of the genome across 27 tissue and cell types. We overlapped our variants with the locations of interacting genomic fragments to find interactions within +/-10 kb of each variant. We then linked each variant with the promoters that each interacting fragment was shown to contact. We repeated this process twice: once to obtain a cell type-specific list, and once to obtain a generic list. For the cell type-specific (stringent) list of locus-gene links, we included only those interactions observed in cell types corresponding to the cell lines used in our lentiMPRA: ESCs, NPCs and mesenchymal stem cells as an approximation for osteoblasts (given that osteoblast Hi-C data is not publicly available to the best of our knowledge, and that osteoblasts differentiate from MSCs). For the generic (non-stringent) list, we used interactions across any of the 27 tissue and cell types analyzed by Jung et al.[69]. 4,688 out of the 14,042 loci overlapped at least one region that interacts with a promoter.

3) Putative enhancers

Lastly, we checked which of our variants were in previously reported putative enhancers. To this end, we downloaded the GeneHancer database[70] V4_12 and searched for putative enhancers within +/- 10kb of each of our variants, linking each variant to the target genes of each putative enhancer within that distance. GeneHancer provides “elite” or “non-elite” status

to their defined enhancer-target gene connections depending on the strength of the evidence supporting each connection. Using this information, we repeated the process twice: once for the elite status and once for all annotations. 5,017 out of the 14,042 loci overlapped at least one putative enhancer

4) Promoters

Promoters were defined as the region 5kb upstream to 1kb downstream of GENCODE[104] v29 GRCh38 TSSs. If a variant fell within this region, we linked it to that TSS's gene. Each variant was assigned to all the promoters it fell within. 1,466 out of the 14,042 loci were found within a promoter.

Overall, 11,207 out of the 14,042 loci were linked to at least one putative target gene, with a median of four target genes per locus. 2,830 of the remaining loci were linked to their closest TSS, regardless of distance. The last 5 without hg38 coordinates for their closest TSS were not linked to a gene. Importantly, these links do not necessarily mean that these target genes are regulated by these loci, but rather they serve as a list of potential target genes for the loci showing a regulatory function through lentiMPRA.

DNA methylation in active and differentially active sequences

The four highest resolution DNA methylation maps for modern and archaic bone samples were taken from Gokhman et al. 2014 [ref [20]] and Gokhman et al. 2020 [ref [12]]. Promoter sequences were defined as sequences within 5 kb upstream to 1 kb downstream of a TSS. CpG-poor promoter sequences were defined as promoter sequences ranking at the bottom half based on

their CpG density. Enhancer sequences were defined as sequences annotated in chromHMM as putative enhancers (i.e., enhancers, genic enhancers, and bivalent enhancer) in osteoblast cells. In putative enhancer sequences we found a slightly weaker link between methylation and activity compared to promoter sequences, with 3% hypermethylation of downregulating sequences and 5% hypomethylation of upregulating sequences. Perhaps in accordance with the much weaker link between enhancer methylation and activity[48], this trend is not significant despite having similar statistical power to the promoter analysis (P = 0.12, paired t-test). To test whether our results might have been affected by CpG density, we compared CpG density in differentially active compared to non-differentially active sequences, and in upregulating compared to downregulating sequences. We found no significant difference in CpG density between these groups (P-values > 0.05, t-test).

The hypermethylation of downregulating sequences in modern compared to archaic humans, and the hypomethylation of upregulating sequences in modern compared to archaic humans is also observed to some extent when testing these sequences in NPCs, but not in ESCs. For example, the top 10 upregulating sequences are hypomethylated by 7% on average in modern compared to archaic humans, top 10 downregulating sequences are hypermethylated by 13% in modern compared to archaic humans. This is in line with previous observations that differentially methylated regions tend to be shared across tissues[105].

Differential transcription factor binding sites

We predicted differences in binding of human transcription factors caused by each of our variants as follows. First, we downloaded the entire set of publicly available human transcription factor binding motifs (7,705 motifs, 6,608 publicly available) from the Catalogue of Inferred Sequence Binding Preferences (CIS-BP) database (http://cisbp.ccbr.utoronto.ca/), and filtered them to include only motifs labeled as *directly determined* (i.e., we filtered out inferred motifs), resulting in 4,351 motifs. Next, to enrich our mapping result for matches covering the variant location, we trimmed each of our oligo sequences containing a single variant to +/- 30 bp around the variant (the length of the longest motif). We did not trim oligos containing >1 variant. We used FIMO[54] to map each remaining motif to both the archaic and modern alleles of each trimmed sequence (or untrimmed, for sequences with >1 variant). A background model was generated using fasta-get-markov using the trimmed (or untrimmed, if >1 variant) sequences. For each motif mapping to both the archaic and modern alleles at the same strand and location, we required that at least one allele had a *q*-value (as supplied by FIMO) ≤ 0.05). Then, we found cases where the FIMO predicted binding score of a motif differed between the archaic and modern alleles. FIMO uses a *P*-value cutoff of 10^{-4} for reporting predicted binding. Therefore, some sequence pairs have a reported score for only one of the alleles. To assign these sequence pairs with a score difference, we used a conservative approach where we assigned the unscored allele with this lowest score reported for that motif, representing a score that is closest to a *P*-value of 10^{-4}. Because the unreported score could be anywhere below the lowest reported score, but could not have been above it, this results in a conservative underestimation of the score difference. Finally, we linked each motif to the transcription

factor (TF) it is most confidently associated with in CIS-BP, thereby generating lists of TFs that showed differential predicted binding for each sequence. For cases in which multiple unique motifs corresponded to the same TF, we used the motif with the largest score difference between alleles. TF enrichment analyses were done on all predicted differential TF binding sites for TFs with a minimum of 10 predicted differential sites. TFs that are not expressed in the cell types we examined in this study (FPKM < 1) were removed from the analyses. For TF expression in ESCs, we used ENCODE RNA-seq data for H1-hESC[75]. For osteoblast expression, data[106] was downloaded from GEO under accession number: GSE57925. For NPC expression, data[107] was downloaded from GEO under accession number: GSE115407. Fisher's exact test was used to compute enrichment of a TF among differentially active sequences compared to other active sequences. *P*-values were FDR-adjusted.

To further test the enrichment of ZNF281, we examined various cutoffs of the number of predicted bound motifs, ranging from 5 to a maximum of 14 (the number of motifs predicted to be differentially bound by ZNF281) in steps of 1. We found that with the exception of the cutoffs of 5 and 6 (where ZNF281 is only slightly above the significance threshold: FDR = 0.058 and FDR = 0.053, respectively), ZNF281 is the only significant TF across all of these cutoffs (FDR ≤ 0.05). We repeated the same test for FPKM cutoffs, ranging from 0.5 to 3 in steps of 0.5, and found that ZNF281 is the only significantly enriched TF (FDR ≤ 0.05) across all of these cutoffs. For the predicted binding vs. expression correlation analysis, a cutoff of 10 sites per TF was used. *P*-values were computed using Pearson's correlation.

Overlapping loci with genomic features

The following datasets were used for the overlap analyses: GENCODE v28 GRCh38 human genome TSSs[108], GTEx v8 eQTLs[47], and broad peaks for the following histone modification marks: H3K27ac, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, and H3K27me, and the histone variant H2A.Z from the Roadmap Project for ESCs, ESC-derived NPCs, and osteoblasts[42]. We overlapped each of these datasets with the lists of inactive and active sequences, and computed enrichment *P*-values using a Fisher's exact test. We repeated this for various RNA/DNA cutoffs (1, 1.5, 2, 2.5, 3 and 3.5). Sex chromosomes were removed from the analyses. *P*-values were FDR-adjusted using the Benjamini-Hochberg procedure.

Sequence conservation within primates was taken from the Altai Neanderthal genome annotation, which used the PhyloP metric[33].

Human-chimpanzee *cis*-regulatory expression changes

We investigated the expression of genes associated with differentially active sequences by analyzing human and chimp RNA-seq data. As the expression changes we report are driven by *cis*-regulatory changes, we used our recently generated RNA-seq data from human-chimp hybrid cells[65] (GEO accession numbers: GSE146481 and GSE144825). In these hybrid cells, the human and chimpanzee chromosomes are found within the same nuclear environment and are exposed to the same trans factors (e.g., transcription factors). Therefore, any differential expression observed between the human and chimpanzee alleles within these hybrid cells is attributed to *cis*-regulatory changes. These cells are hybrid human-chimpanzee induced pluripotent stem cells (iPSCs), and we therefore investigated whether genes associated with

upregulating sequences in our ESC lentiMPRA data tend to be upregulated in the hybrid iPSCs, and vice versa. It is important to note that differential expression between humans and chimpanzees reflects ~12 million years of evolution (i.e., changes that emerged along the human as well as along the chimpanzee lineages since their split from their common ancestor ~6 million years ago). However, our lentiMPRA data was done on sequences that changed along the modern human lineage (~550-765 thousand years). Therefore, the human-chimpanzee differences span an evolutionary time that is ~20-fold longer than the modern human lineage, and the effect of modern-derived variants on gene expression between humans and chimpanzees is expected to be largely diluted by the many other changes that accumulated along the rest of this time. Indeed, we observe a very slight, but significant correlation between differential expression observed in the lentiMPRA data and differential expression observed in the human-chimp hybrid data (P = 0.017, Pearson's r = 0.1, **Figure S2.5g**).

Phenotype enrichment analyses

Body part enrichment analyses were conducted using Gene ORGANizer v13. The analyses were conducted on sequences driving increased expression, sequences driving decreased expression, and all differentially active sequences. This was done in each of the three cell types. We conducted these analyses using various $\log_2$(fold-change) thresholds: 0, 0.5, and 0.75, on the non-stringent locus-gene associations, and using a cutoff of 5 genes per term. Analyses were done against the active sequences as background, and using the ORGANizer tool with the *confident* option. *P*-values were FDR-adjusted using the Benjamini-Hochberg procedure. For

osteoblasts, non-skeletal organs were removed from the analyses. For NPCs, non-neuronal organs were removed.

For the HPO analyses, we used HPO[72] build 1268 (08 November, 2019), analyzing gene lists identical to the Gene ORGANizer analyses, with the exception of using a cutoff of 3 genes per term, because fewer genes are linked to HPO terms than to Gene ORGANizer terms. Lists of phenotypes from HPO were generated for each variant through its linked genes. Hypergeometric test *P*-values were computed per phenotype and FDR-adjusted. Similarly to the Gene ORGANizer analysis, we removed non-skeletal phenotypes from the osteoblast results, and non-neuronal phenotypes from the NPC results.

Gene Ontology, Gene ORGANizer and HPO analyses were also done on the full set of genes linked to the 14,042 fixed variants using the same parameters described above (**Supplementary File 6**). Importantly, unlike the analyses of differentially active sequences, which can be compared against a non-differentially active sequences background to control for potential biases, the full set of sequences cannot be compared against a background set. Therefore, these results may be affected by different confounders such as GC content, the ability to call SNPs, DNA degradation patterns, and it is still to be determined to what extent these results reflect true evolutionary trends.

SATB2 phenotypic analysis was done as previously described in Gokhman et al[14]. In short, we used HPO[72] build 1268 (08 November, 2019) to link phenotypes to *SATB2*. In addition, we

conducted a literature search to expand gene-phenotype links to include studies that did not appear on HPO (**Supplementary File 5**). We used only skeletal directional phenotypes, i.e., phenotypes that could be described on a scale (e.g., smaller/larger hands), as these could be examined against the fossil record. This resulted in 34 phenotypes that are the result of *SATB2* heterozygous loss-of-function (LOF) (**Supplementary File 5**). Phenotypes that are included in another phenotype (e.g., *Prominent nasal bridge* and *Prominent nose*) were merged, and contradicting phenotypes (e.g., *Broad nose* and *Thin/small nose*) were removed. This resulted in a final list of 17 phenotypes (**Supplementary File 5**). Given that the mechanism underlying these phenotypes is a decrease in the dosage of SATB2, and that *SATB2* is possibly downregulated in modern humans, we sought to investigate if similar phenotypes exist between modern human patients with *SATB2* heterozygous LOF and archaic humans. For each phenotype, we determined if it is divergent between the modern and archaic humans based on previously published annotation[14]. Then, for remaining divergent phenotypes, we tested if the direction between patients and healthy individuals matches the direction between modern and archaic humans. The significance of directionality match was computed using a binomial test, with a random probability of success $p = 0.5$. To compute the significance of the overall number of phenotypes that are divergent and match in direction, we compared the overall number of annotated divergent phenotypes to the number of divergent phenotypes associated with *SATB2* using a hypergeometric test. Out of a total of 696 annotated phenotypes between modern and archaic humans[14], 434 are annotated as divergent, and the direction of 50% of them (217 phenotypes) is expected to match by chance.

2.6 SUPPLEMENTARY

All supplementary materials can be accessed through eLife where this article has been published: https://doi.org/10.7554/eLife.63713

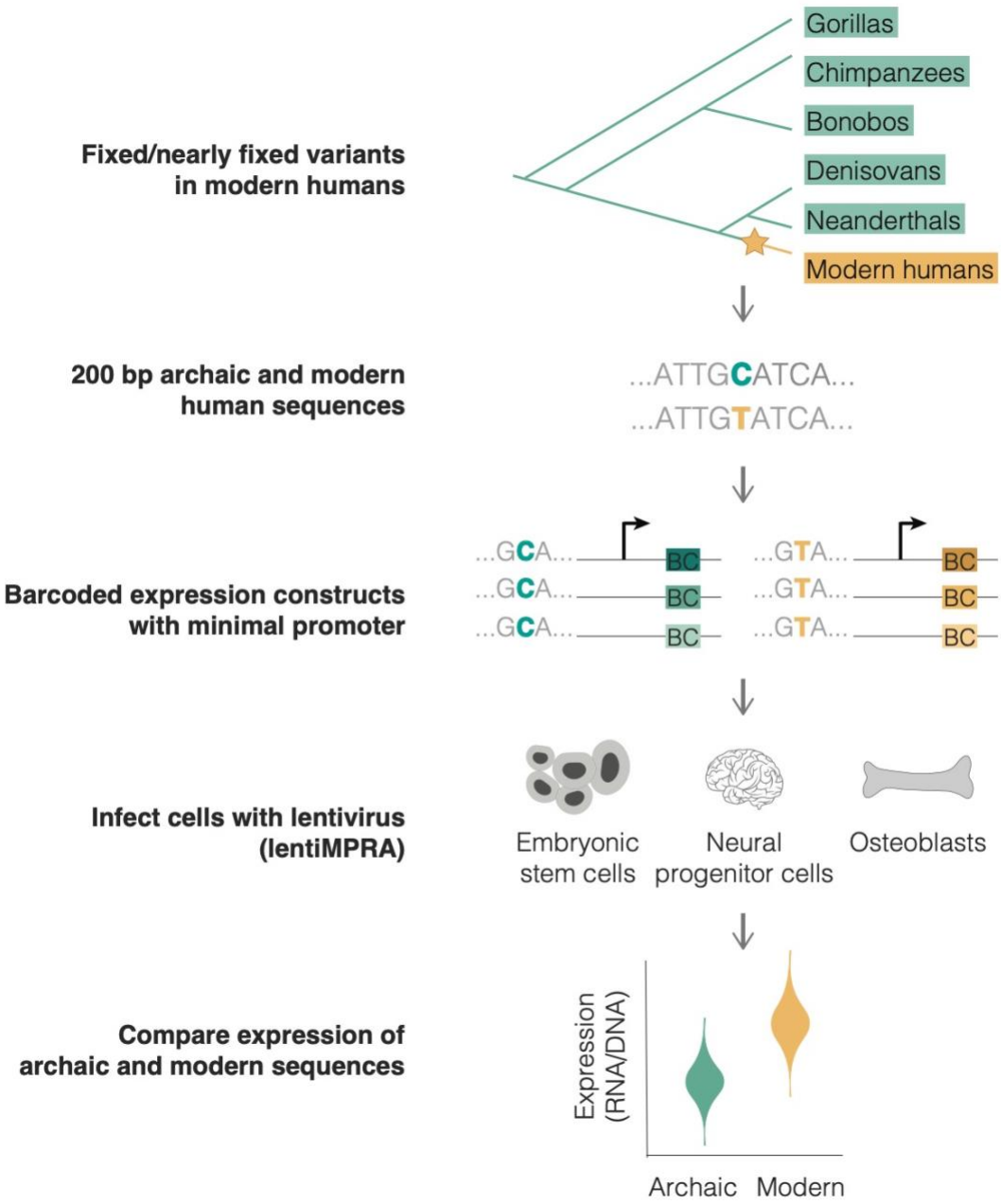

Figure 2.1: Using lentiMPRA to identify variants driving differential expression in modern humans

We analyzed variants that likely emerged and reached fixation or near fixation along the modern human lineage (yellow) and that were not polymorphic in any other ape or archaic genome (green) (top). The modern and archaic human variants and their surrounding 200 bp were synthesized, cloned into barcoded expression constructs and infected in triplicates into three human cell lines using a chromosomally integrating vector, following the lentiMPRA protocol[32] (see methods). We compared the activity (RNA/DNA) of the modern and archaic human constructs to identify variants promoting differential expression using MPRAnalyze[40] (bottom).

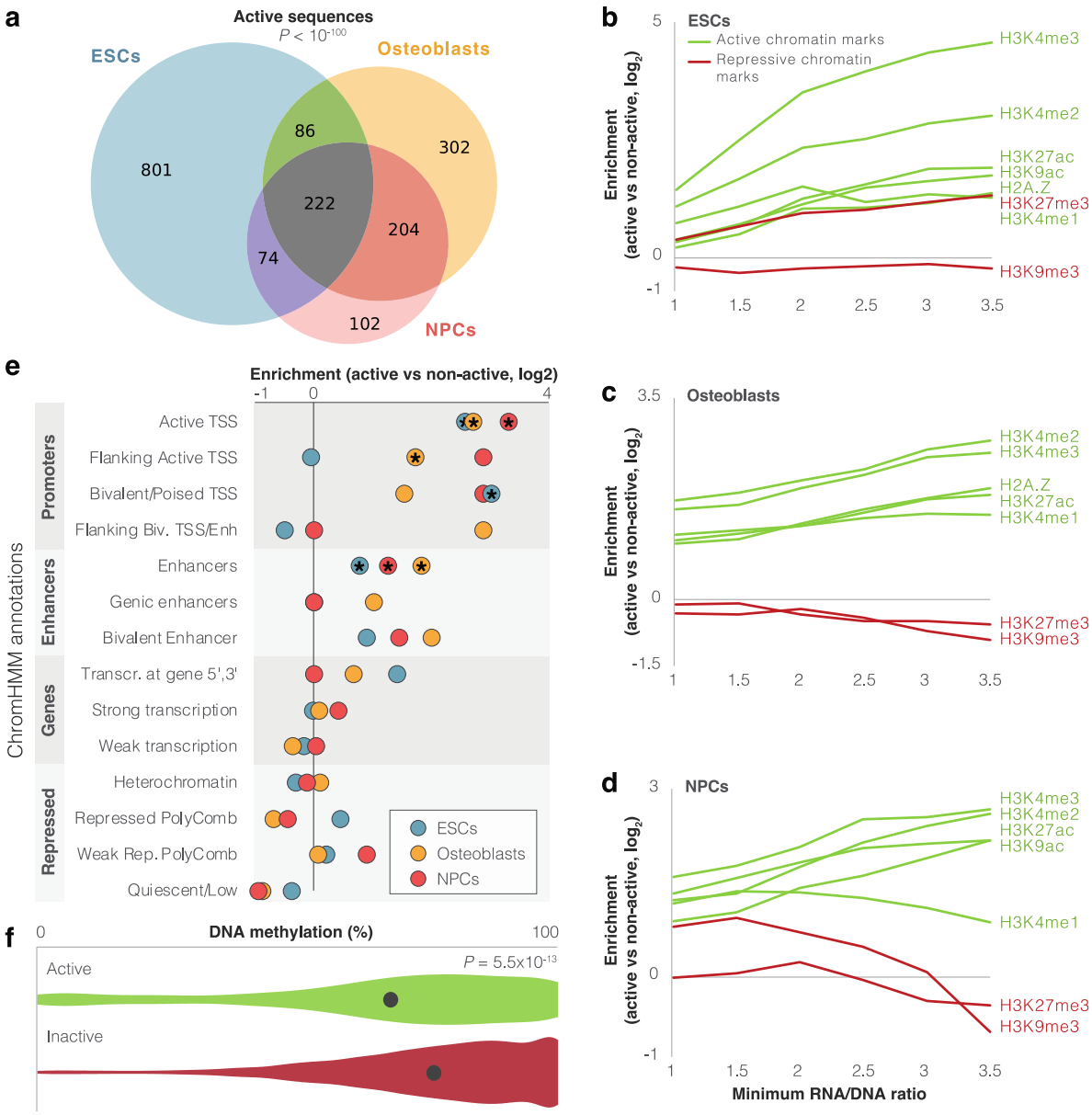

Figure 2.2: Identification of modern human sequences promoting expression in lentiMPRA
a. Overlap between cell types of active sequences. Super Exact test *P*-value is shown for the overlap of the three groups. b-d. Enrichment levels of active and repressive histone modification marks within active sequences. Enrichment is computed compared to inactive sequences. The enrichment of H3K27me3 in ESCs possibly reflects the presence of this mark in bivalent genes, which become active in later stages of development[49]. For confidence intervals see Supplementary File 2. e. Enrichment of differentially active sequences in various chromatin-based genomic annotations. Missing circles reflect no differentially active sequences in that category. Stars mark significant enrichments (FDR < 0.05). f. Violin plots of DNA methylation levels for active (green) vs. inactive (red) sequences in osteoblasts. Methylation levels per sequence were computed as the mean methylation across all modern and archaic human bone

methylation samples. The circle marks mean methylation across all sequences in each group. *t*-test *P*-value is shown.

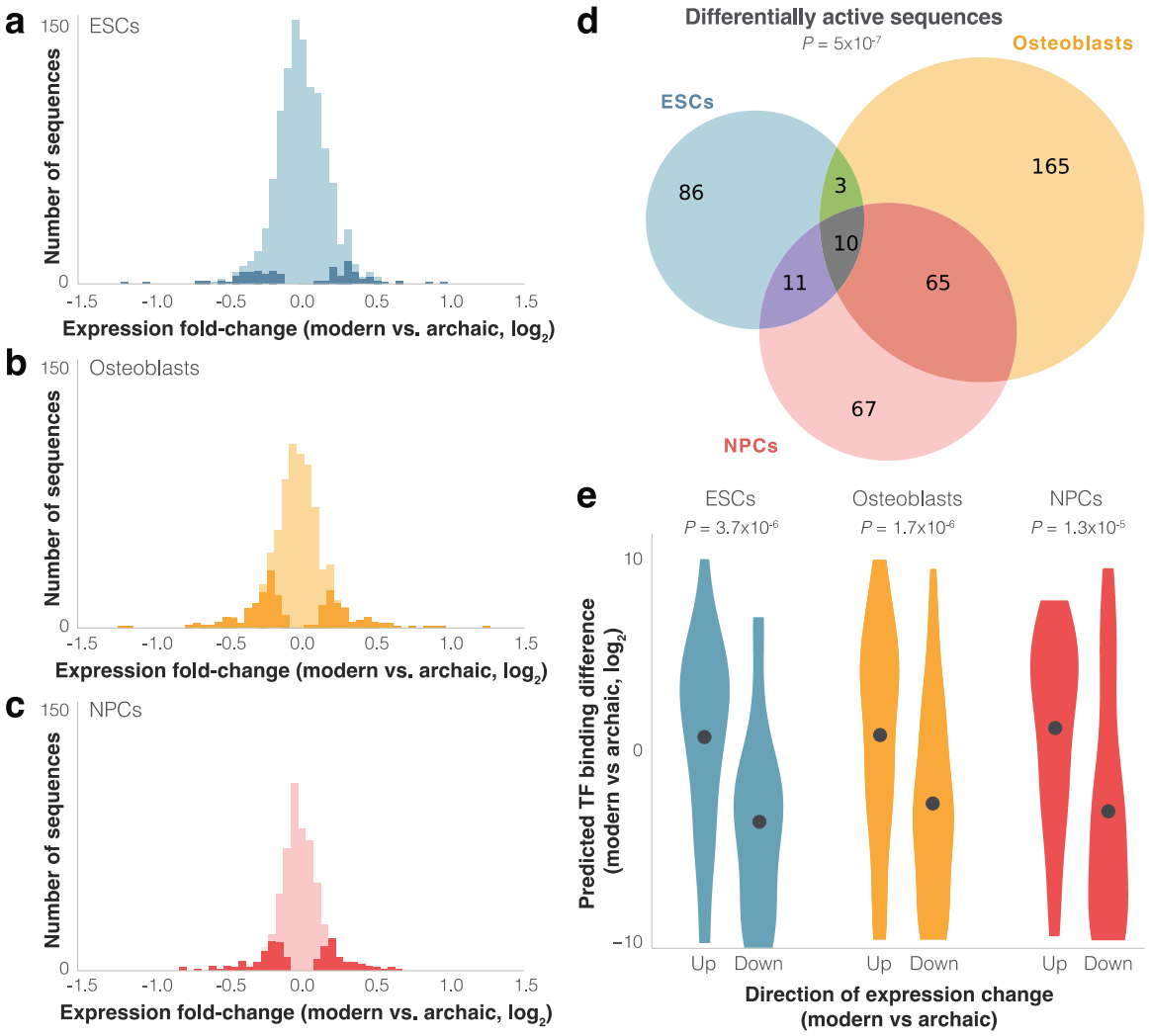

Figure 2.3: Differential activity of derived modern human sequences

a-c. Distributions of expression fold-changes (RNA/DNA) of active (light) and differentially active (dark) sequences in each cell type. d. Overlap of differentially active sequences between cell types. Super Exact test *P*-value is presented for the overlap of the three groups compared to active sequences. In the 10 sequences that were differentially active across all three cells types, the direction of fold-change was identical across all cell types ($P = 1.9 \times 10^{-3}$, Binomial test). e. Violin plots of predicted TF binding score difference between modern and archaic sequences. Positive scores represent increased binding in the modern sequence. Points show mean.

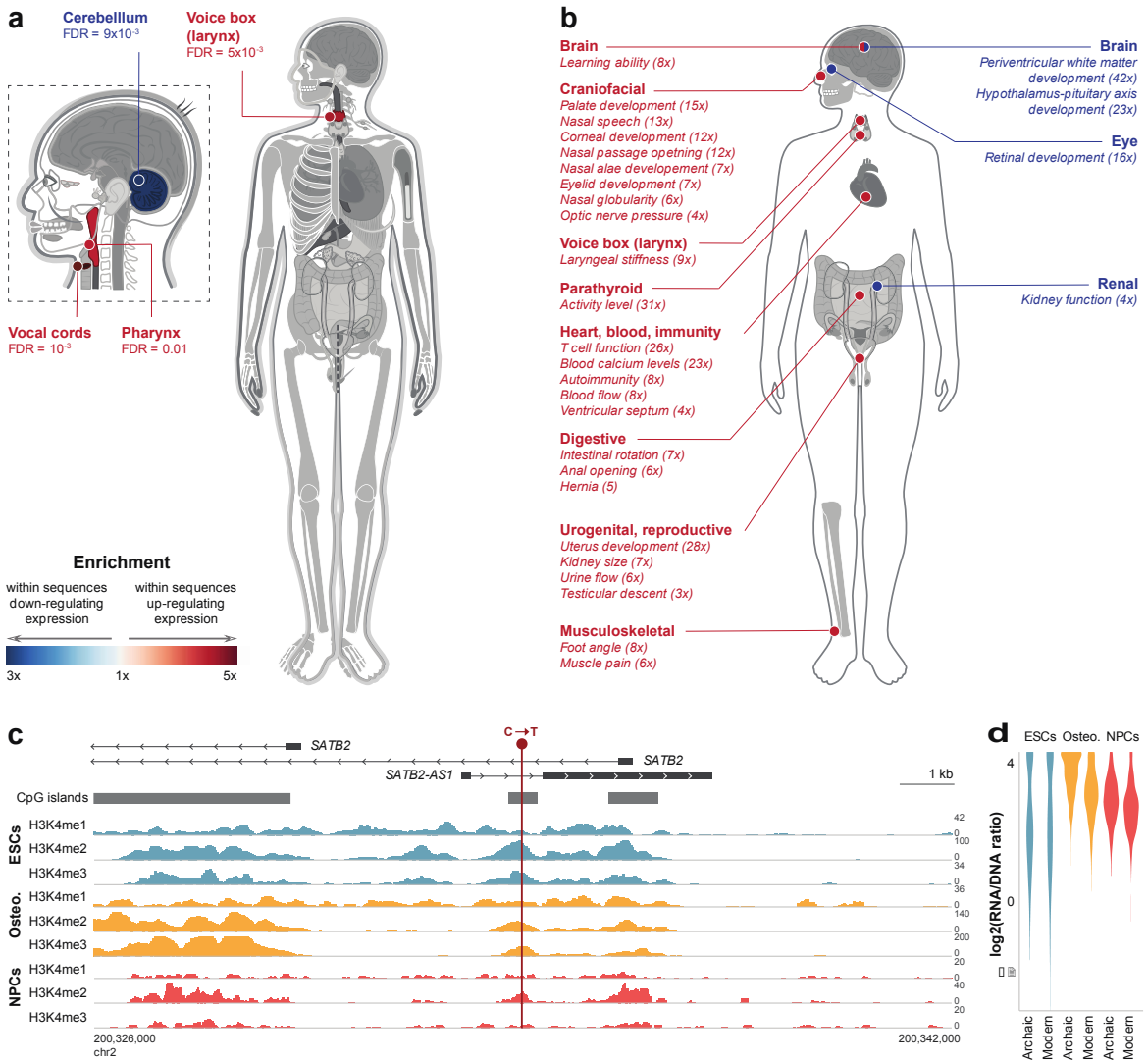

Figure 2.4: Differentially active sequences are linked to genes affecting the vocal tract and brain

a. Gene ORGANizer enrichment map showing body parts that are significantly over-represented within genes linked to differentially active sequences (FDR < 0.05). Organs are colored according to the enrichment scale. See Supplementary File 4 for cell types. b. HPO phenotypes significantly enriched (FDR < 0.05) within differentially active sequences. Fold-enrichment is shown in parentheses. See Supplementary File 4 for cell types. c. CpG islands and read density of active histone modification marks[42] around the differentially active sequence in *SATB2* (GRCh37 genome version). d. Violin plots of archaic vs. modern activity of the differentially active sequence in *SATB2*.

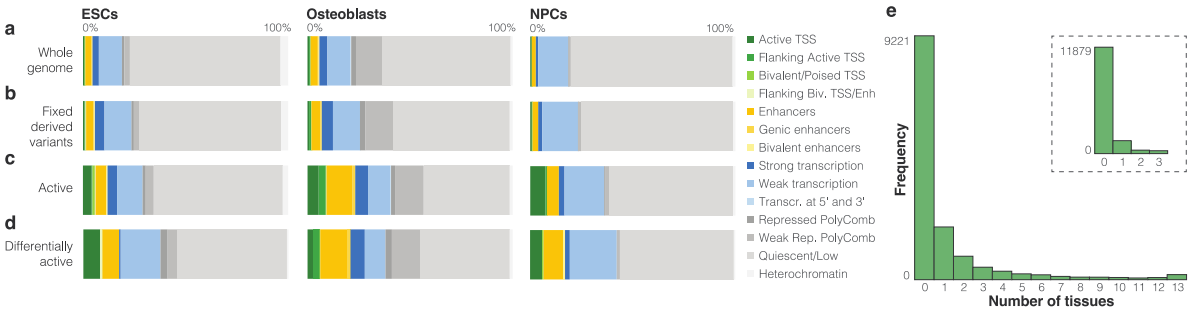

Figure S2.1: Classification of chromHMM annotations for different groups of variants
Relative percentage of bases in each chromHMM[41,42] category throughout the entire genome (a), in fixed or nearly fixed modern human-derived variants (b), in active sequences (c) and in differentially active sequences (d), per cell type. See Discussion for cell-type specificity and enhancer enrichment. e. Histogram of the number of tissues and number of sequences with TSS- or enhancer-related chromHMM marks for all 14,042 sequences. Tissues and cell types investigated include ESCs, osteoblasts, NPCs, mesenchymal stem cells, monocytes, skin fibroblasts, brain hippocampus, skeletal muscle, heart left ventricle, sigmoid colon, ovary, fetal lung, and liver. Inset shows data for ESC, osteoblast and NPC only.

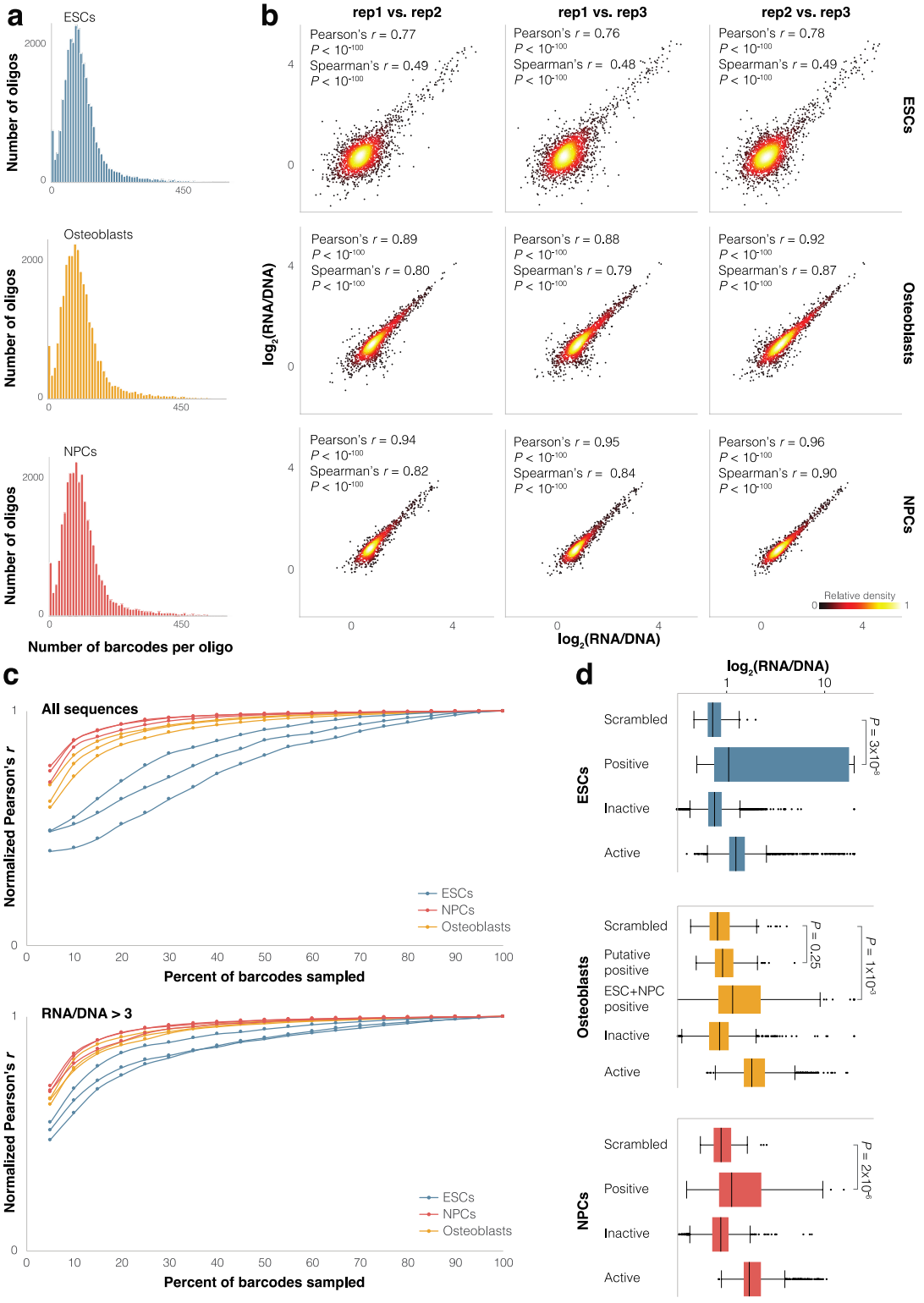

Figure S2.2: Reproducibility of lentiMPRA data

a. Distribution of number of barcodes per each sequence. b. Replicate-by-replicate correlation of expression (RNA/DNA). Each point represents an active sequence. c. Simulations of barcode

down-sampling showing Pearson's correlation of expression (RNA/DNA) between replicates. Upper panel shows all sequences and lower panel shows sequences with higher expression (RNA/DNA > 3). Pearson's *r* values are normalized to maximum Pearson's *r* observed for each pair of replicates. d. Box plots of scrambled, positive control, inactive and active sequences. One-sided *t*-test *P*-values are shown. Boxes show interquartile range (IQR), black line within box shows median, whiskers show 1.5xIQR from box borders, points show outliers.

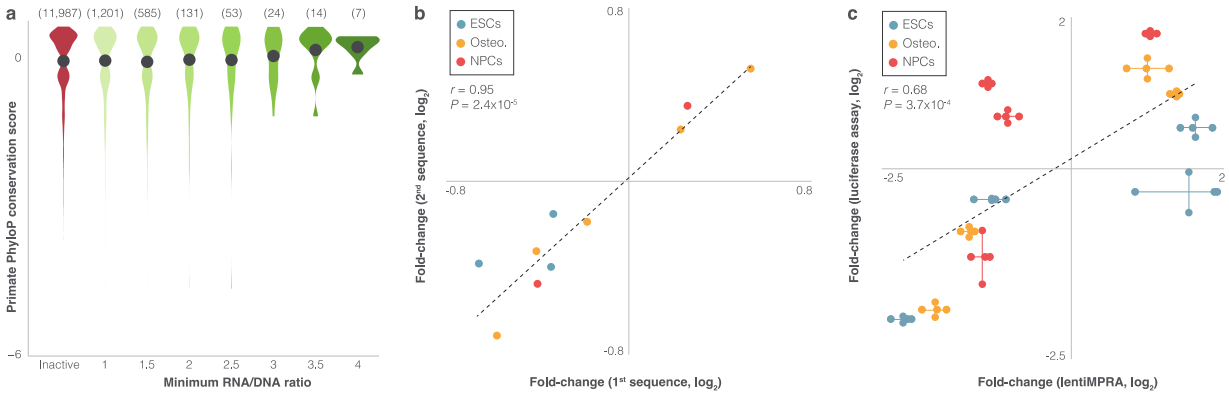

Figure S2.3: Differential expression is replicated across overlapping sequences and in a reporter assay validation

a. Primate PhyloP conservation scores in inactive sequences and active sequences with increasingly higher RNA/DNA ratios (maximum RNA/DNA across the three cell types). Dots signify mean conservation per bin. Numbers in parentheses show number of sequences per bin. b. Expression fold-change of overlapping pairs of sequences. Pearson's *r* and *P*-value are presented. c. Expression fold-change of lentiMPRA vs luciferase assay. Each pair of points connected by a vertical line represents two replicates in the luciferase assay. Each triplet of points connected by a horizontal line represents three lentiMPRA replicates. Pearson's *r* and *P*-value are presented.

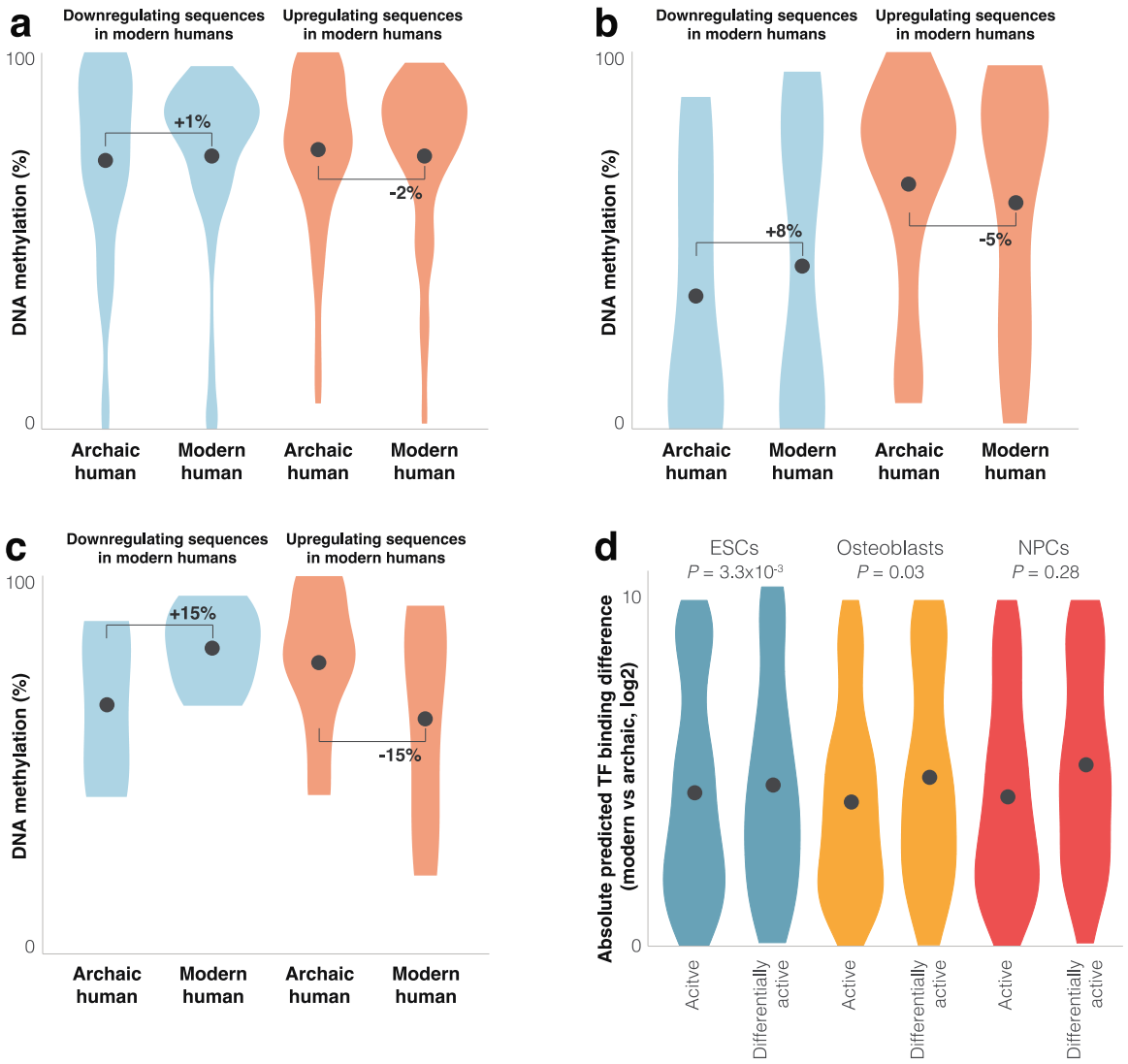

Figure S2.4: Differential activity is associated with differential DNA methylation and TF binding

a-c Violin plots of DNA methylation levels in modern and archaic human bone methylation samples, for differentially active (a), promoter differentially active (b), and CpG-poor promoter differentially active (c) sequences in osteoblasts. Promoter sequences are sequences between 5 kb upstream to 1 kb downstream of a TSS. CpG-poor promoter sequences were defined as the bottom 50% promoter sequences. d. Violin plots of absolute predicted TF binding score difference between modern and archaic sequences. Points show mean.

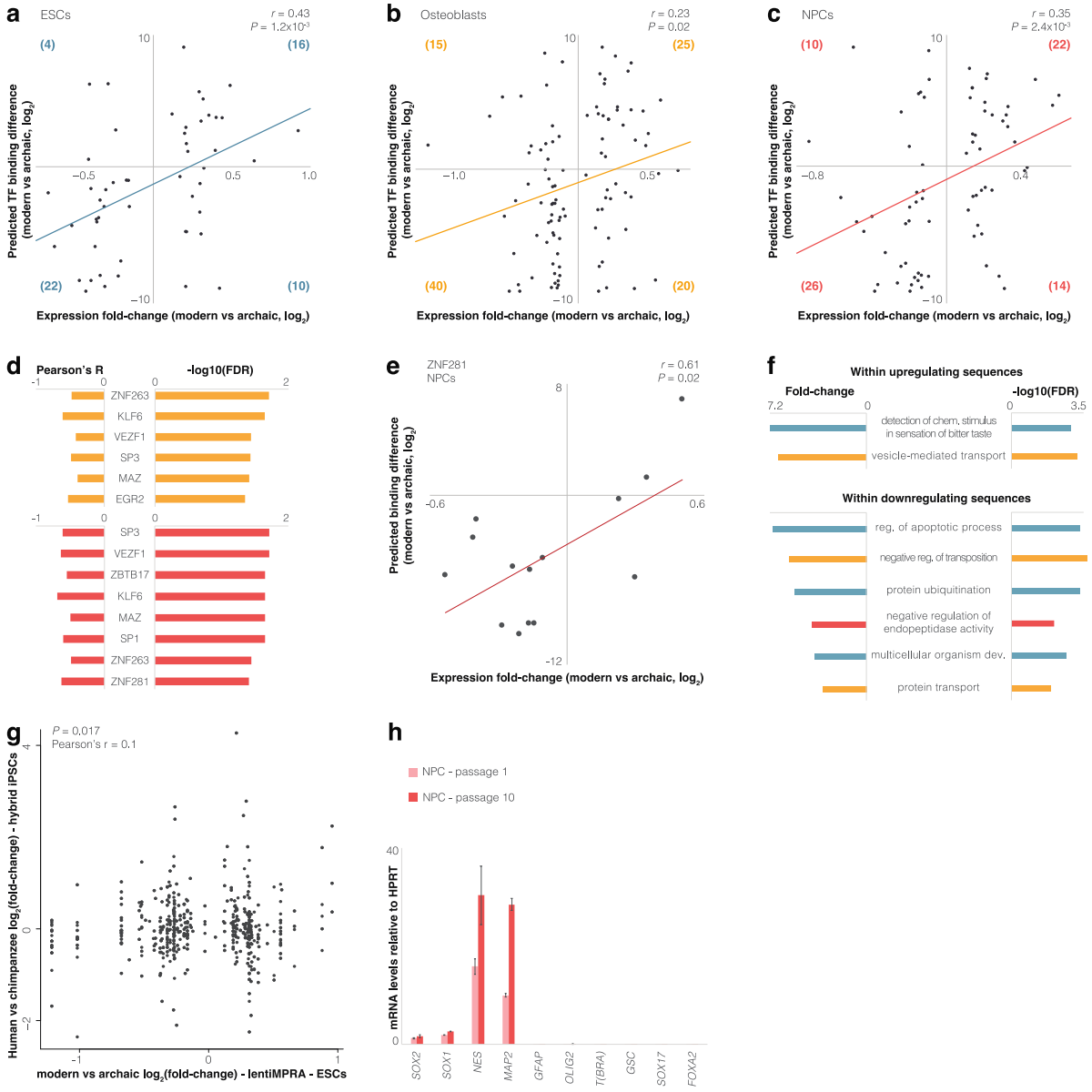

Figure S2.5: Predicted TF binding is correlated with differential activity

a-c. Expression fold-change vs predicted TF binding fold-change for each sequence. Positive scores represent increased binding in the modern sequence. Parentheses show number of points in each quadrant with a score difference > 0. d. Pearson's correlation between differential expression and predicted differential binding affinity. Only significant TFs (FDR <= 0.05, Supplementary File 3) are shown for osteoblasts (yellow) and NPCs (red). e. Expression fold-change vs predicted TF binding fold-change for ZNF281 in NPCs. Pearson's *r* and *P*-value are shown. f. Enriched Gene Ontology terms for ESCs (blue), osteoblasts (yellow) and NPCs (red). g. Expression fold-change of differentially active sequences compared to the *cis*-regulatory expression fold-change between human and chimpanzee of genes associated with these sequences. *cis*-regulatory expression changes were taken from hybrid human-

chimpanzee induced pluripotent stem cells (iPSCs)[65]. h. RT-qPCR validation of NPCs at passage 1 (pink) and passage 10 (red). Expression levels are normalized to *HPRT* expression.

2.7 REFERENCES

1. Neubauer, S., Hublin, J. J. & Gunz, P. The evolution of modern human brain shape. *Sci. Adv.* (2018). doi:10.1126/sciadv.aao5961
2. Gunz, P. *et al.* Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Curr. Biol.* (2019). doi:10.1016/j.cub.2018.10.065
3. Aiello, L. & Dean, C. *An Introduction to Human Evolutionary Anatomy*. (Elsevier, 2002).
4. Kochiyama, T. *et al.* Reconstructing the Neanderthal brain using computational anatomy. *Sci. Rep.* (2018). doi:10.1038/s41598-018-24331-0
5. Yan, S. M. & McCoy, R. C. Archaic hominin genomics provides a window into gene expression evolution. *Curr. Opin. Genet. Dev.* **62**, 44–49 (2020).
6. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* (1971). doi:10.1086/406830
7. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
8. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Res.* (2014). doi:10.1101/gr.164822.113
9. Fraser, H. B. Gene expression drives local adaptation in humans. *Genome Res.* **23**, 1089–1096 (2013).
10. McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* **168**, 916-927.e12 (2017).
11. Petr, M., Pääbo, S., Kelso, J. & Vernot, B. Limits of long-term selection against Neandertal

introgression. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1814338116

12. Gokhman, D. *et al.* Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun.* **11**, 1189 (2020).

13. Colbran, L. L. *et al.* Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* (2019). doi:10.1038/s41559-019-0996-x

14. Gokhman, D. *et al.* Reconstructing Denisovan Anatomy Using DNA Methylation Maps. *Cell* **179**, 180-192.e10 (2019).

15. Dannemann, M. & Racimo, F. Something old, something borrowed: admixture and adaptation in human evolution. *Curr. Opin. Genet. Dev.* **53**, 1–8 (2018).

16. Weyer, S. & Pääbo, S. Functional analyses of transcription factor binding sites that differ between present-day and archaic humans. *Mol. Biol. Evol.* (2016). doi:10.1093/molbev/msv215

17. Vespasiani, D. M., Jacobs, G. S., Brucato, N., Cox, M. P. & Romero, I. G. Denisovan introgression has shaped the immune system of present-day Papuans. *bioRxiv* 2020.07.09.196444 (2020). doi:10.1101/2020.07.09.196444

18. Grogan, K. E. & Perry, G. H. Studying human and nonhuman primate evolutionary biology with powerful in vitro and in vivo functional genomics tools. *Evolutionary Anthropology* (2020). doi:10.1002/evan.21825

19. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements, Major Drivers of Human Disease. *Annu. Rev. Genomics Hum. Genet.* (2017). doi:10.1146/annurev-genom-091416-035537

20. Gokhman, D. *et al.* Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science* **344**, 523–7 (2014).

21. Gokhman, D., Meshorer, E. & Carmel, L. Epigenetics: It's Getting Old. Past Meets Future in Paleoepigenetics. *Trends Ecol. Evol.* **31**, 290–300 (2016).

22. Barker, H. R., Parkkila, S. & Tolvanen, M. E. E. Evolution is in the details: Regulatory differences in modern human and Neanderthal. *bioRxiv* (2020). doi:doi.org/10.1101/2020.09.04.282749

23. Batyrev, D., Lapid, E., Carmel, L. & Meshorer, E. Predicted Archaic 3D Genome Organization Reveals Genes Related to Head and Spinal Cord Separating Modern from Archaic Humans. *Cells* (2019). doi:10.3390/cells9010048

24. Pedersen, J. S. *et al.* Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* **24**, 454–466 (2014).

25. Silvert, M., Quintana-Murci, L. & Rotival, M. Impact and Evolutionary Determinants of Neanderthal Introgression on Transcriptional and Post-Transcriptional Regulation. *Am. J. Hum. Genet.* (2019). doi:10.1016/j.ajhg.2019.04.016

26. Moriano, J. & Boeckx, C. Modern human changes in regulatory regions implicated in cortical development. *BMC Genomics* **21**, 304 (2020).

27. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* (2015). doi:10.1016/j.ygeno.2015.06.005

28. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* (2016). doi:10.1016/j.cell.2016.04.027

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

30. Ryu, H. *et al.* Massively parallel dissection of human accelerated regions in human and chimpanzee neural progenitors. *bioRxiv* (2018). doi:10.1101/256313

31. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc. Natl. Acad. Sci. U. S. A.* (2021). doi:10.1073/pnas.2007049118

32. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* (2020). doi:10.1038/s41596-020-0333-5

33. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–9 (2014).

34. Prüfer, K. *et al.* A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science (80-.).* **358**, 655–658 (2017).

35. Mafessoni, F. *et al.* A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci.* (2020). doi:10.1073/pnas.2004944117

36. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–6 (2012).

37. De Manuel, M. *et al.* Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science (80-.).* (2016). doi:10.1126/science.aag2602

38. Auton, A. *et al.* A global reference for human genetic variation. *Nature* (2015). doi:10.1038/nature15393

39. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* (2001). doi:10.1093/nar/29.1.308

40. Ashuach, T. *et al.* MPRAnalyze: Statistical framework for massively parallel reporter

assays. *Genome Biol.* (2019). doi:10.1186/s13059-019-1787-z

41. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods* (2012). doi:10.1038/nmeth.1906

42. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

43. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* (2017). doi:10.1101/gr.212092.116

44. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* (2020). doi:10.1038/s41592-020-0965-y

45. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* (2019). doi:10.1038/s41467-019-11526-w

46. Wang, M., Zhao, Y. & Zhang, B. Efficient Test and Visualization of Multi-Set Intersections. *Sci. Rep.* (2015). doi:10.1038/srep16923

47. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).

48. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–92 (2012).

49. Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics* (2020). doi:10.1016/j.tig.2019.11.004

50. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* (2015). doi:10.1016/j.cell.2015.01.006

51. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* (2009). doi:10.1038/nature08514

52. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* (2011). doi:10.1038/nature10716

53. Schlesinger, F., Smith, A. D., Gingeras, T. R., Hannon, G. J. & Hodges, E. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome Res.* (2013). doi:10.1101/gr.157271.113

54. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr064

55. Suske, G. NF-Y and SP transcription factors — New insights in a long-standing liaison. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (2017). doi:10.1016/j.bbagrm.2016.08.011

56. Frey-Jakobs, S. *et al.* ZNF341 controls STAT3 expression and thereby immunocompetence. *Sci. Immunol.* (2018). doi:10.1126/sciimmunol.aat4941

57. Bruderer, M., Alini, M. & Stoddart, M. J. Role of HOXA9 and VEZF1 in endothelial biology. *Journal of Vascular Research* (2013). doi:10.1159/000353287

58. Frietze, S., Lan, X., Jin, V. X. & Farnham, P. J. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J. Biol. Chem.* (2010). doi:10.1074/jbc.M109.063032

59. Song, J. *et al.* Transcriptional regulation by zinc-finger proteins Sp1 and MAZ involves

interactions with the same cis-elements. *International journal of molecular medicine* (2003). doi:10.3892/ijmm.11.5.547

60. Zhu, C., Chen, G., Zhao, Y., Gao, X. M. & Wang, J. Regulation of the development and function of B cells by ZBTB transcription factors. *Frontiers in Immunology* (2018). doi:10.3389/fimmu.2018.00580

61. Ji, W., Mu, Q., Liu, X. Y., Cao, X. C. & Yu, Y. ZNF281-miR-543 Feedback Loop Regulates Transforming Growth Factor-β-Induced Breast Cancer Metastasis. *Mol. Ther. - Nucleic Acids* (2020). doi:10.1016/j.omtn.2020.05.020

62. Morita, K. *et al.* Emerging roles of Egr2 and Egr3 in the control of systemic autoimmunity. *Rheumatol. (United Kingdom)* (2016). doi:10.1093/rheumatology/kew342

63. Syafruddin, S. E., Mohtar, M. A., Nazarie, W. F. W. M. & Low, T. Y. Two sides of the same coin: The roles of KLF6 in physiology and pathophysiology. *Biomolecules* (2020). doi:10.3390/biom10101378

64. Pieraccioli, M. *et al.* ZNF281 inhibits neuronal differentiation and is a prognostic marker for neuroblastoma. *Proc. Natl. Acad. Sci. U. S. A.* (2018). doi:10.1073/pnas.1801435115

65. Gokhman, D. *et al.* Human-chimpanzee fused cells reveal cis-regulation underlying skeletal evolution. *Nat. Genet.* in press (2021).

66. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* (2019). doi:10.1016/j.cell.2018.11.029

67. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* (2019). doi:10.1038/s41588-019-0538-0

68. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* (2020). doi:10.1038/s41576-019-0209-0

69. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0494-8

70. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford).* (2017). doi:10.1093/database/bax028

71. Gokhman, D. *et al.* Gene ORGANizer: Linking genes to the organs they affect. *Nucleic Acids Res.* **45**, W138–W145 (2017).

72. Köhler, S. *et al.* The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42**, (2014).

73. Zarate, Y. A. & Fish, J. L. SATB2-associated syndrome: Mechanisms, phenotype, and practical recommendations. *American Journal of Medical Genetics, Part A* (2017). doi:10.1002/ajmg.a.38022

74. Liu, S. H. *et al.* A novel antisense long non-coding RNA SATB2-AS1 overexpresses in osteosarcoma and increases cell proliferation and growth. *Mol. Cell. Biochem.* (2017). doi:10.1007/s11010-017-2953-9

75. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

76. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1120

77. Claes, P. *et al.* Modeling 3D Facial Shape from DNA. *PLoS Genet.* **10**, e1004224 (2014).

78. Zarate, Y. A., Kaylor, J. & Fish, J. SATB2-Associated Syndrome. in (eds. Adam, M. P. et al.) (1993).

79. Gigek, C. O. *et al.* A molecular model for neurodevelopmental disorders. *Transl. Psychiatry* (2015). doi:10.1038/tp.2015.56

80. Qian, Y. *et al.* Paternal Low-Level Mosaicism-Caused SATB2-Associated Syndrome. *Front. Genet.* **10**, 630 (2019).

81. Li, Y. *et al.* Satb2 Ablation Impairs Hippocampus-Based Long-Term Spatial Memory and Short-Term Working Memory and Immediate Early Genes (IEGs)-Mediated Hippocampal Synaptic Plasticity. *Mol. Neurobiol.* (2017). doi:10.1007/s12035-017-0531-5

82. Zhang, Q., Huang, Y., Zhang, L., Ding, Y. Q. & Song, N. N. Loss of satb2 in the cortex and hippocampus leads to abnormal behaviors in mice. *Front. Mol. Neurosci.* (2019). doi:10.3389/fnmol.2019.00033

83. Dobreva, G. *et al.* SATB2 Is a Multifunctional Determinant of Craniofacial Patterning and Osteoblast Differentiation. *Cell* (2006). doi:10.1016/j.cell.2006.05.012

84. Mattioli, K. *et al.* Cis and trans effects differentially contribute to the evolution of promoters and enhancers. *Genome Biol.* **21**, 210 (2020).

85. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science (80-.).* (2008). doi:10.1126/science.1159974

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

87. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annual Review of Cell and Developmental Biology* (2002). doi:10.1146/annurev.cellbio.18.020402.140619

88. Telis, N., Aguilar, R. & Harris, K. Selection against archaic hominin genetic variation in regulatory regions. *Nat Ecol Evol* (2020). doi:10.1038/s41559-020-01284-0

89. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**, 68–84 (2015).

90. Lieberman, P. The Evolution of Human Speech: Its Anatomical and Neural Bases. *Curr. Anthropol.* **48**, 39–66 (2007).

91. Mariën, P. *et al.* Consensus paper: Language and the cerebellum: An ongoing enigma. *Cerebellum* (2014). doi:10.1007/s12311-013-0540-5

92. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *RepeatMasker Open-3.0* (1996).

93. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell* (2019). doi:10.1016/j.stem.2019.09.010

94. Hojo, H., Ohba, S., He, X., Lai, L. P. & McMahon, A. P. Sp7/Osterix Is Restricted to Bone-Forming Vertebrates where It Acts as a Dlx Co-factor in Osteoblast Specification. *Dev. Cell* (2016). doi:10.1016/j.devcel.2016.04.002

95. Meyer, M. B., Benkusky, N. A., Onal, M. & Pike, J. W. Selective regulation of Mmp13 by 1,25(OH)2D3, PTH, and Osterix through distal enhancers. *Journal of Steroid Biochemistry and Molecular Biology* (2016). doi:10.1016/j.jsbmb.2015.09.001

96. Khalid, A. B. *et al.* GATA4 represses RANKL in osteoblasts via multiple long-range

enhancers to regulate osteoclast differentiation. *Bone* (2018). doi:10.1016/j.bone.2018.07.014

97. Khalid, A. B. *et al.* GATA4 Directly Regulates Runx2 Expression and Osteoblast Differentiation . *JBMR Plus* (2018). doi:10.1002/jbm4.10027

98. Loots, G. G. *et al.* Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res.* (2005). doi:10.1101/gr.3437105

99. Fukami, M., Kato, F., Tajima, T., Yokoya, S. & Ogata, T. Transactivation function of an ~800-bp evolutionarily conserved sequence at the SHOX 3' region: Implication for the downstream enhancer. *American Journal of Human Genetics* (2006). doi:10.1086/499254

100. Kawane, T. *et al.* Runx2 is required for the proliferation of osteoblast progenitors and induces proliferation by regulating Fgfr2 and Fgfr3. *Sci. Rep.* (2018). doi:10.1038/s41598-018-31853-0

101. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkl822

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

103. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* (2014). doi:10.1101/gr.173518.114

104. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* (2012). doi:10.1101/gr.135350.111

105. Hernando-Herraez, I. *et al.* Dynamics of DNA Methylation in Recent Human and Great Ape Evolution. *PLOS Genet* **9**, e1003763 (2013).

106. Moriarity, B. S. *et al.* A Sleeping Beauty forward genetic screen identifies new genes and pathways driving osteosarcoma development and metastasis. *Nat. Genet.* **47**, 615–24 (2015).

107. Lu, L. *et al.* Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. *Mol. Cell* (2020). doi:10.1016/j.molcel.2020.06.007

108. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky955

CHAPTER 3: Identification and exploration of human specific gained and lost CTCF sites

3.1 SUMMARY

Gene regulation is predicted to be the driving force behind modern human speciation and the primary reason for divergent phenotypes between humans and chimpanzees. Exactly how 3D genome structure functions to facilitate gene regulation is still unknown and little is understood about how this form of regulation has influenced modern human evolution. CTCF is a highly conserved ubiquitous transcription factor that can function as an activator, repressor, insulator, and has a well-documented role in genome looping. In this study, we focus on identifying CTCF sites that are uniquely gained or lost in humans. We've identified CTCF sites gained and lost in humans as compared to other primates and identified a subset of these sites that have been gained compared to extinct archaic species, Neanderthal and Denisovan. Using our pipeline, we found many more human specific gained sites (2,230) than lost sites (24) as compared to non-human primates, and only 24 CTCF sites gained in humans as compared to Neanderthal and Denisovan. Notably, these results are likely impacted by using many more human CTCF datasets compared to non-human primate in our analysis. We find an enrichment of our identified human gained sites at TAD boundaries, and for the genes within TADs that harbor a human gained CTCF site there is an enrichment for genes related to cognition and chondrocyte differentiation. Additionally, I deleted one CTCF human gained site of particular interest near the *ZNF589* gene which is expressed in the brain and has been implicated in intellectual disability. I created stable neuron cell lines with this single CTCF site deleted and measured the effect the deletion had on gene expression in the region. Although we did not see a change in gene expression for nearby genes, we cannot rule out the importance of other human specific

CTCF sites in gene regulation. This study provides a list of uniquely gained and lost CTCF sites in modern humans which provides candidates for future experiments to determine if and how CTCF sites have influenced gene regulation in modern humans.

3.2 INTRODUCTION

Ninety-eight percent of the human genome is made up of noncoding sequences. Within these sequences reside transcriptional gene regulatory elements that instruct genes when, where, and at what levels to transcribe. These elements include promoters, enhancers, and others. There is a growing body of work suggesting that these sequences could have a major impact on human evolution. For example, over 30% of human accelerated regions (HARs), which are evolutionarily conserved sequences that have undergone rapid changes in humans, are predicted to be developmental enhancers[1], and are implicated in schizophrenia[2] and autism spectrum disorder[3]. With the availability of genomic sequences from extinct hominin species, Neanderthal[4] and Denisovan[5], our ability to further pinpoint modern human regulatory changes has increased. For example, utilizing comparisons to archaic genomes in combination with massively parallel reporter assay (MPRA), a subset of single nucleotide changes specific to the modern human lineage have been shown to functionally alter regulatory activity and could play a role in modern human specific gene regulation[6]. These studies and others have aided in our understanding of how gene regulatory elements have shaped modern humans, but our knowledge of how other regulatory factors contributed to modern human evolution is still lacking.

It is generally believed that long range chromatin interactions facilitate gene transcription by bringing enhancers closer to their target promoters. However, we know little about how this 3D genome architecture is associated with modern human evolution. The genome is partitioned into topologically associated domains (TADs), typically hundreds of kilobases in size, wherein

chromosomal interactions are enriched within domains rather than between domains. An important feature of TADs are their boundaries. These insulating genomic sequences flank a given TAD and reduce inter-TAD contacts. Although most TAD boundaries are conserved, a subset have been found to be species specific[7]. In a recent study, researchers found that human specific TADs have defining features that imply a functional role in gene regulation that could represent a mechanism by which modern humans have developed unique gene regulatory patterns[8]. Furthermore, this study found that human specific genome loops are enriched for enhancer-enhancer interactions which are predicted to be part of multi-enhancer regulatory networks active in the developing brain. This study suggests that chromatin structures could affect the evolution of human specific gene regulation.

CCCTC-binding factor (CTCF) is a highly conserved architectural protein[9,10] known to facilitate various 3D chromatin structures. For example, CTCF binding enrichment has been observed at TAD boundaries where CTCF is thought to function as an insulator[11]. In addition to its insulator function, CTCF can also promote gene expression by facilitating long-range enhancer-promoter interactions through genome looping[12]. CTCF is thought to help generate chromatin loops via the loop extrusion model in which the circular cohesion complex extrudes DNA until stalled by two convergently oriented and bound CTCF motifs[13]. This looping brings sequences that are far apart in linear genomic space into closer proximity, allowing them to interact. Whether CTCF functions to hinder or promote genomic interactions is highly context dependent, but nonetheless CTCF remains a key component of 3D genome architecture.

Alteration of CTCF binding is known to lead to various phenotypes and thus could be important for modern human specific evolution. A striking example of this is a deletion of three CTCF sites, which reside about one megabase away from the gene Sonic Hedgehog (*SHH*), that causes altered interactions between the primary *SHH* limb enhancer and its promoter resulting in acheiropodia (limb truncation) in humans[14]. Although there are several examples of how perturbations to CTCF sites or clusters can result in aberrant gene expression[15–17], linking CTCF variation to modern human specific gene expression patterns remains limited. Given the critical role of CTCF in 3D genome architecture, CTCF binding is an intriguing target to study in the context of human specific gene regulation.

In this study, we capitalize on the availability of archaic genomic datasets to identify modern human specific CTCF binding sites that could have led to distinctly human regulatory phenotypes. Our work expands on previously performed comparisons of CTCF sites across mammalian species[18,19], and recent work using machine learning to reconstruct archaic 3D genomes[20]. We performed CTCF CUT&Tag experiments on phenotypically relevant cell types from humans and non-human primates which, combined with publicly available CTCF ChIP-seq datasets and human specific variant lists, allowed us to identify CTCF binding sites that are uniquely gained and lost in the human lineage. We overlayed these sequences with archaic human variants allowing us to further identify differential CTCF binding sites between modern and archaic humans. Our pipeline produced a list of gained and lost CTCF sites in humans compared to non-human primates, and separately, to extinct archaic humans. We analyzed where these CTCF sites are in the genome and found enrichment for human specific gained

sites at TAD boundaries. Additionally, we performed a gene ontology analysis for all human specific gained and lost CTCF sites that were within TADs with protein coding sequences. We found the genes within these TADs were enriched for terms related to cognition and chondrocyte differentiation. Finally, I created a human neuron cell line in which I deleted one human specific gained CTCF site and measured the effect of the deletion on gene expression in the locus. Although we did not see a change in gene expression in these cells, we cannot rule out the possibility that other human gained CTCF sites are crucial for proper gene regulation. Together, our study identifies CTCF binding sites gained and lost in humans and provides a framework to further explore the role of CTCF in human specific gene regulation.

3.3 RESULTS

3.3.1 Identification of human specific gained and lost CTCF sites

We generated a pipeline to identify CTCF sites that have been gained or lost specifically in the human lineage. We used a combination of CTCF binding data from humans, chimpanzee, and orangutan as input for our pipeline. For humans, we used coordinates of CTCF binding sites from Search Candidate *cis*-Regulatory Elements by ENCODE (SCREEN-ENCODE)[21]. This dataset encompasses a list of candidate *cis*-regulatory elements (cCREs) identified through the ENCODE consortium that have high CTCF signal, as determined by ChIP-seq data, and includes information from over 1,500 human cell types. This dataset and an explanation of how it was analyzed can be found at screen.encodeproject.org. Although this dataset does not represent all CTCF sites in the human genome, it does prioritize CTCF sites that are likely functional, in that these are sites categorized as cCREs that are also bound by CTCF. This is an important distinction when considering our identified human specific lost CTCF sites, discussed more below.

We compared the human CTCF sites to publicly available CTCF ChIP-seq data from chimpanzee and orangutan lymphoblast cells (see Methods) to determine which CTCF sites were specific to human (human specific gained sites), and which were only found in chimpanzee and orangutan samples (human specific lost sites), see Methods. To increase our confidence in these candidate CTCF sites, we used a human derived variant list (single nucleotide variants and indels specific to humans since the human-chimpanzee split) and overlayed them with our preliminary human specific gained and lost CTCF sites, further curating the list to only include CTCF sites that

contained a human derived variant. Notably, because we used the SCREEN-ENCODE data as our input human CTCF binding data, we cannot be completely confident about the identified human specific CTCF lost sites since these sites may be present in humans just not overlapping a cCRE.

To add specificity to our pipeline, we incorporated phenotypically relevant cell type information from humans, chimpanzee, and orangutan. To this end, we performed CTCF CUT&Tag experiments on human and chimpanzee neural progenitor cells (NPCs) and on NPCs dissociated from week five brain organoids from human, chimpanzee, and orangutan (referred to as *organoid sample* for simplicity). A schematic of this complete pipeline can be seen in **Figure 3.1**. Combining both NPC replicates, we identified 28,938 CTCF peaks in human and 35,842 peaks in chimpanzee. Due to scarcity of input material, only one replicate could be performed for each organoid sample. For the organoid samples, we found 20,135 CTCF peaks in human, 24,584 in chimpanzee, and 17,232 in orangutan. These counts are small in comparison to the raw number of peaks generally observed in a ChIP-seq experiment. However, we believe our calls represent the most robust CTCF peaks for these cells. Although we have not done a direct comparison, others have done side-by-side comparisons and found that CUT&Tag experiments produce a subset of peaks representing the strongest hits from an equivalent ChIP-seq experiment[22]. **Figure 3.2a** shows how the CUT&Tag data compares across samples, showing an expected pattern of conservation. We used these CUT&Tag datasets to further filter the human specific gained and lost CTCF candidates (see Methods). **Figure 3.2b** shows a combination of all peaks from all datasets broken down by species, again showing an expected pattern of conservation.

In total we found significantly more human specific gained sites (2,230) than human specific lost sites (24), although these numbers are likely skewed by the fact that we have much more CTCF binding data for human than for non-human primates (**Figure 3.2b**).

3.3.2 Identification of recent human specific gained CTCF sites

Since we were particularly interested in modern human specific evolution, we incorporated archaic human data into our analysis. As input, we used the previously mentioned lists of human specific gained and lost CTCF sites then overlayed them with a variant list containing human derived variants specific to modern humans since the split from Neanderthal and Denisovan. This allowed us to categorize which CTCF sites were recent human specific gained or lost. Not surprisingly, this greatly reduced the candidate CTCF sites in each category. In total, we only found 24 recent human specific gained CTCF sites and zero recent human specific lost sites. A breakdown of each category can be found in **Table 3.1**.

3.3.3 Position Weight Matrix (PWM) analysis

We wanted to test if the variant or variants within a candidate CTCF site were predicted to cause a change in CTCF binding and if those predictions matched how we had categorized the sites (either human specific gain, human specific lost, or recent human specific gain). To this end, we used the tool FABIAN-variant[23] to perform a position weight matrix (PWM) analysis comparing the human variant to either the great ape or archaic human variant (see Methods), depending on the category. We found that over 50% (1,162 of 2,230) of the human specific gained CTCF sites predicted an increase in binding affinity with the human variant as compared

to the great ape variant, thus agreeing with the categorization as human specific gained CTCF sites. The remaining CTCF sites were either predicted to have no change (13) or decreased (1,055) binding affinity. 22 of the 24 human specific lost sites were predicted to have decreased binding affinity with the human variant versus the great ape variant, one was predicted to have no change, and the one was predicted to have increased binding affinity. Finally, for the recent human specific gained CTCF sites, nearly 60% (14 of 24) were predicted to have higher binding affinity with the modern human variant as compared to the archaic humans. The other 10 sites were predicted to have a decreased binding affinity with the modern human variant.

3.3.4 Locations of candidate CTCF sites

The location of a given CTCF site can have a major impact on how the site functions in the region. CTCF sites within TAD boundaries often function as insulators where they create genome loops that help define the boundary[13]. CTCF sites within TADs can function differently, for example, CTCF binding sites upstream of the *Shh* gene are capable of driving gene expression[24]. We therefore thought it was important to analyze the location of the identified candidate CTCF sites in each category. CTCF sites are known to be enriched at TAD boundaries[11], so we first looked there. We found that none of the human specific lost sites were located within TAD boundaries, but there was an enrichment of human gained CTCF sites at TAD boundaries (53 of 2,230; permutation test, p-value<0.0001), see Methods. TAD boundary calls used in this analysis is unpublished data generously gifted from Dr. Lucia Carbone. In brief, Hi-C was performed on human EBV-transformed lymphoblastoid B-cell lines (LCL) and boundary calls were made using the hicFindTADs command from HiCExplorer. CTCF

sites not located in a TAD boundary can also be influencing gene regulation, albeit through different mechanisms. The target gene of a CTCF site is generally restricted to the genes within the same TAD, therefore, we identified which CTCF sites in each category are within TADs that have protein coding sequences. We found 1,514 human specific gained sites, 16 human specific lost sites, and 16 recent human gained sites are within TADs containing genes. All these categories and numbers are summarized in **Table 3.1**.

3.3.5 Enrichment of genes near candidate CTCF sites

Considering the CTCF sites that share TADs with protein coding sequences, we did a gene enrichment analysis using the Gene Ontology Enrichment Analysis and Visualization Tool (GOrilla)[25,26]. For this analysis, we used all genes within a given TAD, not just the closest to the CTCF site (see Methods). For human specific sites, we found the highest enrichment for *keratinization* (Exact mHG p-value= 1.75e-5) and *embryonic skeletal system morphogenesis* (Exact mHG p-value= 1.09e-4). Interestingly, we also saw enrichment for *cognition* (Exact mHG p-value= 8.04e-4). For human specific lost sites, the top enriched terms were *negative regulation of extrinsic apoptotic signaling pathway* (Exact mHG p-value= 3.31e-5) and *regulation of lymphocyte apoptotic process* (Exact mHG p-value= 3.75e5). Finally, for the recent human specific gained category, the top enriched terms were *chondrocyte differentiation* (Exact mHG p-value: 1e-4) and proline *transmembrane transport* (Exact mHG p-value= 1.42e4). All GOrilla enrichment results can be found in **Table 3.2**. Specifically, the enrichment for *cognition* and *chondrocyte differentiation* are of particular interest given our interest in modern human evolution.

3.3.6 Deletion of a single CTCF site in the *ZNF589* TAD shows no change in gene expression

To determine if a single human gained CTCF site is capable of changing gene expression I chose one candidate CTCF site that was gained specifically in the human lineage, including extinct archaic humans. Importantly, this CTCF site was identified as a human specific gained site in an earlier iteration of our pipeline and did not come up in our final analysis (likely because we used different human CTCF binding data as input). This CTCF site is located downstream of the *ZNF589* gene and contains three different sequence changes between humans and non-human primates (**Figure 3.3**), which made it particularly interesting. Additionally, *ZNF589* is expressed in the brain and has been implicated in intellectual disability[27], which made it an even more intriguing candidate from a biological perspective. I created stable induced pluripotent stem cell (iPSC) lines in which this particular CTCF site was deleted (see Methods). I chose to do these experiments in WTC11-Ngn2 cells which are human iPSCs that have been engineered with an inducible mouse neurogenin 2 (*Ngn2*) transgene that allows for a simplified two step neuron differentiation protocol[28]. Given the association of *ZNF589* with cognitive function, we believed testing subsequent phenotypic analysis in neurons was an appropriate choice. I created one homozygous and two heterozygous knockout iPSC lines that were then differentiated into neurons for the following experiments (**Figure 3.4**).

I first determined if there were any changes in gene expression caused by the deletion. I performed three replicates of RNA-sequencing on each cell line alongside three replicates of wildtype cells for comparison. When comparing the wildtype cells to the homozygous line, I found no significant change in gene expression for the genes within the TAD that harbored the

deletion (**Figure 3.5**). All genes within the TAD showed marginal expression decrease (excluding *ZNF589*, which showed no change), but none were significant. We did see other genes that were up- or down-regulated in the deletion line versus the wildtype, but it is not immediately clear how these changes were facilitated.

To confirm these RNA-sequencing results, I performed qPCR analysis on the homozygous deletion line for all genes within the TAD and compared to wildtype cells (**Figure 3.6**). All genes, except for *NME6*, showed no significant change in expression. *NME6* had slightly increased expression (1.2-fold over wildtype) that was barely significant with a p-value of 0.0451, two-tailed t-test. Notably, *NME6* has been implicated in stem cell renewal[29], which could be an interesting find given the potential human specificity of the CTCF site. However, since the qPCR results do not agree with the RNA-sequencing results, additional work will be required to be certain if there are any meaningful changes to gene expression within this TAD caused by the deletion.

To analyze the heterozygous lines, we utilized the fact that the WTC11-Ngn2 genome was previously phased[30], which allowed us to determine allele specific expression within the heterozygous cell lines I created (see Methods). Using this method requires phased SNPs to be present within exons of the genes of interest. All genes except two (*NME6* and *PLXNB1*) had SNPs within their exons (**Figure 3.3**) and gene expression could therefore be determined in an allele specific manner. This method of measuring gene expression gave us the same insignificant results as the RNA-sequencing results from the homozygous line. Combining our

RNA-sequencing and qPCR results, there appears to be no meaningful change in gene expression caused by the deletion of this particular CTCF site.

3.4 DISCUSSION

Our work shows high conservation of CTCF sites across human, chimpanzee, and orangutan in on our CUT&Tag data (**Figure 3.2a**) and this conservation is similarly observed when it is incorporation with the publicly sourced ChIP-seq data (**Figure 3.2b**). High conservation of CTCF sites is also suggested by the small number of CTCF sites we've identified for each CTCF category (**Table 3.1**). This trend is corroborated by other studies that see a similar level of conservation of CTCF sites across mammalian species[18,19]. TAD boundaries are also conserved across evolution[7] and CTCF plays a major role in insulation at TAD boundaries[31]. This relationship suggests there could be evolutionary pressure to maintain CTCF sites in order to ultimately maintain TAD structures. Our results analyzing the location of candidate CTCF sites also support this theory; human specific lost sites are not enriched at TAD boundaries, but the human specific gained sites are. Additionally, it could be theorized that new CTCF sites within TADs may be selected against as they could interrupt crucial long-range gene regulation within the TAD. Therefore, new CTCF sites would be best tolerated next to already existing CTCF sites at TAD boundaries.

Although there is evidence for a biological reason for conservation, we cannot rule out the possibility that our results have been skewed for technical reasons. Notably, our human CTCF dataset is overpowered compared to the great ape datasets, in terms of number of CTCF datasets and peaks. This pipeline would benefit from the incorporation of CTCF binding information in additional cell types from chimpanzee and orangutan to increase confidence in categorized CTCF sites. Having more peaks for these great apes could theoretically increase the

number of observed human and recent human specific lost sites and decrease the number of false positives in the human specific gained categories. Using the PWM analysis to further filter CTCF sites is another way to reduce the number of false positives in the human gained category.

Although the conservation of CTCF is highly evident in our data and supported by the literature, the effect of the human specific gained and lost CTCF sites on gene expression is more elusive. CTCF has been shown to be important, in some scenarios, for maintaining proper gene expression[14], which could serve as another reason for the observed conservation of CTCF sites. However, this could also be a mechanism by which new gene expression patterns have emerged in humans. Our gene ontology (GO) analysis reveals some interesting GO terms that implicate these sites in crucial gene regulatory networks. Advanced cognitive abilities have long been considered a human specific phenotype, so it is very promising that we also see this term enriched in our human specific gained category. It has also been suggested, given the differences in skeletal morphology between modern and archaic humans[31], that cartilage formation is likely diverged as well. Again, our results of chondrocyte differentiation being a top enriched term for recent human specific gained sites proves promising. With additional follow up experiments, it will be possible to determine if these particular CTCF sites are important for human specific phenotypes.

The CTCF deletion cell line I created showed no change in gene expression for the genes within the TAD. This, however, does not entirely rule out the possibility that this CTCF site is involved

in gene expression in the region, as it may be functioning in combination with other nearby elements. The literature suggests that the absence of multiple CTCF sites can alter gene expression. At the *SHH* locus, the deletion of three CTCF sites causes acheiropodia (extreme limb and digit malformations) in humans[14]. This example suggests there could be an observed change in expression if our target CTCF site were deleted in combination with others nearby. As **Figure 3.3** shows, there are several additional CTCF sites in the region, so determining which sites are likely interacting with our target CTCF site would be crucial for determining which additional regions to delete to test this hypothesis. The activity of CTCF sites can also be skewed by more drastic perturbations. In the *Sox9-Kcnj2* locus, altered expression was only observed when all CTCF sites within the TAD and its boundaries were deleted[33]. Therefore, even deleting additional CTCF sites in this region may not be enough to show a gene expression phenotype. Additionally, it is possible that nearby CTCF sites may be compensating for the loss of the deleted site thus maintaining proper gene expression. CTCF CUT&Tag experiments in this deletion line would determine if CTCF binding is completely lost in this region, or if compensation is occurring.

Although none of the genes within the TAD showed a notable change in gene expression, the RNA-sequencing results did show a change in expression of other genes outside of this TAD in the homozygous deletion line. The top three up-regulated and down-regulated genes have been noted on **Figure 3.5**. Additional work will be required to determine if these genes are relevant or reflective of a real change in expression caused by the CTCF deletion in neurons. Due to gene expression being highly cell-type dependent, it is possible that this perturbation

may be more impactful in a different cell type. Analysis of gene expression data in different cells would be informative for determining if a different cell type would be more apt for this experiment.

Allele specific expression analysis for the heterozygous deletion lines also did not show a change in gene expression caused by the CTCF deletion. However, this analysis was underpowered since only one SNP was used per gene. In the future, multiple phased SNPs per gene should be used (where possible) in addition to targeted sequencing after RT-qPCR. This would provide more power and ultimately higher confidence in the analysis.

Most notably, the CTCF site we targeted was identified as a human specific gained site in an earlier iteration of our pipeline but did not appear in the category in our final pipeline. This means our target CTCF site may not be a true human specific gained site. Additional experiments will be required to determine if the newly identified human gained CTCF sites are important for gene regulation in their respective loci.

In the future, it would be best to refine this pipeline (by incorporating more ChIP-seq data, as mentioned previously) and to perform additional filtering to prioritize CTCF candidates for additional follow-up experiments. Also as mentioned previously, an ideal first step would be to use the PWM scores to determine which sites contain variants predicted to change binding in such a way that agrees with our categorization and to prioritize those with the highest predicted change. It would also be beneficial to incorporate other datasets to better link a given

CTCF site to its target gene and to predict if an effect might be observed by its deletion. Hi-C data shows the interactions between genome sequences and can be used to link a particular CTCF site to a potential target gene. This could be done using human Hi-C data (in the case of human specific gained sites) and in chimpanzee Hi-C data (in the case of human specific lost sites) to link our categorized CTCF sites to genes. It is also possible to link a CTCF site to a target gene if it is in an already characterized promoter or enhancer.

Once the sites are linked to a potential target gene, assessing if there is species-specific gene expression for a given gene will be crucial in selecting a strong candidate for follow-up. It would also be important to determine if the candidate gene or genes are phenotypically interesting in terms of human evolution. Well characterized phenotypic divergence inferred from skeletal differences between modern and archaic humans include skeletal[31] and brain structures[33,34] and would be a good filter for phenotypic relevance. Additionally, our previous work using massively parallel reporter assay (MPRA) to assess the regulatory potential of single fixed or nearly fixed derived changes in the modern human genome inferred soft tissue phenotypic differences between modern humans and extinct archaic humans including heart, digestive tract, and renal function[6]. This is an additional resource to utilize when determining if a gene function is phenotypically relevant enough to warrant additional experiments. Finally, identifying what tissue a gene is expressed in will be crucial in determining which cell type to perform additional experiments in.

Once top candidates are determined, a superior experiment would be to use CRISPR Prime editing[35] to create a human cell line that reflects the archaic version of the sequence. This would be an improvement on my method of deleting a single CTCF site for two reasons. First, making a sequence change would allow for the use of CTCF ChIP-seq to determine if the edited sequence alters CTCF binding at that site, as the pipeline predicts. Second, this technique could be used to investigate both human specific gained and human specific lost sites in a human cell line.

The pipeline and datasets we've created provides a framework for studying human specific gained and lost CTCF sites. Our work has taken an important first step in dissecting how and to what extent CTCF site specificity has functioned in modern human evolution.

3.5 METHODS

CUT&Tag

CTCF CUT&Tag was carried out using the CUT&Tag-IT™ Assay Kit (Active Motif, cat. no. 53160) following the manufacturer's protocol. 100,000-200,000 freshly harvested cells were used as input for each reaction for human and chimpanzee NPCs. Human NPCs were harvested at passage 15 for both replicates, and chimpanzee NPC cells were harvested at passage 27 and 23 for replicate 1 and replicate 2, respectively. Frozen NPCs dissociated from week five organoids were thawed and counted before beginning the protocol. For human and chimpanzee organoid dissociated cells, ~500,000 cells were used per reaction. Due to a limited amount of material, only ~55,000 cells were used per reaction for orangutan organoid dissociated cells. Two CUT&Tag reactions were performed for each sample, one using an IgG antibody (Cell Signaling Technologies, cat. no. 2729S) and the other using a CTCF antibody (Active Motif, cat. no. 61932). A left side bead clean-up was performed after library preparation to remove adapter dimers for the IgG reactions for human and chimpanzee cells dissociated from organoids. Both reactions using IgG and CTCF antibodies for the orangutan organoid dissociated cells required a left side bead clean-up in addition to 10 more cycles of PCR, followed by a 0.9X bead clean-up, in order to obtain sufficient material for sequencing. All samples were sequenced on a HiSeq4000 using PE100 and PE150 for replicate 1 and replicate 2, respectively. Two replicates per sample were performed on the human and chimpanzee NPC samples. Due to lack of material, only one replicate was done for the three samples that were dissociated from organoids. CTCF peaks were called using *Partek® Flow®* software, v10.0, following their ChIP-seq analysis protocol.

Candidate CTCF identification pipeline

Coordinates for likely CTCF bound candidate regulatory elements in humans were downloaded from the screenENCODE database (https://screen.encodeproject.org/) and were lifted over to the human genome build GRCh37 using the liftOver tool from UCSC Genome Browser. Great ape CTCF ChIP-seq datasets were downloaded from the EMBL's European Bioinformatics Institute database for *chimpanzee troglodytes* (chimpanzee) lymphoblast cells (https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1511/E-MTAB-1511.processed.18.zip/do1285_CTCF_LCL_07729upstate_ptr18359_CRI01.fq.sam.bam) and *Pongo pygmaeus pygmaeus* (orangutan) lymphoblast cells (https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-1511/E-MTAB-1511.processed.21.zip/do1256_CTCF_LCL_07729upstate_ppyEB185JC_CRI01.fq.sam.bam). CTCF peaks were called with MACS2, using the suggested parameters for regular peak calling. Peak coordinates for chimpanzee and orangutan were lifted over to the human genome build GRCh37 using the liftOver tool. *BEDtools intersect*, from BEDtools, was used to determine overlap between the human and great ape CTCF data including the novel CTCF Cut&Tag data (production described above). Sites were determined to be either human specific CTCF gained sites, where a peak is present in all human data but none of the great ape data, or human specific CTCF lost sites, where a peak is present in all great ape data but none of the human data. These lists were further refined by overlaying with a variant list encompassing all human derived fixed single nucleotide variants (SNVs) and insertions and deletions (indels) since the human chimpanzee split (https://krishna.gs.washington.edu/download/CADD-development/v1.6/training_data/GRCh37/). An additional filtering was done to determine

recent human gained and lost CTCF sites by overlapping with a variant list encompassing variants derived in humans since the split with Neanderthal and Denisovans (http://ftp.eva.mpg.de/neandertal/altai/catalog/HumanCatalog/).

Position weight matrix (PWM) analysis

CTCF candidates that overlapped a variant were assessed for the likelihood of that variant causing a change in CTCF binding affinity. To this end, FABIAN-variant tool from GeneCascade (https://www.genecascade.org/fabian/) was used (CITE). This tool estimates the impact of a variant on CTCF binding affinity using a PWM-based model separately applied to four TF-binding motif databases (JASPAR2022, SwissRegulon, Jolma2013, and HOCOMOCO11). In this study, we used the mean score from all four models for each candidate CTCF site.

Candidate CTCF site intersection with TAD boundaries

We used *BEDtools intersect* to determine if candidate CTCF sites were within TAD boundaries using coordinates for boundaries generously given to us by collaborators in the Carbone lab at Oregon Health Sciences University (unpublished data). We then determined if there was enrichment of these sites within TAD boundaries by performing a 1000-iteration permutation test, in which the number of overlaps between CTCF-binding regions and TAD boundaries was compared to the overlap of the same boundaries with a random set of regions of the same size and on the same chromosome as the CTCF sites. This process was done for all CTCF categories (human gain, human lost, and recent human gain).

Analysis of candidate CTCF sites near coding sequences

To determined which CTCF sites were within TADs containing protein coding sequences, we intersected TADs called from neural progenitor cells at a 50kb resolution downloaded from TADKB[36] with gene locations (ftp://ftp.ensembl.org/pub/release-96/gff3/homo_sapiens/Homo_sapiens.GRCh38.96.chr.gff3.gz). The coordinates for TADs containing genes were then overlapped with our candidate CTCF sites. All genes within TADs containing candidate CTCF sites were inputted into Gene Ontology enRIchment anaLysis and visuaLizAtion tool (GOrilla) (http://cbl-gorilla.cs.technion.ac.il/). A background set of genes was derived from TADs that do not contain candidate CTCF sites. The output of this query is what is reported in **Table 3.2**.

Creation of CTCF deletion cell lines

For each locus, guide RNAs were designed using the Custom Alt-R® CRISPR-Cas9 guide RNA tool from IDT (https://www.idtdna.com/site/order/designtool/index/CRISPR_CUSTOM). Four total guides were designed, two on each side of the target (**Supplement Table 3.1**). WTC11 cells with a doxycycline-inducible mouse *Ngn2* transgene (WTC11-ngn2[1]) were a generous gift from Li Gan (Gladstone Institute) and were used for these assays. Established WTC11 maintenance protocols were followed when culturing WTC11-ngn2 cells; namely maintaining cells in mTeSR™1 media (STEMCELL Technologies, cat. no. 85850) with daily media changes and the addition of Rock Inhibitor (Selleckchem, cat. no. S1049) when cells were plated. For transfection, cells were seeded at a density of 300,000 cells per 6-well in mTeSR™1 media plus Rock Inhibitor and cultured for one day. Cells were transfected with 800ng of each of the four

sgRNAs, 6250ng of TrueCut Cas9 Protein v2 (Invitrogen, cat. no. A36498), and 500ng of MSCV Puro-SV40:GFP plasmid (Addgene #68483) using Lipofectamine CRISPRMAX Cas9 Transfection Reagent (Thermo Scientific, cat. no. CMAX00003) following the manufacturer's protocol. Passage number at time of transfection was passage 20 post introduction of *Ngn2* transgene to the cell line. Cells were cultured for two days with daily media changes then washed once with 1X PBS, dissociated using Accutase (STEMCELL Technologies, cat. no. 07920), quenched with 1X PBS, and spun at 800RPM for 3 minutes. To prepare cells for FACS sorting, the cell pellet was resuspended in a buffer consisting of 1X PBS, 0.5M EDTA (Neta Scientific, cat. no. 324506), 1M HEPES PH7.0 (Neta Scientific, cat. no. H0887), 1% fetal bovine serum, and Rock Inhibitor and then filtered through a 35µm cell strainer. Single GFP positive cells were sorted on a BD FACSAria Flow Cytometer using a 100-micron nozzle into 96-well plates containing mTeSR™1 media supplemented with Rock Inhibitor, 1% Penicillin-Streptomycin (ThermoFisher, cat. no. 15140122), and 10% CloneR2 (STEMCELL Technologies, cat. no. 100-0691). Three days after sorting, 150uL of mTeSR™1 media was added to the 96-well plates and 6 days after sorting half the media was changed for fresh mTeSR™1 media. Following day 7 after sorting, individual colonies were observed everyday and expanded incrementally once wells reached confluency. AllPrep DNA/RNA Mini kit (Qiagen, cat. no. 80204) was used to extract DNA from a subset of cells from each single colony. Using KOD One PCR Master Mix (DiagnoCine, cat. no. KMM-201), genotyping of each colony was performed. Primer sets and expected band sizes can be found in **Supplemental Table 3.1**.

WTC11-ngn2 iPSC to neuron differentiations

Cells were differentiated into neurons for 14 days following a previously described protocol from Wang, et al.[28]. Briefly, cells were maintained in Pre-differentiation Media (defined in Wang, et al.) with daily media changes for three days. Cells were dissociated on the third day, counted, and plated in Differentiation Media (defined in Wang, et al.) supplemented with 2ug/mL doxycycline according to recommended seeding densities. Seven days later, cells were given a partial media change using Differentiation Media, and on day 14 DNA and RNA were extracted from cells using AllPrep DNA/RNA Mini kit (Qiagen, cat. no. 80204). Identity of neurons were confirmed by qPCR for neuron specific marker genes (see below).

qPCR for deletion line expression analysis and neuron identification

cDNA was synthesized from extracted RNA using SuperScript™ III Reverse Transcriptase (Invitrogen, cat. no. 18080044) following the manufacturer's protocol. cDNA was diluted 1:10 and used for qPCR using SsoFast EvaGreen Supermix (BioRad, cat. no. 1725205). Primers for each gene target are listed in **Supplemental Table 3.1**. qPCR reactions were done in triplicate and normalized to the housekeeping gene *GAPDH.* Additionally, a comparison of *PPIA* expression between wildtype cells and deletion lines was performed to exclude the possibility that variation in differentiations caused global expression changes.

RNA-sequencing

RNA was extracted from differentiated neurons at day 14 as described above. RNA was submitted to Novogene for library preparation and sequencing. RNA libraries were sequenced

using an Illumina NovaSeq PE150. Gene expression analysis of the homozygous deletion line was performed using *Partek® Flow®* software, v10.0, following the RNA-seq tutorial.

Allele specific RNA-sequencing analysis

SNPs were associated with each allele (deletion or wildtype) in each of the heterozygous deletion lines using the phased genome[30] for WTC11-Ngn2 cell line (NCBI GEO GSE113481). This was done by performing PCR using KOD One PCR Master Mix (DiagnoCine, cat. no. KMM-201) on each cell line using primers targeting the deletion (**Supplemental Table 3.1**). The target deletion spans a phased SNP in the WTC11-Ngn2 cell line, therefore only the wildtype band was extracted using the QIAquick Gel Extraction Kit (Qiagen, cat. no. 28706X4). The SNP associated with the wildtype allele was then identified via sanger sequencing, and the deletion allele at this location was deduced. This initial identification allowed for the association of exon SNPs for each gene within the TAD for both the deletion and wildtype allele. All but two genes had phased SNPs within exons, therefore these two genes (*NME6* and *PLXNB1*) were not included in the analysis. A complete list of exon associated SNP locations and sequences for each gene in each cell line can be found in **Supplemental Table 3.2**. Heterozygous deletion lines were differentiated and sequenced through Novogene as described above. High quality of the reads was verified with FastQC. Counts of the relevant phased SNPs were tabulated with GATK's ASEReadCounter function, and the counts of the deletion-associated SNPs were normalized by the total number of reads covering that region. The binomial test was then used to determine if this proportion was significantly different from the allele-specific expression observed in the wildtype samples.

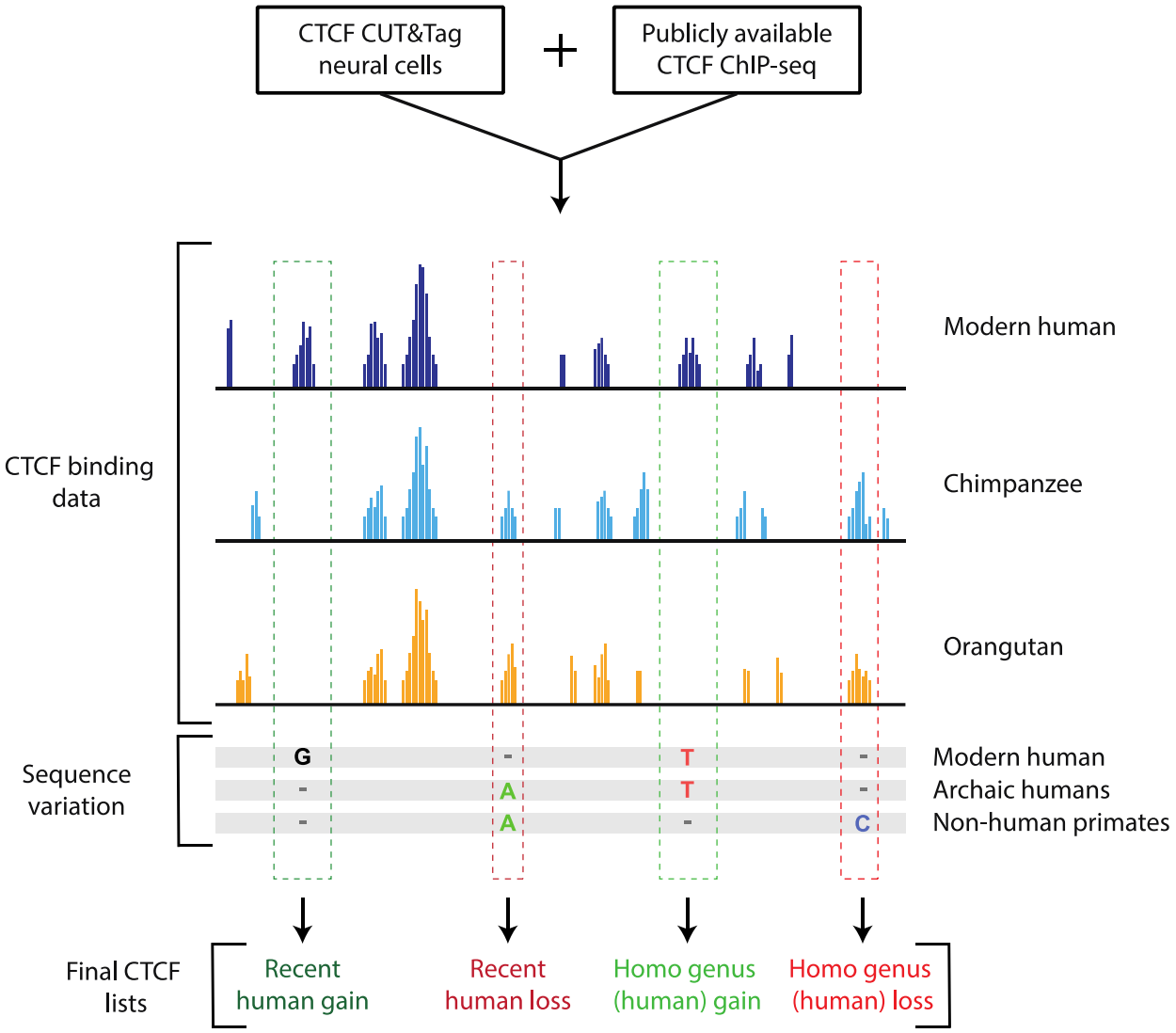

Figure 3.1: Simplified schematic of the CTCF identification pipeline
This shows a simplified version of our pipeline in which we use CTCF CUT&Tag and ChIP-seq data from human, chimpanzee, and orangutan in combination with sequence variation in modern humans, archaic humans, and non-human primates to determine the different categories of CTCF sites: recent human gain, recent human loss, human gain, and human loss.

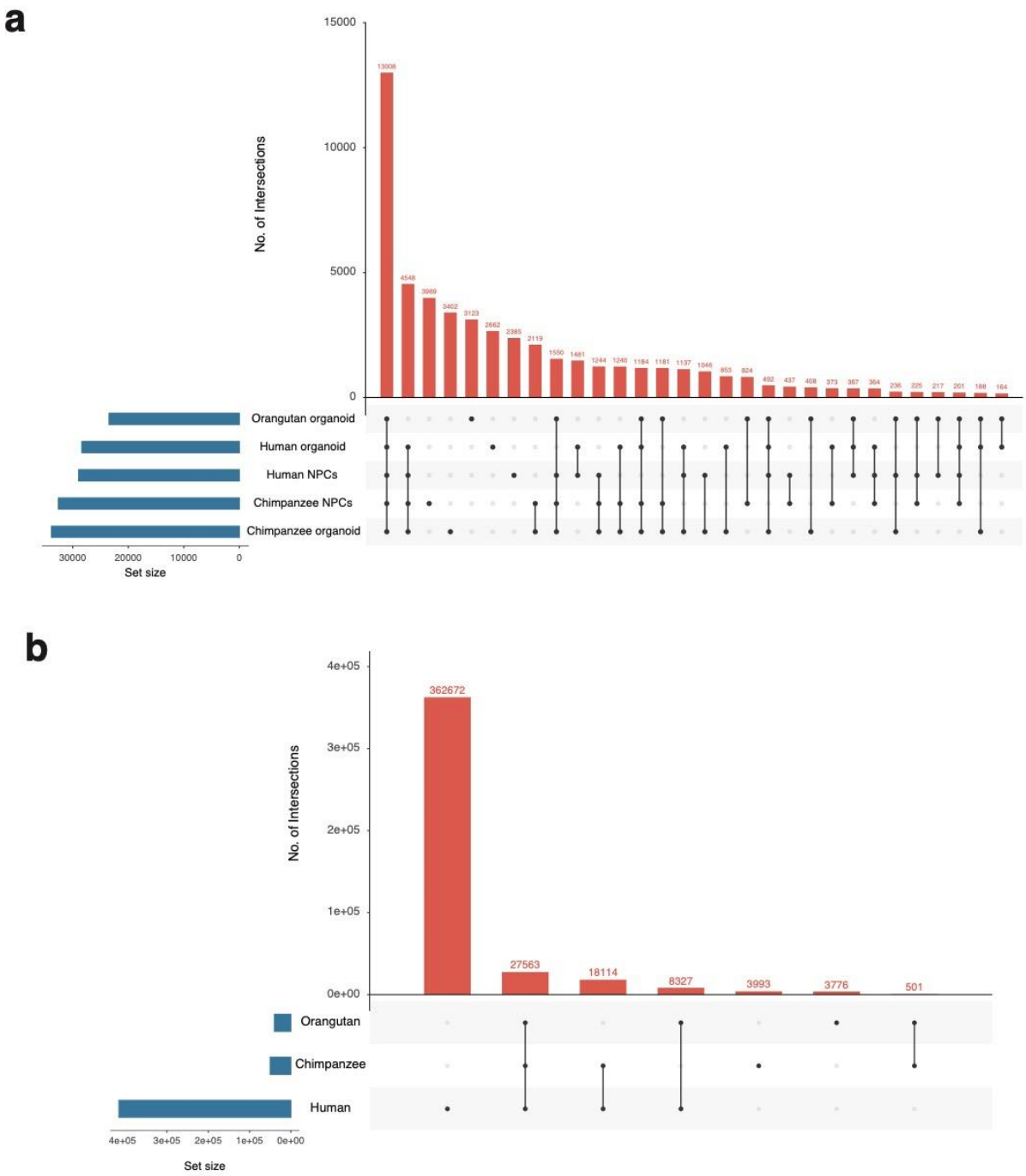

Figure 3.2: Comparisons of CTCF binding data
(a) Comparison of CTCF CUT&Tag data broken down by sample. (b) Comparison of all peaks from all datasets (CTCF CUT&Tag and ChIP-seq) broken down by species.

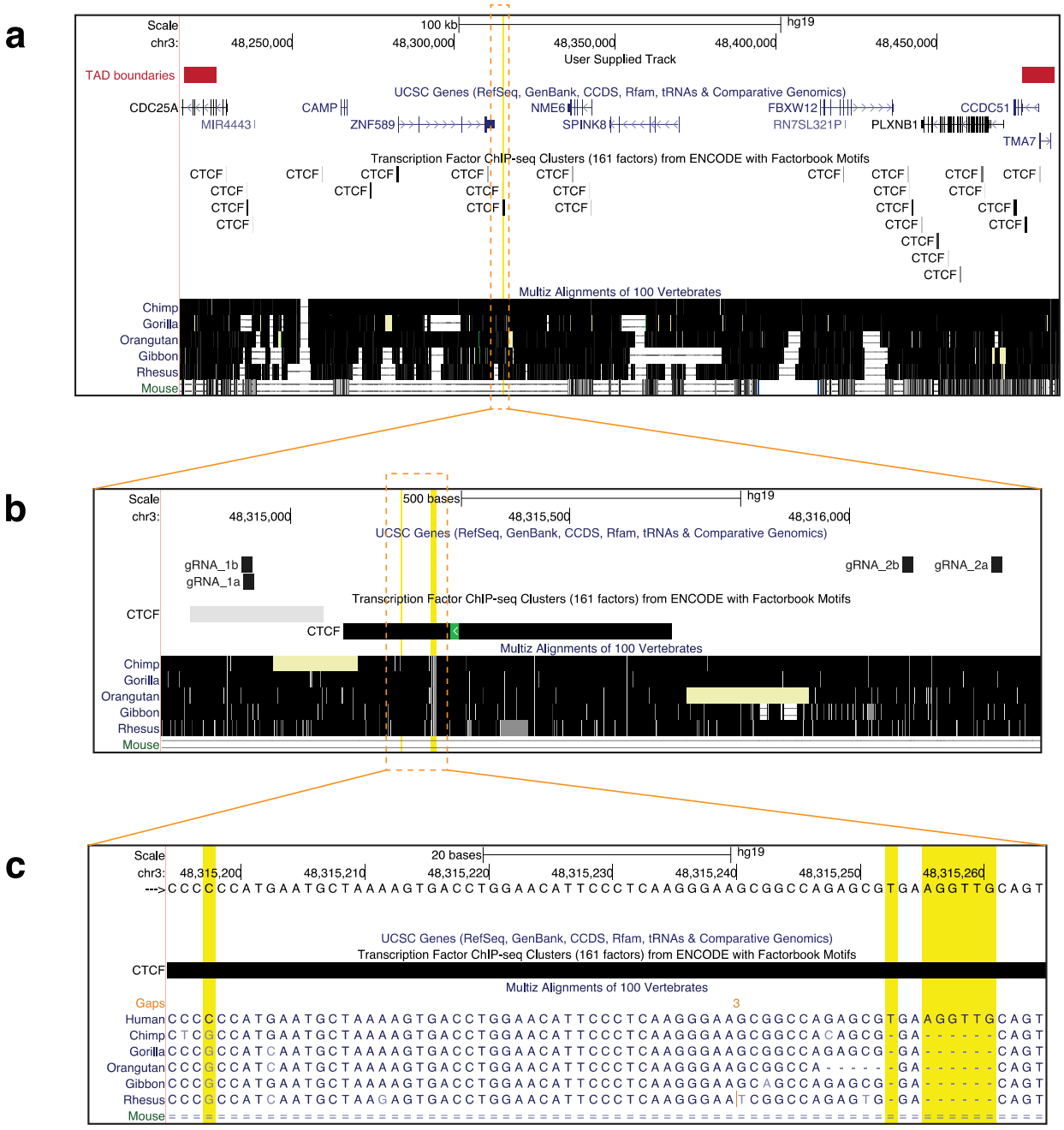

Figure 3.3: Region encompassing target CTCF for deletion cell line experiments
Genome browser image showing increasing granularity of the region targeted for deletion. (a) The TAD encompassing the target CTCF site (highlighted in yellow) with all nearby genes and other CTCF sites. (b) A closer view of the CTCF site targeted for deletion. This panel also shows the gRNAs used to cause the deletion when creating the deletion cell lines. (c) An even closer view showing the nucleotide changes between modern humans and non-human primates.

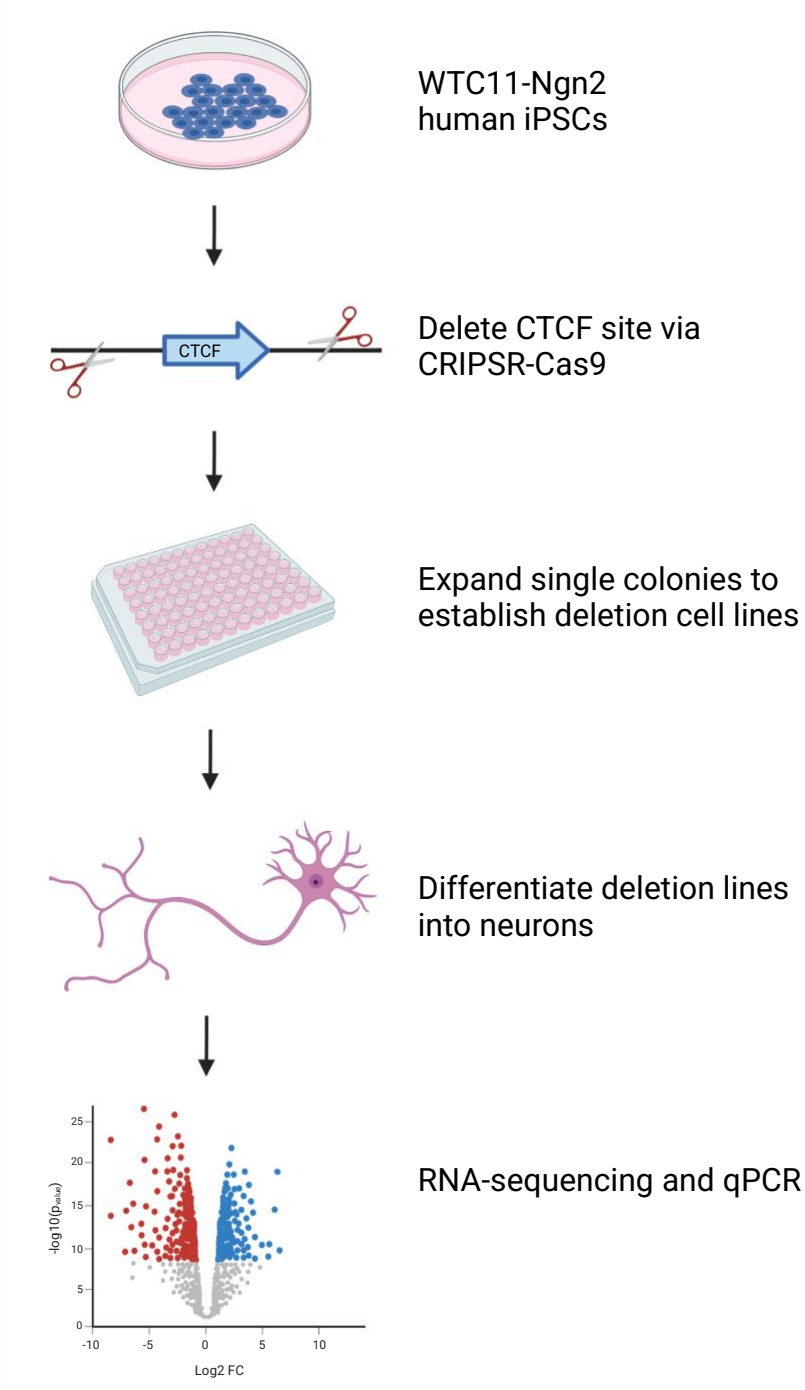

Figure 3.4: CTCF deletion cell line creation workflow
A simplified schematic of the workflow used to create the CTCF deletion cell line, differentiation, and downstream experiments.

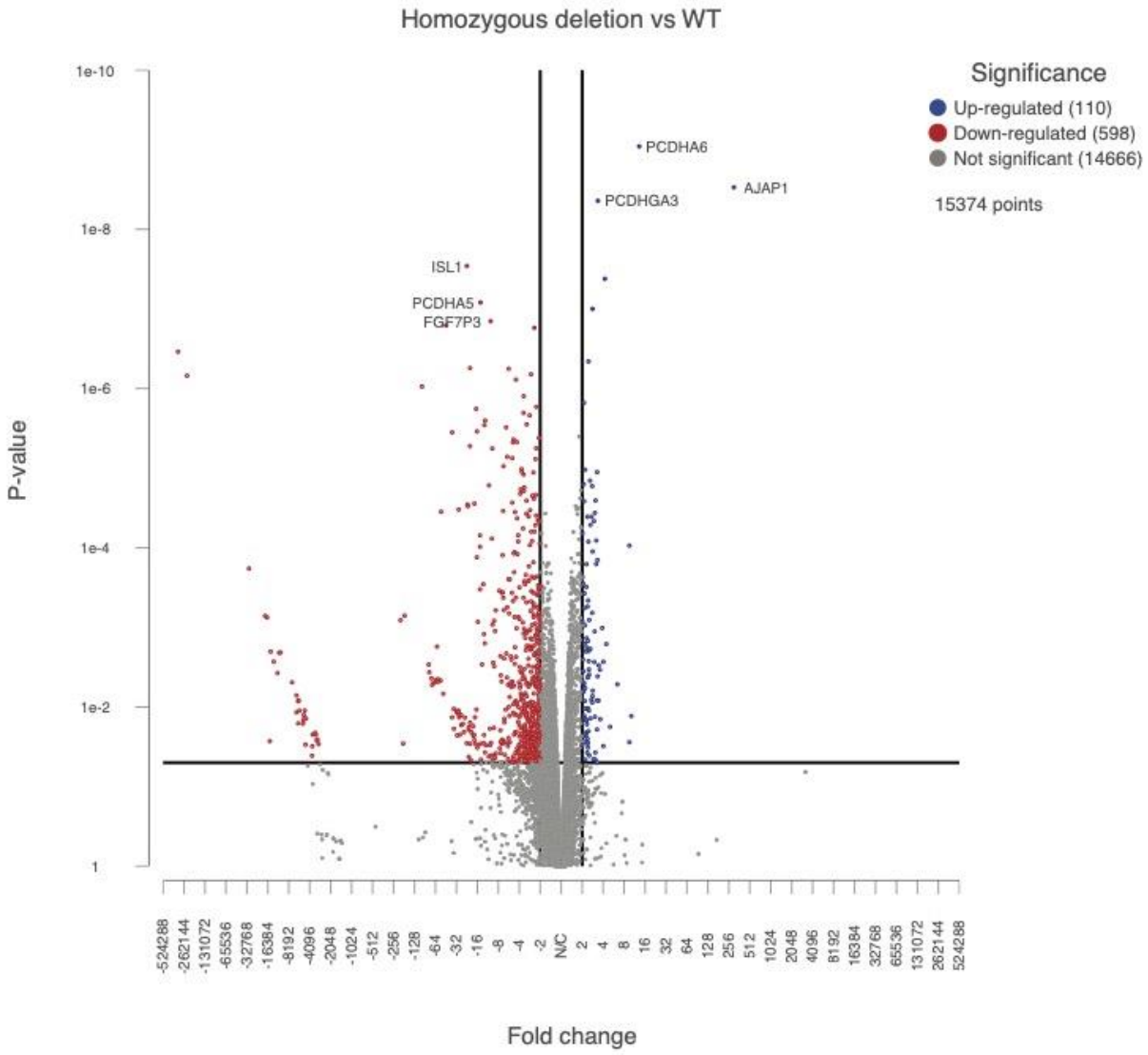

Figure 3.5: RNA-sequencing results from homozygous deletion line
Bulk RNA-sequencing results for the homozygous deletion line compared to wildtype cells. Genes that are up-regulated are represented by blue dots and down-regulated genes are represented by red dots. Grey dots depict genes that showed no significant change in expression determined by a cut-off of either a p-value>= 0.05 or a fold change of less than +/- 2.

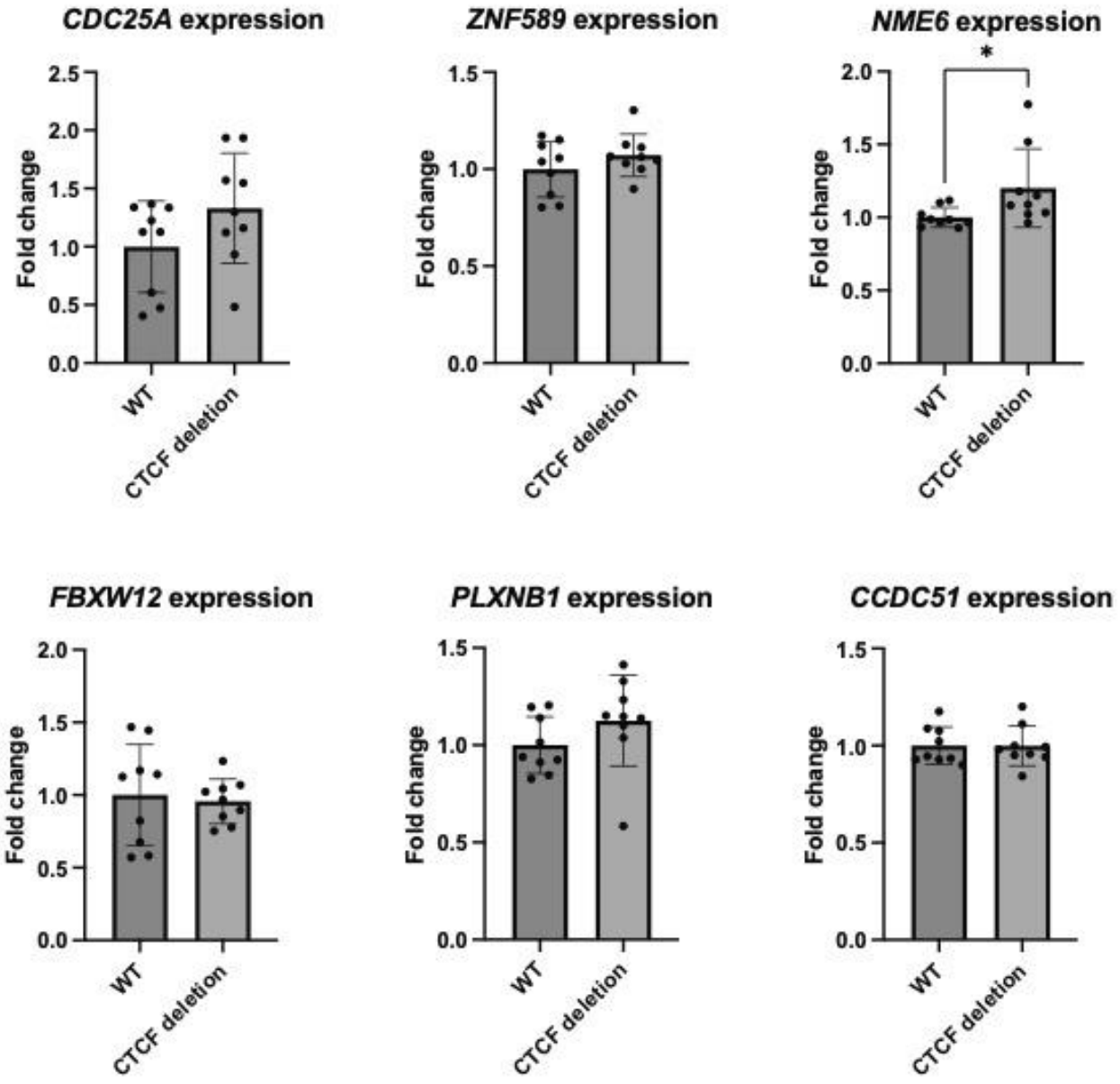

Figure 3.6: qPCR results from homozygous deletion line
Observed fold change in gene expression compared to wildtype for each gene within the TAD encompassing the deleted CTCF site. Two genes (*CAMP* and *SPINK8*) are not shown here because they are not endogenously expression in neurons. All changes in gene expression are not significant except where indicated for *NME6* (two-tailed t-test, p-value= 0.0451).

Table 3.1: Number of categorized CTCF sites

In addition to indicating number of CTCF sites in each category, this table also includes a count of how many CTCF sites overlap TAD boundaries or are within a TAD with protein coding sequences.

Category	Number identified sites	Number of CTCF sites within a TAD boundary	Number of CTCF sites within a TAD with protein coding sequences
Human specific gained sites	2230	29	1514
Human specific lost sites	24	0	16
Recent human specific gained sites	24	53	16

Table 3.2: Results of gene ontology enrichment analysis

Table includes category of CTCF site and output from GOrilla. Full descriptions and definition of each category can be found at http://cbl-gorilla.cs.technion.ac.il/.

Category	GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
Human specific gained	GO:0031424	keratinization	1.75E-05	2.56E-01	1.39 (14426,123,7018,83)
Human specific gained	GO:0048704	embryonic skeletal system morphogenesis	1.09E-04	7.99E-01	1.47 (14426,67,7018,48)
Human specific gained	GO:0006959	humoral immune response	1.13E-04	5.53E-01	1.29 (14426,170,7018,107)
Human specific gained	GO:0006952	defense response	1.15E-04	4.22E-01	1.12 (14426,914,7018,499)
Human specific gained	GO:0060674	placenta blood vessel development	1.28E-04	3.74E-01	1.85 (14426,20,7018,18)
Human specific gained	GO:0006356	regulation of transcription by RNA polymerase I	1.59E-04	3.88E-01	1.74 (14426,26,7018,22)
Human specific gained	GO:0060716	labyrinthine layer blood vessel development	1.75E-04	3.66E-01	2.06 (14426,12,7018,12)
Human specific gained	GO:0006955	immune response	2.63E-04	4.80E-01	1.13 (14426,778,7018,426)
Human specific gained	GO:0033135	regulation of peptidyl-serine phosphorylation	2.66E-04	4.33E-01	1.35 (14426,110,7018,72)
Human specific gained	GO:0033139	regulation of peptidyl-serine phosphorylation of STAT protein	2.71E-04	3.96E-01	1.78 (14426,22,7018,19)
Human specific gained	GO:0098542	defense response to other organism	3.00E-04	3.99E-01	1.19 (14426,336,7018,195)
Human specific gained	GO:0046688	response to copper ion	4.24E-04	5.17E-01	1.64 (14426,30,7018,24)
Human specific gained	GO:0019883	antigen processing and presentation of endogenous antigen	4.85E-04	5.46E-01	1.76 (14426,21,7018,18)
Human specific gained	GO:0061844	antimicrobial humoral immune response mediated by antimicrobial peptide	6.46E-04	6.75E-01	1.44 (14426,60,7018,42)
Human specific gained	GO:0033138	positive regulation of peptidyl-serine phosphorylation	7.06E-04	6.88E-01	1.36 (14426,86,7018,57)
Human specific gained	GO:0050890	cognition	8.04E-04	7.36E-01	1.22 (14426,215,7018,128)
Human specific gained	GO:0046364	monosaccharide biosynthetic process	8.23E-04	7.09E-01	1.49 (14426,47,7018,34)
Human specific gained	GO:0042742	defense response to bacterium	8.60E-04	6.99E-01	1.24 (14426,182,7018,110)
Human specific lost	GO:2001237	negative regulation of extrinsic apoptotic signaling pathway	3.31E-05	4.85E-01	6.44 (14422,80,224,8)
Human specific lost	GO:0070228	regulation of lymphocyte apoptotic process	3.75E-05	2.74E-01	9.42 (14422,41,224,6)
Human specific lost	GO:0035524	proline transmembrane transport	1.95E-04	9.53E-01	24.14 (14422,8,224,3)
Human specific lost	GO:0090023	positive regulation of neutrophil chemotaxis	2.75E-04	1.00E+00	12.26 (14422,21,224,4)
Human specific lost	GO:0015808	L-alanine transport	2.90E-04	8.48E-01	21.46 (14422,9,224,3)
Human specific lost	GO:2000106	regulation of leukocyte apoptotic process	3.57E-04	8.72E-01	6.33 (14422,61,224,6)
Human specific lost	GO:1902624	positive regulation of neutrophil migration	3.98E-04	8.32E-01	11.20 (14422,23,224,4)
Human specific lost	GO:2001236	regulation of extrinsic apoptotic signaling pathway	4.01E-04	7.33E-01	4.52 (14422,114,224,8)
Human specific lost	GO:0001516	prostaglandin biosynthetic process	4.09E-04	6.66E-01	19.32 (14422,10,224,3)
Human specific lost	GO:0015816	glycine transport	4.09E-04	5.99E-01	19.32 (14422,10,224,3)

Category	GO term	Description	P-value	FDR q-value	Enrichment (N, B, n, b)
Human specific lost	GO:0015824	proline transport	4.09E-04	5.45E-01	19.32 (14422,10,224,3)
Human specific lost	GO:0046457	prostanoid biosynthetic process	4.09E-04	4.99E-01	19.32 (14422,10,224,3)
Human specific lost	GO:0071624	positive regulation of granulocyte chemotaxis	4.72E-04	5.31E-01	10.73 (14422,24,224,4)
Human specific lost	GO:0070229	negative regulation of lymphocyte apoptotic process	4.72E-04	4.93E-01	10.73 (14422,24,224,4)
Human specific lost	GO:0032268	regulation of cellular protein metabolic process	5.61E-04	5.47E-01	1.59 (14422,1984,224,49)
Human specific lost	GO:0070232	regulation of T cell apoptotic process	7.51E-04	6.87E-01	9.54 (14422,27,224,4)
Human specific lost	GO:0090022	regulation of neutrophil chemotaxis	7.51E-04	6.46E-01	9.54 (14422,27,224,4)
Human specific lost	GO:0042325	regulation of phosphorylation	8.02E-04	6.52E-01	1.76 (14422,1241,224,34)
Human specific lost	GO:2000145	regulation of cell motility	9.20E-04	7.09E-01	2.04 (14422,725,224,23)
Human specific lost	GO:0030224	monocyte differentiation	9.42E-04	6.90E-01	14.86 (14422,13,224,3)
Human specific lost	GO:1903131	mononuclear cell differentiation	9.42E-04	6.57E-01	14.86 (14422,13,224,3)
Recent human specific gained	GO:0002062	chondrocyte differentiation	1.00E-04	1.00E+00	10.55 (14422,34,201,5)
Recent human specific gained	GO:0035524	proline transmembrane transport	1.42E-04	1.00E+00	26.91 (14422,8,201,3)
Recent human specific gained	GO:1902953	positive regulation of ER to Golgi vesicle-mediated transport	1.93E-04	9.43E-01	71.75 (14422,2,201,2)
Recent human specific gained	GO:0015808	L-alanine transport	2.11E-04	7.71E-01	23.92 (14422,9,201,3)
Recent human specific gained	GO:0015816	glycine transport	2.98E-04	8.72E-01	21.53 (14422,10,201,3)
Recent human specific gained	GO:0015824	proline transport	2.98E-04	7.26E-01	21.53 (14422,10,201,3)
Recent human specific gained	GO:0032328	alanine transport	8.67E-04	1.00E+00	15.38 (14422,14,201,3)

Supplemental Table 3.1: gRNAs and primers used in the cell line deletion experiments
Top panel shows the gRNAs used to create the deletion cell lines. The bottom panel shows all primers used for genotyping (with expected bands), qPCR, and SNP phasing.

gRNA1a	gRNA1b	gRNA2a	gRNA2b
CCAGGTTTATTCGTCACAAC	GGTTTATTCGTCACAACAGG	TTGCCAAGGCCAAGTAATAT	CCTCCCATAGTACACCCCAG

Forward primer	Forward primer sequence	Reverse primer	Reverse primer sequence	Use	Size with deletion	Size without deletion
CTCF_del_genotyping_fwd	GGTCCAGGCAGTCATAGGAA	CTCF_del_genotyping_rev	TAACACTGGCCTGACCACAA	CTCF deletion genotyping	800-970bp	2,150bp
ZNF589_qPCR_fwd	TGGCTGTGCTTTTCACTGAGGC	ZNF589_qPCR_rev	AAGGGCAGGTATGGACTTCTGG	qPCR expression analysis of *ZNF589*	N/A	N/A
NME6_qPCR_fwd	TGCCAGAGGTTTTACCGAGAGC	NME6_qPCR_rev	TCCAGAGCTGGATGGCATCCTT	qPCR expression analysis of *NME6*	N/A	N/A
FBXW12_qPCR_fwd	TTCAGCATCACTGGCTTCCTGC	FBXW12_qPCR_rev	CAGAGTTCTCGGATGTGGTGAG	qPCR expression analysis of *FBXW12*	N/A	N/A
PLXNB1_qPCR_fwd	TCTCACCCTGAATGGCTCCAAG	PLXNB1_qPCR_rev	CTGGTCTCACACCGCAGTTGTT	qPCR expression analysis of *PLXNB1*	N/A	N/A
CDC25A_qPCR_fwd	TCTGGACAGCTCCTCTCGTCAT	CDC25A_qPCR_rev	ACTTCCAGGTGGAGACTCCTCT	qPCR expression analysis of *CDC25A*	N/A	N/A
CCDC51_qPCR_fwd	GAGGACTTGGAAGTTCACCAGG	CCDC51_qPCR_rev	TCTTCTGCACGCAGATAGGCTG	qPCR expression analysis of *CCDC51*	N/A	N/A
GAPDH_qPCR_fwd	GGCCATCCACAGTCTTCTG	GAPDH_qPCR_rev	TCATCAGCAATGCCTCCTG	qPCR housekeeping gene (*GAPDH*)	N/A	N/A
MAP2_qPCR_fwd	AGGCTGTAGCAGTCCTGAAAGG	MAP2_qPCR_rev	CTTCCTCCACTGTGACAGTCTG	Neuron identification via qPCR	N/A	N/A
TUBB3_qPCR_fwd	TTTGGACATCTCTTCAGGCC	TUBB3_qPCR_rev	TTTCACACTCCTTCCGCAC	Neuron identification via qPCR	N/A	N/A
PPIA_qPCR_fwd	GGCAAATGCTGGACCCAACACA	PPIA_qPCR_rev	TGCTGGTCTTGCCATTCCTGGA	Control gene for qPCR	N/A	N/A
CTCF_del_SNP_fwd	AAAACAAGAGCCCAGTCAGC	CTCF_del_SNP_rev	GTTGCGGTGGATCCTTGATG	SNP phasing	6,737bp	none

Supplemental Table 3.2: Heterozygous SNP phasing

Cell line	Gene	Exon SNP location (hg19)	Deletion allele SNP	WT allele SNP
Heterozygous deletion line #1	*ZNF598*	chr3:48,309,424	G	C
Heterozygous deletion line #1	*NME6*	no SNP in exon	--	--
Heterozygous deletion line #1	*FBXW12*	chr3:48,419,897	T	C
Heterozygous deletion line #1	*PLXNB1*	no SNP in exon	--	--
Heterozygous deletion line #1	*CDC25A*	chr3:48,199,754	G	T
Heterozygous deletion line #1	*CCDC51*	chr3:48,476,431	G	C
Heterozygous deletion line #2	*ZNF598*	chr3:48,309,424	G	C
Heterozygous deletion line #2	*NME6*	no SNP in exon	--	--
Heterozygous deletion line #2	*FBXW12*	chr3:48,419,897	T	C
Heterozygous deletion line #2	*PLXNB1*	no SNP in exon	--	--
Heterozygous deletion line #2	*CDC25A*	chr3:48,199,754	G	T
Heterozygous deletion line #2	*CCDC51*	chr3:48,476,431	G	C

3.6 REFERENCES

1. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130025 (2013).
2. Xu, K., Schadt, E. E., Pollard, K. S., Roussos, P. & Dudley, J. T. Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. *Mol. Biol. Evol.* **32**, 1148–1160 (2015).
3. Doan, R. N. *et al.* Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. *Cell* **167**, 341-354.e12 (2016).
4. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
5. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
6. Weiss, C. V. *et al.* The cis-regulatory effects of modern human-specific variants. *Elife* **10**, (2021).
7. Lazar, N. H. *et al.* Epigenetic maintenance of topological domains in the highly rearranged gibbon genome. *Genome Res.* **28**, 983–997 (2018).
8. Luo, X. *et al.* 3D Genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell* **184**, 723-740.e21 (2021).
9. Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. & Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17507–17512 (2012).

10. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* **17**, 520–527 (2001).
11. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
12. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
13. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
14. Ushiki, A. *et al.* Deletion of CTCF sites in the SHH locus alters enhancer-promoter interactions and leads to acheiropodia. *Nat. Commun.* **12**, 2282 (2021).
15. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
16. Hanssen, L. L. P. *et al.* Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* **19**, 952–961 (2017).
17. Khoury, A. *et al.* Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat. Commun.* **11**, 54 (2020).
18. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
19. Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–348 (2012).

20. McArthur, E. *et al.* Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence. *bioRxiv* 2022.02.07.479462 (2022) doi:10.1101/2022.02.07.479462.

21. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

22. Hu, D. *et al.* CUT&Tag recovers up to half of ENCODE ChIP-seq peaks. *bioRxiv* 2022.03.30.486382 (2022) doi:10.1101/2022.03.30.486382.

23. Steinhaus, R., Robinson, P. N. & Seelow, D. FABIAN-variant: predicting the effects of DNA variants on transcription factor binding. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac393.

24. Paliou, C. *et al.* Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12390–12399 (2019).

25. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).

26. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.* **3**, e39 (2007).

27. Agha, Z. *et al.* Exome sequencing identifies three novel candidate genes implicated in intellectual disability. *PLoS One* **9**, e112687 (2014).

28. Wang, C. *et al.* Scalable Production of iPSC-Derived Human Neurons to Identify Tau-Lowering Compounds by High-Content Screening. *Stem Cell Reports* **9**, 1221–1233 (2017).

29. Wang, C.-H. *et al.* A shRNA functional screen reveals Nme6 and Nme7 are crucial for embryonic stem cell renewal. *Stem Cells* **30**, 2199–2211 (2012).

30. Song, M. *et al.* Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.* **51**, 1252–1262 (2019).

31. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).

32. Sawyer, G. J. & Maley, B. Neanderthal reconstructed. *Anat. Rec. B New Anat.* **283**, 23–31 (2005).

33. Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).

34. Kochiyama, T. *et al.* Reconstructing the Neanderthal brain using computational anatomy. *Sci. Rep.* **8**, 6296 (2018).

35. Neubauer, S., Hublin, J.-J. & Gunz, P. The evolution of modern human brain shape. *Sci Adv* **4**, eaao5961 (2018).

36. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

37. Liu, T. *et al.* TADKB: Family classification and a knowledge base of topologically associating domains. *BMC Genomics* **20**, 217 (2019).

CHAPTER 4: Conclusion

My thesis work encompasses two projects with the broad goal of identify noncoding gene regulatory sequences that may have impacted modern human specific gene expression and ultimately human speciation. Chapter 2 is a published article[1] that investigates all single nucleotide changes, fixed or nearly fixed, in the modern human lineage as compared to extinct archaic humans, Neanderthal and Denisovan. In this study we tested the archaic and modern versions of these sequences in the same experiment through MPRA and found some of the sequences to have regulatory potential. A subset of these active sequences also showed differential activity between the archaic and modern sequences. The differentially active sequences were subsequently linked to genes involved in brain and vocal cord function, among other phenotypes. This research provides a catalog of potential modern human specific enhancers and gene regulatory elements that could prove to be, with further interrogation, important for the development of human specific traits.

Chapter 3 of my thesis represents an ongoing project identifying human specific CTCF sites and determining their role, if any, in gene regulation. Through novel CTCF CUT&Tag data and computational tools, we've identified human gained and lost CTCF sites as compared to great apes and separately to extinct archaic humans. These lists provide a framework in which to further study human specificity of these sites. We will continue to refine this dataset and our pipeline by incorporating additional relevant data, such as additional ChIP-seq data for great apes, Hi-C datasets, and species-specific gene expression. These datasets and others will refine

our pipeline and by extension, our identified human specific gained and lost CTCF sites, providing additional targets for future research.

The experiment in which I deleted a single human gained CTCF site in neurons produced an informative, although negative, result. I did not see a change in gene expression caused by the deletion, and although this site could function in combination with other elements to regulate genes, these results speak to the greater state of the field when it comes to CTCF function and role in gene expression. There have been many studies aimed at dissecting the rules that govern CTCF function[2–5] which have built on one another, slowly increasing our knowledge of how CTCF works. However, my research and others suggest we are far from fully elucidating the exact role of CTCF in gene regulation and it will take additional carefully calculated research to discern a more comprehensive understanding of this topic.

My thesis work has focused on exploring how sequence changes in the genome may have shaped gene expression and phenotypes in modern humans. This is an expansive topic encompassing many questions that have yet to be answered. My work chips away at the larger question of “what makes us human” and opens new avenues for future work.

4.1 REFERENCES

1. Weiss, C. V. *et al.* The cis-regulatory effects of modern human-specific variants. *Elife* **10**, (2021).
2. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions. *Nature* **604**, 571–577 (2022).
3. Despang, A. *et al.* Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
4. Paliou, C. *et al.* Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12390–12399 (2019).
5. Hanssen, L. L. P. *et al.* Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* **19**, 952–961 (2017).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Lana Harshman
2ED10A842D6D438...
Author Signature

11/17/2022
Date