Massively parallel characterization of enhancers in evolution and disease

Jason C. Klein

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Jay Shendure, Chair

James Thomas

R. David Hawkins

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Massively parallel characterization of enhancers in evolution and disease

Jason C. Klein

Chair of the Supervisory Committee:

Professor Jay Shendure, MD, PhD

Genome Sciences

On average, protein-coding sequence is over 99.9% identical between humans, yet some individuals develop disease while others do not. Similarly, protein-coding sequences are over 99% identical between human and chimpanzees despite large phenotypic differences. Our genome makes several different phenotypes using the same set of instructions (genes) by regulating the relative timing and expression level of genes throughout development. One way in which our genome does this is through enhancers.

Enhances were first discovered in 1981 as pieces of DNA that are able to increase gene expression independent of their relative position and orientation to the transcriptional start site (TSS). Despite enhancers being implicated in a variety of evolutionary phenotypes and disease,

the field still lacks a comprehensive understanding of how they function. Massively parallel reporter assays (MPRAs) have served as an indispensable tool to screen short pieces of DNA for enhancer activity. MPRAs test the ability of thousands to hundreds of thousands of sequences to increase expression of a reporter gene on a plasmid in a single experiment. However, the assay relies on testing short pieces of DNA (<200bp), which may contribute to the assay's relatively low sensitivity. In chapter two of this dissertation, I discuss a protocol I developed to assemble libraries of short (<230bp) DNA fragments into large (up to 700bp) sequences. This protocol has allowed our group and others to screen large synthetic sequences for enhancer activity as well as for protein folding for the first time.

In chapters three and four, I apply MPRAs to characterize regulatory changes over primate evolution and identify mutations within enhancers that may contribute to a patient's risk for Osteoarthritis. Enhancers have previously been suggested to play a unique role in evolution and speciation due to their decreased pleiotropy and penetrance compared to protein-coding changes. In chapter three, I synthesized and screened present day and ancestral orthologs for 348 liver enhancers in order to characterize each enhancer's evolutionary-functional trajectory throughout 40 million years of primate evolution. We identified groups of enhancers with similar trajectories, most of which could be explained by one or two mutational events. We identified and quantified the correlation between sequence and functional divergence of naturally evolving sequence and implicate cytosine deamination within CpGs as a potential driver in enhancer evolution.

In chapter four, I applied the same assay to identify common polymorphisms that are associated with Osteoarthritis, which alter enhancer activity. From 35 lead GWAS variants, I identified 1605 SNPs associated with Osteoarthritis in European populations. I then synthesized

196bp around the major and minor alleles for each SNP, and screened each sequence for enhancer activity. We confidently measured enhancer activity of both alleles for 753 SNPs, and found six that drove differential enhancer activity at an FDR of 5%. Our lead SNP increases expression of an isoform of HBP1 (a negative regulator of the Wnt-beta-catenin pathway) with an alternative TSS in an osteosarcoma cell line, CRISPR-edited chondrosarcoma cell line, and cartilage from knee Osteoarthritis patients.

Chapter five discusses ongoing and future projects to systematically compare different enhancer assays to improve sensitivity and specificity of enhancer screens. It also suggests future applications of these assays to better understand the role of gene regulation in evolution and disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many people were instrumental in preparing me for graduate school as well as in supporting and teaching me throughout the past four years. First, I would like to acknowledge my mentor, Jay Shendure, for allowing me the opportunity to train in his lab with some of the brightest and most helpful colleagues I could ask for. Jay has provided amazing insight and encouragement for my research projects and career development, and has served as a role model for my future career. His enthusiasm and optimism are contagious, and continue to push myself and others forward.

I would also like to acknowledge my previous mentors. Dave McClay was my first scientific mentor in undergraduate, and really taught me how to conduct good science and enjoy it. During undergraduate, Eric Spana and Francois Schweisguth also served as scientific mentors, who helped to cultivate my interest in genetic research.

Everyone in the Shendure Lab has also contributed to my growth as a scientist. In particular, several members of the lab directly contributed to the projects described within this dissertation, notable Jerrod Schwartz, Martin Kircher, Vikram Agarwal, Beth Martin, and Aidan Keith. Several other members of the lab have continuously provided insight, assistance and advice, notably Choli Lee, Molly Gasperini, Seungsoo Kim, Jes Alexander, Aaron McKenna, Greg Findlay, and Lea Starita. Aidan Keith, an undergraduate who began working with me in February 2017 deserves an additional acknowledgement. It has been a great experience to teach and work with him over the past year, and his work and enthusiasm has helped to carry many of my projects to fruition.

Outside of the lab environment, I would like to acknowledge Scott Brothers, who taught me much of what I know computationally. On multiple occasions, he has helped me identify and apply algorithms and heuristics from outside the biological sciences towards my questions.

DEDICATION

This work is dedicated to my parents, Edward and Olga Klein, as well as my brothers, Evan, Billy, and Adam Klein, for their continuous support of my education.

Chapter 1. INTRODUCTION

1.1 THE DISCOVERY AND CHARACTERIZATION OF ENHANCERS

Although all cells of a multicellular organism share the same genes, they differentiate to myriad cell types. Moreover, while the coding sequences between humans and chimpanzees are more than 99% identical, many phenotypic differences exist (King and Wilson, 1975). For decades, the field has asked how a similar set of genes can encode for such a different output. Most evidence points towards the relative timing and levels of gene expression contributing to the diverse phenotypes we observe. The spatiotemporal control of gene expression is influenced by distal DNA sequences known as enhancers. Enhancers were first defined in 1981 as short sequences of DNA that are able to increase the expression of a gene independent of their relative position or orientation to the transcriptional start site (TSS) (Banerji et al., 1981; Moreau et al., 1981).

Banerji *et al.* identified enhancers by testing expression of a 4.7kb sequence of the rabbit genome surrounding the *β-globin* gene with a transient expression assay in HeLa cells. He noticed minimal RNA when the 4.7kb fragment was cloned into a vector without SV40 DNA, and 200x more RNA when the vector contained SV40 DNA. He showed with further experiments that at least one copy of a 72bp repeat within SV40 was necessary for increased transcription, the SV40 had to be in *cis* with the gene, the relative orientation of SV40 to the transcriptional start site was not important, and SV40 could increase transcription from multiple positions in the vector.

Soon after Banerji's characterization of the SV40 enhancer, several groups identified the first mammalian enhancers – the immunoglobulin heavy chain enhancer (Banerji et al., 1983;

Gillies et al., 1983; Neuberger, 1983) and the immunoglobulin κ gene enhancer (Queen, 1983). These findings validated the existence of enhancers in mammalian cells, as well as their ability to act at a distance from the promoter. Moreover, both Banerji and Gillies noticed enhancer activity in lymphoid cells but not in other cell types. This was the first suggestion that enhancers act in a cell-type or tissue-specific manner.

In 1985, Nowock *et al.* showed that the TGGCA protein from HeLa cells binds to multiple viral enhancers (Nowock et al., 1985) and Piette *et al.* showed that proteins from mouse 3T6 cells bind the polyoma B enhancer using EMSAs and DNase I footprinting (Piette et al., 1985). Together, these were the first two findings to suggest that enhancers interact with endogenous proteins. Serfling *et al.* proposed that enhancers are composed of short motifs, responsive to different signals, which together contribute to the element's activity (Serfling et al., 1985). Many of these motifs were found to be binding sites for transcription factors (AP-1 (Lee et al., 1987), AP-2 (Mitchell et al., 1987) and NF-κB binding to the SV40 enhancer (Sen and Baltimore, 1986)), providing a mechanism for how enhancers act in both a stage-specific and tissue-specific manner (Atchinson, 1988).

The definition of enhancers from the 1980s is still used today. Modified from Atchinson (Atchinson, 1988), enhancers are defined as segments of DNA that:

a) increase transcription of heterologous promoters in *cis*
b) function independent of their relative orientation to the promoter
c) can function over long distances and at different positions relative to the promoter
d) can act in a time and tissue-specific manner

1.2 TRADITIONAL METHODS TO SCREEN FOR ENHANCERS

As described above, the immunoglobulin heavy chain enhancer was the first cellular enhancer identified. Banerji, Gillies and Neuberger concurrently identified this enhancer, located in the intron of the upstream constant region, by cloning segments of the IgH locus into a plasmid with a promoter and reporter gene (SV40 early promoter and rabbit *β-globin* gene (Banerji et al., 1983), either SV40 early promoter or V_H promoter with the *immunoglobulin heavy chain* gene (Gillies et al., 1983), and V_{NP} promoter, SV40 early promoter, and thymidine kinase promoter with the mouse *immunoglobulin mu* gene and *thymidine kinase* (Neuberger, 1983)). The concept of cloning a sequence of interest into a plasmid with a minimal promoter and reporter gene has become a gold standard for identifying enhancer elements. Many additional enhancers have been identified by testing the ability of candidate sequences to activate expression of a luciferase or *lacZ* reporter gene on an episomal construct (Arnone et al., 2004).

In 1986, Hamada modified the traditional reporter assay to screen for enhancers in the genome, in what is now considered an enhancer-trap. In one study, he transformed mouse L cells with a plasmid containing a *chloramphenicol acetyltransferase* gene linked to the SV40 early promoter or enhancer (Hamada, 1986a). He noted that the construct without the SV40 enhancer was actively transcribed, suggesting that it was being activated by endogenous enhancers. In the same issue, he also published a screen transforming an enhancerless plasmid with the *xanthine-guanosine phosphoribosyltransferase* (*gpt*) gene with the SV40 promoter into HeLa cells (Hamada, 1986b). He cloned DNA sequence from transformants with active *gpt* expression and characterized two elements capable of increasing transcription on a transient plasmid.

O'Kane and Gehring described the first *in situ* screen for enhancers (O'Kane and Gehring, 1987). O'Kane transformed a construct with the *lacZ* gene downstream of the P-

element promoter into the germ-line of *Drosophila*. Since the P-element promoter drives minimal expression, the construct alone will not yield detectable β-galactosidase. However, the authors noted that 70% of fly lines expressed *lacZ* in a tissue-specific manner, most frequently in the nervous system. Soon after, several groups modified O'Kane's screen, using transposons instead of transformation (Bellen et al., 1989; Bier et al., 1989; Grossniklaus et al., 1989; Wilson et al., 1989). Enhancer traps provided additional evidence for cellular enhancers and identified regions of the genome with enhancer activity. However, since enhancers can act from long distances, it proved an additional challenge to identify the enhancer activating the integrated reporter (Skarnes, 1990). Therefore, while "genome-wide," enhancer traps are still not scalable to annotate enhancer elements within a genome.

1.3 BIOCHEMICAL SCREENS FOR ENHANCERS

In recent years, two high-throughput approaches have emerged to identify or validate enhancers. The first approach is to survey genomes for biochemical marks that are associated with enhancer activity, and will be described in this section. The second approach, massively parallel reporter assays, is described in Chapter 1.4.

While the primary sequence of DNA is the same between cell types, the chromatin landscape changes. Several of these changes are associated with function, and have been used as markers for regulatory elements within the genome. The ENCODE Project has characterized biochemical marks associated with enhancers in numerous cell lines (Project Consortium, 2004, 2007, 2011, 2012; Mouse, 2012). These marks include, H3K4me1 ChIP-seq, P300 ChIP-seq, H3K27ac ChIP-seq, DNase I hypersensitivity (DHS), and others (Shlyueva et al., 2014).

In 2007, Heintzman *et al.* performed chromatin immunoprecipitation paired with a DNA microarray (ChIP-chip) for several histone modifications as part of the ENCODE project to

identify modifications associated with different regulatory elements (Heintzman et al., 2007). In particular, they mapped H3K9/14ac, H4K5/8/12/16ac, H3K4me1, H3K4me2, H3K4me3, RNA polymerase II, TBP-associated factor 1, and the transcriptional coactivator p300 within 44 loci covering 30Mb of the human genome. This study identified H3K4me1 as a global marker for enhancers. However, H3K4me1 is not specific to enhancers as it is also found around the TSS. In a later study, Heitzman showed that H3K4me1 in the absence of H3K4me3 was specific for enhancers (Heintzman et al., 2009).

P300 and CREB binding protein (CBP) are histone acetyltransferases and transcriptional coactivators critical for embryonic development (Eckner et al., 1994; Merika et al., 1998; Yao et al., 1998). P300 is a ubiquitous part of the enhancer-associated protein complex, and has been shown to co-localize with enhancers using P300 ChIP-chip and ChIP-seq (Heintzman et al., 2007; Xi et al., 2007; Visel et al., 2009). Recently, our group verified P300 as the marker most associated with enhancer activity based on functional data from a lentiviral-based reporter assay in HepG2 cells (Inoue et al., 2016).

One role of P300 is to acetylate lysine 27 on histone 3 (H3K27ac) (Tie et al., 2009; Jin et al., 2011). In 2010, Creyghton *et al.* showed that H3K27ac distinguishes active enhancers from inactive or poised enhancers (which are also marked by H3K4me1) (Creyghton et al., 2010). However, similar to H3K4me1, H3K27ac is also found at active promoters, and therefore is not specific for enhancer elements. One common approach to search for enhancers today is to look for H3K27ac peaks distant to any TSS and/or lacking a nearby H3K4me3 peak (a mark for Pol II-bound promoters).

In order to interact with transcriptional co-activators, DNA must be accessible. DNaseI hypersensitivity assays identify regions of open (accessible) chromatin (Gross and Garrard,

1988; Xi et al., 2007). Over the years, DNase hypersensitivity mapping has progressed from a low-throughput readout - southern blotting (Wu et al., 1979; Wu, 1980; Gross and Garrard, 1988), to a mid-throughput readout – microarrays (Crawford et al., 2006a), to a high-throughput readout – sequencing (Crawford et al., 2006b; Boyle et al., 2008). Our group recently adapted a similar method that also tests for open chromatin, ATAC-seq (Buenrostro et al., 2013), for analysis of accessible chromatin in single cells (Cusanovich et al., 2015). While scalable, these assays are fundamentally descriptive, and do not directly measure enhancer activity.

1.4 MASSIVELY PARALLEL REPORTER ASSAYS

Chapter 1.4 is adapted with minimal modifications from:

> Klein, J.C., Chen, W., Gasperini, M., and Shendure, J. (2018). Identifying novel enhancer elements with CRISPR-based screens. ACS Chemical Biology *13* (2), 326-332.

The second approach, massively parallel reporter assays (MPRAs), was developed in our lab and builds on the *in vitro* reporter assays described in Chapters 1.1 and 1.2. However, instead of measuring reporter-gene activity of a single candidate, MPRAs rely on sequencing-based quantification of thousands of RNA barcodes in parallel, each associated with a different candidate enhancer. Patwardhan *et al.* first applied the approach to promoter sequences (Patwardhan et al., 2009), and she and others soon adapted it to screen for enhancers by utilizing a plasmid with a minimal promoter (Melnikov et al., 2012; Patwardhan et al., 2012). This approach has allowed researchers to scale traditional reporter assays to ask genome-wide questions.

MPRAs, and a recent derivative, STARR-seq (Arnold et al., 2013), have been used for saturation mutagenesis of known promoters and enhancers (Patwardhan et al., 2009, 2012; Melnikov et al., 2012), dissecting enhancer logic (Smith et al., 2013; Nguyen et al., 2016),

identifying functional bases and motifs within enhancers (Kheradpour et al., 2013; Ernst et al., 2016), testing the effects of variants on enhancer function (Vockley et al., 2015; Tewhey et al., 2016; Ulirsch et al., 2016), genome annotation (Arnold et al., 2013, 2014), and validating biochemically-predicted enhancer elements (Kwasnieski et al., 2014; Inoue et al., 2016).

While MPRAs directly measure enhancer activity and are scalable, they traditionally test short DNA fragments on a plasmid with a minimal promoter, therefore missing the extended sequence context, including its chromatin landscape and any pairing of the candidate enhancer with its endogenous promoter across a long distance. Our lab recently collaborated to describe a strategy ('lentiMPRA') wherein libraries are integrated randomly to the genome (Inoue et al., 2016), addressing some but not all of the limitations of MPRAs.

During my thesis, I have taken two approaches to address the lack of sequence context in MPRAs. This is a particularly strong concern as many biochemical marks associated with enhancers are several kilobases in length, while most MPRAs screen sequences <200 bases. The first approach is to tile across a putative element with overlapping short sequences. When an enhancer is in fact <200 bases, this approach identifies the active components (Chapter 3). However, some enhancers are likely larger than 200 bases, and cannot be recapitulated by short DNA sequences. To overcome this, I have developed a method for assembling short array-synthesized DNA oligonucleotides into libraries of larger sequences, up to 678bp each (Chapter 2).

1.5 CRISPR-BASED SCREENS FOR ENHANCERS

Chapter 1.5 is adapted with minimal modifications from:

> Klein, J.C., Chen, W., Gasperini, M., and Shendure, J. (2018). Identifying novel enhancer elements with CRISPR-based screens. ACS Chemical Biology *13* (2), 326-332.

Rather than testing candidate enhancer elements for their positive activity when removed from context, a complementary approach is to perturb these same sequences in their endogenous genomic locations. Until recently, these types of experiments in the laboratory have been limited by the difficulty of engineering targeted perturbations. In 2013, two groups adapted the bacterial CRISPR/Cas9 innate immune system to selectively create targeted double-stranded breaks in mammalian genomes (Cong et al., 2013a; Shalem et al., 2014). Due to the ease of cloning the guide sequences, which target the CRISPR complex to specific positions in the genome, Shalem et al. applied CRISPR as a genome-wide screen for essential genes in 2014 (Shalem et al., 2014). This chapter will focus on a new wave of literature implementing the CRISPR/Cas9 system to perturb and screen thousands to millions of genomic bases for enhancer function in their endogenous sequence contexts.

The relatively short guide sequences used to target Cas9 to its desired target sites facilitates large scale knock-out screens. Unlike previous genome engineering techniques such as zinc fingers and TALENs, which are laborious to synthesize, tens-of-thousands of guide RNA (gRNA) sequences can be synthesized in parallel by array-based oligonucleotide synthesis. These gRNAs are then cloned into a library of lentiviral vectors, which each deliver one gRNA into Cas9-expressing cells (Shalem et al., 2014). Each cell receives its own specific gRNA and resulting programmed mutation; when a functional selection is applied on the diverse pool of cells, the relative abundance of each gRNA, before and after selection, can be readily quantified by sequencing.

In 2015, Canver et al. tiled individual gRNAs across a non-coding region to look for *cis*-regulatory elements (Canver et al., 2015). A single Cas9 cleavage followed by non-homologous end joining (NHEJ) results in a spectrum of insertions and deletions (indels) at the target site.

These individual gRNA tiling experiments assume that short indels, when occurring at a critical location, will disrupt enhancer function. Canver et al. tiled 3 DHS sites within a previously described *BCL11A* composite enhancer, totaling 3,917 nucleotides, with 1,130 gRNAs. Since reduction of *BCL11A* results in an increase in HbF, the group sorted cells on HbF level, and examined the prevalence of each gRNA in high versus low HbF populations. The vast majority of gRNAs showed neither enrichment nor depletion, but the approach revealed discrete genomic loci within DHSs that carry clusters of enriched gRNAs in the HbF high population, indicating that when disrupted, these loci reduce *BCL11A* enhancer activity.

Shortly after, several similar screens were published, utilizing up to 18,000 gRNAs to examine 715kb of genomic sequence (Diao et al., 2016; Korkmaz et al., 2016; Rajagopal et al., 2016; Sanjana et al., 2016). A nuance of each study is how it enriches for gRNAs impacting enhancer function. Canver et al. sorted cells based on HbF expression and studied an enhancer, that when diminished, increased HbF. Korkmaz et al. focused on identifying enhancers regulated by p53 and ERα and selected based on oncogene-induced senescence and ERα expression (Korkmaz et al., 2016). Sanjana et al. focused on identifying enhancers surrounding *NF1, NF2* and *CUL3* and selected based on Vemurafenib resistance (Sanjana et al., 2016). All of these assays are limited to only screen for enhancers affecting a particular gene or pathway. Rajagopal et al. and Diao et al. developed more generalizable strategies by creating fluorescently-tagged reporter cell lines and screening for enhancers that modulate expression of the reporter (Diao et al., 2016; Rajagopal et al., 2016; Klann et al., 2017).

However, there are several limitations of screening for functional enhancers with short indels created by individual gRNAs. First, the small 1-20bp deletions introduced through NHEJ (Tsai et al., 2015; McKenna et al., 2016) may be insufficient to disrupt enhancer function. Both

comparisons of enhancer conservation between species (Taher et al., 2011) and reporter assays with synthetic sequences (Smith et al., 2013) have supported a "billboard" model, where at least for some enhancers, a collection of transcription factor binding sites can drive activity in different orders or orientations (Arnosti and Kulkarni, 2005). As such, small deletions may be insufficient to knock down function of some bona fide enhancers. Second, some critical sites may simply not be targeted, as current gRNA design is constrained by the location of PAM sites and other considerations. Finally, because the screens are noisy, these studies aggregate counts from multiple gRNA targets within a sliding window, reducing resolution and power.

Whereas introducing an individual gRNA can create short indels through NHEJ, two groups demonstrated in 2013 that introducing two gRNAs in close proximity can result in a drop-out of the intervening sequence (Cong et al., 2013b; Yang et al., 2013). By pairing two gRNAs on the same lentiviral vector, libraries of gRNA-pairs can create large deletions of programmed size (Aparicio-Prat et al., 2015; Vidigal and Ventura, 2015). Zhu et al. relied on this for a screen in 2016 that used paired gRNAs to program deletions of 700 human long non-coding RNAs, identifying 51 which can regulate human cancer cell growth (Zhu et al., 2016). In 2017, Diao et al. applied this type of screen to search for enhancers around the *POU5F1* locus (Diao et al., 2017), while we and colleagues applied it to scan the *HPRT1* locus (Gasperini et al., 2017).

In 2017, Diao et al. tested a library of paired gRNAs on the GFP-tagged OCT4 cell line published in 2016 (Diao et al., 2017). The study tiled 2Mb of genomic DNA in human embryonic stem cells with kilobase-sized deletions and identified 45 *cis*-regulatory elements. Gasperini et al. scanned a 206kb region around *HPRT1* with kilobase-sized deletions and selected for loss of HPRT function with 6-thioguanine (Gasperini et al., 2017). The HPRT screen deviated from the previous designs in two potentially important ways. First, by utilizing

overlapping deletions, the screen disrupted each base a median of 27 times, providing additional strength and confidence in calls. Second, in addition to sequencing the gRNAs themselves, Gasperini et al. used long-read sequencing to directly sequence editing events as part of post-screen validation. The study found that *HPRT1* was largely robust to non-coding deletions, concluding that proximal regulatory sequence was sufficient for *HPRT1* expression and direct sequencing of selected edits is important for reducing false positives in CRISPR-based screens of non-coding sequence.

A highly related approach to deletion scanning is to instead use CRISPR to modify the epigenetic landscape around candidate enhancer sequences. Since Cas9 can be targeted to almost any region of the genome, Qi *et al.* developed a catalytically inactive version of Cas9 (dCas9), which can function as an RNA-guided DNA recognition platform (Qi et al., 2013). Several groups have since fused dCas9 to repressor and activator domains to modulate expression of target genes. Here, we only focus on the effectors that have been used to target enhancers, including two repressors (KRAB, LSD1) and two activators (VP64 and p300).

The Krüppel-associated box (KRAB) domain is the most commonly used repressor for dCas9 experiments. KRAB recruits cofactors that repress transcription through histone methylation and deacetylation (Schultz et al., 2002; Sripathy et al., 2006; Groner et al., 2010; Reynolds et al., 2012; Thakore et al., 2015). Gilbert et al. targeted dCas9-KRAB fusions to promoters in order to repress gene expression (Gilbert et al., 2013). Since then, several groups have targeted dCas9-KRAB to putative enhancers in order to validate regulatory function in the genome. These studies used dCas9-KRAB to target previously described enhancers of *NANOG* (Gao et al., 2014), *OCT4* (Gao et al., 2014; Kearns et al., 2015), *TBX3* (Kearns et al., 2015) and hemoglobin subunit genes (Thakore et al., 2015). Kearns et al. compared the roles of KRAB and

lysine-specific histone demethylase1 (LSD1) when fused to *Neisseria meningitides* dCas9 to target regulatory sequences. LSD1 is a chromatin regulator that has been proposed to silence enhancers during embryonic stem cell differentiation by demethylating histone H3 on Lysine 4 or Lysine 9 (Whyte et al., 2012). While dCas9-KRAB repressed expression when targeted to promoters, proximal enhancers and distal enhancers, dCas9-LSD1 only repressed gene expression when targeted to distal enhancers. Ultimately, utilizing both of these orthogonal proteins and more may add additional sensitivity and specificity to future screens.

While the previous papers targeted known enhancers, Fulco et al. used dCas9-KRAB to scan 1.29 Mb of genomic sequence with 98,000 gRNAs around *GATA1* and *MYC* in K562 erythroleukemia cells (Fulco et al., 2016). Since *GATA1* and *MYC* affect proliferation of these cells, gRNAs that disrupt enhancer function would be depleted after cell proliferation. Through this screen, they identified 2 and 7 distal regulatory elements for *GATA1* and *MYC*, respectively. Similar to many of the deletion scans, however, this functional assay is limited to identifying regulators of genes important in proliferation, and therefore not generalizable to all loci.

In an attempt to create a more generalizable assay, Klann et al. labeled their genes of interest with a mCherry-tagged reporter, similar to Rajagopal et al. and Diao et al.'s use of GFP-tagging in individual gRNA scanning. While a fluorescent tag allows selection on a gene without a known selectable function, it is not scalable for annotating several genes, as a new cell line has to be created for each gene of interest in each cell line of interest. For example, to study all genes in five cell lines, one would need over 100,000 genetically engineered cell lines.

Xie et al. attempted to create a more scalable, generalizable assay by utilizing single cell RNA-sequencing to phenotype enhancer perturbations. Mosaic single-cell analysis by indexed CRISPR sequencing (MOSAIC-seq) uses single cell RNA-sequencing as a global measurement

of differential gene expression (Xie et al., 2017). The group designed 241 gRNAs targeting dCas9-KRAB to 71 constituent enhancers from 15 super-enhancers - large regions composed of multiple predicted enhancers. Surprisingly, only one to two enhancers within each super-enhancer significantly reduced target gene expression. The enhancers that decreased target gene expression were significantly enriched for RNA polymerase II and p300 binding and only showed moderate enrichment for enhancer-associated histone modifications, H3K4me1 and H3K27ac. While MOSAIC-seq is much more generic than other assays, the detection limit and cost of single-cell sequencing may reduce its utility in practice to genes that are highly expressed.

Activators have also been fused to dCas9 in order to increase target gene expression. Similar to KRAB, dCas9-VP64 was first used to target promoter regions. Both Gao et al. and Hilton et al. targeted distal enhancers with dCas9-VP64 and showed moderate gene activation (Gao et al., 2014; Hilton et al., 2015). Simeonov et al. recently screened for novel enhancers of *CD69* and *IL2RA* using dCas9-VP64. The group tiled 135kb around *CD69* with over 10,000 gRNAs and 178kb around *IL2RA* with over 20,000 gRNAs and identified several dCas9-VP64-responsive elements (Simeonov et al., 2017). Hilton et al. also fused p300, the catalytic core of human acetyltransferase, to dCas9 (dCas9-$p300^{Core}$) to target enhancers (Hilton et al., 2015). dCas9-$p300^{Core}$ significantly enhanced target gene expression, potentially by means of acetylation on histone H3 on Lysine 27 (H3K27ac). dCas9-$p300^{Core}$ was highly specific and able to induce robust activation with only one guide RNA, making it a candidate tool for genome-wide screening. Recently, Klann et al. developed CRISPR-Cas9-based epigenomic regulatory element screening (CERES), which combines dCas9-$p300^{Core}$ and dCas9-KRAB to obtain both gain and loss of function information by targeting the same regions with a repressor and an

activator (Klann et al., 2017). They targeted a 4-Mb region including 433 DHSs surrounding *HER2* with a library of 12,189 gRNAs and measured *HER2* expression using immunofluorescence staining. Loss and gain of function assays were performed in two different cell lines - A431 epidermoid carcinoma cells with moderate *HER2* expression and HEK293T cells with low *HER2* expression, respectively. With the same gRNA library, results from A431 and HEK293T cells were highly correlated with mirrored effects, same trend but opposite direction, providing additional confidence, which is critical for regions with small effects on transcription.

Unfortunately, we are not quite ready to apply these screens to entire genomes or to multiple cell-types, which each provide a unique *trans* environment for enhancers to act. First, we need a better understanding of the relative sensitivities and specificities of the different CRISPR-based screens. However, for a true comparison, we need studies that screen the same locus or loci with multiple assays. Although it will be a valuable dataset for this nascent field moving forward, such a systematic comparison has yet to be conducted or published. Moreover, in order to reach these goals, non-coding CRISPR screens must become more generalizable, higher throughput and less expensive. The GFP and mCherry-tagging screens described above could in theory be applied to every gene, but would require the generation of thousands of different engineered cell lines followed by thousands of independent experiments. Coupling CRISPR-based modifications or perturbations to single-cell RNA sequencing, as performed in MOSAIC-seq, provides a widely applicable assay, but has its own limitations. Due to the relatively low depth of current single cell RNA sequencing protocols, assaying genes that are not highly expressed would require deep sequencing of many cells. Methods such as Drop-seq and single cell combinatorial indexing RNA-seq (sci-RNA-seq) are making single cell RNA readouts

more scalable (Macosko et al., 2015; Cao et al., 2017). As these techniques reduce cost and increase scalability, methods like MOSAIC-seq will become more feasible to perform genome-wide or on multiple cell types.

We emphasize that CRISPR screens (which try to knock out or modulate native sequences, in context) and MPRAs (which test large numbers of sequences for positive activity, independent of context) are complementary rather than competing. While CRISPR-based screens have the advantage of detecting enhancers in their endogenous locations, they lack the resolution needed to routinely screen large numbers of sequences or sequence variants for their isolated/independent effects on expression. CRISPR-screens do not remove sequences being tested from their broader genomic context, which is either a disadvantage or advantage, depending on the question that one is asking. There are likely currently underexplored opportunities to synergize MPRAs and CRISPR-based screens, *e.g.* to quantify the effects of variants within CRISPR-verified enhancers, or to understand the extent to which enhancer effects are dependent or independent of the broader sequence context. As such, MPRA methods remain indispensable for testing synthetic sequences and sequence variants, which either do not occur in the genome or would be difficult to introduce in high-throughput by genome engineering.

The scalability of CRISPR has opened the door for novel screens and assays to dissect the non-coding genome. Together with MPRAs, we predict that CRISPR screens will play a critical role in elucidating how and when genes are expressed, which remains one of the most important and difficult questions in genomics. In the near term, these assays will improve our ability to annotate the genome for regulatory elements, a critical challenge in the wake of ENCODE as we are currently awash in unverified noncoding regulatory elements based on

biochemical marks. The validation of these annotations (or their failure to validate), ideally through some combination of CRISPR and MPRA-based assays, is a critical next step for the field of regulatory genomics. Of note, CRISPR based screens in cell lines, although 'in context' in terms of genomic sequence, remain 'out of context' in terms of endogenous development. *In vivo* studies – also enabled by CRISPR but still low-throughput – will remain a critical tool as well.

1.6 ENHANCERS IN EVOLUTION

Over 40 years ago, both Britten and Davidson as well as King and Wilson proposed that changes in gene regulation account for a greater proportion of phenotypic evolution in higher organisms than changes in protein sequence (Britten and Davidson, 1971; King and Wilson, 1975). Since then, researchers have characterized the roles of several enhancers in evolution through systematic analysis of single enhancers. For example, Wittkopp *et al.* showed that expression of the *yellow* and *ebony* genes correlate with a *Drosophila biarmipes* male-specific pigmentation pattern on the wing (Wittkopp et al., 2002). Shortly after, the group demonstrated that the evolution of the spot resulted from modifications of an ancestral enhancer of *yellow*, which had gained multiple transcription factor binding sites in *Drosophila biarmipe* (Gompel et al., 2005). Modification of *yellow* expression in the wing resulted in a novel evolutionary phenotype, which affects how animals interact with the environment.

Two other examples demonstrate the utility of enhancers as a mechanism for phenotypic novelty. Both pelvic fin reduction in sticklebacks and lactase persistence in humans occurred independently in different populations. Chan *et al.* showed recurrent deletions in a tissue-specific enhancer of the *Pituitary homeobox transcription factor 1 (Pitx1)* gene in pelvic-reduced populations (Chan et al., 2010). The locus showed positive selection in populations undergoing

pelvic reduction, indicating that the modulation of the *Pitx1* enhancer was an efficient mechanism for natural selection to reduce the pelvic fin, and allow freshwater sticklebacks to avoid grasping insects. Similarly, multiple groups have identified recurrent mutations within the intron of *minichromosome maintenance 6 (MCM6*), which increase transcription from the *lactase* promoter resulting in lactase persistence (Bersaglieri et al., 2004; Tishkoff et al., 2007)

Recent studies have attempted a more global approach to characterizing the role of enhancer evolution in speciation. Kunarso *et al.* identified species-specific transposable elements that altered OCT4 and NANOG binding in human and mouse embryonic stem cells (Kunarso et al., 2010). Mikkelsen *et al.* performed ChIP-seq on human and mouse preadipocytes and adipocytes and showed turnover of enhancers and transcription factor motifs (Mikkelsen et al., 2010). Cotney *et al.* compared H3K27ac (a mark for active promoters and enhancers) in human, rhesus and mouse embryonic limb, showing that 11% of enhancers gained H3K27ac in humans since the split from rhesus (Cotney et al., 2013). Villar *et al.* and Trizzino *et al.* both used H3K27ac ChIP-seq to show large-scale turnover of enhancers in primate liver (Villar et al.; Trizzino et al., 2017). Arnold *et al.* took a different approach, screening fragmented regions of the genomes of five different Drosophila species for enhancer activity in a massively parallel reporter assay, STARR-seq, and also showed large-scale turnover (Arnold et al., 2014).

1.7 ENHANCERS IN DISEASE

While the genes underlying over 4,000 Mendelian phenotypes have been discovered (Chong et al., 2015), most inherited human diseases do not result from classical Mendelian inheritance. To date, characterizing the functional genetic variation leading to complex traits has remained one of the largest hurdles to genomic medicine. One common approach to address this

question is the Genome-Wide Association Study (GWAS), in which a large number of patients and controls are genotyped to identify common variants associated with a trait. To date, over 1,600 genome-wide association studies have linked common genetic variation to many traits (McCarthy, Mark I et al., 2008). Currently, the GWAS catalog contains more than 18,000 unique risk alleles for hundreds of complex traits and common diseases (Welter et al., 2014).

However, identifying causal variants from GWAS remains a challenge for two main reasons:

a) many variants fall in non-coding DNA (approximately 96%), and it remains a challenge to predict the effect of non-coding variants (Zhang and Lupski, 2015)
b) Significant SNPs are often in tight linkage with neighboring polymorphisms, leading to hundreds of possibly causal SNPS for each GWAS hit

There is growing evidence that some of these non-coding GWAS hits may be functional by disrupting enhancer activity. In fact, multiple enhancer mutations have been linked to disease (Noonan and Mccallion, 2010). More recently, several studies have identified enrichment of non-coding GWAS variants in cell-type specific enhancers (Ernst et al., 2011; Maurano et al., 2012) and have shown that SNPs in regulatory DNA contribute to variation in gene expression (Degner et al., 2012). Some of these alleles have been functionally validated (Edwards et al., 2013). Vockley *et al.* screened hundreds of polymorphic regions from human populations in a massively parallel reporter assay in order to identify common polymorphisms altering enhancer activity (Vockley et al., 2015). Tewhey *et al.* also screened over 32,000 variants from cis-expression quantitative trait loci (eQTLs) for differential enhancer activity in a reporter assay, identifying over 800 with differential activity (Tewhey et al., 2016). These results suggest that regulatory mutations may contribute towards many complex traits.

1.8 TOPICS IN THIS DISSERTATION

The following chapters of this dissertation describe three projects I have worked on to synthesize and test libraries of enhancer elements. Chapter 2 describes a method to assemble libraries of long DNA sequences, termed Multiplex Pairwise Assembly (MPA), and a recent derivation, Hierarchical Multiplex Pairwise Assembly (HMPA). This technology will help to overcome one limitation of several enhancer assays by allowing researchers to test more sequence context around the region of interest. The next two chapters describe applications of Massively Parallel Reporter Assays (MPRAs), in which I screen thousands of putative enhancer elements. Chapter 3 utilizes computational sequence reconstruction and MPRAs to functionally characterize how enhancer elements evolved throughout the primate lineage. Chapter 4 utilizes several Genome Wide Association Studies on osteoarthritis, along with MPRAs, to prioritize functional polymorphisms that may increase a patient's susceptibility to osteoarthritis. All chapters have been modified from manuscripts in preparation (Chapter 4), under review (Chapter 3), or published (Chapter 2).

Chapter 2. MULTIPLEX PAIRWISE ASSEMBLY OF ARRAY-DERIVED DNA OLIGONUCLEOTIDES

Chapter 2 is adapted with modifications from:

> Klein, J.C., Lajoie, M.J., Schwartz, J.J., Strauch, E., Nelson, J., Baker, D., and Shendure, J. (2015). Multiplex pairwise assembly of array-derived DNA oligonucleotides. Nucleic Acids Research *44*, e43.

2.1 ABSTRACT

While the cost of DNA sequencing has dropped by five orders of magnitude in the past decade, DNA synthesis remains expensive for many applications. Although DNA microarrays have decreased the cost of oligonucleotide synthesis, the use of array-synthesized oligos in practice is limited by short synthesis lengths, high synthesis error rates, low yield, and the challenges of assembling long constructs from complex pools. Towards addressing these issues, we developed a protocol for multiplex pairwise assembly of oligos from array-synthesized oligonucleotide pools. To evaluate the method, we attempted to assemble up to 2,271 targets ranging in length from 192-252 bases using pairs of array-synthesized oligos. Within sets of complexity ranging from 131-250 targets, we observed error-free assemblies for 90.5% of all targets. When all 2,271 targets were assembled in one reaction, we observed error-free constructs for 70.6%. While the assembly method intrinsically increased accuracy to a small degree, we further increased accuracy by using a high throughput "Dial-Out PCR" protocol, which combines Illumina sequencing with an in-house set of unique PCR tags to selectively amplify perfect assemblies from complex synthetic pools. This approach has broad applicability to DNA assembly and high-throughput functional screens.

2.2 INTRODUCTION

Traditionally, DNA has been synthesized by solid-phase phosphoramidite chemistry (Beaucage and Caruthers, 1981). Column-based synthesis generates up to 200-mers with error rates of about 1 in 200nt and yields of 10 to 100nmol per product (Kosuri and Church, 2014). Column-based DNA synthesis is limited in throughput to 384-well plates (Kosuri and Church, 2014), and oligos cost from $0.05 to $1.00/bp depending on length and yield (Kong et al., 2007; Kosuri et al., 2010; Kosuri and Church, 2014). The commercialization of inkjet-based printing of nucleotides with phosphoramidite chemistries (Blanchard et al., 1996; Hughes et al., 2001; Saaem et al., 2010) (Agilent) and semiconductor-based electrochemical acid production arrays (Ghindilis et al., 2007) (CustomArray) have increased throughput and decreased the cost of oligo synthesis. These oligos range from $0.00001-0.001/bp in cost, depending on length, scale and platform (Kosuri and Church, 2014). However, these platforms are limited by short synthesis lengths, high synthesis error rates, low yield, and the challenges of assembling long constructs from complex pools.

Many methods have recently addressed the high error rates of array-synthesized oligos, with a trade-off between cost and fidelity. Low-cost methods include proteins such as MutS (Carr et al., 2004; Wan et al., 2014), polymerases (Smith and Modrich, 1997; Young and Dong, 2004; Binkowski et al., 2005; Fuhrmann et al., 2005; Bang and Church, 2008), and other proteins that bind and cut heteroduplexes (Kosuri et al., 2010; Dormitzer et al., 2013). However, as these methods rely on identifying mismatches and require the majority of sequences to be identical, they are not always compatible with complex libraries (Matzas et al., 2010; Schwartz et al., 2012) and therefore must be performed after individual gene assemblies. Furthermore, as these methods retain error rates as high as 1 per 1,000 bases, further screening is required to confirm

the correct sequence. More recent methods such as Dial-Out PCR rely on DNA sequencing followed by retrieval of sequence-verified constructs, achieving error rates as low as 10^{-7} (Matzas et al., 2010; Schwartz et al., 2012; Kim et al., 2018). While these methods can work on complex oligo pools and yield very low error rates, they are costly, time-intensive, and do not always recover targeted molecules.

Despite their high error rates, inexpensive oligo pools cleaved from microarrays have recently enabled high-throughput analysis of promoter (Patwardhan et al., 2009; Schlabach et al., 2010; Sharon et al., 2012) and enhancer function (Melnikov et al., 2012; Smith et al., 2013), providing novel insight into the vocabulary of these regulatory elements. They have also been used in deciphering the role of genetic variants in protein function (Findlay et al., 2014). However, these studies were all limited by short synthesis lengths—about 160bp for CustomArray and 230bp for Agilent.

To our knowledge, Tian et al. was first to perform gene synthesis from pools of array-derived oligos (Tian et al, 2004). Since array synthesis only provides yields of 1-10fmol per sequence (Quan et al., 2011), Tian et al. amplified all oligos with a common set of primers. However, the study limited synthesis to 21 genes in order to circumvent high synthesis error rates and the challenges of assembling constructs from complex pools (Tian et al., 2004; Zhou et al., 2004; Linshiz et al., 2008; Borovkov et al., 2010). To address this, Kosuri *et al.*, demonstrated pre-amplifying subsets of the oligo-pool involved in specific assemblies, to reduce the spurious cross-hybridization observed in large-scale assemblies (Kosuri et al., 2010). The study relied on amplifying fragments for each gene separately, which was successful but limited throughput to the assembly of 47 genes. In 2012, Kim et al. described shotgun synthesis on 228 array-derived oligos spanning the penicillin biosynthetic gene cluster (Kim et al., 2018). Similar

to Tian et al., Kim amplified oligos with universal primers, and removed adaptor sequences with two restriction enzymes. After assembling all oligos via PCR, they selected for fragments between 300-500bp. While successful, only 3% of sequenced products were error-free. In order to retrieve error-free constructs, they barcoded and sequenced their pool, identifying accurate fragments covering 88% of their targets. They then ordered primers corresponding to the barcodes to retrieve the fragments of interest.

Short synthesis lengths and high error rates present bottlenecks to the use of array-derived oligos for both functional assays and gene assembly. Here we describe a method to assemble thousands of array-derived oligos into targets approaching length estimates of cis-regulatory elements (Levine, 2003; Kristiansson et al., 2009) and protein domains (Xu and Nussinov, 1998). Compared to existing methods, our method does not limit sequence space by using restriction enzymes, it is high throughput, and it offers an efficient way to retrieve error-free assemblies.

2.3 RESULTS

2.3.1 *Assembling targets in sets of 131-250*

We designed *in-silico* 2,271 targets ranging from 192-252 bases (156-216 of unique sequence) to assemble from array-derived oligos. All targets consisted of a unique sequence flanked by the same 18bp 5' and 3' common adapters. Each target sequence was split into two fragments, A and B, containing an overlap region with a Tm >56°C. The 2,271 target sequences were split into 10 sets of 131-250 targets, and each set received unique adapters flanking the 3' end of the A fragments and the 5' end of the B fragments designed for uracil incorporation

(**Figure 2.1**). The corresponding oligos (160-mers with buffer sequence) were synthesized by CustomArray in duplicate to reduce oligo dropout and increase uniformity.

We first amplified each pool of oligos off the array with a sub pool specific primer (A fragment uniqueF or B fragment uniqueR) on one end and a common primer (YF/YR) on the other (**Table 2.1**). Sequencing of the oligo library showed good uniformity, with an interquartile range of 5.5 (**Figure 2.3A**).

The oligo pools provided by CustomArray were then amplified using either Uracil-containing A fragment primer and YF or Uracil-containing B fragment primer and YR (**Table 2.1**), and the corresponding specific adapters were removed with Uracil Specific Excision Reagent (USER). We tested amplifying oligos with either one or two unique primer sites, and observed no difference in assembly composition or uniformity. The corresponding A and B fragments were mixed for each set of targets and assembled through five cycles of annealing with extension and approximately 25 cycles of amplification with Kapa HiFi. In all cases, the correct size band was observed. Each assembled set was barcoded and sequenced.

For each set, we identified error-free assembled constructs for 72.7-96.4% of targets at a sequencing depth of 90,000 reads (**Figure 2.3B**). For each target, we examined the number of error-free reads for the corresponding A and B oligos (out of 1.2 million reads). Of the 223 targets that failed to yield error-free assemblies, 55 (24.7%) fell in the bottom 10^{th} percentile of limiting oligo concentration (<6 error-free reads out of 1.2 million) and 97 (43.5%) fell in the bottom 20^{th} percentile of limiting oligo counts (<11 error-free reads out of 1.2 million). **Figure 3C** shows higher yield (% of targets with at least one perfect assembled sequence) for targets assembled from better-represented limiting oligos in the array pool, suggesting that increasing oligo uniformity would likely improve the yield of full-length designs. We next looked at the

composition of the raw oligo pools and the assembled target libraries (**Figures 2.3D and 2.3E**). 23.8% of molecules represented error-free assemblies, 36.2% contained indel-free assemblies, and 53.4% contained small indels (<5bp). An additional 2.3% contained large indels (>5bp), 4.8% contained chimeras, 2.1% contained truncated constructs and 0.6% unmapped reads. Within each set, six of ten sets had <15-fold difference in the interquartile range. While this may be an issue for some applications, the uniformity is tight enough to use the sets directly for some downstream screening applications, such as functional protein screens. Uniformity plots are shown in **Figure 3F**.

In summary, of the 2,271 targets synthesized in individual sets, we assembled error-free constructs for 2,055 (90.5%). Much of the drop-out appears to be due to poor representation of the corresponding oligos in the array pool (**Figure 2.3C**). Additionally, the majority of errors identified in the assembled sets are likely from the array-synthesis, since similar error profiles are identified in the oligo pools (**Figure 2.3D and 2.3E**). Chimeric assembly (assembly of the wrong A and B fragments) is rare.

2.3.2 *Multiplex assembly of 2,271 pairs of fragments*

To test the limitations of our assembly protocol, we increased complexity by adding one additional set at a time, up to a complexity of 2,271 designs. At a complexity of 2,271 we assembled error-free constructs for 70.6% of targets (**Figures 2.4A and 2.4B**). We observed an even greater correlation between yield and representation of the limiting oligo in the array pool compared to the smaller sets (**Figure 2.4C**).

We were also interested in whether increasing complexity would affect the composition of assembled libraries. While the two lowest complexity sets (250 and 462 targets) show the highest percentage of perfect and indel-free reads, it is likely due to the fact that these two sets

are composed of sets 2 and 3, which individually showed high percentages of perfect and indel-free reads (**Figure 2.3E**). The remaining libraries all share similar compositions. For all complexity levels, 11.8-31.3% of reads represented perfect constructs, 10.0-18.7% represented constructs with mismatches only, 41.4-48.5% represented small indels, 2.6-3.5% represented large indels, 3.7-21.5% represented chimeras, 2.5-4.9% represented truncations and 0.1-0.7% unmapped reads (**Figure 2.4D**). Within each set, there was a 10 to 34-fold difference in the interquartile range. Uniformity plots are shown in **Figure 4E**.

2.3.3 *Error correction of assembled targets*

Oligo pools were sequenced and aligned to a reference of intended target sequences. For error analysis, we chose to examine one set of 250 targets, each 237 bases long (set 5). We calculated average nucleotide accuracy from bases with aligned reads having quality mapping score >20. We identified a 98.68% average nucleotide accuracy of oligos after amplification off the first array. Since our assembly process relies on two priming sites and an overlap region, we hypothesized that assembly might intrinsically increase accuracy in these regions. Indeed, we found that the average nucleotide accuracy of all aligning molecules in the 250-plex reaction was 99.02% (Poisson rate ratio 95% CI 1.36-1.38), showing the highest accuracy around the two priming sites and overlap region (**Figure 2.5A**). In particular, the average nucleotide accuracy for the overlapping region increased from 98.53 to 99.44% (Poisson rate ratio 95% CI 2.64-2.77).

While we see a significant increase in accuracy at the nucleotide level (p~4.9e-324), we were still limited to a maximum of 37% perfect reads in an assembled set. For downstream applications relying on accurate molecules, such as gene assembly, we were interested in

retrieving perfect assemblies from our assembled sets. To do so, we modified the Dial-Out PCR protocol (Schwartz et al., 2012) to incorporate a set of in-house static Dial-Out tags to allow for cost-efficient PCR retrieval of sequence-verified constructs.

We designed primers that append M13F and M13R during the assembly reaction for targets from sets 2 and 6 (each 250 targets). The assembled libraries were then tagged with the static Dial-Out tags, and sequenced for verification. We first analyzed the distribution of tag pairs, and found that 84.0% and 85.6% of all molecules in assembled and tagged sets 2 and 6 containing a unique, retrievable tag pair (out of 1.3mil reads for set 2 and 1.6 mil reads for set 6) (**Figure 2.5B**). 98.4% and 95.6% of targets had a sequence-verified assembly with a unique tag pair.

From set 2, we chose 25 targets to retrieve, each of which was represented in at least 5 out of 1.3 million reads. All 25 targets amplified, and we evaluated retrieval accuracy by pooling all 25 retrieval reactions together and sequencing them with 1 million reads. All 25 targets were sequenced between 8,600 and 62,000 times, revealing error-free reads to the detection limit of Illumina sequencing chemistry, which is more quantitative than Sanger sequencing (**Figure 2.5C**). 78% of all sequencing reads aligned to one of the 25 targets, suggesting some background amplification of the deep sequencing library.

2.3.4 *MPA using longer, more accurate oligos*

Since we noticed an improvement in uniformity from duplicating oligonucleotides during synthesis, we hypothesized that higher quality oligonucleotides would yield a higher quality assembly. To test this, we ordered the same set of oligos for the 2271-plex assembly from Twist, another array-synthesized oligonucleotides manufacturer. We noticed an improved uniformity of oligos from Twist (**Figure 2.6A**), as well as a higher fraction of perfect and indel-free constructs

(**Figure 2.6A**). After assembly, we were able to synthesize 2160 out of the 2271 targets (95.1%) in a single reaction with the Twist oligos, compared to 1603 (70.6%) from our original Custom Array oligos. We achieved much better uniformity of assembled constructs (**Figure 2.6C**) as well as a higher percentage of perfect and indel-free assemblies (**Figure 2.6D**).

Currently, the longest array-synthesized oligos are 230mers offered by Agilent. In order to demonstrate that we could assemble longer constructs, and that our protocol is robust to manufacturer, we designed a set of 2336 354bp target sequences and ordered them as split 230mers in duplicate from Agilent. We assembled all 2,336 targets in a single reaction, identifying error-free assemblies for 2188 out of the 2336 targets (93.66%) with only 100,000 sequencing reads. Again, we obtained great uniformity, with an interquartile range of 5.9 (**Figure 2.6E**).

2.3.5 *Hierarchical MPA*

As our assembly uniformities with Twist and Agilent are similar to the oligo uniformity from Custom Array, we wanted to test whether we could perform our assembly protocol in a hierarchical fashion. In order to do so, we modified our protocol such that each target is split into 4 fragments – A, B, C, and D. We then assembled A and B, while incorporating uracils on the 3' end, and C and D, while incorporating uracil nucleotides on the 5' end. We then repeated our assembly by performing a USER digest, End Repair, and Assembly, achieving libraries of 678bp targets.

For a pilot, we designed oligos for 2,064 putative enhancers, each 678bp, in HepG2. We split the 2,064 designs into 12 sets, each with 172 targets, and ordered the A, B, C, and D fragments for each set of targets on an Agilent 230-mer array. For each set of 172 target enhancers, we synthesized 86.0-96.5%% of targets with an interquartile range from 21.17-39.45

(**Figure 2.7**). Pooled together *in silico* (combining equal number of reads from each set), we assembled 95.0% of 2,064 targets with an interquartile range of 30.08.

2.4 DISCUSSION

We sought to develop a protocol to overcome the limitations of array-derived oligonucleotides for library generation and gene assembly. Our method relies on multiplex pairwise assembly and Dial-Out molecular tagging. This method produced sequence-verified, individually retrievable 192-252-mers from CustomArray and Twist oligonucleotides and 354-mers from Agilent oligonucleotides. It also produced 678-mers from Agilent oligonucleotides after a hierarchical assembly.

First, we tested multiplex pairwise assembly in sets of 131 to 2,271. In addition to perfect sequences, the composition of the assembled sets consisted of mostly small indels and mismatches, similar to the raw array-derived oligo pools. The composition of sets did not change noticeably with complexity beyond 712 targets, suggesting that increasing the number of targets per reaction does not strongly alter the amount of resulting chimeras or error-containing assemblies. While we observed reduced yield with increasing complexity, we were still able to assemble 70.6% of all targets in a 2,271-plex reaction. By parallelizing similar complexity reactions in a 96-well plate, we could theoretically assemble a set of 200,000 constructs with 70% yield. If the experiment relies on representation of all targets, our data suggests that uniformity can be improved by performing assembly in sets of 250, to achieve >90% yield.

The main limitations in our protocol currently are the relatively high DNA synthesis error rate (e.g., mismatches and indels), moderate DNA assembly error rate (e.g., chimeras), and low uniformity. Low uniformity of input oligos impairs target uniformity in assembled sets. This is apparent in **Figure 8C**, as well as a separate array in which oligos were not duplicated (**Figure**

2.8). We therefore suggest that for increased yield and uniformity, all oligos be duplicated during synthesis. We were further able to demonstrate this point by ordering the same oligos from a different vendor (Twist), with higher uniformity and lower errors. The resultant assembly yielded over 95% of targets, compared to 70%. When assembling 2336 targets from even longer Agilent oligos, we were able to assembly over 96% of targets in a single reaction.

High-throughput functional screens would benefit from highly accurate and uniform assemblies. However, in many applications, error-containing molecules can be filtered in the analysis stage, or may provide additional diversity for directed evolution. The spread in uniformity may also be accounted for with a post-hoc analysis by normalizing a post-selection sample to a pre-selection sample. For gene assembly requiring very high accuracy, we implemented Dial-Out PCR to isolate perfect gene sequences. However, for hierarchical assembly, yield is a concern, as every fragment must be represented in order to assemble larger constructs. In applications for hierarchical gene assembly, constructs should be assembled in smaller sets, as we are able to achieve yields up to 96% in sets of 250.

We believe that with the exception of chimeras, both the high error rate and lack of uniformity are due to our input reagents, and not the multiplex pairwise assembly protocol. The error profiles of the assembled sets match closely with the profile of the raw oligos (**Figure 2.3**). In fact, we saw an increase in accuracy at priming and assembly sites from our assembly protocol. Moreover, we assembled at least one error-free sequence for each target with high representation of both oligos, suggesting that much of the dropout and uniformity issues are due to poor uniformity in array synthesis. Therefore, using a higher-fidelity and more uniform array should also reduce these limitations.

Our protocol inherently is prone to producing chimeras. While these can be filtered out in most downstream applications, they may cause issues in more complex reactions by diluting the designed library. We were able to minimize chimeras, to a maximum of 21.5%, by utilizing a custom script that examines all possible cross-hybridizations. In a separate experiment without the script, we identified chimera rates as high as 42% (**Figure 2.6**). However, since the designs were different, we cannot make a direct comparison of chimera rates.

Our protocol for multiplex pairwise assembly of array-derived DNA oligonucleotides provides a method for inexpensive, sequence-verified, oligonucleotide assembly from array synthesis. To our knowledge, this is the first study to assemble thousands of array-derived oligos in multiplex, and to use a static set of PCR tags to retrieve sequence-verified molecules. We suggest the applicability of this protocol for both complex library generation and gene synthesis. Creating a library of 3,118 such 200-mers would be ~38-fold less expensive than column-based synthesis methods (~0.84USD/target). Retrieving individual sequence-verified assemblies for each of the 3,118 would still be 17-fold less expensive with in-house Dial-Out tags and retrieval primers, and 4-fold less expensive including the one-time costs of the Dial-Out tag and retrieval primer libraries (**Table 2.2**). While column-based synthesis is limited to 200 bases, our protocol synthesized 252-mers at 0.84USD/target (0.0042USD/base) with the similar efficiency as 200-mers. With the advent of next-generation sequencing, high-throughput functional screens of DNA have shed light on the mechanisms of gene regulation (Patwardhan et al., 2009; Schlabach et al., 2010; Melnikov et al., 2012; Sharon et al., 2012; Smith et al., 2013) and the classification of variants of uncertain significance (Findlay et al., 2014). The ability to synthesize defined libraries at an unprecedented cost will allow researchers to address these questions using precisely designed sequences rather than relying on biased mutagenesis methods. Moreover,

gene synthesis has contributed to novel pharmaceuticals and a better understanding of genome organization, and we expect that increasing the length of DNA assemblies that can be produced with low-cost, high complexity DNA synthesis will provide new opportunities for synthetic biology.

2.5 METHODS

2.5.1 *Target designs*

Target sequences range from 156-216 bases of unique sequence and were split into ten sets. Each target was fragmented into two pieces (A and B) using a custom python script that determines overlaps with the least chance of cross-hybridization. Briefly, we automated the following procedure using python: bases for the overlap region were dynamically added starting from the midpoint-7 position until the melting temperature was >56°C (Allawi and Santalucia, 1997) . The overlap fragment was then checked against all sequences in the set and accepted if <15 consecutive bases aligned to any other sequence. To quickly evaluate alignments against all sequences in a given set, we utilized a simple sliding algorithm, which scores the longest consecutive alignments (Nguyen-Dumont et al., 2013). If the overlap sequence failed these conditions, we swapped out up to 6 codons at random within this sequence region, and if the melting temperature was still >56°C, we repeated the alignment step. If conditions still were not met, the starting position for the overlap region was shifted and the procedure was repeated. A window of 6 bases around the starting position was explored. A common 18bp adapter was appended to the 5' end of A fragments and 3' end of B fragments. Two adenines were appended to the 3' end of A fragments, and two thymines were appended to the 5' end of B fragments. Finally, depending on length, either one or two pool-specific primers site(s) were added to all

oligo designs, and random bases were added on the 3' side to reach 160 bases for each oligo design (**Figure 2.1**). The pools of oligos were then synthesized by CustomArray or Twist in duplicate to decrease oligo dropout and increase uniformity.

We also designed oligos to assemble 2236 354bp putative regulatory elements. We chose to synthesize 2,236 genomic sequences previously studied as 171-bp sequences in a lentiMPRA (Inoue et al., 2016). We also included the top 50 and bottom 50 haplotypes, averaging 409-bp, from a screen conducted in the STARR-seq vector (Vockley et al., 2015). For the 2,336 total sequences, we designed libraries of 192nt, 354nt, and 678nt, centered at the center position of the previously tested design. We extracted genomic sequence using Bedtools Getfasta. To the 192nt library, we added the HSS-F-ATGC adapter to the 5' end (5'-TCTAGAGCATGCACCGGATGC -3') and the HSS-R-clon adapter to the 3' end (5'-TCGACGAATTCGGCCGG- 3').

For the 354nt library, we split each sequence into two overlapping fragments, A and B. Fragment A included positions 1-190 and fragment B included positions 161-354. To fragment A, we appended the HSS-F-ATGC adapter to the 5' end and the DO_15R_U adapter (5'-AAGCTTACGGGCAACATGG -3') to the 3' end. To fragment B, we appended the DO_5F_U adapter (5'- GCGAAGTCCCTTACCCTTT -3') to the 5' end and the HSS-R-clon adapter to the 3' end.

For the 678nt library, we only designed the 2236 sequences from Inoue et al. We split the sequences into 13 different sets of 172 sequences each. We then split each sequence into 4 fragments. Fragment A included positions 1-190, fragment B included positions 161-352, fragment C included positions 323 to 514, and fragment D included positions 485-678.

2.5.2 *Pairwise oligonucleotide assembly*

Targets were separated into sets of complexity ranging from 131-250. Each pool of A and B fragments was amplified off of the array using one common primer and one pool-specific uracil-containing primer with the Kapa HiFi HotStart Uracil+ Readymix. Quantitative PCR (qPCR) was performed in 25 µl reactions with SYBR Green on a MiniOpticon Real-Time PCR system (Bio-Rad) with 2.5ng template. Each pool was pulled from the thermocycler one cycle before plateauing, purified with 1.8x AMPure XP beads and eluted in 20uL. 2uL NEB USER enzyme was mixed with the purified PCR pools, and incubated at 37°C for 15 minutes, followed by 15 minutes at room temperature. The pools were then treated with NEBNext End Repair Module per manufacturer's protocol to remove adapter sequences. The pools were purified and concentrated in 10uL using Zymo DNA Clean and Concentrator.

Corresponding A and B fragment libraries were assembled using Kapa Hifi Hotstart Readymix (Kapa Biosystems) using qPCR with a total of 1.5ng of the purified, corresponding input DNA pools and 7.5 x 10^{-12} moles of each outer primer (YF-pu1L and YR-pu1R) using the following protocol:

(i) 95°C for 2 min, (ii) 98°C for 20 s, (iii) 65°C for 15 s, (iv) 72°C for 45 s, (v) repeat steps ii-iv 35 times and (vi) 72°C for 5 min. Outer amplification primers were not added until the 5^{th} cycle of PCR, allowing for initial annealing and extension of the template strands before amplification. Reactions were removed from the thermocycler one cycle before plateauing, purified with 1.8x AMPure XP beads and eluted in 20uL.

2.0ng of the purified reaction was used in another real-time PCR with Kapa HiFi Hotstart Readymix with Pu1L_Flowcell and Pu1R_Flowcell primers. Reactions were pulled from the cycler one cycle before plateauing, purified with 1.8x AMPure XP beads, and sequenced on an

Illumina MiSeq with paired end 155bp reads with Pu1_Sequencing_F, Pu1_Sequencing_R and Pu1_Sequencing_I (**Table 2.1**). For complex sets of up to 2,271 targets, input DNA from the corresponding sub pools were mixed together, maintaining the same total amount of 1.5ng input DNA.

2.5.3 *Hierarchical pairwise oligonucleotide assembly*

All libraries were amplified off the array with KAPA HiFi HotStart Uracil+ ReadyMix PCR Kit as described above. During the first round of assembly, fragments A and B were assembled with HSS_F_ATGC and 31_PU_R and fragments C and D were assembled with 8_PU_F and HSS_R. Assembled libraries were then purified with a 0.65x Ampure cleanup following the manufacturer's protocol, and eluted in 20ul. 2ul of USER enzyme was added to the purified assembly reactions and incubated at 37°C for 15 minutes followed by 15 minutes at room temperature, and then repaired using the NEBNext End Repair Module, following manufacturer's protocol, and purified using the Zymo DNA Clean and Concentrator 5 and eluted in 10uL. All libraries were then quantified using the Thermo Fisher Scientific Qubit dsDNA HS Assay kit and eluted to 0.75ng/ul. Assemblies AB and CD were then assembled together following the protocol described in Klein *et al.* After the second assembly, libraries were purified using a 0.6x AMPure cleanup and eluted in 30uL. We then amplified 1uL of each assembly with HSS-F-ATGC-pu1F and HSS-R-clon-pu1R to add flow cell adapters and indexes.

2.5.4 *In silico design of Static Tag Library*

We generated random 13-mer sequences and screened them for several properties: no homoguanine or homocytosine stretches >5bp, no homoadenine or homothymine stretches >8bp and GC content between 45% and 65%. 13-mers passing this filter were added to a potential set

if the last 10 bases had <90% nucleotide identity with any other forward, reverse, complement, or reverse complement already in the list. This pipeline was repeated several times, ultimately with 1.2 million iterations, to generate a library of 7,411 13-mers.

The Gibbs free energy of every possible primer pair was calculated using Unafold (Markham and Zuker, 2008) with the following settings: --NA=DNA, --run-type=html, --Ct=0.000001, --sodium=0.050, --magnesium=0.002. All 13-mer pairs with dG > -9kcal/mol were indexed and added to a MatrixMarket Matrix. The maximum library of 13-mers with all pairwise dG > -9kcal/mol was then identified using the Parallel Maximum Clique Library (arXiv:1302.6256). The indexed 13-mers were converted back to their corresponding sequences, and an additional step was applied to remove any primers with potential homodimers. This left a set of 4,637 13-mers, which was split into a forward library of 2,318 tags and a reverse library of 2,319 tags, with a total tag complexity of 5,444,982 (**Figure 2.2**).

To the forward 13-mers: 5'-CGACAGTAACTACACGGCGA-3' was added to the 5' end as a bridge for the flow cell adapter, and M13 (5'-GTTTTCCCAGTCACGAC-3') was added to the 3' end as the Dial-Out seed sequence. To the reverse 13-mers: 5'-GTAGCAATTGGCAGGTCCAT-3' was used as the bridge and M13R (5'-CAGGAAACAGCTATGAC-3') was used as the seed sequence.

2.5.5 *Design and synthesis of Dial-Out Retrieval Primers*

For each 13-mer, the Tm was calculated using: $T_m = 81.5 + 16.6 \times \log_{10}[Na^+] + 41 \times (\%GC) - \frac{600}{n}$ (Sambrook et al., 1989). Primer sequences were determined by recursively adding 2bp from the bridge sequence to the 5' end of the primer until the Tm was between 58°C and 61°C. After this procedure, all primers were 17nt or 19nt long, with Tm between 58.2°C and 60.6°C. Primers were ordered from IDT in 96-well plate format with standard desalting.

2.5.6 *Static Tag Library Synthesis and Preparation*

The 4,637 tags were synthesized using CustomArray's semiconductor electrochemical process in duplicate. Forward and reverse tag sets were amplified in 24 parallel 50uL reactions from 1.25 x 10^{-14} moles template/reaction using FP: 5'- CGACAGTAACTACACGGCGA -3' and RP: 5'- GTCGTGACTGGGAAAAC -3' with Kapa Hifi Hotstart Readymix for 17 cycles. 10nmol PCR products were digested with NEB lambda exonuclease following manufacturer's protocol. 113ng sample was mixed with equivolume Novex TBE Urea Sample Buffer and heated at 70°C for 3 minutes, then chilled on ice. Samples and ladder were run on a Novex TBE Urea Gel, and the corresponding 50bp band was cut. The bands were diced and spun through a 600mL Eppendorf with a hole from a 22 gauge needle. The slurries were incubated with TE buffer at 65°C for 2 hours and purified on a Spin-X column (Corning). Purified DNA was treated with the Qiagen nucleotide removal kit per manufacturer's protocol.

2.5.7 *Tagging of assembled targets*

Several concentrations of tags and input were tested for optimal tagging with several different polymerases (**Table 2.3**). We identified that 8.5 x 10^{-14} moles of tags with 3 ng input (a 10:1 tag:input molecular ratio) with Kapa HiFi HotStart Readymix, yielded optimal performance. During the assembly process, we amplified targets with primers containing M13F and M13R, following the assembly protocol above. Libraries were purified with 1.8x AMPure XP beads and eluted in 20uL. 3ng of purified assembly library was tagged with 8.5 x 10^{-14} moles of dial-out tags (Dial-Out Tags F and Dial-Out Tags R) using Kapa HiFi HotStart Readymix using qPCR and the following cycling conditions: (i) 95°C for 2 min, (ii) 98°C for 20 s, (iii) 65°C for 15 s, (iv) 72°C for 45 s, (v) repeat steps ii-iv 30 times and (vi) 72°C for 5 min and (vi) 72°C for 5 min. After the first 5 cycles, the reaction was paused, and 1.5 x 10^{-11} moles of

barcoded forward and reverse flow-cell primers (Dial-Out_Flow_Cell_F and Dial-Out_Flow_Cell_R) were added. The tagged libraries were removed from the cycler one cycle before plateauing, and purified using 1.8x AMPure XP beads.

2.5.8 *Sequence-verification of Dial-Out tagged targets*

The tagged library was sequenced on an Illumina MiSeq with PE 155bp reads using Dial-Out_Sequencing_F, Dial-Out_Sequencing_R and Dial-Out_Sequencing_I primers. Reads were merged with PEAR using default settings, and tag pairs for all reads were identified (Zhang et al., 2014). Using a custom python script, we identified all reads containing sequence-verified constructs, and their corresponding tag pairs. One correctly-assembled molecule per target meeting the following criteria was randomly selected for retrieval: (i) containing a unique tag set not identified on any other molecule and (ii) represented in at least 5 sequencing reads.

2.5.9 *Dial-Out retrieval*

Selected oligonucleotides were retrieved via PCR with Kapa HiFi Hotstart Readymix using real-time PCR with 0.135ng template and 1.5 x 10^{-11} moles each of the corresponding forward and reverse dial-out retrieval primer with the following conditions: (i) 95°C for 3 min, (ii) 98°C for 20 s, (iii) 65°C for 15 s, (iv) 72°C for 40 s, (v) repeat steps ii-iv 34 times and (vi) 72°C for 5 min. Reactions were removed from the cycler just before plateauing, purified with 1.8x Ampure, and quantified using a Qubit (Invitrogen). Equal concentrations of each retrieval reaction were mixed for sequencing.

2.5.10 *Analysis of average nucleotide accuracy*

All sequencing reads were aligned to a reference of intended target sequences using BWA v.0.7.3. The average nucleotide accuracy was calculated from bases with aligned reads

with base and quality mapping score >20. To compare accuracy rates between experiments, we analyzed error rates for set 5 before and after assembly. We performed Exact Poisson Tests on the 15,935,028 bases of the assembled set and 9,325,493 bases of the corresponding oligo pools passing our quality cutoffs. We also performed the test on the 1,546,665 bases of the overlapping region in the assembled set and 1,617,760 in the oligo pools.

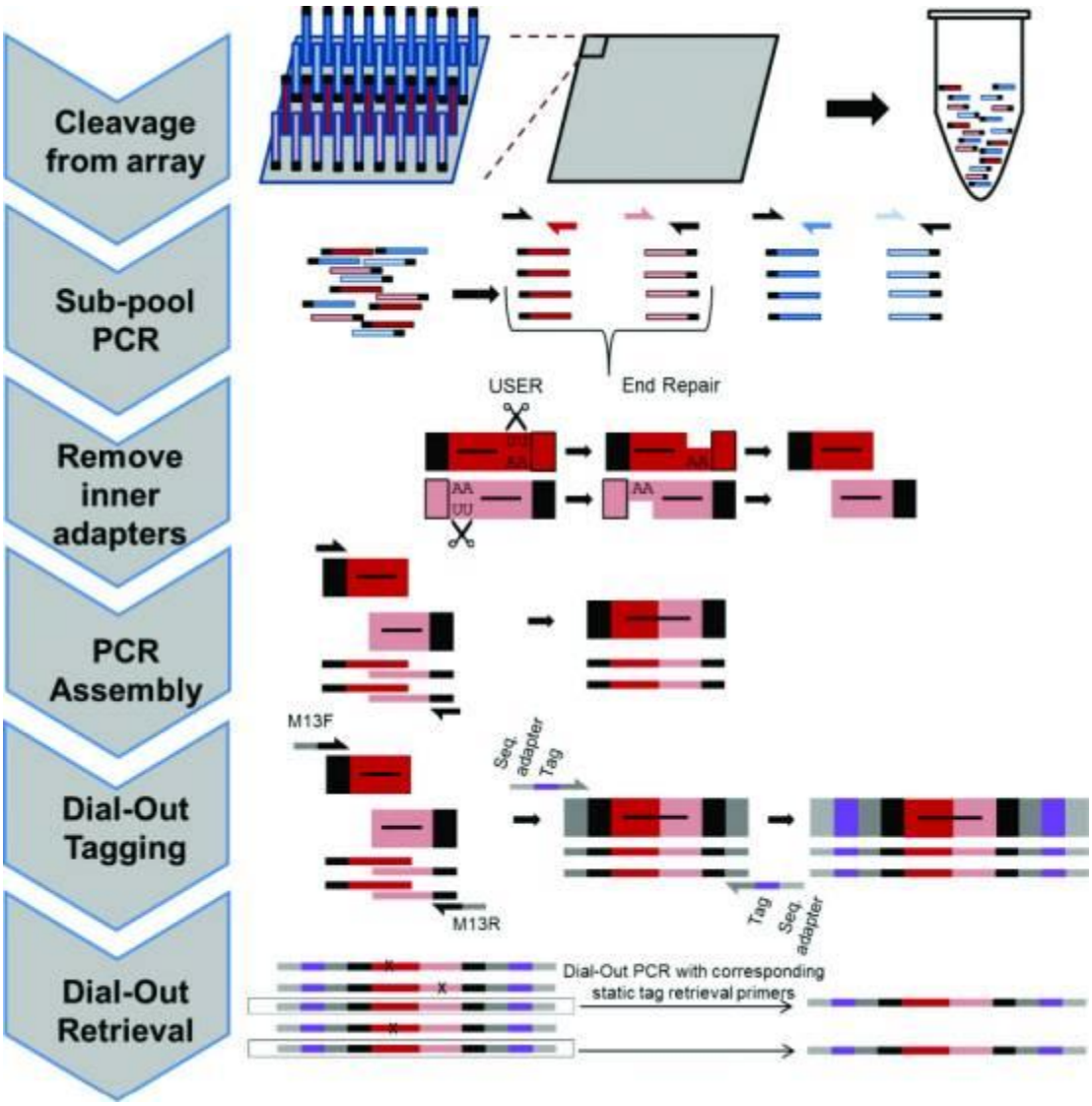


Figure 2.1. Multiplex Pairwise assembly.

A total of 2271 targets were separated into 10 sets of 131–250 genes. Each gene was split into A and B fragments with overlapping sequences providing >56°C melting temperature (Tm) for PCR-mediated assembly. All oligos were cleaved off the array into one tube. We then amplified each sub-pool with one common and one uracil-containing pool-specific primer. The pool-specific primer was then removed with Uracil Specific Excision Reagent (USER) followed by New England BioLabs End Repair kit. During PCR assembly, corresponding sub-pools were allowed to anneal and extend through 5 cycles of PCR, before adding a set of common, outer primers for amplification. During PCR assembly, M13F and M13R sequences can be introduced to the constructs in order to allow for Dial-Out Tagging and retrieval of sequence-verified constructs. In this study, we assembled up to 252-mers from 160-mer CustomArray oligos.

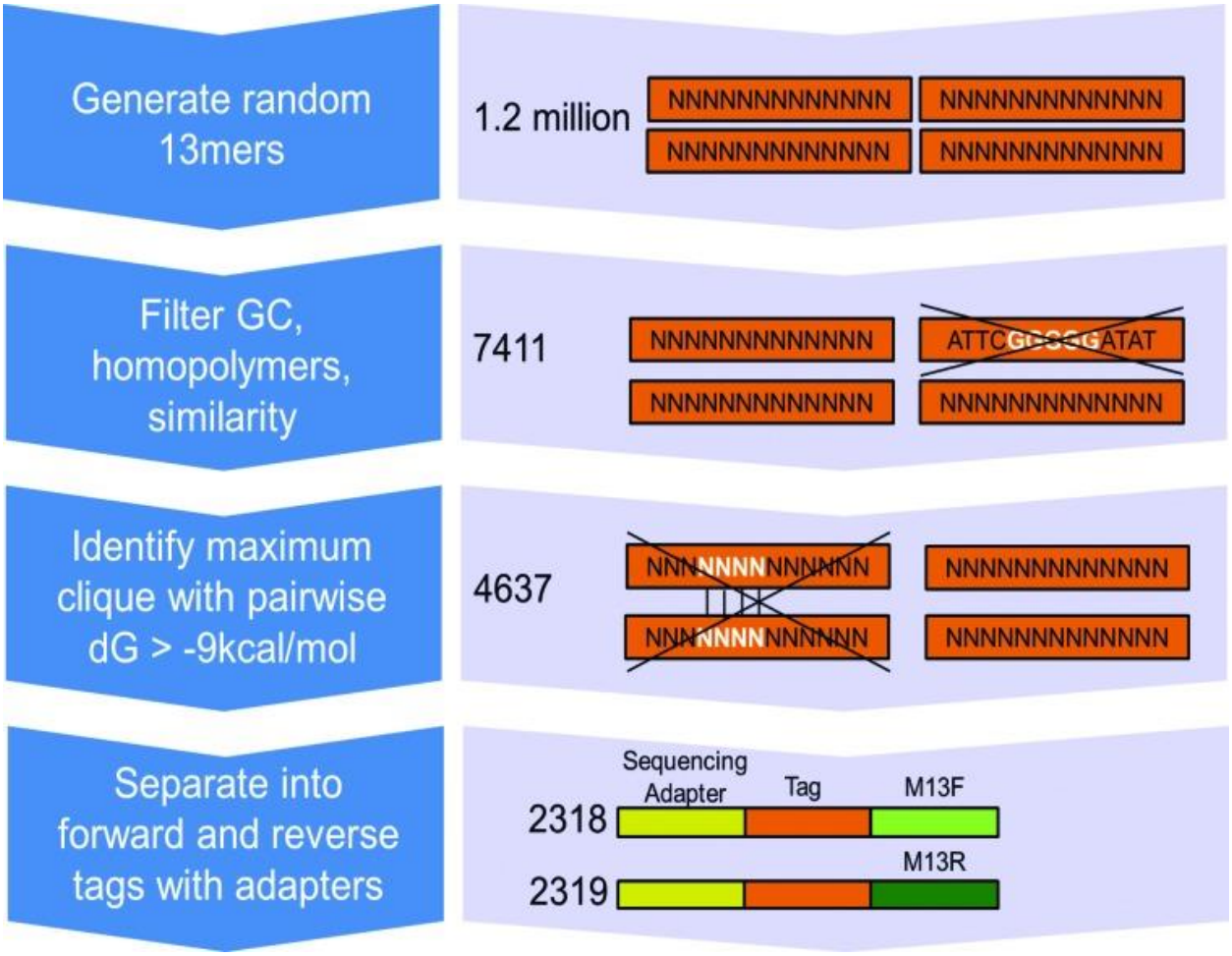
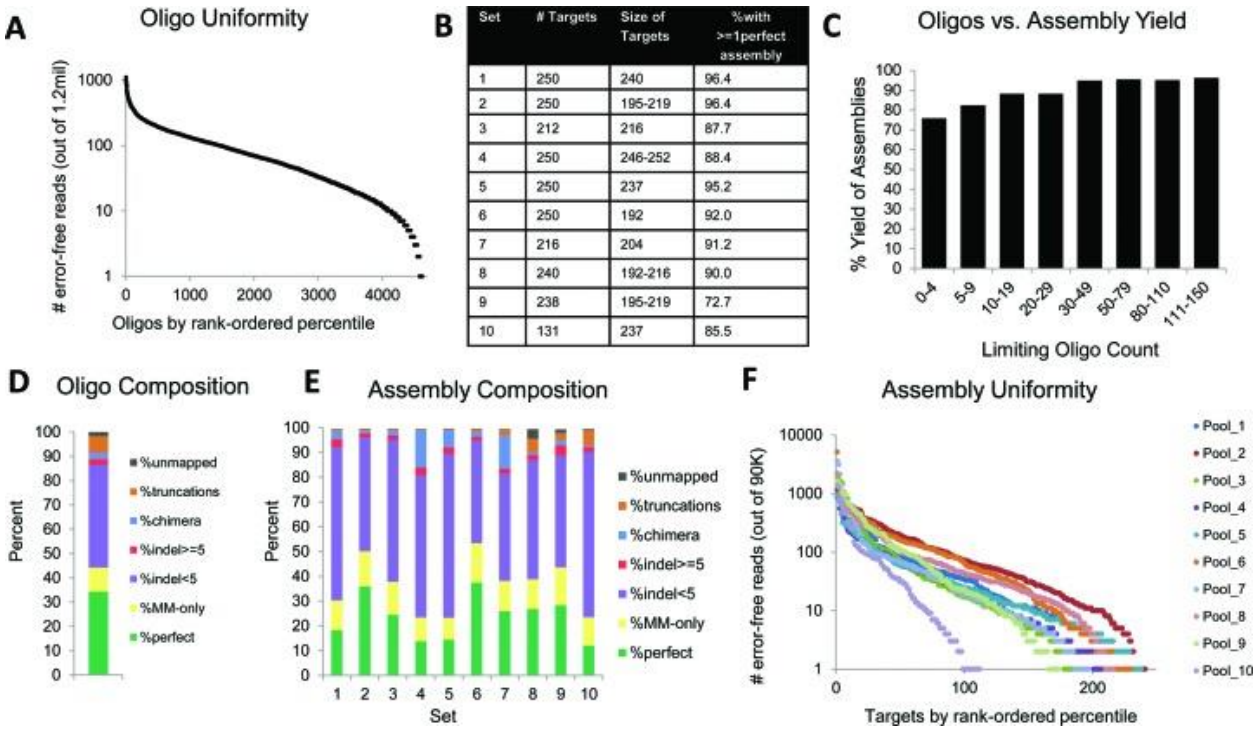


Figure 2.2. Pipeline for generation of static tag library.

We generated 1.2 million random 13-mers, and screened them for no homoguanine or homocytosine stretches >5 bp, no homoadenine or homothymine stretches >8 bp and GC content between 45% and 65%. We also screened for <90% nucleotide identity in the last 10 bp, which generated a set of 7411 13-mers. From this set of 7411 sequences, we calculated every pairwise Gibbs free energy, and identified the maximum number of sequences such that no two members had a dG ≤ -9 kcal/mol. This left a set of 4637 sequences, which were split into a set of 2318 forward tags and 2319 reverse tags.



Set	# Targets	Size of Targets	%with >=1perfect assembly
1	250	240	96.4
2	250	195-219	96.4
3	212	216	87.7
4	250	246-252	88.4
5	250	237	95.2
6	250	192	92.0
7	216	204	91.2
8	240	192-216	90.0
9	238	195-219	72.7
10	131	237	85.5

Figure 2.3. Assembling targets in sets of 131-250.

(**A**) Uniformity plot of error-free array-derived oligos by rank-ordered percentile for all 2271 targets. (**B**) Number and size of targets, and error-free yield for each target set. (**C**) Each target is placed into a bin based on the limiting oligo count, which is the number of error-free reads out of 1.2 million that are limiting for its corresponding target. The %Yield of assemblies is the percentage of targets in that bin with at least one perfect assembly. (**D**) The percentage of perfect, mismatch only, small indel (<5 bp), large indel (≥5 bp), truncations and unmapped reads for all oligos. (**E**) The percentage of perfect, mismatch only, small indel (<5 bp), large indel (≥5 bp), chimeras, truncations and unmapped reads for each assembled library. (**F**) Uniformity of each set of targets. Note that set 10 only has 131 targets.

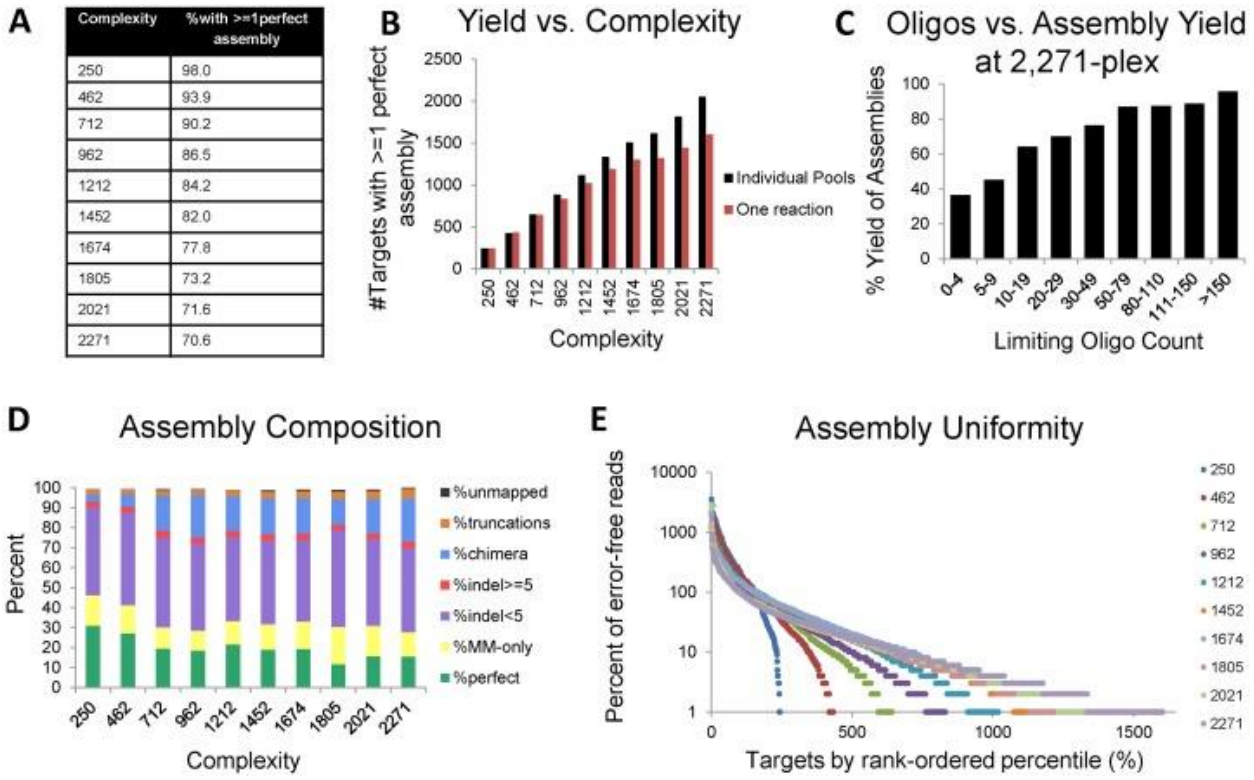


Figure 2.4. Effect of complexity on assembly performance.

(**A**) Percentage of targets with at least one error-free assembly for each level of complexity. (**B**) Yield (number of targets with at least one perfect read) versus complexity. Red bars show the total number of targets with error free assemblies at each level of complexity. Black bars show the number of targets from the corresponding sets with error-free assemblies, which were individually assembled in sets of complexity ranging from 131–250. (**C**) Each target is placed into a bin based on the limiting oligo count, which is the number of error-free reads (out of 1.2 million), that are limiting for its corresponding target.The %Yield of assemblies is the percentage of targets in that bin with at least one perfect assembly. (**D**) Percentage of perfect, mismatch only, small indels (<5 bp), large indels (≥5 bp), chimeras, truncations and unmapped reads in sets of increasing complexity. (**E**) Uniformity of each set of targets.

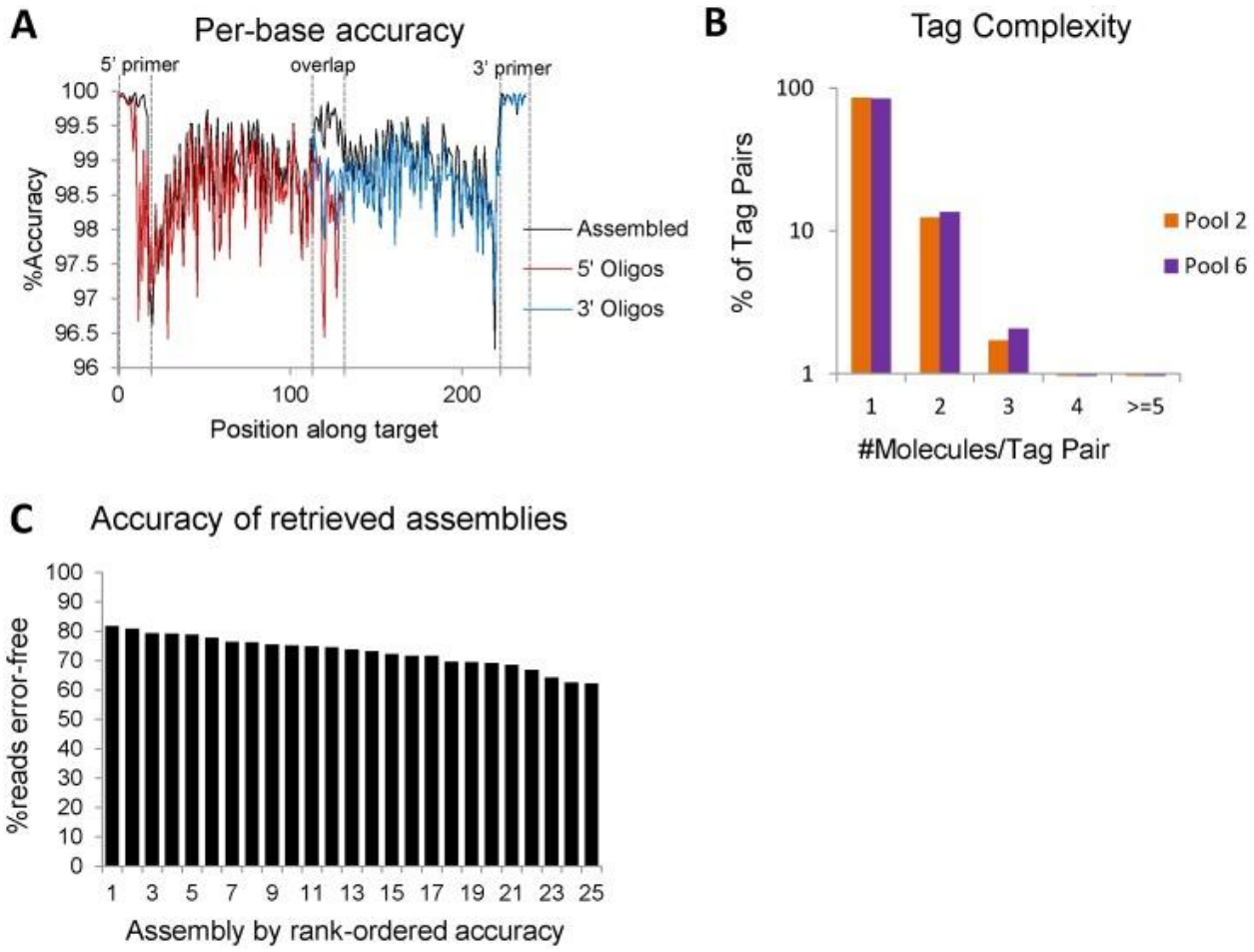


Figure 2.5. Error correction of assembled constructs.

(**A**) The per base accuracy of assembled constructs in black and their corresponding oligos in red and blue. Increased accuracy is seen at both priming sites and the overlap region. (**B**) Bar graphs for the percentage of tags identified on only one, two, three, four or at least 5 different molecules in the sequenced library. Orange and purple bars are two different assembly sets, each with 250 targets. (**C**) The percentage of aligning reads that contain no errors for each of the 25 retrieved assemblies.

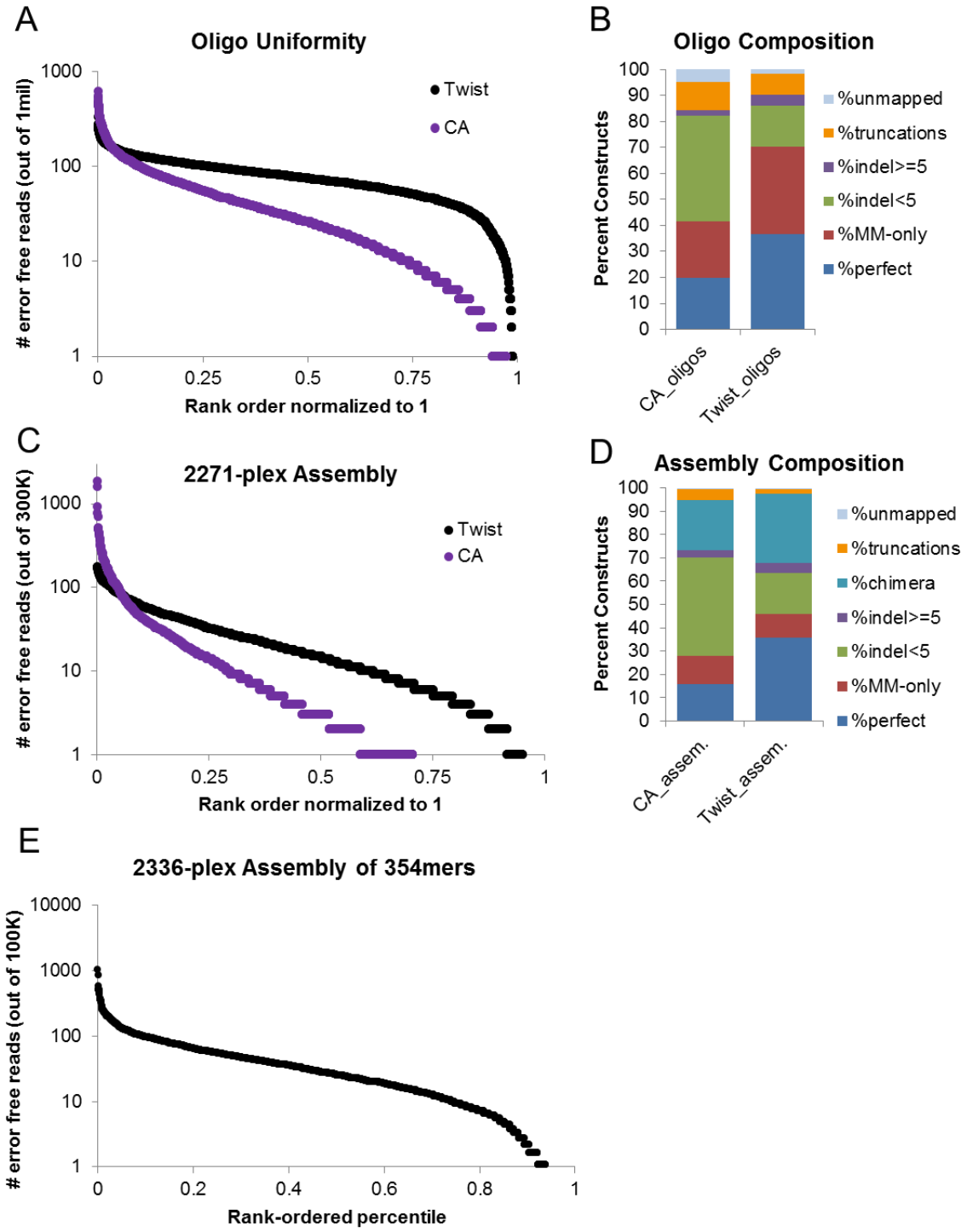
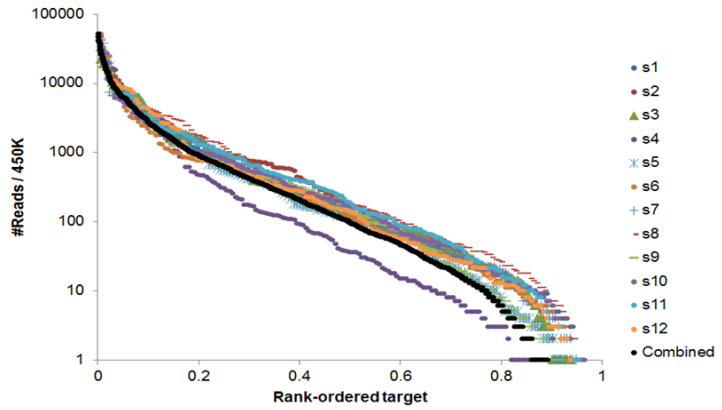


Figure 2.6. Higher Quality Oligos yield Higher Quality Assembly.

He, we synthesized the same library as in figures 2.3 and 2.4. (**A)** Comparison of oligo uniformity between Twist and Custom Array, for the same set of 4542 oligos in duplicate. **(B)** Comparison of the composition of oligos between Twist and Custom Array. **(C)** Comparison of

assembly uniformity between Twist and Custom Array for a 2271-plex assembly reaction. **(D)** Comparison of assembly composition for a 2271-plex assembly between Twist and Custom Array. **(E)** Assembly uniformity of a 2336-plex assembly of 354mers from 230mer Agilent oligonucleotides.

A

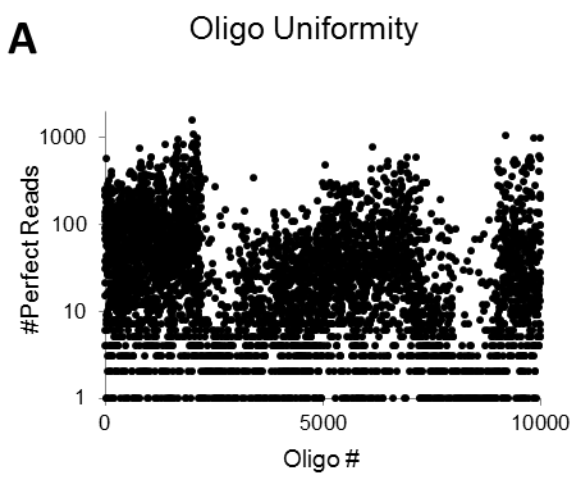


B

set	#assembled	%assembled	IQR
1	165	95.93	21.17
2	160	93.02	33.11
3	161	93.60	27.46
4	148	86.05	39.45
5	158	91.86	27.82
6	161	93.60	25.75
7	163	94.77	29.43
8	165	95.93	29.95
9	157	91.28	32.88
10	166	96.51	29.15
11	164	95.35	24.58
12	161	93.60	24.25
Combined	1960	94.96	30.08

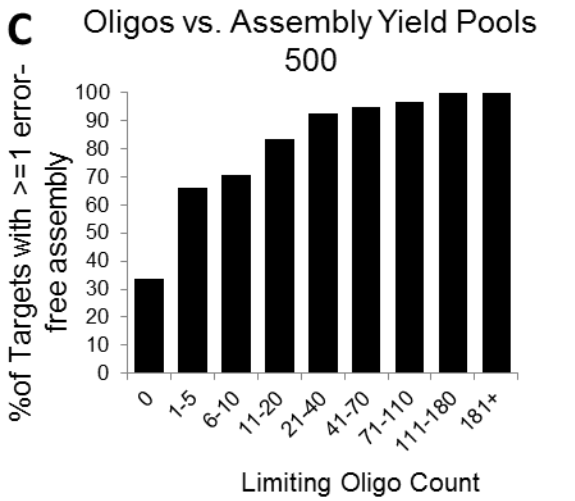
Figure 2.7. Hierarchical Multiplex Pairwise Assembly.

(A) Assembly uniformity for sets of 172 target 678-bp putative enhancers composed of four fragment oligos each. Twelve independent sets, as well as an *in silico* combined set, in which an equal number of reads from each set were pooled and analyzed. **(B)** The percent yield and interquartile range for each set of 172 targets and the *in silico* combined set.



B

Genes	Oligos	%with >=1perfect assembly	%with >=1 perfect, MM, or small indel
1-500	1-1000	83.6	96.4
501-1K	1001-2000	78.0	94.0
1001-1.5K	2001-3000	24.4	36.0
1501-2K	3001-4000	28.8	59.0
2001-2.5K	4001-5000	45.0	72.6
2501-3K	5001-6000	39.8	70.0
3001-3.5K	6001-7000	34.2	63.8
3501-4K	7001-8000	22.8	58.2
4001-4.5K	8001-9000	33.6	62.8
4501-5K	9001-10000	50.0	75.0



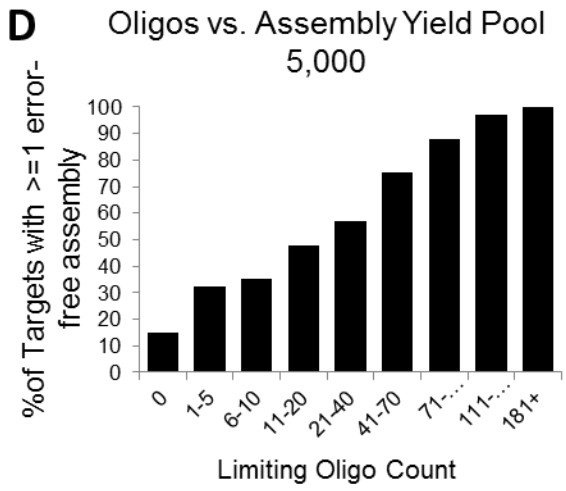


Figure 2.8. Assembly without duplicated oligos.

(A) Oligo uniformity across 10,000 oligos corresponding to 10 sub-pools of targets. **(B)** Assembly yield of sets of 500 targets. **(C)** Aggregate data from all pools of 500. Each target is placed into a bin based on the limiting oligo count, which is the number of error-free reads (out of 525K), that are limiting for its corresponding target. %Yield of assemblies is the percentage of targets in that bin with ≥1 perfect assembly. **(D)** Data from pool of 5,000.

Primer	Sequence
Uracil-containing A fragment primer	GCGA**N13**UU
Uracil-containing B fragment primer	CCAT**N13**UU
A fragment uniqueF	CCAT**N13**
B fragment uniqueR	GCGA**N13**
YF	GTTTTCCCAGTCACGAC
YR	CAGGAAACAGCTATGAC
Dial-Out_Tags_F	CGACAGTAACTACACGGCGA**N13**GTTTTCCCAGTCACGAC
Dial-Out_Tags_R	GTAGCAATTGGCAGGTCCAT**N13**CAGGAAACAGCTATGAC
Dial-Out_Flow_Cell_F	AATGATACGGCGACCACCGAGATCTACACACGTAGGCCGA CAGTAACTACACGGCGA
Dial-Out_Flow_Cell_R	CAAGCAGAAGACGGCATACGAGATNNNNNNNNNGACCGT CGGCGTAGCAATTGGCAGGTCCAT
Dial-Out_Sequencing_F	ACGTAGGCCGACAGTAACTACACGGCGA
Dial-Out_Sequencing_R	GACCGTCGGCGTAGCAATTGGCAGGTCCAT
Dial-Out_Sequencing_I	ATGGACCTGCCAATTGCTACGCCGACGGTC
YF-pu1L	CTAAATGGCTGTGAGAGAGCTCAGGTTTTCCCAGTCACGAC
YF-pu1R	ACTTTATCAATCTCGCTCCAAACCCAGGAAACAGCTATGAC
Pu1L_Flow_Cell	AATGATACGGCGACCACCGAGATCTACACACGTAGGCCTA AATGGCTGTGAGAGAGCTCAG
Pu1R_Flow_Cell	CAAGCAGAAGACGGCATACGAGATNNNNNNNNNGACCGT CGGCACTTTATCAATCTCGCTCCAAACC
Pu1_Sequencing_F	ACGTAGGCCTAAATGGCTGTGAGAGAGCTCAG
Pu1_Sequencing_R	GACCGTCGGCACTTTATCAATCTCGCTCCAAACC
Pu1_Sequencing_I	GGTTTGGAGCGAGATTGATAAAGTGCCGACGGTC

Table 2.1. Primers used in protocol.

The uracil-containing primers and Dial-Out tags both include *in-silico* designed 13-mer barcodes, represented as **N13** in this table. These were used for amplifying sub-pools from the array, as well as for tagging assembled constructs for Dial-Out PCR.

	Multiplex Pairwise Assembly	Column-based synthesis
Raw oligo cost	$2,400	$99,776
Oligo Pool amplification with Kapa HiFi Uracil+	$24	-
USER treatment + End Repair	$36	-
Assembly PCR	$12	-
Sequence Verification	$150	-
Total Cost	**$2,622**	**$99,776**
Dial-Out Tag Library- one-time cost	$1,800	-
Retrieval Primer Library- one-time cost	$17,118	-
Dial-Out: Total one-time cost	**$18,918**	
Dial-Out Tagging and Sequencing	$150	
Dial-Out Retrieval	$3,118	
Total Cost with Dial-Out Retrieval	**$5,890**	-

Table 2.2. Cost breakdown for 3,118 200-mers.

Raw oligo cost for multiplex pairwise assembly is based on duplicating all oligos and filling one 12,472-array from CustomArray with 160mers. Column-based oligo cost is based on IDT price of 384-well sub-nanomole plates. Note that IDT cannot synthesis oligos longer than 200bp by this method, whereas we demonstrate the synthesis of 252-mers. The rest of the steps for multiplex pairwise assembly are based on separating targets into six pools. Sequencing costs are based on a MiSeq v2 300 cycle spike-in (2 million reads). For Dial-Out Tagging, there is a one-time cost of the tag and PCR retrieval libraries. The total cost with Dial-Out Retrieval does not include the one-time cost.

Polymerase	Tags in moles	%Yield (Targets with perfect assemblies)	% of molecules with unique dial-out tag combination
Kapa HiFi	8.5E-14	91.3	81.5
Kapa HiFi	4.25E-13	91.3	85.7
Kapa HiFi	8.5E-13	84.6	89.3
Kapa HiFi	1.0E-12	90.3	-
Kapa 2G Robust	8.5E-14	64.4	-
Kapa 2G Multiplex	8.5E-14	85.5	-

Table 2.3. Optimizing polymerase and tag concentration.

We first tested the effect of several different tag concentrations on assembly yield with Kapa HiFi polymerase. For this dataset, the M13 sequences were present on the oligos, and tags were introduced during assembly. We found that we obtained the greatest yield with approximately a 10:1 molar ratio of tag:template, without a large loss in the percentage of unique tag pairs. We then tested this ratio with two different polymerases, Kapa 2G Robust and Kapa 2G Multiplex.

Chapter 3. FUNCTIONAL CHARACTERIZATION OF ENHANCER EVOLUTION IN THE PRIMATE LINEAGE

Chapter 3 is adapted from an unpublished manuscript under review as of April, 2018:

> Klein, J.C., Keith, A., Agarwal, V., Durham, T., and Shendure, J. (2018). Functional characterization of enhancer evolution in the primate lineage. *Submitted*.

3.1 ABSTRACT

Enhancers play an important role in morphological evolution and speciation by controlling the spatiotemporal expression of genes. Due to technological limitations, previous efforts to understand the evolution of enhancers in primates have typically studied many enhancers at low resolution, or single enhancers at high resolution. Although comparative genomic studies reveal large-scale turnover of enhancers, a specific understanding of the molecular steps by which mammalian or primate enhancers evolve remains elusive.

We identified candidate hominoid-specific liver enhancers from H3K27ac ChIP-seq data. After locating orthologs in 11 primates spanning ~40 million years, we synthesized all orthologs as well as computational reconstructions of 9 ancestral sequences for 348 “active tiles” of 233 putative enhancers. We concurrently tested all sequences (20 per tile) for regulatory activity with STARR-seq in HepG2 cells, with the goal of characterizing the evolutionary-functional trajectories of each enhancer. We observe groups of enhancer tiles with coherent trajectories, most of which can be explained by one or two mutational events per tile. We quantify the correlation between the number of mutations along a branch and the magnitude of change in

functional activity. Finally, we identify 57 mutations that correlate with functional changes; these are enriched for cytosine deamination events within CpGs, compared to background events.

We characterized the evolutionary-functional trajectories of hundreds of liver enhancers throughout the primate phylogeny. We observe subsets of regulatory sequences that appear to have gained or lost activity at various positions in the primate phylogeny. We use these data to quantify the relationship between sequence and functional divergence, and to identify CpG deamination as a potentially important force in driving changes in enhancer activity during primate evolution.

3.2 INTRODUCTION

Despite seemingly large phenotypic differences between species across the primate lineage, protein-coding sequences remain highly conserved. Britten and Davidson as well as King and Wilson proposed that changes in gene regulation account for a greater proportion of phenotypic evolution in higher organisms than changes in protein sequence (Britten and Davidson, 1971; King and Wilson, 1975). A few years later, Banerji and Moreau observed that the SV40 DNA element could increase expression of a gene independent of its relative position or orientation to the transcriptional start site (Banerji et al., 1981; Moreau et al., 1981). This finding led to the characterization of a new class of regulatory elements, enhancers.

Several aspects of enhancers make them ideal substrates for evolution. Enhancers control the location and level of gene expression in a modular fashion (Wray, 2007). While a coding mutation will disrupt function throughout an organism, a mutation in an enhancer may only affect the expression of a gene at a particular time and in a particular location. This modularity of regulatory elements may facilitate the development of novel phenotypes, *e.g.* by decreasing pleiotropy (True and Carroll, 2002). Enhancers also commonly exist in groups of redundant

elements, referred to as shadow enhancers, which provide phenotypic robustness (Frankel et al., 2010; Levine, 2010; Wittkopp and Kalay, 2011). Therefore, mutations within enhancers generally exhibit lower penetrance than mutations in coding sequences, facilitating the accumulation of variation.

Researchers have studied the role of enhancers in evolution through two main methods: high-resolution, systematic analysis of single enhancers, or low-resolution, genome-wide analysis of many enhancers. Examples of the former include fruitful investigations of how specific enhancers underlie phenotypic changes, *e.g.* cis-regulatory changes of the *yellow* locus affecting Drosophila pigmentation (Wittkopp et al., 2002; Gompel et al., 2005), recurrent deletions of a *Pitx1* enhancer resulting in the loss of pelvic armor in stickleback (Chan et al., 2010), and recurrent SNPs in the intron of *MCM6*, resulting in lactase persistence in humans (Bersaglieri et al., 2004; Tishkoff et al., 2007).

Low-resolution, genome-wide approaches for discovering candidate enhancers via biochemical marks, when applied to multiple species, have identified large-scale turnover of enhancers between human and mouse embryonic stem cells (Kunarso et al., 2010), human and mouse preadipocytes and adipocytes (Mikkelsen et al., 2010), mammalian limb bud (Cotney et al., 2013) and vertebrate and mammalian liver (Villar et al., 2015; Trizzino et al., 2017). STARR-seq has also been used to characterize enhancer evolution within five Drosophila species, providing functional evidence of large-scale turnover (Arnold et al., 2014).

High-resolution studies can provide clear insights into the evolution of individual enhancers, but the findings may not be broadly generalizable. Low-resolution studies have the advantage of characterizing thousands of enhancers at a time, but fail to pinpoint functional variation. Massively parallel reporter assays (MPRAs) may offer an opportunity to bridge the

insights offered by low- and high-resolution studies. MPRAs have enabled high-resolution functional dissection of enhancers by testing the effects of naturally occurring and synthetic variation on regulatory activity (Patwardhan et al., 2009, 2012; Melnikov et al., 2012; Vockley et al., 2015; Tewhey et al., 2016; Ulirsch et al., 2016).

Here we set out to concurrently study the evolutionary-functional trajectories of hundreds of enhancers with MPRAs. We identified potential hominoid-specific liver enhancers based on genome-wide ChIP-seq and then functionally tested all of these in parallel. After identifying "active tiles" of these candidate enhancers, we tested eleven primate orthologs and nine predicted ancestral reconstructions of each active tile for their relative activity. Normalizing to the activity of the reconstructed sequences of the common ancestor of hominoids and Old World monkeys, we identify several subsets of active tiles that appear to have gained or lost activity along specific branches of the primate lineage; only some of these patterns are consistent with ChIP-seq-based expectations. We also use these data to examine how mutational burden impacts enhancer activity across the phylogeny, quantifying the correlation between sequence divergence and functional divergence. Finally, we examine the set of mutations that appear to drive functional changes, and find enrichment for cytosine deamination within CpGs.

3.3 RESULTS

3.3.1 *Identification of candidate hominoid-specific enhancers*

From a published ChIP-seq study in mammals (Villar et al., 2015), we identified 10,611 H3K27ac peaks (associated with active promoters and enhancers) that were present in humans and absent from macaque to tasmanian devil, and that were not within 1 kilobase (Kbp) of a H3K4me3 peak (associated with active promoters). We considered this set of peaks as potential

hominoid-specific enhancers (active within the clade from gibbon to human). We narrowed this to a subset of 1,015 candidate enhancers overlapping ChromHMM strong-enhancer annotations in human HepG2 cells (Ernst et al., 2011) that also had orthologous sequences in the genomes of species from human to marmoset. On average, the intersection between the hominoid-specific H3K27ac peak and HepG2 ChromHMM call was 1,138bp (**Figure 3.6A**). In order to identify active subregions of each candidate enhancer, we designed 194nt sequences tiling across the length of each, overlapping by 93-100bp (**Figure 3.1A**).

We synthesized and tested all 10,544 tiles for enhancer activity in a massively parallel reporter assay. Specifically, we used the STARR-seq vector (Arnold et al., 2013), in which candidate enhancers are cloned into the 3' UTR of an episomal reporter gene, in human HepG2 cells in triplicate. After extracting, amplifying and sequencing DNA and RNA corresponding to the enhancer region from transfected cells, we calculated an enrichment score for each tile as the ratio of RNA to DNA (rho for pairs of replicates between 0.581 and 0.676). We defined "active tiles" as elements with a mean enrichment score at least two standard deviations above the mean enrichment score of 125 scrambled control sequences. While most of the 1,015 candidate enhancers contained no active tiles, we identified 697 active tiles (out of 10,544, or 6.6%), occurring within 34% of the candidate enhancers (**Figure 3.6B**). While we chose a strict cutoff for active tiles to increase specificity, we do note a significant shift towards more positive enrichment scores for our tiles as compared to scrambled control sequences (mean z-score 0.504 v 0.000, $p<1e-5$, t-test for two independent means) (**Figure 3.6C**).

3.3.2 *Computationally predicting the activity of ancestral and orthologous sequences*

A goal of this study was to characterize how the number and spectrum of mutations relate to the functional divergence of enhancer activity in primates. We used eleven high-quality

primate genomes (human, chimpanzee, gorilla, orangutan, gibbon, rhesus, crab-eating macaque, baboon, vervet, marmoset and squirrel monkey) to locate similarly-sized orthologs of each of our 697 active human tiles. We were able to find orthologs in all eleven species for 348 of the 697 active human tiles. Since these species are separated by only ~40 million years, they retain high nucleotide identity. We sought to take advantage of this to ask whether we could computationally prioritize the sequence changes that underlie apparent functional differences between orthologous sequences within primates. Of note, we had not yet measured the functional activity of orthologs of active tiles. Rather, we were assuming that previously observed patterns of gain/loss in H3K27 acetylation were reflective of whether particular tiles were active or inactive in each primate.

We first examined turnover of DNA-binding motifs known to be associated with enhancer activity in HepG2: FosL2 and JunD (Inoue et al., 2016). We focused on comparing the human ortholog to the marmoset ortholog, the furthest outgroup with ChIP-seq data. We identified a modest enrichment of the AP-1 consensus motif, the motif for JunD and FosL2 binding, in the human ortholog compared to marmoset ($p=0.012$, Fisher's exact). However, AP-1 site turnover could only explain 5% of the gain-of-function events predicted by H3K27ac ChIP-seq.

As a different approach, we built a computational model for predicting enhancer activity in HepG2 cells, and then sought to apply that model to the active tiles and their orthologs. Specifically, we trained a gapped k-mer support vector machine (gkm-SVM), a sequence-based classifier based on the abundance of gapped k-mers in positive and control training data, on an independent massively parallel reporter assay experiment in HepG2 cells (Ghandi et al., 2014; Inoue et al., 2016). We evaluated the model by predicting the enrichment scores from our tiling

experiment on human orthologs, which the model had not seen during training. Although the original data was based on an entirely different MPRA assay ('lentiMPRA'), the scores for each tile predicted from the gkm-SVM model correlated reasonably well with our enrichment scores obtained through STARR-seq in HepG2 cells (Spearman rho=0.453, p-value <1e-10) (**Figure 3.2A**). We then used the model to predict regulatory activity for the rhesus, vervet and marmoset orthologs, all of which did not have H3K27ac peaks. We expected to find lower predicted activity for these three orthologs compared to human. However, the predicted activity for the human vs. rhesus, vervet or marmoset orthologs were not significantly different (p=0.10, two-sample t-test), although it did trend in the right direction for all three comparisons (**Figure 3.2B**).

With the goal of increasing our power to detect mutations underlying gains or losses in enhancer activity, we reconstructed nine ancestral sequences of the 11 primate orthologs using FastML, a maximum-likelihood heuristic (**Figure 3.1B**) (Ashkenazy et al., 2012). We then applied the gkm-SVM model to predict regulatory activities for the 20 orthologs (11 from present-day primate genomes, 9 reconstructed ancestral sequences) of the human-active tiles. To characterize evolutionary trajectories, we performed hierarchical clustering on the vectors of predicted activity for each tile, and identified a group of 82 enhancer tiles that show decreased predicted activity in rhesus, vervet and marmoset compared to human, following the pattern predicted by H3K27ac ChIP-seq (**Figure 3.2C**). However, this was a clear minority of all tiles evaluated with this computational model (82/348 or 24%).

3.3.3 *Functional characterization of ancestral and orthologous sequences*

We were surprised that only a quarter of our computational predictions were concordant with ChIP-seq predictions. This could be due to limitations either in interpreting patterns in H3K27ac gain/loss, or of the computational models that we are applying to predict the relative

activities of orthologs, or both. To investigate this further, we synthesized and functionally tested all 20 versions of each of the 348 active tiles with the STARR-seq vector in HepG2 cells. With the goal of improving accuracy and reproducibility, we added degenerate barcodes adjacent to each sequence of interest while cloning the library, so that we could distinguish multiple independent measurements for each element. Furthermore, we performed three biological replicates, which correlated well (independent transfections; Spearman rho between 0.773 and 0.959) (**Figure 3.7A-C**). We took the average enrichment score of all barcodes over all three replicates and filtered out any element with less than six independent measurements. On average, this set had 31 independent measurements per element (**Figure 3.7D**).

The resulting dataset included enrichment scores for 5,426 of the 6,960 sequences tested (78.0%), corresponding to 344 of the 348 human-active enhancer tiles (98.9%). As expected, the average pairwise correlation between species was higher within clades (hominoid, Old World monkeys and New World monkeys) than between clades (**Figure 3.3A**). For our initial analyses, we normalized the enrichment scores for all non-human orthologs to the enrichment score of the human ortholog, given that these tiles were first identified on the basis of the human ortholog exhibiting activity.

We identified 219 enhancer tiles for which we successfully assayed activity for the human ortholog and at least 14 other orthologs. For each of these orthologs, we asked how well the experimental measurements correlated with the gkm-SVM predictions from **Figure 3.2C**. Specifically, we asked whether the gkm-SVM model predicted functional differences between closely related orthologs by comparing our scores of model predictions vs. functional data (all scores normalized to the human ortholog). There was no correlation between the predicted vs. experimental normalized scores (**Figure 3.3B**; Spearman rho= -0.002, p-value=0.892).

Therefore, while the kmer-based model performed well at characterizing relative activities of diverse elements (**Figure 3.2A**), it did not predict the relative activities of closely related sequences as measured here (**Figure 3.3B**).

We next performed hierarchical clustering on the vectors of experimentally measured activity for each tile (*i.e.* where each vector consists of the set of activities experimentally measured for all orthologs or ancestral reconstructions of a human-active tile, normalized against the activity of the human ortholog; **Figure 3.3C**). We identified a group of 50 enhancer tiles with relatively higher activity in either humans or hominoids (50/219 or 23%) (**Figure 3.3C**, green group on dendrogram), a group of 32 enhancer tiles with relatively lower activity in the Old World monkey lineage (14.6%) (**Figure 3.3C**, orange group on dendrogram), and a group of 43 enhancer tiles with relatively higher activity in the Old World monkey lineage (19.6%) (**Figure 3.3C**, red group on dendrogram). As a negative control, when we permuted species' ids for each tile (*i.e.* shuffling raw scores represented within each column of **Figure 3.3C**, renormalizing, and hierarchical clustering), we no longer observe coherent clustering of activity patterns by clades (**Figure 3.8**).

The first group, *i.e.* the subset of 50 enhancer tiles (23% of 219 tested) with greater activity in humans or hominoids relative to other primates, corresponds to the pattern predicted by the ChIP-seq data, and a similar proportion to the 82 tiles (24% of 348 tested) whose computationally predicted activity was concordant with the pattern predicted by the ChIP-seq data (**Fig. 3.2C**). However, only 14 enhancer tiles overlapped between these groups, which is not more than expected by chance ($p=0.48$, Fisher's exact test). This was consistent with the lack of correlation between the experimental and predicted relative scores shown in **Figure 3.3B**.

For several reasons, we chose to move forward with the experimentally measured activities of primate enhancer tile orthologs. First, we believe that experimental measurements are preferable to computational predictions when available. A condition for this preference is that the experimental measurements are reproducible, which in this case they are (**Figure 3.7A-C**). Second, the computational model used here is predicting the likelihood of a sequence belonging to an active vs. inactive group, while the experimental data measure the relative activity of each sequence. Experimental data is therefore better suited for quantifying differences in activity, which is the attribute that we would like to correlate with sequence divergence. Third, the differences in experimentally measured activity between orthologs relative to human were generally much greater in magnitude than the computational predictions (*e.g.* compare **Figure 3.2C vs. Figure 3.3C**, which use the same color scale), and furthermore in patterns that were consistent with the phylogeny relating those sequences to one another (**Figure 3.3C vs. Figure 3.8**, which is the same data permuted).

3.3.4 *Evolutionary-functional trajectories for hundreds of enhancer tiles across the primate phylogeny*

We had originally normalized enhancer tile activity to the human ortholog with the assumption that most enhancer tiles would be hominoid-specific based on patterns in H3K27ac ChIP-seq data. While our largest group did agree with the ChIP-seq data, it only represented 23% of the tested tiles. Given that the groups that we did observe were relatively coherent in relation to the lineage tree (**Figure 3.3C**), we turned to asking whether we could quantify the enhancer activity of various orthologs relative to their common ancestor.

For this, we normalized the enhancer tile activity scores for all orthologs to the most recent common ancestor (MRCA) of Catarrhines (N2; common ancestor of hominoids and Old

World monkeys). We then performed hierarchical clustering on the 206 enhancer tiles with scores for N2 and at least 14 additional orthologs. The resulting heatmap is shown in **Figure 3.4A**. We observed several subsets of enhancer tiles that exhibited gains or losses in activity as measured by STARR-seq, relative to the experimentally measured activity of the reconstructed sequence of their common ancestor (**Table 3.1**). Many of these subsets were coherent in relation to the lineage tree, meaning that more closely related orthologs exhibited consistent changes in activity in relation to one another.

A first group (purple; **Figure 3.4B**) contains enhancer tiles that maintain activity in all orthologs except for Old World monkeys, which have consistently decreased activity relative to N2. This group contains 29 enhancer tiles (14%). A parsimonious explanation is that this group comprises tiles in which loss-of-activity events occurred on the branch between N2 and the MRCA to Old World monkeys.

A second group (green; **Figure 3.4C**) contains enhancer tiles with decreased activity in hominoids and Old World monkeys, relative to New World monkeys and N2. This group contains 15 enhancer tiles (7%). A potential explanation is that this group, the smallest of the five that we highlight in **Figure 3.4**, comprises tiles in which loss-of-activity events that occurred on both the branch between N2 and the MCRA of hominoids as well as on the branch between N2 and the MRCA of Old World monkeys.

A third group (yellow; **Figure 3.4D**) contains enhancer tiles with decreased activity restricted to the outgroup of New World monkeys. This group contains 32 enhancer tiles (16%), and is consistent with a single gain-of-activity event occurring between the MRCA to Simiformes (hominoids, Old World monkeys and New World monkeys) and N2, or alternatively a single loss-of-activity event on the branch leading to the New World monkeys.

A fourth group (orange; **Figure 3.4E**) contains 22 enhancer tiles (11%) with decreased activity in hominoids (lesser apes, great apes and humans) or only in hominids (great apes and humans) relative to N2. The most parsimonious explanation is a single loss-of-activity event along the branch from N2 to the MRCA of hominoids in some of these tiles and along the branch from the MRCA of hominoids to the MRCA of hominids in others (**Figure 3.4E**). This group is particularly interesting in that the human enhancer tiles, which are active based on ChIP-seq and our initial tiling experiment, have lower activity than some ancestral sequences as well as Old and New World monkeys. Looking more broadly, there are 35 enhancer tiles (17%) for which the human sequence exhibits significantly lower activity than the reconstructed N2 ortholog ($p<0.05$, 2-sample t-test). This bias towards reductions in activity relative to the ancestral N2 ortholog is not unique to human orthologs. Across the full dataset, 774 orthologs showed a significant reduction in activity compared to the N2 ortholog while only 694 showed a significant increase ($p=0.023$, z-test two population proportions). This result suggests that the ancestral forms of regulatory sequences queried here tended to have greater activity than descendant sequences.

A fifth group (gray; **Figure 3.4F**) contains 25 enhancer tiles (12%) with decreased activity in N2 relative to all other orthologs. Given that this pattern would require 3 independent gains/losses for each enhancer, a simpler explanation may be that these represent low-quality reconstructions of the N2 sequence. As the most distant sequences from present-day orthologs, we expect N2 orthologs to be more likely to contain errors in reconstruction, which would in turn be expected to be biased towards reduction in activity. To test this, we extracted the confidence of each reconstruction for each of our ancestral sequences from FastML. Overall, N10 (the MRCA to New World monkeys) had the lowest average marginal probability of 0.335, followed

by N2 with an average marginal probability of 0.547 (**Figure 3.9**). The remaining reconstructions all had marginal probabilities greater than 0.8. Next, we looked at the marginal probabilities for N2 from this group of 25 enhancers compared to the remaining 323 enhancers. This group had a significantly lower average marginal probability for the N2 reconstructions compared with the other enhancers (0.396 v 0.558, p=0.018, t-test for two independent means). This is consistent with the interpretation that the pattern observed for these 25 enhancers shown in **Figure 3.4F** is consequent to poor N2 reconstructions.

We examined whether tiles derived from the same enhancer peaks tend to fall within the same groups defined above. The 348 human-active enhancer tiles for which we tested additional orthologs derived from 233 candidate enhancers. Of these 233, 75 contained multiple tiles in our set, 8 of which had pairs of tiles that both fell within one of the five groups, which is significantly greater than expected by chance (p<1e-5, permutation test). Three of these eight pairs of enhancers were overlapping tiles, which can potentially narrow down the location of the causal mutation.

3.3.5 *Characterizing molecular mechanisms for enhancer modulation*

We next explored the relationship between the sequence versus functional evolution in enhancer activity across the primate phylogeny. As a starting point, we asked whether there was a correlation between the accumulation of sequence variation and the magnitude of change in functional activity for enhancer tiles. For every branch along the tree, we calculated the number of mutations between the mother and daughter nodes and the change in activity between the nodes. There was a significant, albeit modest, correlation between the number of mutations accumulated along a branch and the absolute change in functional activity (Spearman rho=0.207,

p=4e-26) (**Figure 3.5A**). On average, each nucleotide substitution was associated with a 6.2% change in functional activity (slope of best fit line, with y intercept fixed to 1).

We first asked whether different subsets of mutations were associated with functional changes. We focused on mutations that disrupt transcription factor (TF) motifs and asked whether these mutations were associated with changes in functional activity. For each enhancer tile, we identified all motifs associated with a TF expressed in HepG2 that were either lost or gained in at least one ortholog. We then ran linear regressions for the presence or absence of each TF motif against the functional scores for all orthologs, testing whether the mean slope of a TF from all enhancers was significantly different than zero using a two-sample t-test (**Table 3.2**). The top two scoring TF motifs were E2F1, which was negatively correlated with activity (*i.e.* the presence of an intact motif is associated with decreased activity), and ATF2, which was positively correlated with activity (*i.e.* the presence of an intact motif is associated with increased activity). Both TFs play important roles in the liver (Wuestefeld et al., 2013; Denechaud et al., 2016), with E2F1 acting as a transcriptional repressor and ATF2 acting as a transcriptional activator. The E2F1 motif was disrupted in 60 enhancer tiles, and the ATF2 motif in 20 enhancer tiles. However, neither correlation was significant after a Bonferroni correction controlling for multiple hypothesis testing.

We next sought to prioritize specific mutations based solely on sequence vs. functional differences across the phylogeny. For each position along a given enhancer tile with a variant in at least one ortholog, we characterized each allele as ancestral (matching the MRCA of human and squirrel monkey, N2) or derived. We then performed Kolomogorov-Smirnov (K-S) tests at each position to test for association between allele status and functional scores, while applying a Bonferroni correction to account for the number of variants tested for each tile. Through this

analysis, we identified a total of 57 mutations that we will refer to as "prioritized variants", which correlate with the functional scores (**Table 3.3**). We also generated a set of "background variants," which did not correlate with functional scores (non-significant by the K-S test). Within the 57 prioritized variants, there was a significant overabundance of C→T and G→A mutations over background (p=0.037, Fisher's exact test, Bonferroni corrected) (**Figure 3.5B**). In order to test whether this effect is due in part to methylation, we looked at the subset of these C→T and G→A mutations, which disrupted a CpG. Cytosine deamination within a CpG accounted for 21% of our prioritized variants, compared to only 10% of background variants (p=0.013, Fisher's exact test). When subtracting CpG deamination events, the enrichment for C→T and G→A mutations is no longer present (p=0.224, Fisher's exact test), suggesting that CpG deamination events explain the observed enrichment.

3.4 DISCUSSION

While genome-wide studies demonstrate large-scale turnover of enhancers, the general molecular mechanisms underlying this turnover remain largely unexplored. In this study, we characterized modulation in the activity of hundreds of enhancer tiles throughout primate evolution, with nucleotide-level resolution. We first tried to characterize functional changes using computational tools, and although our tools were able to differentiate enhancers with low nucleotide identity (**Figure 3.2A**), they did not correlate well with our ChIP-seq-based predictions (**Figure 3.2C**) and performed poorly at predicting functional changes between evolutionarily similar sequences (**Figure 3.3B**). We therefore decided to test all sequences using STARR-seq, a reporter assay that experimentally measures regulatory activity for a library of sequences.

By testing all elements in the same *trans* environment (a single cell type), our experimental approach provided quantitative and directly comparable measurements, allowing us to measure functional differences between closely related sequences. However, this experimental approach assumes conserved trans-environments throughout the primate lineage. Previous studies have indeed noted that both the specificity of transcription factors for DNA and coactivators has remained highly conserved over much longer evolutionary time scales (Dowell, 2010; Zheng et al., 2011; Nitta et al., 2015; Long et al., 2016).

From both our computational predictions and functional scores, we note a low concordance with ChIP-seq based predictions (24%-37%). These numbers are similar to previous attempts to replicate biochemical predictions with high-throughput reporter assays (Kheradpour et al., 2013; Kwasnieski et al., 2014; Inoue et al., 2016), and there are plausible explanations for the difference. The ChIP-seq predictions were based on experimental data from primary liver samples from three individuals per species. Although most of the liver is composed of a single cell type, hepatocytes, there is still more diversity in such primary tissue than in the cell culture system we used for STARR-seq. Moreover, while we maintain a single *trans* environment for all of our orthologs, it is not an exact replica of primary liver. HepG2 cells are derived from a hepatocellular carcinoma, and likely have acquired changes during cancer development and immortalization. However, the fact that our enhancer tiles are both active in HepG2 cells (ChromHMM and STARR-seq) and in primary liver from humans (H3K27ac ChIP-seq) adds to our confidence that we are characterizing bona-fide enhancers.

Through hierarchical clustering of enhancer tiles normalized to human, we identified several functional groups. The largest group matched our ChIP-seq based predictions, with increased activity in humans and/or hominoids compared to other primate orthologs. We also

identified a large group with decreased activity in Old World monkeys (concordant with three of the four ChIP-seq based predictions) and a third group with increase increased activity in Old World monkeys, or decreased activity in humans. The third group is the opposite of what we expected based on our ChIP-seq predictions. There are several possible explanations for this discordance in addition to the ones listed above. These regions may not have had ChIP-seq signal in rhesus, vervet, and marmoset if they were only active in a subset of liver cells (and therefore missed in a bulk-assay) or act as redundant (shadow) or poised-enhancer (and therefore lack active enhancer marks but are capable of activating transcription based on their sequence).

We next characterized evolutionary-functional trajectories for 206 of the enhancer tiles by normalizing all orthologs to the furthest ancestor, the MRCA between hominoids and Old World monkeys. We grouped these trajectories using hierarchical clustering, and identified several common patterns of modulation throughout the primate phylogeny. The most common patterns were tiles with a single loss of activity in Old World monkeys (n=29, **Figure 3.4B**), a single gain of activity in Catarrhini (or loss on the branch to New World monkeys) (n=32, **Figure 3.4D**), a single loss of activity in hominoids or hominids (n=22, **Figure 3.4E**), and two independent loss of activity events in hominoids and Old World monkeys (n=15, **Figure 3.4C**). We also identified a group of tiles with a decrease of activity unique to N2, without a clear parsimonious explanation (n=25, **Figure 3.4F**). Based on the lower marginal probabilities for the N2 reconstructions within this group, we propose that this subset of enhancer tiles may be due to incorrect reconstructions of the N2 sequence.

The group of enhancer tiles with decreased activity in hominoids may indicate sub-optimization or fine-tuning of enhancers. In total, 17% of our tiles showed a significant reduction of activity in human compared to N2, suggesting that reductions, without complete loss, of

activity may in fact be a common phenomenon in primate enhancers. To determine whether sub-optimization was a general trend across the phylogeny (Farley et al., 2015), we calculated the number of enhancer tiles with significant increase or decrease of activity relative to N2. We identified significantly more tiles with decreases relative to N2 than increases. If we remove the cluster with reduced expression solely in N2, which we believe to be an artifact, the trend becomes more significant (768 vs 503, $p<1e-5$, z-test). All of these findings are concordant with high ancestral activity of present-day enhancers with subsequent loss to fine-tune activity along the phylogeny, at least for the enhancers that we chose to characterize here, which may be biased by the manner in which they were selected.

Ultimately, we wanted to look for general trends between sequence and functional divergence of enhancers throughout evolution. First, we looked at how the number of mutations accumulated along any branch on the tree correlates with the functional divergence along the branch. We found a modest, but significant correlation between sequence and functional divergence (Spearman rho=0.207, p=4.2e-26). Previous studies have associated naturally occurring genetic variation to evolutionary changes in expression (Arnold et al., 2014) and population variation in expression (Vockley et al., 2015). They have also related synthetic variation to changes in reporter activity (Patwardhan et al., 2009, 2012; Melnikov et al., 2012; Smith et al., 2013; Farley et al., 2015). However, this study was unique in that it quantified the relationship between single nucleotide changes occurring during neutral evolution in closely related species and experimentally-measured functional differences.

To further characterize mechanisms of mutations important in enhancer evolution, we utilized the high nucleotide identity between orthologs and reconstructed ancestral sequences to prioritize several variants, which were likely causal for functional divergence. While we first

tried to prioritize variants based on transcription factor motif turnover, we did not find any significant motifs. Both ATF2 and E2F1 showed high correlations in our analysis, but neither was significant after multiple testing. Instead, we relied on prioritizing variants solely based on sequence content and functional scores, resulting in a list of 57 potentially causal variants. Of note, these 57 variants were enriched for cytosine deamination, particularly within CpGs, compared to variants that were not significantly associated with functional scores. Especially within closely related species, CpG deamination is a promising source of evolutionary novelty. Since spontaneous deamination of 5-methylcytosine (5mC) yields thymine and G-T mismatch repair is error prone, 5mC has a mutation rate four to fifteen-fold above background (Cooper et al., 2010).

Besides its increased rate of mutation, there are multiple mechanisms by which CpG deamination may play a significant role in enhancer modulation. One mechanism is by introducing novel transcription factor binding sites or disrupting existing binding sites. In fact, Zemojtel *et al.* suggested that CpG deamination creates TF binding sites more efficiently than other types of mutational events (Zemojtel et al., 2011). CpG deamination may also alter enhancer activity by modifying methylation. Enhancer methylation has been correlated with gene expression, most frequently in cancer patients but also in healthy individuals (Aran and Hellman, 2013). Notably, enhancer methylation is both correlated with increased and decreased gene expression, possibly explaining why we see an enrichment of CpG deamination in both gain and loss of function events (Long et al., 2017).

In this study, we aimed to characterize general molecular mechanisms that underlie enhancer evolution. In order to do so, we conducted a large-scale screen of enhancer modulation with nucleotide-level resolution by combining genome-wide ChIP-seq with STARR-seq of many

orthologs. We functionally characterized evolutionary-functional trajectories for hundreds of enhancer tiles, demonstrating a significant correlation between sequence and functional divergence along the phylogeny. We identify that many present-day enhancers actually have decreased activity relative to their ancestral sequences, supporting the notion of sub-optimization. We prioritized 57 variants, which correlated with functional scores, and found enrichment for cytosine deamination within CpGs among these potentially functional events. We propose that CpG deamination may have acted as an important force driving enhancer modulation during primate evolution.

3.5 METHODS

3.5.1 *Identification of potential hominoid-specific enhancers*

We downloaded processed H3K27ac and H3K4me3 peak calls from Villar *et al.* (Villar et al., 2015). Within each species, we called enhancers as H3K27ac peaks with a mean fold change ≥ 10 that were not within 1000bp of an H3K4me3 replicated peak. We converted all replicated H3K27ac peaks in rhesus, vervet and marmoset to hg19 coordinates using the UCSC liftover tool with a minimum match of 0.5. Villar *et al.* called the vervet peaks using the rhesus genome as a reference. We identified potential hominoid gain of function enhancers as predicted enhancers that did not have orthologous H3K27ac enrichment within 1kb from the summit in rhesus, vervet or marmoset. We converted the 10,611 gain of function enhancers back to the marmoset and rhesus genome with a minimum match of 0.9, with 6,862 having orthologs in the three genomes. We intersected our 6,862 GOF enhancers with ChromHMM strong enhancer calls in HepG2 using bedtools (Quinlan and Hall, 2010), resulting in a final set of 1,015 potential hominoid gain of function enhancers predicted to be active in HepG2.

3.5.2 *Design and synthesis of tiles*

For each potential hominoid gain of function enhancer, we defined end points by using the intersection of the H3K27ac peak and HepG2 ChromHMM strong enhancer call. For any intersections less than or equal to 200nt, we designed a 194bp tile around the center. For intersections with 200 ≤ length ≤ 400, we split the sequence into 3 overlapping fragments. For intersections > 400nt, we used 100bp sliding windows. We created negative controls from 800 tiles using uShuffle to create 200 dinucleotide shuffles each (Jiang et al., 2008), and then picked the shuffled sequence with the fewest 7mers present in the original tile. We then synthesized all 10,544 tiles and 800 negative sequences as part of a 244K 230mer array from Agilent. The library was amplified from the Agilent array using the HSS_cloning-F (5'-TCTAGAGCATGCACCGG-3') and HSS_cloning-R (5'-CCGGCCGAATTCGTCGA-3') primers and cloned into the linearized human STARR-seq plasmid using NEBuilder HiFi DNA Assembly Cloning Kit (Arnold et al., 2013). The library was transformed into NEB 3020 cells and midi-prepped using the ZymoPURE Plasmid Midiprep Kit (Zymo Research).

3.5.3 *Identification of active tiles*

We transfected 5ug of our tiling library and 2.5ug of a puromycin expressing plasmid into three 60mm dishes, each with approximately 1.5 million HepG2 cells using Lipofectamine 3000 (ThermoFisher) according to manufacturer's instructions. Twenty-four hours post-transfection, we selected cells with 1ng/mL puromycin for 24 hours. Forty-eight hours post transfection, we extracted DNA and RNA from the cells using the Qiagen AllPrep DNA/RNA Mini Kit (Qiagen). We treated RNA with the TURBO DNA-free Kit (ThermoFisher) and performed reverse transcription with SuperScript III Reverse Transcriptase (ThermoFisher). We amplified the cDNA using NEBNext High-Fidelity 2x PCR Master Mix with 5ul of RT reaction

with primers HSS-F and HSS-R-pu1 in a 50ul reaction for three cycles with a 65°C annealing temperature. PCR reactions were cleaned with 1x Agencourt AMPure XP and eluted in 19ul (Beckman Coulter). We then performed a nested PCR using the whole purified cDNA reaction with primers HSS-NFpu1 (5'-CTAAATGGCTGTGAGAGAGCTCAGGGGCCAGCTGTTGGGGTGTCCAC-3') and pu1R (5'-ACTTTATCAATCTCGCTCCAAACC-3'). DNA was amplified in one reaction using HSS-NFpu1 and HSS-R-pu1 (5'-ACTTTATCAATCTCGCTCCAAACCCTTATCATGTCTGCTCGAAGC-3') with 1-2ug of DNA in a 50ul reaction and purified with 1.8x AMPure. We added barcodes and Illumina adaptors using Kapa HIFI HotStart Readymix in 50uL reactions with 1ul of previous PCR product with a 65°C annealing temperature and primers Pu1F-idx (5'-AATGATACGGCGACCACCGAGATCTACACACGTAGGCCTAAATGGCTGTGAGAGAG CTCAG-3') and Pu1R-idx (5'-CAAGCAGAAGACGGCATACGAGATNNNNNNNNNGACCGTCGGCACTTTATCAATCT CGCTCCAAACC-3') and sequenced on a 300 cycle NextSeq 500/550 Mid Output v2 kit with PE150bp reads. We aligned sequencing reads to the input library using BWA mem (Li, 2013). We then calculated RNA/DNA ratio for each sequence and defined active tiles as ones at least two standard deviations above the average negative sequence.

3.5.4 *Design of orthologs and ancestral sequences*

We identified all orthologs using the UCSC liftover tool with a minimum match of 0.9. For each sequence, we determined the longest ortholog, and set it to 194bp around the center. We then used LiftOver to identify the end points in other species. 348 of the 697 sequences were present through squirrel monkey, and we decided to use squirrel monkey as our outgroup moving

forward. For ancestral reconstruction, we trimmed the hg38 phyloP 20way tree to the 11 species of interest and ran the Fastml heuristic (Ashkenazy et al., 2012). We aligned each sequence with ClustalO to obtain a multiple sequence alignment (Sievers et al., 2011), and then ran FastML (v3.1) with default settings on that alignment and the phyloP tree to create ancestral reconstructions.

3.5.5 *Prediction of tiling and evolutionary results*

We trained the gksvm-1.2 from Ghandi et al using the 500 top- and bottom-scoring lenti-MPRA sequences from Inoue et al as the positive and negative training sets, respectively, with default settings (Ghandi et al., 2014; Inoue et al., 2016). We used this model to predict scores for all tiles, and calculated the Spearman rho with our functional data. We next predicted scores for our positive human tiles and predicted negative orthologs from rhesus, vervet and marmoset and performed a two-sample t-test for each comparison. We calculated delta gkm-SVM scores by subtracting the predicted score of each ortholog from the predicted score of the human ortholog. We then predicted all eleven orthologs and nine ancestral nodes for all 348 enhancers.

3.5.6 *Functional testing of orthologs and ancestral sequences*

All orthologs and ancestral sequences were synthesized as part of an Agilent 230mer 244K array. We appended 5bp degenerate barcodes to each sequence by amplifying off the array with JK_R48_5N_HSSR (5'-CCGGCCGAATTCGTCGANNNNNCCATTGAGCACGACAGC-3') and HSS_cloning-F (5'-TCTAGAGCATGCACCGG-3'). We then cloned the library into the STARR-seq vector in NEB 3020 cells, transfected into HepG2 cells, and prepared sequencing libraries as described above. Since some orthologs have very similar sequences, we aligned sequencing reads to our reference

and only extracted error-free matches. We then calculated RNA/DNA ratios, and averaged across all barcodes for a given ortholog.

3.5.7 *Molecular characterization*

We first looked to see whether the turnover of any transcription factor motifs correlated with functional scores. We ran FIMO to identify TF motifs from HOCOMOCO v9 that were lost or gained in at least one ortholog for each enhancer (Grant et al., 2011; Kulakovskiy et al., 2012). For one enhancer at a time, we ran a linear regression for the presence or absence of each TF motif against the functional scores of all orthologs tested. For each TF, we then tested whether the mean slope across all enhancers was equal to zero using a two-sample t-test and filtered for transcription factors with an FPKM $\geq$ 10, for a list of 134 motifs. We further filtered the list for TFs that turned over in at least 20 enhancers, further narrowing the list to 115 motifs.

We next looked to see whether any sequence mutations in an enhancer correlated with functional scores of the orthologs. For each enhancer, we performed a multiple sequence alignment using ClustalO. For each site along the enhancer (skipping the first to avoid alignment artifacts), we characterized the allele as ancestral or derived. For each site with a singleton derived allele in at least one ortholog, we performed a K-S test to see whether the allele associated with the functional scores. We then corrected the p-values for the number of sites along the enhancer that had derived alleles. For each site with only a single derived allele present in at least one species, we characterized the nucleotide change and summed the number of events over all enhancers. We calculated the Fisher's exact p-value for each type of mutational event, using Bonferroni's correction to adjust for multiple hypothesis testing. We then looked to see what fraction of C→T and G→A mutations disrupted CpGs, and calculated the Fisher's exact p-values.

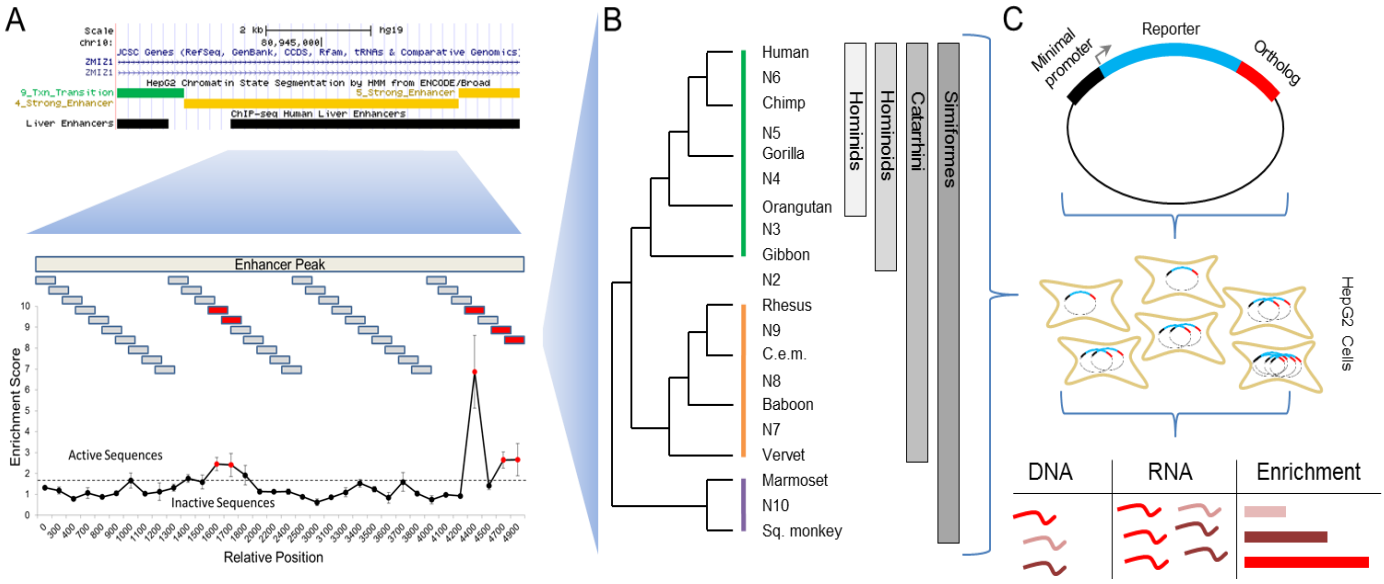


Figure 3.1. Schematic of experimental design.

(A) We identified potential hominoid-specific enhancers by intersecting hominoid-specific ChIP-seq predicted enhancers from primary human liver with ChromHMM-predicted strong enhancers in HepG2 cells (screenshot from *http://genome.ucsc.edu*) (Kent et al., 2002). We then tiled across each candidate enhancer using 194nt sequences and identified 697 tiles that were active in the STARR-seq reporter assay in HepG2 cells. **(B)** We located orthologous sequences in 11 primates and computationally reconstructed 9 ancestral sequences for 348 of the active tiles. **(C)** We then cloned all 20 present-day or ancestral orthologs per tile and performed STARR-seq again in HepG2 cells.

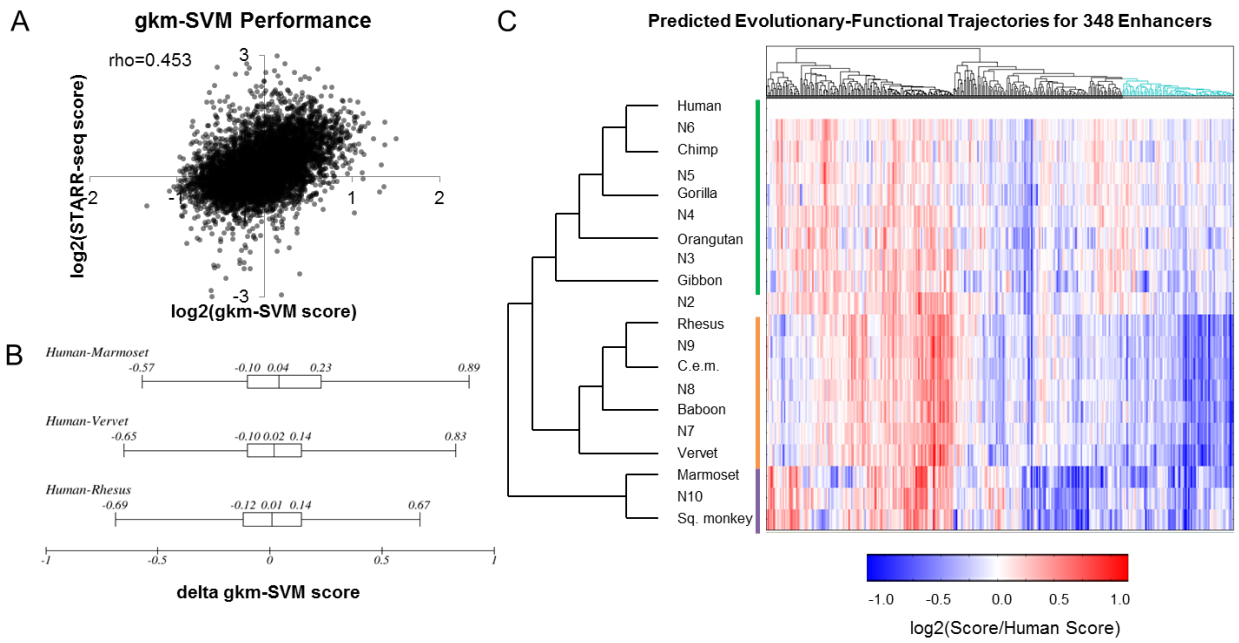


Figure 3.2. Performance of Computational Predictions.

(A) We trained the gapped-kmer support vector machine classifier (gkm-SVM) on an independent reporter assay experiment conducted in HepG2 cells. We then predicted the functional activity of all of our human sequence tiles and found a modest correlation with our functional data. **(B)** The distributions of differences (delta gkm-SVM score) in predicted score between the human vs. marmoset, vervet, or rhesus ortholog for all active human tiles. **(C)** Predicted scores for all orthologs of the 348 human-active enhancer tiles, normalized to the human ortholog. Clades are denoted by colored lines (green: hominoid, orange: Old World monkeys, purple: New World monkeys). Cyan clade of dendrogram denotes a group of 82 enhancer tiles that follows expectations for hominoid-specific enhancers as predicted by ChIP-seq comparative genomics.

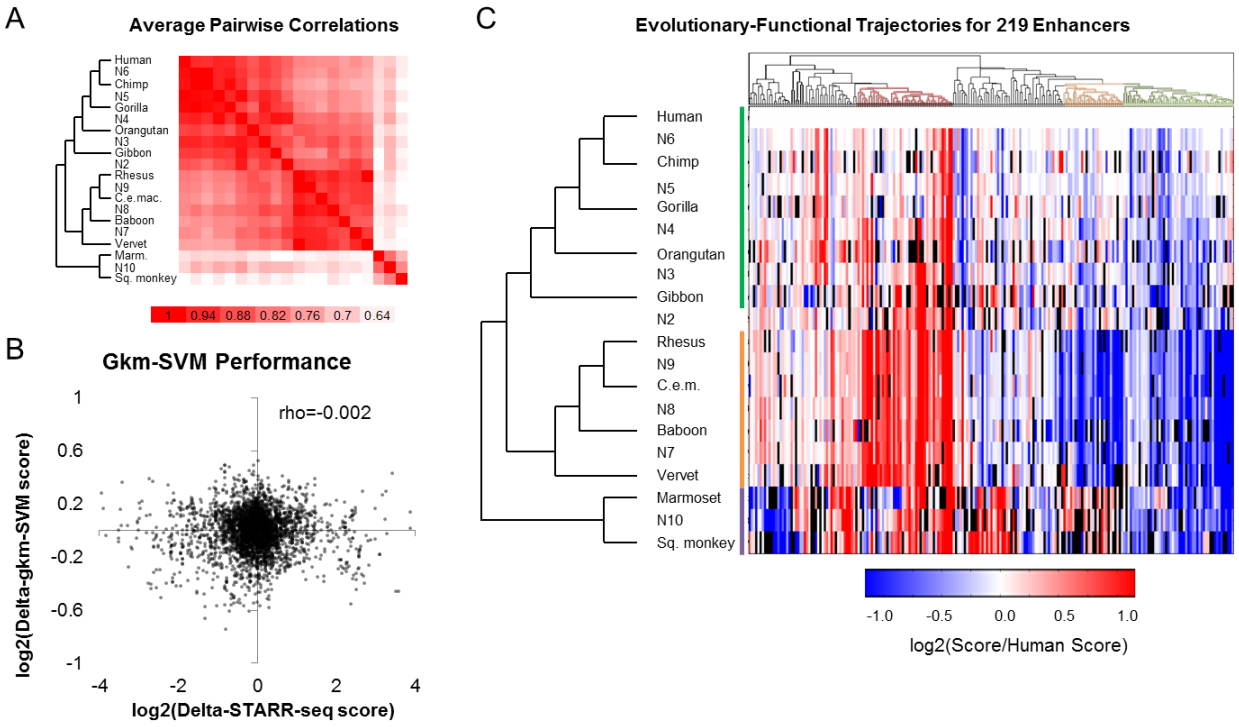


Figure 3.3. Functional scores for orthologs and ancestral sequences.

(A) The average pairwise Spearman correlation of functional scores between any two orthologs across all enhancer tiles tested. **(B)** Correlation between the STARR-seq enrichment scores, normalized to the enrichment score of its human ortholog (log2[non-human score/human score]), and gkm-SVM predicted scores, similarly normalized to the predicted score of its human ortholog (log2[non-human prediction/human prediction]). **(C)** Functional scores normalized to human for all orthologs of the 219 enhancer tiles. Black bars represent missing data. Clades are denoted by colored lines (green: hominoid, orange: Old World monkeys, purple: New World monkeys). Groups are color-coded in the dendrogram; red: relatively higher activity in in Old World monkeys, orange: relatively lower in Old World monkeys, green: relatively higher activity in either humans or hominoids.

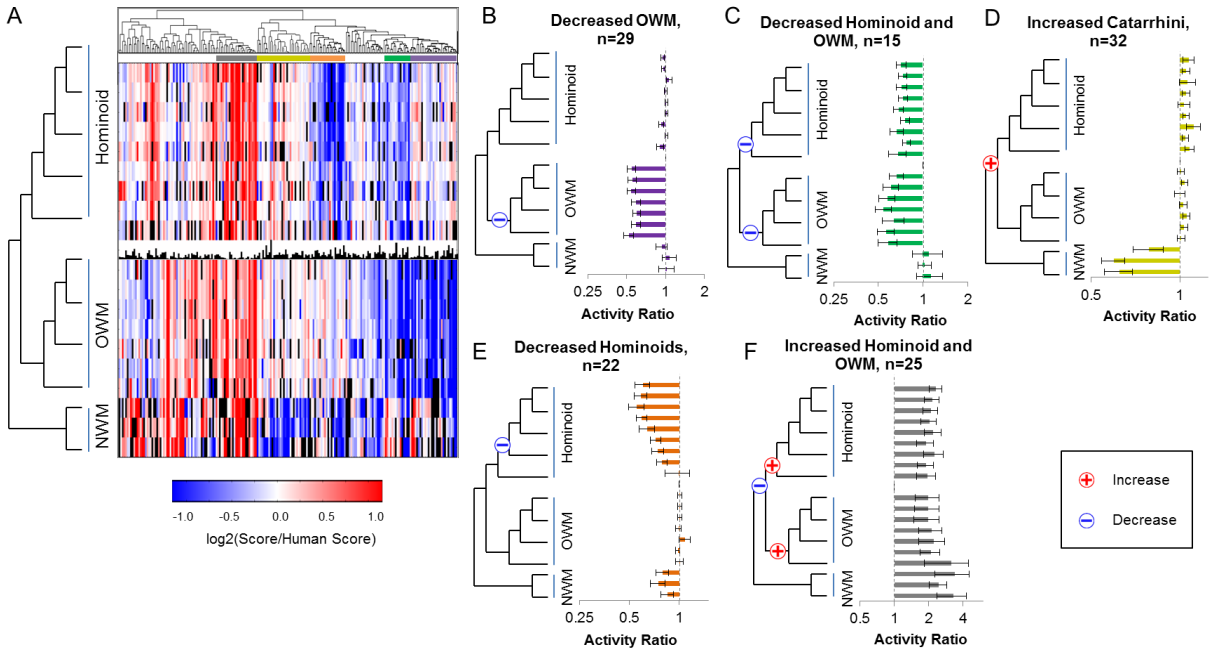


Figure 3.4. Common Patterns of Enhancer Modulation over the Primate Phylogeny.

(A) Functional scores for all enhancer tiles normalized to the MRCA to hominoids and Old World monkeys (N2). Black bar graph in the center contains the N2 score for each tile. Color bars above the heatmap indicate subsets of enhancer tiles exhibiting coherent patterns with respect to gain/loss of activity across the primate phylogeny, including: increased in human and hominoid (grey), increased in Catarrhini (yellow), decreased in hominoids (orange), decreased in hominoids and Old World monkeys (green), decreased in Old World monkeys (purple). **(B)** The average score normalized to N2 for each species across the group of 29 enhancer tiles with decreased activity in Old World monkeys. Blue "-" indicates the timing of a loss of activity event. Error bars are one standard error. Dashed grey line at a ratio = 1. **(C)** Same as Figure 4B for a group of 15 enhancer tiles with decreased activity in hominoids and Old World monkeys. **(D)** Same as Figure 4B for a group of 32 enhancer tiles with increased activity in catarrhini. Red "+" indicates the timing of a gain of activity event. **(E)** Same as Figure 4B for a group of 22 enhancer tiles with decreased activity in hominoids. **(F)** Same as Figure 4B for a group of 25 enhancer tiles with decreased activity only in N2.

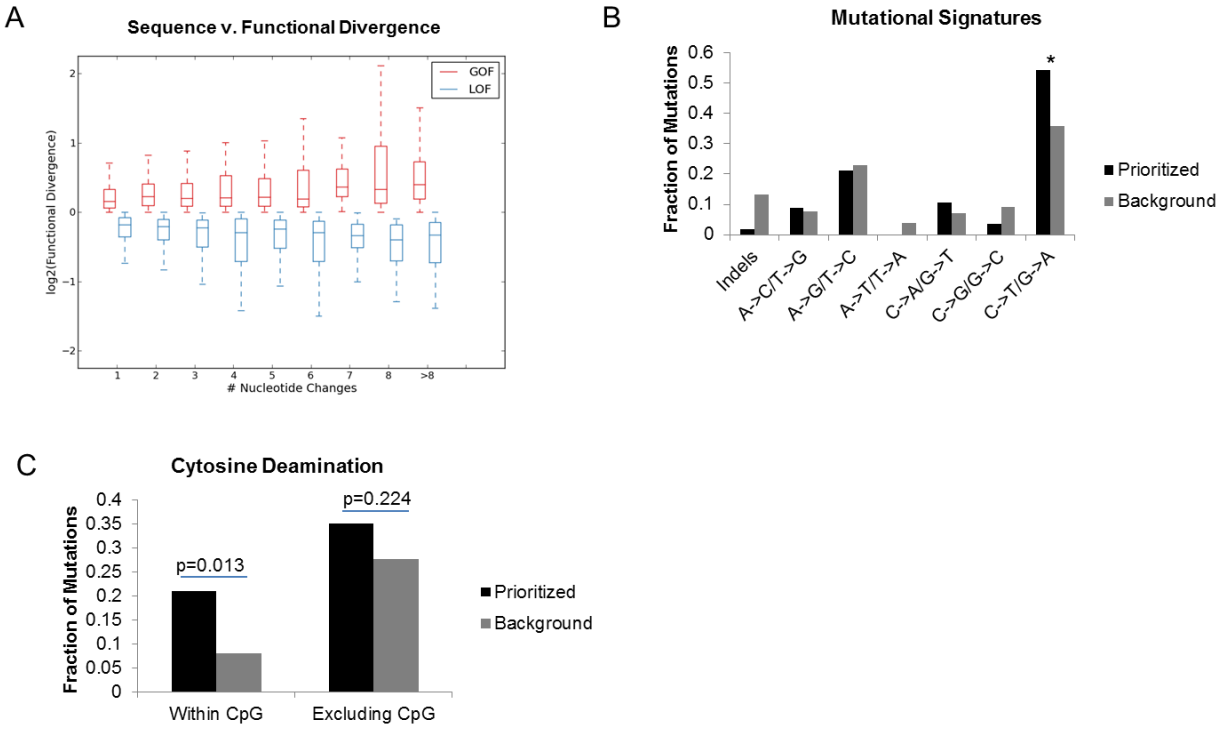


Figure 3.5. Molecular characterization of enhancer modulation.

(A) For every branch along the tree, we calculated the nucleotide and functional divergence. The number of nucleotide changes is on the x-axis and the log2 of the ratio of the daughter to ancestral functional score is on the y-axis. **(B)** The fraction of indels, A→C, T→G mutations, A→G, T→C mutations, A→T, T→A mutations, C→A, G→T mutations, C→G, G→C mutations and C→T, G→A mutations in our set of 57 prioritized mutations (those associated with a significant functional difference) in black and 2,766 background mutations (those associated with a non-significant functional difference) in grey. Asterisk represents a Bonferroni-corrected p-value < 0.05 (Fisher's exact test). **(C)** The fraction of prioritized (black) and background (grey) mutations that are cytosine deamination events within CpGs and C→T , G→A mutations not disrupting a CpG.

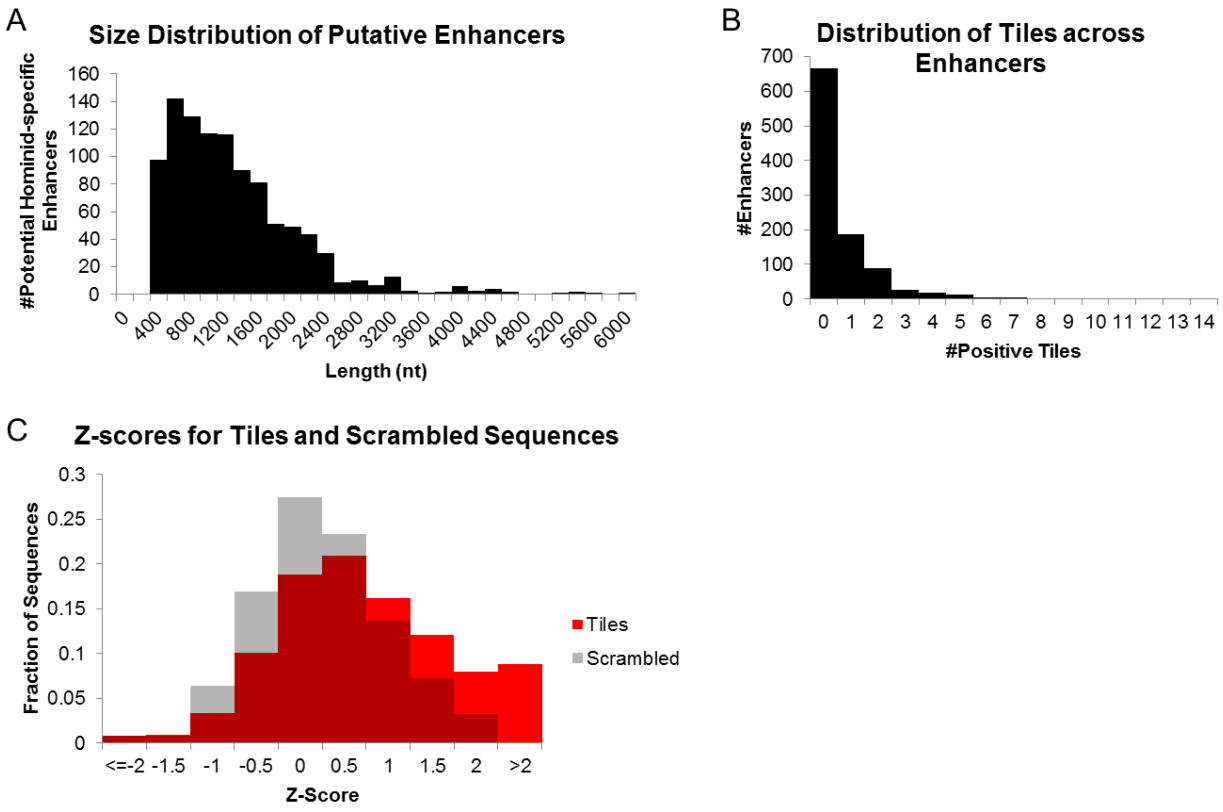


Figure 3.6. Tiling Across Large Enhancer Regions.

(A) Histogram of the size for each putative hominoid-specific gain-of-function enhancer defined by the intersection of H3K27ac ChIP-seq from primary tissue and HepG2 ChromHMM strong-enhancer calls. **(B)** Histogram of the number of positive tiles (defined as a score two-standard deviations above the average negative control) per putative enhancer. **(C)** Histogram of Z-scores for 6,724 tiles and 124 scrambled sequences.

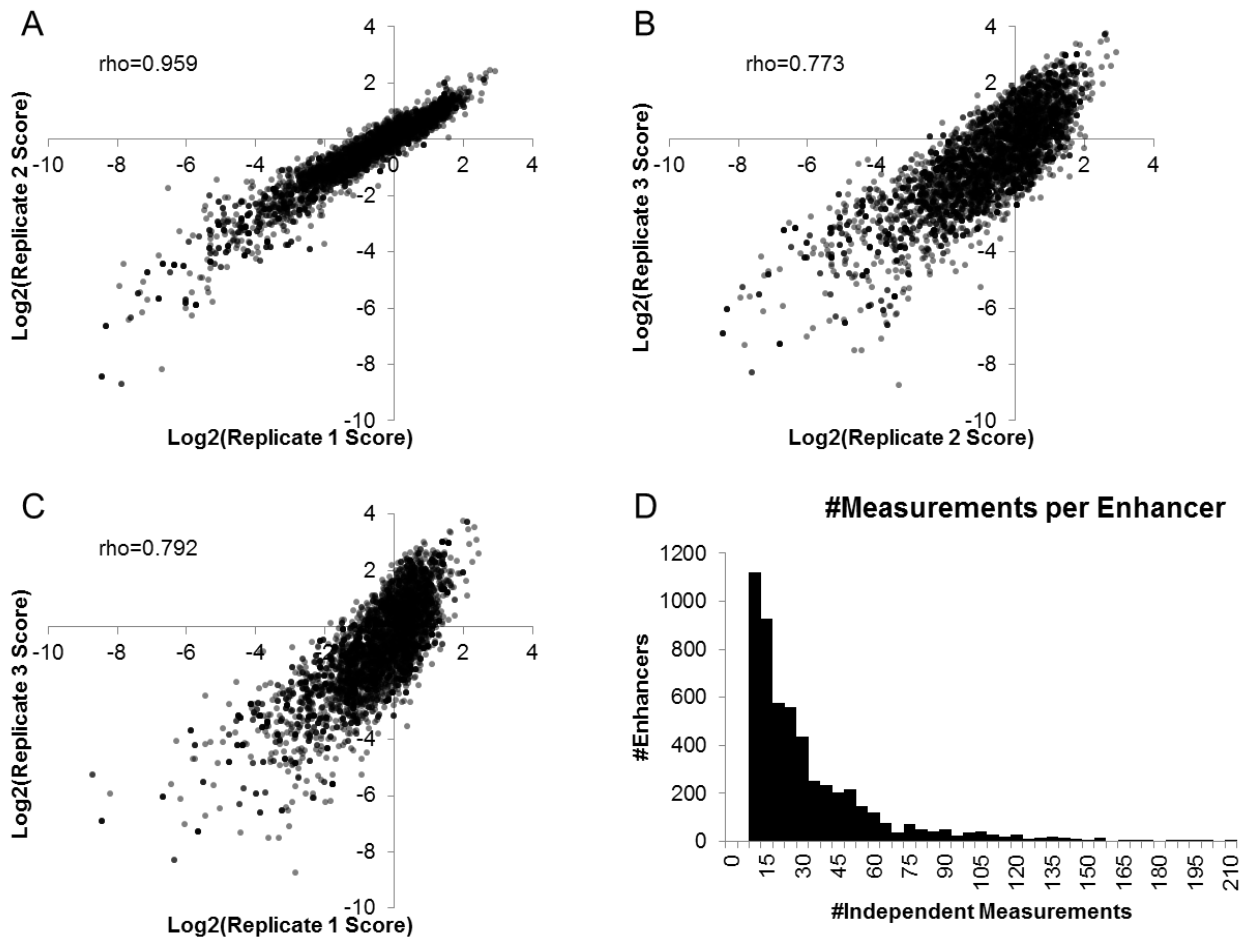


Figure 3.7. Reproducibility of Functional Scores.

(A) Spearman correlation between $\log_2$(normalized RNA/DNA) for biological replicates 1 and 2. **(B)** Spearman correlation between replicates 2 and 3. **(C)** Spearman correlation between replicates 1 and 3. D) Histogram of the number of independent measurements (barcodes) for each enhancer summed across all three replicates.

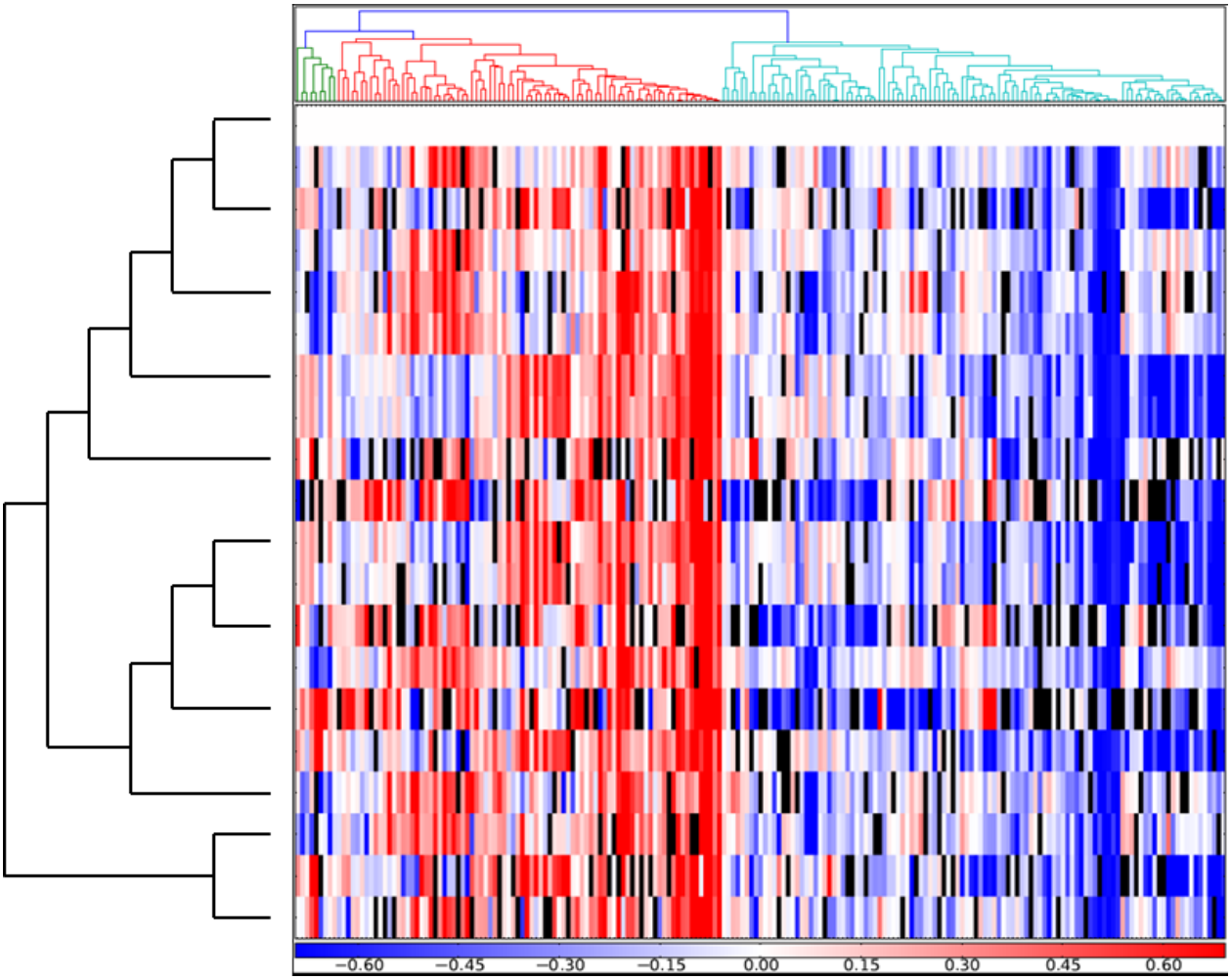


Figure 3.8. Permuted Species' IDs.

Species' IDs were randomly permuted (seed=40) before normalization. Orthologs were hierarchically clustered and the same heatmap from Figures 2C and 3C was created.

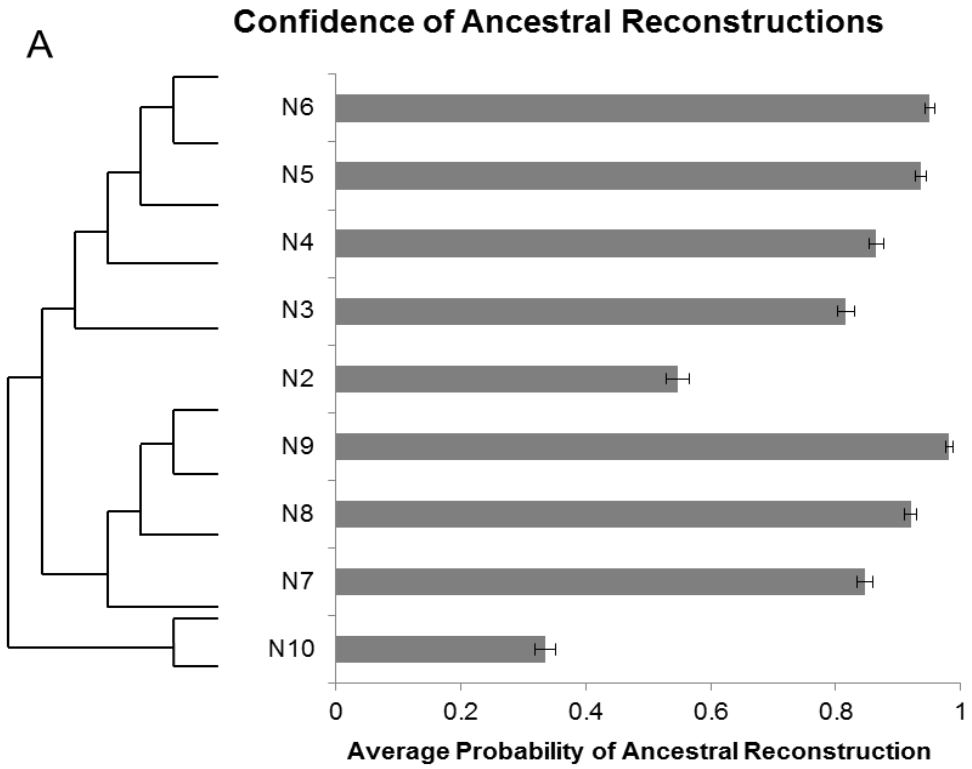


Figure 3.9. Confidence of Ancestral Reconstructions.

The average marginal probability for each ancestral node across all 348 enhancers. Error bars indicate one standard error around the average probability for each node.

Lost OWM (4B)	Lost Both (4C)	Decreased in NWM (4D)	Lost Hominoids (4E)	Decreased in N2 (4F)
1:229283881-229284067	15:74737651-74737837	20:48937797-48937983	16:9221603-9221789	22:37898661-37898841
20:62868160-62868346	10:73102300-73102482	16:9221503-9221689	7:1004980-1005162	X:47060260-47060446
16:88570803-88570989	11:67914034-67914208	22:36902314-36902500	20:49013197-49013382	10:72123798-72123984
20:30309735-30309921	20:30300661-30300843	21:43098235-43098421	15:76016249-76016435	19:47600169-47600344
22:19948507-19948687	7:47612782-47612963	5:135459111-135459297	18:46423007-46423190	12:57495743-57495929
11:72868556-72868742	17:55601407-55601591	2:135215437-135215618	1:17086917-17087103	3:193429310-193429496
16:29273104-29273289	22:19879504-19879690	22:37897361-37897542	5:1563904-1564089	1:181104188-181104373
14:102197955-102198133	22:43201660-43201846	22:51139238-51139424	1:243669382-243669567	10:73746298-73746484
1:234775081-234775267	15:72040050-72040236	20:10647806-10647992	16:88268247-88268428	10:102101014-102101200
7:2797578-2797763	11:73041956-73042142	18:11994705-11994892	20:48937998-48938183	20:10647109-10647290
2:235772065-235772251	17:48188805-48188991	18:3257113-3257288	1:36947817-36948003	1:234951371-234951556
19:7463704-7463890	1:153517680-153517866	20:49407297-49407483	16:27482121-27482306	8:126286122-126286308
1:235010682-235010868	5:139023120-139023305	22:19949905-19950089	12:6481478-6481664	2:232566760-232566946
6:168068354-168068541	14:68844251-68844437	22:51137644-51137818	7:98970268-98970454	8:341766-341949
16:29271415-29271594	2:87785589-87785775	20:11668304-11668490	8:140981329-140981508	9:138899788-138899963
6:36095326-36095511	1:19372125-19372301	7:47611279-47611465	20:48755297-48755483	20:47002497-47002683
19:39919165-39919349		20:48937400-48937580	2:10425760-10425932	7:1506685-1506860
15:39726916-39727100		17:48739205-48739390	19:46144967-46145153	22:21240004-21240190
11:62321028-62321214		1:112275481-112275667	1:150975281-150975464	22:19879610-19879796
16:88266497-88266676		22:30609311-30609491	1:110438782-110438965	11:67914328-67914514
1:112275982-112276167		19:47599566-47599748	16:70780703-70780889	10:11917402-11917580
20:30311437-30311620		2:47301901-47302086	1:17087417-17087603	1:212769887-212770068
19:51855593-51855778		2:43521102-43521284		1:45121217-45121403
20:49011097-49011283		14:65414851-65415037		10:80944098-80944284
1:229283982-229284167		1:94533819-94534005		20:25520004-25520189
3:49561102-49561285		7:657316-657502		
3:49561003-49561182		9:116880443-116880628		
6:30756525-30756711		16:68824703-68824889		
11:67382032-67382217		15:86251102-86251284		
		12:122884051-122884237		
		15:76016555-76016741		
		20:48937897-48938083		

Table 3.1. Groups normalized to N2.

List of all enhancer tiles falling within each group of Figure 3.4. Coordinates are in hg19.

TF	Average Beta	Uncorrected p-value	#Enhancers	HepG2 FPKM
E2F1	-0.12501642	0.004387572	60	19.45
ATF2	0.260074918	0.011278218	20	70.09
STAT2	0.152115566	0.011486208	38	64.12
CEBPZ	-0.244088533	0.012861103	23	11.97
HIF1A	-0.140920359	0.033555269	40	31.65
DDIT3	-0.128364132	0.035059869	25	38.41
IRF3	0.125092184	0.053234218	42	47.21
BACH1	-0.102835442	0.07181549	43	12.65
SREBF2	-0.056772605	0.075829977	125	55.25
RXRB	0.100675782	0.081236664	53	42.44
NFYB	-0.138013239	0.090303964	21	14.29
STAT3	-0.109265054	0.093455712	36	47.6
FOXM1	0.091697255	0.099959015	26	13.01
HINFP	0.099751699	0.104541266	29	16.31
NFE2L2	0.104874769	0.134628284	55	162.79
NR5A1	-0.082320383	0.14559166	46	23.09
SMAD3	-0.048709795	0.170609126	73	44
USF1	0.108732464	0.172727673	41	20.5
RORA	0.08743307	0.172852158	23	30.04
JUN	-0.078705667	0.173182245	26	13.87
CREM	0.145326791	0.186295219	23	16.25
ATF6	-0.092015106	0.187374609	44	11.06
FUBP1	0.076722336	0.211290514	23	74.56
TCF3	0.060924361	0.214244973	65	39.39
YBX1	0.096311395	0.215138977	38	257.86
ATF1	0.110900505	0.216042024	23	39.51
XBP1	-0.092590559	0.217676347	23	317.2
SP3	-0.045067665	0.227143667	146	40.91
CEBPG	-0.097727974	0.272635307	22	19.56
CREB1	0.096182729	0.283715072	23	39.36
STAT1	-0.082661877	0.292031899	28	16.61
NFYC	-0.053787973	0.302276251	41	56.15
ELF1	-0.058163903	0.314456407	58	19.41
TFDP1	-0.036095975	0.33519783	20	29.46
TBX3	-0.077666007	0.342412244	50	32.25
CBFB	-0.060717191	0.366439517	50	16.75
SMAD4	-0.027288112	0.38984213	55	53.35
USF2	0.039965266	0.39397386	62	81.14
ENO1	0.057144221	0.394992005	50	2269.3
RBPJ	0.078023883	0.396791614	41	25.29
MYC	-0.054313333	0.399877845	45	50.48
RELA	-0.055791876	0.40217337	59	30.86

ARID5B	0.052840075	0.412644773	20	16.03
NR2F1	-0.033067421	0.41423659	68	45.42
CXXC1	0.052068677	0.41575877	33	29.57
BHLHE40	-0.0506766	0.462899371	32	38.84
MBD2	-0.041614311	0.48265278	55	14.47
TEAD3	-0.055039819	0.495951593	29	12.78
ETV5	0.033764193	0.524361188	46	45.8
NR2F1	-0.024062221	0.539780801	41	45.42
PPARG	-0.029600634	0.547017973	40	43.85
E2F4	-0.036970437	0.549697184	45	69.26
HSF1	-0.02394861	0.554332336	94	26.67
E2F6	-0.021204479	0.562860981	110	20.77
HES1	-0.031563936	0.573133815	79	11.74
TEAD4	-0.054503418	0.584072401	21	52.9
NFATC3	-0.053407602	0.585479629	26	48.15
ETV4	-0.036041661	0.605011675	62	252.05
ESRRA	0.028289232	0.608667617	49	13.79
PPARD	0.018866428	0.609992697	71	10.09
NR4A1	0.027204796	0.613225812	56	12.04
PPARA	0.018146699	0.614745301	93	14.03
IRF2	0.049307868	0.615493405	27	21.93
MLXIPL	-0.025300451	0.617117225	87	45.66
YY1	0.022547712	0.620568597	35	22.09
CTCF	-0.018943227	0.638977275	121	12.45
ARNT	-0.026936007	0.652342139	25	19.76
NR6A1	-0.030327809	0.657992131	37	22.48
KLF6	-0.021414357	0.658842664	113	55.28
HNF4G	0.015820195	0.659703314	85	13.71
SREBF1	0.02466531	0.668754616	66	29.35
MAFG	0.046312349	0.669079862	22	15.2
NFE2L1	0.046312349	0.669079862	22	178.51
IRF9	0.043714585	0.67149518	23	37.76
MAZ	-0.019138477	0.672160711	113	154.22
SP1	-0.016377724	0.677296736	138	20.34
ZNF219	0.017417192	0.681074237	89	26.24
ETS2	-0.022106314	0.681472739	67	49.46
EPAS1	0.02102088	0.68677748	51	136.86
TEAD1	-0.01853228	0.690791322	61	14.03
PPARA	-0.036052352	0.705885163	23	14.03
ZNF143	-0.013922137	0.712873706	116	27.29
ZFX	-0.015703312	0.713556874	130	10.56
ELF2	0.019492063	0.71525941	70	21.98
TGIF1	0.015305928	0.721913924	74	57.83

FOS	-0.026626576	0.722518772	37	12.15
HSF2	-0.011416407	0.732473402	144	19.95
ELF3	0.017528286	0.751487516	53	41.94
MAX	-0.020478599	0.761765348	45	22.52
KLF3	0.015168988	0.763587762	68	29.05
FOXA1	-0.018064811	0.766125746	27	33.03
SP1	-0.009672419	0.793405199	136	20.34
BRCA1	-0.021699271	0.823226696	23	10.02
SMAD1	0.006724351	0.842185526	137	12.64
FOXO3	-0.012482857	0.855770018	25	16.99
NFKB1	-0.008355756	0.868690297	68	11.71
BCL6	0.014788318	0.879643431	22	41.22
HNF4A	-0.006387617	0.885027303	75	61.72
STAT6	-0.013707517	0.890780854	22	47.35
SMARCC1	0.008376169	0.896301821	48	36.52
NR2F6	-0.005106394	0.897452134	71	23.55
BPTF	0.009387172	0.903568887	28	56.23
NFAT5	-0.009068745	0.918287916	36	41.25
PPARG	0.003316714	0.922819159	97	43.85
JUND	-0.005448099	0.927396287	31	45.63
TFAP4	-0.004254026	0.930397081	78	11.92
NR2C1	-0.003995714	0.93344117	49	54.81
FOXA3	0.004500891	0.938526907	34	36.28
NR1H4	0.00414001	0.948696641	41	21.26
RXRA	-0.002121571	0.954889402	90	14.37
FOSL2	0.00235208	0.964277822	32	13.77
TCF12	-0.000678025	0.986873282	77	36.94
UBP1	-0.000893478	0.991358464	31	61.86
SMAD2	0.000153663	0.996494339	84	30.79
ZBTB7B	6.22E-05	0.998803219	145	13.23

Table 3.2. Motifs associating with functional scores.

List of transcription factor binding motifs associated with functional scores, ordered by significance.

Enhancer	Position	Ancestral	Derived	Ancestral Context	Derived Context
1:112275982-112276167	34	T	C	GTA	GCA
1:112275982-112276167	84	C	T	GCC	GTC
1:211790081-211790267	53	C	T	CCG	CTG
1:229283881-229284067	66	G	A	GGC	GAC
1:229283881-229284067	80	G	A	AGC	AAC
1:36947817-36948003	9	A	G	CAC	CGC
10:102101014-102101200	9	C	T	ACA	ATA
10:102101014-102101200	126	G	A	CGA	CAA
10:102101014-102101200	135	C	A	GCT	GAT
10:102101014-102101200	175	G	A	CGG	CAG
10:72123798-72123984	36	C	T	CCG	CTG
10:72123798-72123984	60	C	T	CCA	CTA
10:72123798-72123984	62	T	C	ATG	ACG
10:72123798-72123984	73	C	T	GCT	GTT
10:72123798-72123984	169	C	A	TCT	TAT
10:80944000-80944182	91	A	G	CAT	CGT
16:27482121-27482306	4	C	T	GCG	GTG
16:27482121-27482306	83	C	A	CCA	CAA
16:52607607-52607785	13	T	C	TTG	TCG
16:52607607-52607785	18	G	T	TGG	TTG
16:52607607-52607785	35	G	A	AGC	AAC
16:52607607-52607785	44	C	T	GCT	GTT
16:52607607-52607785	87	A	G	TAG	TGG
16:52607607-52607785	135	C	T	CCT	CTT
16:52607607-52607785	170	G	A	TGG	TAG
16:70780703-70780889	66	G	T	GGT	GTT
16:88266497-88266676	40	G	C	CGC	CCC
16:88266497-88266676	93	T	C	GTG	GCG
16:88266497-88266676	143	G	A	CGC	CAC
16:88570803-88570989	3	G	A	GGG	GAG
16:9221603-9221789	43	C	T	GCA	GTA
16:9221603-9221789	72	T	G	TTG	TGG
18:3257113-3257288	45	T	G	ATG	AGG
18:3257113-3257288	107	C	T	CCA	CTA
18:3257113-3257288	158	C	T	CCA	CTA
19:51855593-51855778	3	C	T	ACC	ATC
2:43521102-43521284	122	G	A	CGA	CAA
20:49011097-49011283	15	G	A	CGA	CAA
20:49011097-49011283	42	G	T	GGG	GTG
20:49011097-49011283	53	C	T	TCT	TTT
20:49011097-49011283	60	A	G	AAG	AGG
20:49011097-49011283	75	G	A	CGC	CAC

20:49011097-49011283	107	T	C	CTA	CCA
22:19879504-19879690	1	C	T	ACG	ATG
22:19879504-19879690	105	G	A	CGG	CAG
5:135402707-135402889	5	T	C	GTT	GCT
5:135402707-135402889	54	C	T	TCT	TTT
5:135402707-135402889	108	A	C	CAT	CCT
5:135402707-135402889	179	T	G	CTG	CGG
5:139023120-139023305	114	G	A	AGG	AAG
7:1004980-1005162	6	G	C	GGC	GCC
7:1004980-1005162	58	C	T	TCG	TTG
7:1004980-1005162	113	T	C	CTC	CCC
7:1004980-1005162	126	A	C	CAC	CCC
7:1004980-1005162	135	-	G	G--	GG-
7:155744643-155744829	74	T	C	GTG	GCG
7:155744643-155744829	144	C	T	GCC	GTC

Table 3.3. Mutations correlating with functional scores.

List of all mutations correlating with functional scores after a Bonferroni correction. Coordinates are in hg19. Positions are zero-indexed.

Chapter 4. FUNCTIONAL SCREENING OF THOUSANDS OF OSTEOARTHRITIS-ASSOCIATED VARIANTS FOR REGULATORY ACTIVITY

4.1 ABSTRACT

The interpretation of variants from Genome Wide Association Studies (GWAS) is confounded for two main reasons: 1) SNPs are often in linkage with dozens or hundreds of other common polymorphisms and 2) we lack accurate tools to predict the effect of non-coding variation. Therefore, while GWAS have identified several SNPs associated with osteoarthritis, we still do not understand the genetic mechanisms leading to the disease. In this study, we identified all SNPs in linkage with 35 different candidates identified through GWAS, and functionally screened all 1605 OA-associated variants for regulatory activity using STARR-seq. With a false discovery rate of 5%, we identified six SNPs with differential enhancer activity. Our most significant hit, rs4730222, drives increased expression of an alternative isoform of *HBP1* in osteosarcoma and chondrosarcoma cell lines, as well as in chondrocytes derived from osteoarthritis patients.

4.2 INTRODUCTION

Through studying families and twins with osteoarthritis, researchers have predicted that a genetic component accounts for approximately 50% of the disease risk (arcOGEN). Recently, with the advent of novel genetic approaches, groups have identified genes and variants associated with the onset of osteoarthritis. However, to date, we still do not have a complete picture of the genetic mechanism behind the disease. Most of the genetic links have been identified through Genome Wide Association Studies

(GWAS). However, many of these variants fall in non-coding DNA and are often in tight linkage with neighboring polymorphisms, making it challenging to identify the causal variant.

Traditionally, researchers have prioritized non-coding variants by testing their ability to activate expression of a reporter gene (*e.g.* luciferase or β-galactosidase). Although this type of assay has been considered a gold standard for measuring regulatory function, it is limited to testing sequences individually, and is therefore not scalable to test thousands of variants. To overcome this limitation, our lab and others have developed the massively parallel reporter assay (MPRA) to increase the throughput of this type of assay (Patwardhan 2009, Patwardhan 2012, Melnikov 2012, Arnold 2013). MPRAs test a library of thousands of independent sequences, cloned with the same reporter gene. However, instead of measuring the direct output of the reporter, MPRAs rely on next-generation sequencing to link each sequence to its effect on the transcription of the reporter gene.

In this study, we focused on a list of 35 candidate GWAS variants correlated with osteoarthritis. 34/35 of these variants are non-coding, and there are a total of 1605 single nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD) with these variants in Europeans, with an r-squared greater than 0.8. We hypothesize that only a fraction of the 1605 variants contribute to disease, and set out to identify the subset that likely contribute for further study. We used STARR-seq, a recent implementation of MPRAs, to measure all variants in a single experiment (Arnold 2013). Using array-based technology, we synthesized 196bp of genomic sequence, centered at the SNP, for both the major and minor allele for each of the 1605 variants. We screened these sequences for regulatory activity in Saos-2 cells, an osteosarcoma cell line, and identified six with significantly different activity between alleles (Benjamini-Hochberg, FDR < 0.05). We further characterized the most significant SNP, rs4730222, and show that it increases expression of an alternative isoform of *HMG-Box Transcription Factor 1* (*HBP1*) in osteosarcoma and chondrosarcoma cell lines, as well as in chondrocytes derived from osteoarthritis patients.

4.3 RESULTS

4.3.1 *Functional characterization of regulatory activity for >2000 alleles*

We compiled a list of 35 candidate variants associated with osteoarthritis in European populations from various studies up until May 2017, with minor allele frequencies >5%. Each SNP represents an independent signal with p<5e-8 (**Table 4.1**). For each of these variants, we identified all SNPs in LD with an $R^2 > 0.8$ in Europeans using rAggr (**Figure 4.1A**). This resulted in a list of 1605 SNPs, each with two alleles. For each of the 3210 alleles, we extracted 196bp of genomic sequence, centered on the SNP, and synthesized them on an array with library-specific adapters as 230bp oligos (**Figure 4.1B**). During amplification from the array, we added 5bp degenerate barcodes, such that each allele would have multiple barcodes (**Figure 4.1C**). By doing so, we were able to measure the effect of each allele several independent times in the same experiment. We then cloned our barcoded library into the human STARR-seq vector, transfected into Saos-2 cells, an osteosarcoma cell line, and calculated enrichment scores (normalized ratio of RNA reads / ratio of DNA reads) for each barcode-allele combination (**Figure 4.1D-1F**). For all alleles with greater than five independent measurements over three biological replicates (separate transfections), we averaged the allele enrichment scores for a single value. We obtained scores for 2318 of 3210 alleles (72.2%), and scores for both alleles from 753 variants (46.9%).

First, we wanted to check whether our enhancer scores correlated with biochemical marks for putative enhancers. We collapsed our 2318 alleles into 1203 unique SNPs and split the SNPs into quintiles based on their expression level in the assay. We then overlapped each quintile with various biochemical datasets marking putative enhancers in cartilage and bone (H3K27ac in bone marrow-derived chondrocytes, H3K27ac in human embryonic limb buds from

E33, E41, E44, and E47, and ATAC-seq in knee OA cartilage) (**Figure 4.3**). We find a significant enrichment for the highest scoring quintile overlapping H3K27ac ChIP-seq peaks in embryonic limb bud from E33, E41, E44, and E47 (chi-square, p-values: 0.0177, 0.0016, 0.0073, 0.0012). The enrichment is not significant in knee OA cartilage ATAC-seq and H3K27ac in bone marrow-derived chondrocytes (chi-square, p-values: 0.129 and 0.0669). These results are in line with our testing sequences in an osteosarcoma cell line and not a cartilage-derived line (**Figure 4.3**). The highest scoring quintile includes 240 unique genomic sequences, 67 of which overlap putative enhancers from at least one dataset. This enrichment provides confidence that we are testing for functionally and biochemically active enhancers and may identify novel enhancers in bone. Several biochemical datasets have been generated in bone and cartilage, but to date, this is the first large-scale functional annotation of enhancers in bone or cartilage.

Next, we wanted to determine whether any alleles drive differential regulatory activity for the 753 SNPs, where we measured both alleles. After correcting for multiple testing with Benjamini-Hochberg with a 5% FDR, we identified 6 SNPs driving differential expression (**Figure 1G**). The most significant SNP, rs4730222, is located in the 5' UTR of several isoforms of *HBP1*, containing an alternative transcriptional start site. Other significant SNPs (in order of significance) include rs2286798 (intronic to *ITIH1*), rs80095766 (intronic to *COG5*), rs11745630 (downstream of *PIK3R1*), rs6976 (3' UTR of *GLT8D1*), and rs1563351 (upstream of SLC30A10). We chose to further characterize rs4730222 since it was the most significant SNP and drove high expression (enrichment score > 1) for one of its alleles.

4.3.2 *rs4730222 increases expression of a truncated HBP1 isoform with an alternative TSS*

rs4730222 falls within the 5' UTR from isoforms of *HBP1* with an alternative transcriptional start site. Moreover, it has several marks associated with active promoters:

H3K27ac peak (mark for active enhancers and promoters) from bone marrow-derived chondrocytes (Roadmap) and human embryonic limb bud at E33, E41, E44 and E47 (Cotney et al., 2013), H3K27ac and H3K4me3 peaks (mark for active promoters) in ENCODE layered data, and ATAC-seq peaks (mark for open chromatin) from articular knee cartilage of OA patients (Liu et al., 2018) (**Figure 4.2A**). We therefore hypothesized that the variant may alter expression of *HBP1*. To test this, we first confirmed that the alternative start site is utilized in Saos2 and Sw1353 cells with qRT-PCR with primers contained within the UTR as well as spanning to the following exon. When trying to amplify the isoform from the alternative TSS to canonical stop, there was no amplification, suggesting that the alternative TSS may belong to a truncated isoform of *HBP1*. After confirming that the SNP is transcribed in osteogenic and chondrogenic cells, we next genotyped several cell lines (Sw1353, Tc28a/2, Saos-2, chondrogenic progenitor cells) for rs4730222, with the hopes of finding a heterozygote line. Since Sw1353 was heterozygote for rs4730222, we decided to test for allelic expression imbalance (AEI) of the transcribed SNP. In all three biological replicates, we found a significant allelic imbalance (Fisher's exact test, $p<1e-5$), with the minor allele showing a 1.313-1.432-fold relative enrichment in RNA/DNA compared to the major allele (compared to a 2.47-fold enrichment in the reporter assay) (**Figure 4.2C**).

However, in Sw1353, rs4730222 is in linkage with many additional SNPs, which may be regulating expression of the isoform. In order to test for causality, we introduced the minor allele for rs4730222 into Saos-2 cells, which are homozygous for the reference allele, through CRISPR-mediated homology directed repair (HDR). We performed four biological replicates, and compared the fraction of indel-free reads in the RNA coming from the minor allele to the fraction of indel-free reads in the DNA coming from the minor allele. Similar to the Sw1353 AEI

(1.313-1.432-fold), we identified a 1.38-1.64-fold relative enrichment in RNA/DNA for the minor allele compared to the major allele (Fisher's exact test, p<1e-5) (**Figure 4.2D**).

Next, we wanted to test whether this phenomenon occurs in osteoarthritis patients. To do so, we tested for AEI in chondrocytes derived from osteoarthritis patients. For each of 10 patients heterozygote for rs4730222, we extracted RNA and DNA from total joint replacements and amplified and sequenced the 5'UTR containing the SNP (three technical replicates per patient) from DNA and cDNA. Despite low cDNA concentrations and low expression of the alternative TSS, we observed an overall AEI in accordance with our reporter assay and cell models (Mann Whitney U Test, p=0.00386). However, we do note that one patient appears to exhibit AEI in the opposite direction (**Figure 4.5**).

4.4 DISCUSSION

We set out to prioritize non-coding variants associated with osteoarthritis for further validation. In order to do so, we identified 1605 SNPs in high LD with 35 lead variants from different GWAS. We then modified the STARR-seq assay by incorporating random barcodes in order to measure the effect of each variant several independent times in a single experiment. Doing so increased our confidence (by averaging across multiple measurements), and allowed us to test for significance by comparing the distribution of scores for both alleles. We identified six SNPs, which each drove differential expression, with an FDR cutoff of 5%.

Our most significant SNP, rs4730222 was particularly interesting for several reasons. First, the SNP is located in the 5' UTR of multiple isoforms of *HMG-Box Transcription Factor 1* (*HBP1*), a transcriptional repressor that negatively regulates the Wnt-beta-catenin pathway. The Wnt pathway has been heavily implicated in osteoarthritis development and progression in both human and mouse models (Luyten et al., 2009). Second, HBP1 also regulates superoxide

production, and oxidative stress has also been shown to contribute to osteoarthritis development and progression (Berasi et al., 2004; Scott et al., 2010). Third, the SNP falls within an H3K27ac peak (mark for active promoters and enhancers) in chondrocytes and human embryonic limb buds and both H3K27ac and H3K4me3 (mark for active promoters) in layered ENCODE data. These marks suggest that rs4730222 falls in an active promoter for *HBP1*. Fourth, the SNP is transcribed in these isoforms, which allowed us to perform AEI in various models. For these reasons, we performed further validation for rs4730222.

First, we looked for AEI in Sw1353, a chondrosarcoma cell line that is heterozygous for rs4730222. We identified a significant AEI, demonstrating that the minor allele of rs4730222 is correlated with increased expression of isoforms of *HBP1* with the alternative TSS. However, to demonstrate causality, we introduced the minor allele as a single point mutation into Saos-2, an osteosarcoma cell line, using CRISPR-mediated HDR. In this experiment, we also identified AEI with the minor allele driving increased expression of the alternative TSS. This provided three lines of functional evidence for the effect of rs4730222 on *HBP1* expression (reporter assay, Sw1353 AEI, and Saos-2 AEI). However, we were particularly interested in whether this regulatory effect occurs in patients. To test this, we screened 10 heterozygote patients for AEI of rs4730222, and identify a general increased expression from the minor allele.

Previously, we have shown that expression of *HBP1* is decreased in osteoarthritis hip cartilage compared to control hip cartilage through qPCR (Raine et al., 2012). In this study, we show that an osteoarthritis-associated risk allele increases transcription of an isoform of *HBP1* containing an alternative TSS. This alternative isoform may disrupt *HBP1* expression in a number of ways. First, alternative transcriptional start sites have been shown to modulate transcript stability and translational efficiency (Floor et al., 2016; Wang et al., 2016) as well as

tissue specificity of genes (Kimura et al., 2006; Wang et al., 2008; Yamashita et al., 2011). In this manner, the isoform may play a disproportionate role in certain tissues. Second, the isoform expressed in the chondrosarcoma cell line is likely a truncated version of *HBP1*. Therefore, it may disrupt endogenous activity of the gene, potentially acting as a dominant-negative. Third, this isoform of *HBP1* may have its own, yet uncharacterized, function in the cell.

In this study, we prioritized six SNPs from a set of 1605 variants associated with osteoarthritis. These six SNPs all showed differential regulatory activity between alleles in our reporter assay. In particular, we focus on a SNP located in the 5' UTR of an alternative isoform of *HBP1*, and provide an additional line of support for a potential role of *HBP1* in osteoarthritis pathogenesis.

4.5 METHODS

4.5.1 *Identification and design of target SNPs*

We selected SNPs that had a minor allele frequency >5% and had been reported as being associated with OA in European populations at a significance level that surpassed or approached the genome-wide threshold of <5e-8. The deadline date for inclusion was May 2017. In total, 35 SNPs were identified, each representing an independent association signal (**Table 4.1**). We ran rAGGr on our list of 35 candidate SNPs to identify all variants with a minimum minor allele frequency >= 0.001 in linkage with an $R^2 > 0.8$ in Europeans (CEU+FIN+GBR+IBS+TSI) based on 1000 Genomes, Phase 3, Oct 2014. We then filtered out any polymorphisms greater than one nucleotide, resulting in a list of 1605 SNPs. For each variant, we extracted 196nt of genomic sequence centered on the SNP using BEDTOOLS getfasta, and edited the SNP to create both the minor and major alleles (3210 sequences). To each 196nt sequence, we appended HSS_clon_F (5' - TCTAGAGCATGCACCGG - 3') to the 5' end and DO_R6 (5'-

GCCGGTCAGAATGATGG -3') to the 3' end. We then ordered the 3210 sequences in duplicate as part of an Agilent 244K 230-mer array.

4.5.2 *Library Generation*

We amplified our sequences off of the Agilent array using HSS_clon_F and R6_5N_HSSR (5'- CCGGCCGAATTCGTCGANNNNNCCATCATTCTGACCGGC -3') using KAPA HiFi HotStart ReadyMix in a 50ul reaction with 0.75ng DNA with SYBR Green on a MiniOpticon Real-Time PCR system (Bio-Rad) and stopped the reaction before plateauing (13 cycles). This reaction amplified our library, added a 5bp degenerate barcode to each sequence, and added both adapters for cloning into the human STARR-seq vector. We purified the PCR product using a 1.5x AMPure cleanup following manufacturer's protocol. We then ligated 6ng of our purified PCR into 25ng of linearized human STARR-seq backbone using the NEBuilder HiFi DNA Assembly Cloning Kit following manufacturer's protocol. We transformed 1.2uL of the ligation product in 50ul of NEB C3020 cells, grew up overnight in 100mL of LB+Amp, and extracted the library using the Zymo Research ZymoPURE Plasmid Midiprep Kit.

4.5.3 *STARR-seq Screen*

We transfected 1.5 million Saos-2 cells with 20ug of our library in triplicate using the Thermo Fisher Scientific Neon Transfection System with resuspension buffer R at 1250V, 40ms, 1 shock with 100uL pipettes, in triplicate. After electroporation, we added the cells to 10cm plates with pre-warmed media (McCoy's 5A with 10% FBS and 1x Pen/Strep). 48 hours post transfection, we extracted both DNA and RNA from each replicate using the Qiagen ALL Prep DNA/RNA Mini Kit. DNA was eluted in 80ul and RNA was eluted in 30ul. RNA was treated with Thermo Fisher Scientific TURBO DNase following manufacturer's protocol and reverse

transcribed using Thermo Fisher Scientific SuperScript III Reverse Transcriptase in a 20ul reaction with 8ul of RNA. For each replicate, we amplified DNA in two reactions, each with 2ug of DNA using NEBNext High Fidelity 2X PCR Master Mix with primers HSS_NF_pu1 (5'-CTAAATGGCTGTGAGAGAGCTCAGGTACAACTGATCTAGAGCATGCACC -3') and HSS_R_pu1.(5'- ACTTTATCAATCTCGCTCCAAACCCTTATCATGTCTGCTCGAAGC -3') with SYBR Green on a MiniOpticon Real-Time PCR system (Bio-Rad) and stopped the reaction before plateauing (15 cycles). After PCR, products were purified with a 1.5x AMPure cleanup, and pooled together. For each replicate, we also amplified cDNA in two reactions, each with 10ul of RT product in 50ul reactions with NEBNext High Fidelity 2X PCR Master Mix with primers HSS_F_pu1 (5'-CTAAATGGCTGTGAGAGAGCTCAGGGGCCAGCTGTTGGGGTGTCCAC-3') and HSS_R_pu1.(5'- ACTTTATCAATCTCGCTCCAAACCCTTATCATGTCTGCTCGAAGC -3') and stopped before plateauing (18-20 cycles). After PCR, products were purified with a 1.5x AMPure cleanup, eluted in 50ul each, and pooled together. For the cDNA samples, we performed a nested reaction using KAPA HiFi HotStart ReadyMix in a 50ul reaction with 1ul of the pooled outer PCR reaction with HSS-NF-pu1 and pu1R (5'-ACTTTATCAATCTCGCTCCAAACC -'3) and stopped before plateauing (7 cycles). Reactions were purified with a 1.5x AMPure cleanup and eluted in 50ul each. Flow cell adapters and indexes were added to all DNA and cDNA reactions through an additional round of PCR using Kapa HiFi HotStart ReadyMix in 50ul reactions with 1ul of the first DNA PCR or 1ul of the inner cDNA PCR using an indexed pu1_P5 primer (5'-AATGATACGGCGACCACCGAGATCTACACNNNNNNNNNNACGTAGGCCTAAATGGC TGTGAGAGAGCTCAG -3') and an indexed pu1_P7 primer (5'-

CAAGCAGAAGACGGCATACGAGATNNNNNNNNNGACCGTCGGCACTTTATCAATCT CGCTCCAAACC -3') and stopped before plateauing (6 cycles). The libraries were sequencing on an Illumina NextSeq 500/550 v2 300 cycle mid-output kit.

4.5.4 *Analysis of STARR-seq Screen*

We aligned all sequencing reads to a reference fasta file of our variants using BWA mem (Li, 2013) and extracted reads from error-free molecules. Each variant contained several different 5bp barcodes added through PCR. We counted the number of reads from each replicate for each variant-barcode combination in the DNA and cDNA pool. If there were at least 10 DNA reads, we calculated an enrichment score as the number of RNA reads from the variant-barcode combination normalized to the total number of RNA reads divided by the number of DNA reads from the variant-barcode combination normalized to the total number of DNA reads. We then combined all variant-barcode scores from each replicate, and for any variant with at least five different measurements, we averaged the score for a final enrichment score for each variant. This resulted in scores for 2138 of the 3210 alleles. 753 of the 1605 variants contained measurements for both alleles. For each of the 753 variants with measurements for both alleles, we tested whether the 2 alleles drove different expression by performing a Mann-Whitney T test for each variant using scipy. We then set performed a Benjamini-Hochberg correction with an FDR = 0.05 to correct for multiple testing.

4.5.5 *Allelic imbalance of rs4730222 in Sw1353 cells*

We first genotyped several osteogenic and chondrogenic cell lines for rs4730222 (Sw1353, Tc28a/2, Saos-2, chondrogenic progenitor cells) using rs4730222_sangerF (5'-TACGCAGTTCGAATGAATGGGCTC -3') and rs4730222_sangerR (5'-

AGCTACAAAAACCTGGCTGTCCAC -3’). PCR products were purified with a 1.5x AMPure cleanup and Sanger sequenced with rs4730222_sangerF.

We then tested for allelic imbalance of rs4730222 in the isoforms expressing the SNP in Sw1352. We performed three independent DNA and RNA extractions using the Qiagen ALLPrep DNA/RNA Mini Kit. DNA was eluted in 80ul and RNA was eluted in 30ul. RNA was treated with TURBO DNase and reverse transcribed with SuperScript III Reverse Transcriptase. We then amplified the 5’UTR around rs4730222 from each DNA and cDNA sample using _F and _R using KAPA HiFi HotStart ReadyMix in a 50ul reaction with 100ng DNA or 5ul cDNA with HBP1_5UTR_F_pu1 (5’-CTAAATGGCTGTGAGAGAGCTCAGAGTCCGGGCTGCGGTCACATGATG -3') and HBP1_5UTR_R_pu1 (5’-ACTTTATCAATCTCGCTCCAAACCAGCTACAAAAACCTGGCTGTCCAC -3') and stopped DNA reactions at 25 cycles and cDNA reactions at 32 cycles. Products were purified using a 1.5x AMPure cleanup, and flow cell adapters and indexes were added using an indexed pu1_P5 primer and an indexed pu1_P7 primer. Libraries were spiked into a Miseq v2 300 cycle run. Reads were aligned to a fasta reference file using BWA mem and the number of perfect reads coming from both alleles was quantified from both DNA and cDNA from each replicate.

4.5.6 *CRISPR knock-in of rs4730222 in Saos-2 cells*

rs4730222 falls within a potential Cas9 PAM site (5’- ACGCGATGAATGGCGAAAGA G**G**G -3’). We therefore designed a guideRNA that would target rs4730222, so that the minor-allele donor would not be recut. We ordered the following oligos from IDT: rs4730222_guideF (5’- CACCGACGCGATGAATGGCGAAAGA -3’) and rs4730222_guideR (5’-

AAACTCTTTCGCCATTCATCGCGTC -3') and followed the Zhang lab protocol to clone them into the px458 plasmid (SpCas9-2A-EGFP and single guide RNA).

We created our donor vector in two steps. First, we amplified a 1,459bp region around rs4730222 with the following primers, which also append 16bp homologous sequence to puc19 onto each side of the amplicon: HBP1_puc19F (5'-TCGGTACCCGGGGATCAAGTAGGAAAGTTTCGGTTGAGGAG -3') and HBP1_puc19R (5'- TCGACTCTAGAGGATCAACTGAACAGATGACCGACTCTACC -3). We then cloned this into a linearized puc19 plasmid using Clontech's In-Fusion HD Cloning Kit following manufacturer's protocol, transformed into Stellar Competent cells, grew up a single colony and extracted plasmid using the Zymo Research ZymoPURE Plasmid Midiprep Kit. We then re-linearized the puc19-HBP1 wild-type plasmid with puc19_HBP1-linF (5'-GTGGGGGATGGACTTGGCGTG -3') and puc19-HBP1-linR (5'-CTCCTCAACCGAAACTTTCCTACTT -3'). We also amplified a small region around rs4730222, while mutating the SNP, using mut_insF (5'-AAGTAGGAAAGTTTCGGTTGAGGAG -3') and mut_insR (5'-CCAAGTCCATCCCCCACGCTCTTTCGCCATTCATCGCG -3'). We then cloned the mutated insert into puc19-HBP1-wt using the In-Fusion HD Cloning Kit and grew up a single colony with the minor allele at rs4730222 flanked by 600-850bp of homology on both sides.

We transfected 1 million Saos-2 cells with 10ug of our px458-rs4730222 guide and 10ug of our donor library containing the minor allele using the Neon Transfection system as described above. 72 hours post transfection, we performed FACS on a BD FACS Aria III to isolate ~150,000 GFP+ cells (transfected with px458), which we then expanded. On day 10 post transfection, we extracted DNA and RNA, performed reverse transcription with Superscript III, and amplified the

region surrounding rs4730222 from both DNA (using HBP1_5UTR_F_pu1 and HBP_DNA_Routside (5'- TAGGTGGGCAATCCTGGGAGAAGGTAC -3')), and RNA (using HBP1_5UTR_F_pu1 and HBP_RNA_Routside (5'- TGCCAGATTCTGACTCACTATTTGC -3')) in 50ul reactions using KAPA HiFi 2x ReadyMix. We then purified the PCR reactions with a 1.5x AMPure cleanup, eluted in 50ul, and used 1ul in a nested reaction with pu1L (5'- CTAAATGGCTGTGAGAGAGCTCAG -3') and HBP1_5UTR_R_pu1. Reactions were purified with a 1.5x AMPure cleanup, and flow cell adapters and indexes were added using an indexed pu1_P5 primer and an indexed pu1_P7 primer. Libraries were spiked into a Miseq v2 300 cycle run. Reads were aligned to a fasta reference file using BWA mem and the number of perfect reads coming from both alleles was quantified from both DNA and cDNA from each replicate.

4.5.7 *Allelic imbalance of rs4730222 in osteoarthritis patients' chondrocytes*

Cartilage tissue samples were obtained from OA patients who had undergone joint replacement surgery at the Newcastle upon Tyne NHS Foundation Trust hospitals. The Newcastle and North Tyneside Research Ethics Committee granted ethical approval for the collection, with each donor providing verbal and written informed consent (REC reference number 14/NE/1212). Our patient ascertainment criterion has been described in detail previously (PMID: 17616513; PMID: 19565498). The cartilage was removed from the joint using a scalpel and was collected distal to the OA lesion. The tissue samples were stored frozen at -80°C and ground to a powder using a Retsch Mixermill 200 (Retsch Limited) under liquid nitrogen. Nucleic acids were then extracted from the ground tissue using TRIzol reagent (Life Technologies) and according to the manufacturer's instructions, with the upper aqueous phase separated for RNA isolation, while the interphase and lower organic phase were used to isolate DNA. RNA was reverse transcribed using the SuperScript First-Strand cDNA synthesis kit

(Invitrogen). Matched DNA and cDNA were amplified in 15ul technical triplicate reactions with either 1uL (~10ng gDNA) or 4uL cDNA each, using HBP1-UTR-F-pu1 (5'-CTAAATGGCTGTGAGAGAGCTCAGAGTCCGGGCTGCGGTCACATGATG-3') and HBP1-UTR-R-pu1(5'-ACTTTATCAATCTCGCTCCAAACCAGCTACAAAAACCTGGCTGTCCAC-3') with KAPA KAPA2G Robust HotStart ReadyMix PCR Kit Robust with Sybr Green on a MiniOpticon Real-Time PCR system (Bio-Rad) and stopped before plateauing.

All samples were purified with a 1.5x AMPure cleanup following manufacturer's instructions and eluted in 15uL Qiagen Elution Buffer. 1uL of purified product was then indexed for Illumina sequencing with using an indexed pu1_P5 primer and an indexed pu1_P7 primer. Libraries were spiked into either a Miseq v2 300 cycle or 150 cycle run. Reads were aligned to a fasta reference file using BWA mem and the number of perfect reads coming from both alleles was quantified from both DNA and cDNA from each replicate. We pooled all technical and biological replicates (29 total), and performed a Mann-Whitney U Test for the minor allele fraction from DNA vs RNA.

4.5.8 *Characterization of HBP1 rs4730222-containing isoforms*

We designed the following set of primers to differentiate between different *HBP1* isoforms: Major (1stExonF: 5'- GTGTGGGAAGTGAAGACAAATCAGATGC -3' and LastExonR: 5'- CTTCCACCTGTCACCAAGGATCACAC -3'), 5' UTR (UTR_qPCR_F: 5'-CAGTCTCCGCCTTTCAACCTATG -3' and UTR_qPCR_R: 5'-ATGAACTCGAGTGTAGAGTGCACAG -3'), Truncated (UTR_qPCR_F and Exon6_R: CCACCTCATTTTCACGGTAAGTAG -3') and Full Len (UTR_qPCR_F and LastExonR). We performed technical triplicates for each qPCR using KAPA Robust 2x Hoststart Readymix with

cDNA from wild-type Sw1353 cells, letting the reaction go for 40 cycles. We then ran products on a gel, and differentiated between ESNT0000049735 and ESNT00000458546 based both on size and Sanger sequencing.

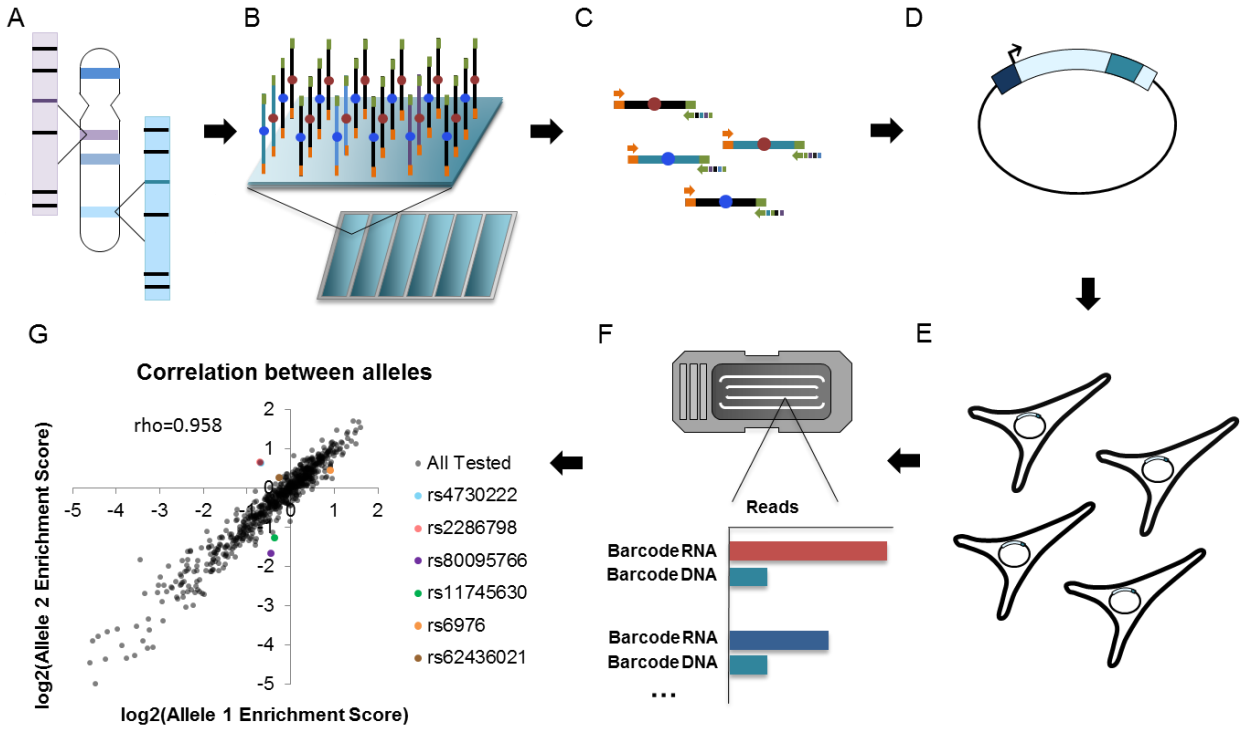


Figure 4.1. Schematic and results from massively parallel reporter assay.

(**A**) For each GWAS lead SNP, we identified all SNPs in LD with r-squared > 0.8 (each color bar on chromosome). Lead SNPs are colored, SNPs in linkage are black. (**B**) For all SNPs, we synthesized 196bp of genomic sequence centered at the minor (red) and major (blue) alleles on an array. (**C**) We amplified our library from the array, while appending 5bp degenerate barcodes and homology to the vector on the 3' end. (**D**) We cloned our barcoded library of all major and minor alleles into the STARR-seq vector. Each putative regulatory region (cyan) is cloned into the 3' UTR of a reporter gene (light blue) with a minimal promoter (dark blue). (**E**) We transfected our library into Saos-2 cells. 48 hours post transfection, we extracted RNA and DNA. (**F**) We determined the abundance of each allele-barcode combination in the RNA and DNA population through sequencing. (**G**). For each allele, we calculated one enrichment score as the average RNA/DNA enrichment across all independent measurements. We identified six SNPs with significantly different activity between the major and minor alleles.

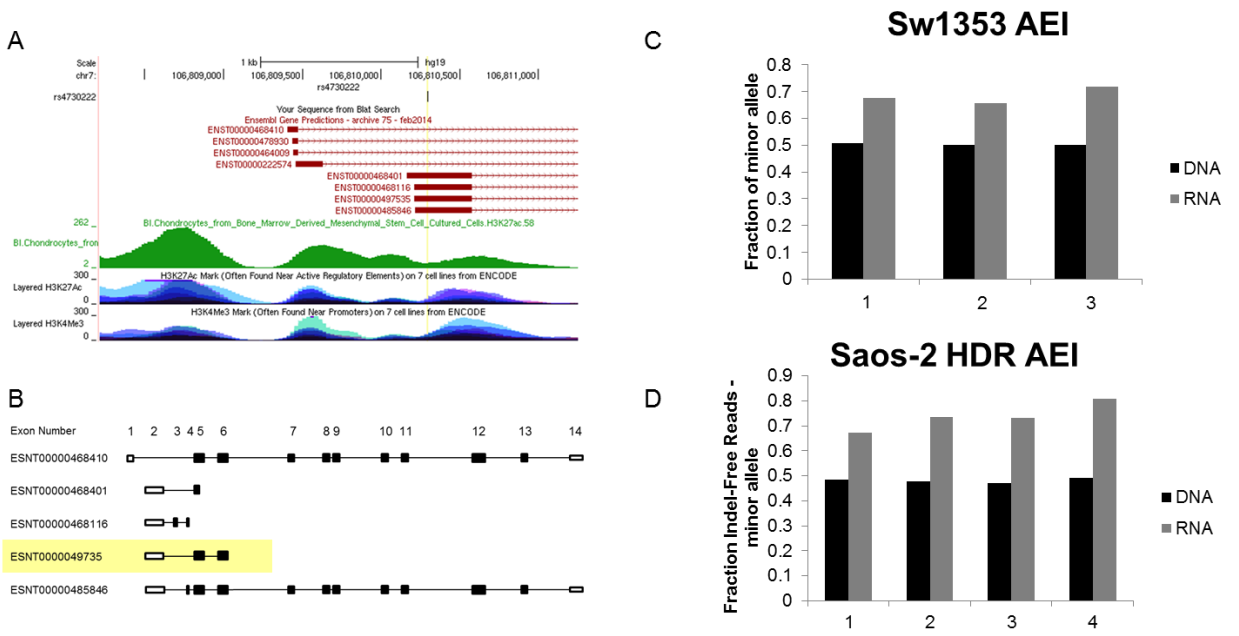


Figure 4.2. Functional validation of rs4730222.

(**A**) UCSC genome browser showing the transcriptional start site for eight different isoforms of HBP1 (screenshot from *http://genome.ucsc.edu*). rs4730222 is indicated by a black line above the gene annotations. The green track is H3K27ac data from chondrocytes derived from cultured bone marrow mesenchymal stem cells (Roadmap). Layered H3K27ac is H3K27ac ChIP-seq (a marker for active enhancers and active promoters) layered from GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK, and NHLF cells. Layered H3K4me3 is H3K4me3 ChIP-seq (marker for active promoters) layered from the same seven ENCODE cell lines. (**B**) Schematic of alternative isoforms of HBP1. Hollow boxes indicate untranslated regions and solid boxes represent coding regions. ENST00000468410 is one of the main isoforms of the gene. The following four isoforms depicted all contain an alternative downstream TSS. ENST00000497353 is actively transcribed in Sw1353 cells and its expression is regulated by rs4730222. It is highlighted in yellow. (**C**) Allelic expression imbalance in Sw1353, a chondrosarcoma cell line heterozygote for rs4730222. Black bars represent the fraction of the minor allele in DNA and grey bars indicate the fraction of the minor allele in cDNA. (**D**) Allelic expression imbalance in Saos-2 cells with the minor allele of rs4730222 introduced through CRISPR-mediated HDR. Bars are the same as in Figure 4.2C.

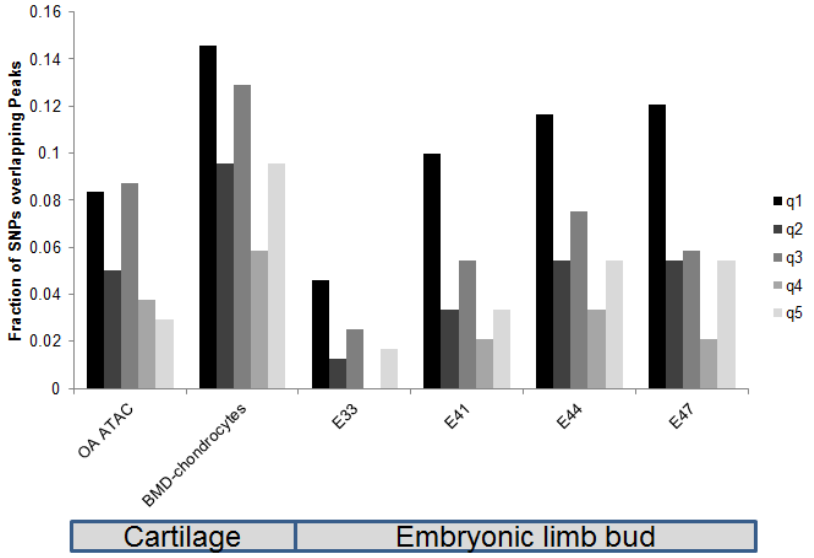


Figure 4.3. Overlap between tested sequences and enhancer marks.

The 1203 SNPs were split into 5 quintiles of 240 sequences each, based on their normalized RNA/DNA enrichment score. Q1 refers to the highest scoring sequences and q5 refers to the sequences with the lowest enrichment scores. We then overlapped each quintile with peaks called from OA ATAC-seq (Liu 2018), BMD-chondrocytes H3K27ac ChIP-seq (Roadmap), and human embryonic limb bud H3K27ac ChIP-seq from e33, e41, e44, and e45. Y-axis is the fraction of the 240 SNPs in each quintile overlapping peaks.

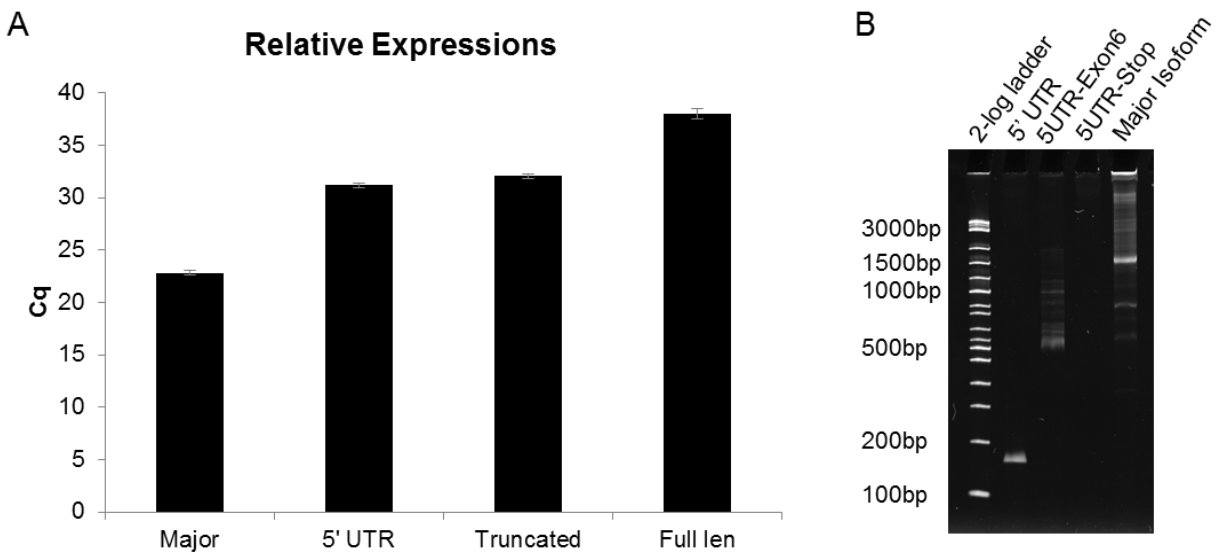


Figure 4.4. Relative Expression of different HBP1 isoforms.

(**A**) Y axis is the Cq value from RT-qPCR from Sw1353 cells. The major allele uses a forward primer at the start codon and reverse primer at the conserved stop codon of the major isoforms. The 5' UTR primer set amplifies a short product, entirely within the alternative 5' UTR. The truncated primer set amplifies both ESNT0000049735 and ESNT00000458546. The full length primer set includes a forward primer in the alternative TSS and reverse primer at the stop codon of the major isoform. We do not identify any full length product utilizing the alternative TSS. (**B**) Gel of qPCR products. First lane is a 2-log ladder. Second lane is the 5' UTR amplification. Expected size is 156bp. Third lane is the truncated ampliciation. The primer set should amplify both ESNT0000049735 (expected size 548bp) and ESNT00000458546 (expected size 846bp). However, ESNT00000458546 contains exon 4 while ESNT0000049735 does not. We ran the PCR product on a gel and Sanger sequenced the purified PCR product, which did not include exon 4. Fourth lane is amplifying from the alternative 5' UTR to canonical stop. There was no amplification product. Fifth lane is the major isoform (amplifying from canonical start to canonical stop. Expected size is 1455bp.

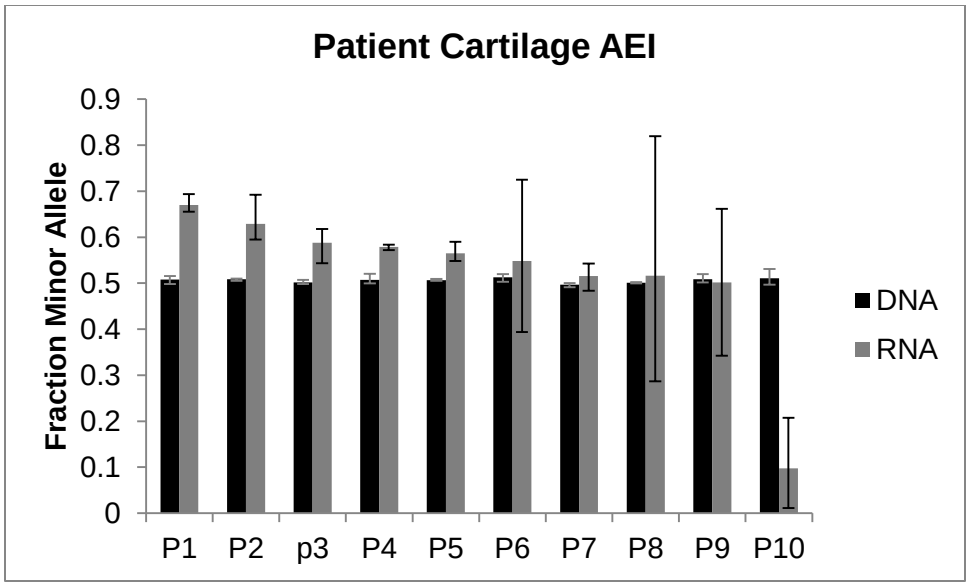


Figure 4.5. rs4730222 AEI in Patient Knee OA cartilage.

For each of ten patients, we quantified the fraction of the minor allele at rs4730222 in gDNA and cDNA for three technical triplicates. Column height indicates the average allele fraction while error bars indicate the minimum and maximum technical replicate allele fraction.

Lead SNP	Nearest protein coding gene	Ref	Annotation	#Linked SNPs tested
rs6976	*GLT8D1*	PMID: 22763110	3’ UTR	280
rs12107036	*TP63*	PMID: 22763110	intronic	1
rs10948172	*SUPT3H*	PMID: 22763110	intronic	213
rs9350591	*FILIP1*	PMID: 22763110	intergenic	73
rs3815148	*HBP1*	PMID: 20112360	intronic	284
rs4836732	*ASTN2*	PMID: 22763110	intronic	2
rs10492367	*KLHDC5*	PMID: 22763110	intergenic	5
rs835487	*CHST11*	PMID: 22763110	intronic	7
rs11842874	*MCF2L*	PMID: 21871595	intronic	8
rs225014	*DIO2*	PMID: 18334578	exonic	19
rs945006	*DIO3*	PMID: 20724312	3’ UTR	1
rs3204689	*ALDH1A2*	PMID: 24728293	3’ UTR	52
rs8044769	*FTO*	PMID: 22763110	intronic	8
rs12982744	*DOT1L*	PMID: 23505243	intronic	21
rs6094710	*NCOA3*	PMID: 23989986	intergenic	11
rs143383	*GDF5*	PMID: 17384641	5’ UTR	98

rs4764133	*MGP*	PMID: 28855172	intergenic	137
rs3850251	*ENPP3*	PMID: 28855172	intronic	6
rs754106	*LRCH1*	PMID: 27974301	intronic	10
rs6766414	*STT3B*	PMID: 27974301	intergenic	29
rs2862851	*TGFA*	PMID: 27701424	intronic	25
rs10471753	*PIK3R1*	PMID: 27701424	intergenic	29
rs2236995	*SLBP*	PMID: 27701424	intronic	4
rs496547	*TREH*	PMID: 27701424	intergenic	3
rs4867568	*LSP1P3*	PMID: 27696742	intergenic	6
rs788748	*IGFBP3*	PMID: 24928840	intergenic	10
rs12901499	*SMAD3*	PMID: 20506137	intronic	28
rs4907986	*COL11A1*	PMID: 24757145	intronic	23
rs1241164	*COL11A1*	PMID: 24757145	intronic	29
rs833058	*VEGF*	PMID: 24757145	intergenic	1
rs10116772	*GLIS3*	PMID: 29436472	intronic	8
rs2820436	*ZC3H11B*	https://doi.org/10.1101/174755	intronic	43
rs11335718	*ANXA3*	https://doi.org/10.1101/174755	intronic	1

rs11780978	*PLEC*	https://doi.org/10.1101/174755	intronic	110
rs2521349	*MAP2K6*	https://doi.org/10.1101/174755	intronic	20

Table 4.1 Lead Osteoarthritis SNPs

List of lead SNPs associated with OA in European populations before May 2017. Table indicates the nearest gene, study in which the SNP was identified, its annotation, and the number of SNPs in linkage in Europeans with an $r^2 > 0.8$.

Chapter 5. CONCLUSIONS AND FUTURE DIRECTIONS

5.1 IMPROVING SENSITIVITY AND SPECIFICITY OF ENHANCER ASSAYS

The canonical definition of an enhancer is a fragment of DNA that can increase transcription of a gene independent of its relative position and orientation to the transcriptional start site. Moreover, enhancers were first identified as fragments of DNA capable of increasing transcription of a reporter gene on a plasmid. Many different assays rely on this definition of an enhancer, by testing sequences in different positions with respect to the TSS. The original MPRA relies on inserting a putative enhancer upstream of a minimal promoter on a plasmid, and linking each putative enhancer to a unique barcode in the 3' UTR of the reporter. A recent derivative, STARR-seq, inserts the putative enhancer itself into the 3' UTR of the reporter, so that it can be directly observed in the RNA transcribed from the reporter gene. However, we still do not have a sense clear sense of the inherent biases from each of these assays. We are currently comparing enhancer activities from nine different reporter assays, using the same putative enhancer library.

This analysis will reveal any differences or biases between reporter assays used in the field, and provide a systematic analysis to suggest which assays should be used for different experimental designs. The analysis will directly compare the effect of having the reporter construct integrated into the genome vs. localized on a plasmid, the putative enhancers being located 5' vs. 3' of the TSS, and having the putative enhancer versus a barcode included within the 3' UTR.

We are also testing for the effect of sequence context on enhancer activity. We have synthesized the same set of >2,000 putative enhancers as 192bp, 354bp, and 678bp sequences

and are testing them in the same MPRA. Preliminary data suggests that the longer sequences are performing better at separating positive and negative controls.

Together, we hope that these projects will improve the field's ability to screen sequences for enhancer activity. By providing a systematic comparison of different assays, researchers can choose the appropriate assay to answer their questions. Through the multiplex and hierarchical multiplex pairwise assembly protocols, researchers will also be able to synthesize long stretches of DNA around their region of interest, changing the paradigm away from screening <200bp sequences.

5.2 A SCALABLE FRAMEWORK TO STUDY REGULATORY SEQUENCE EVOLUTION

In Chapter 3, I characterize evolutionary-functional trajectories for 348 primate liver enhancers. The advantage of my approach over previous studies was the integration of genome-wide ChIP-seq along with higher resolution MPRAs to quantify changes in enhancer activity throughout evolution. This integration of high throughput techniques provides a framework to address enhancer modulation at the scale of the entire genome. This particular project was limited in scope by prioritizing putative hominoid-specific gains from ChIP-seq data. However, given the discordance in trajectories between ChIP-seq from primary liver and the MPRA of synthetic sequences, we propose expanding this study to all human ChIP-seq peaks in a relevant cell line. This will allow for a more powered analysis of enhancer evolution.

For example, using the same ChIP-seq datasets from Chapter 3, there are 6,604 predicted enhancers in primary human liver (H3K27ac peaks not within 1kb of an H4K4me3 peak) that overlap with strong enhancer predictions in HepG2 from ChromHMM. Following the same script described in Chapter 3, after identifying the smallest overlap between the H3K27ac ChIP-seq peaks and ChromHMM calls, the enhancers could be tiled with 122,943 sequences for the

first round of MPRAs. Assuming 6% of tiles are active as before, we would have 7,376 active tiles. If a similar number have orthologs in all eleven primates, we would anticipate that we could characterize evolutionary-functional trajectories for over 2,000 enhancer tiles, each with 20 orthologs (40,000 sequences in MPRA round 2).

Moreover, this design can be applied to other tissues and phylogenies as well. It only requires a relevant, transfectable cell line with enhancer annotations, and preferably ChIP-seq from primary tissue from the same species as the cell line being used to test sequences. Due to large efforts such as ENCODE as well as imputation methods, we have enhancer predictions in hundreds of human cell lines (Ernst and Kellis, 2015; Durham et al., 2018), making this approach applicable to almost any tissue. Moreover, such data and annotations also exist for other phylogenies thanks to efforts in various model organism communities, as highlighted by modENCODE.

A more thorough examination of how naturally occurring genetic variation contributes to regulatory changes is necessary to our understanding of phenotypic diversity and evolution. Such a powerful dataset as proposed here would allow validation of the mechanisms identified in Chapter 3, as well as others, where our analysis was underpowered.

5.3 A SCALABLE FRAMEWORK TO PRIORITIZE GWAS VARIANTS

The identification of genes underlying over 4,000 Mendelian phenotypes in laboratories has led to several new diagnostic and therapeutic approaches in the clinic. However, most diseases act as complex traits involving multiple genes and have been harder to map and understand. After the Human Genome Project and International HapMap Project, researchers have utilized common polymorphisms identified in the population as markers for GWAS. While many of these studies have improved our understanding of complex traits, over 95% of

associated polymorphisms fall in non-coding DNA. Our interpretation of these variants is limited by two factors: 1) multiple polymorphisms can occur in linkage disequilibrium, meaning that the most-associated polymorphisms are not necessarily causal and 2) despite several efforts, we do not have reliable methods to predict the effects of non-coding variation. Accurate interpretation of these variants is a necessary step towards precision medicine and to understanding the underlying physiology behind hundreds of diseases.

In Chapter 4, I discuss a screen to prioritize variants associated with osteoarthritis. A similar screen, focusing on eQTLs has been conducted on lymphoblastoid cell lines (Tewhey et al., 2016). In my screen, we synthesized and screened all SNPs in linkage disequilibrium with each lead osteoarthritis GWAS variant, with an $R^2 > 0.8$ in an osteosarcoma cell line using a modified version of STARR-seq. The framework applied to this study is easily adaptable to prioritize variants from thousands of GWAS studies.

The GWAS catalog contains ~18,000 unique variants in non-coding DNA. Despite efforts to interpret the effects of non-coding variation (Maathuis et al., 2000; Ng and Henikoff, 2003; Cooper et al., 2005; Kircher et al., 2014), we are unable to predict causality for these thousands of GWAS variants. Here, I propose extending my screen from Chapter 4 to functionally annotate as many non-coding GWAS hits as possible in a minimal number of highly-informative experiments. Of all unique variants in the GWAS catalog, 96% fall within non-coding DNA. As performed in Chapter 4, I identified all SNPs in linkage disequilibrium with non-coding GWAS hits, with an $R^2 > 0.8$ in Europeans.

As a quality control metric, I overlapped these SNPs with enhancer annotations in nine difference ENCODE cell lines (Nhek, K562, Nhlf, Hmec, Hepg2, H1hesc, Hsmm, Gm12878 and HUVEC) (Hoffman et al., 2012; Ernst and Kellis, 2013). In accordance with previous studies

(Corradin et al., 2014), I found an enrichment of auto-immune related variants in the set of GM12878 lymphoblastoid enhancers (p=0.022). Similarly, there was an enrichment of liver-function related variants in the set of HepG2 hepatocellular carcinoma cells (p=0.016), breast-related variants in Hmec mammillary epithelial enhancers (p=8.7E-4) and lung function-related variants in Nhlf (p=0.03). This pattern suggests that I am capturing biological meaningful interactions in our nine test cell-lines.

Screening this library in all nine cell lines and identifying differentially active SNPs will help to prioritize causal variants from thousands of GWAS hits, in a cell-type specific manner, for the first time. By using the same library in all cell lines, this experiment is very scalable following the protocol from Chapter 4.

5.4 FINAL REMARKS

It is clear that much of the genetic variation leading to evolution and disease is not encoded in our exome, but rather falls within regulatory DNA, notably, enhancers. The application of next generation sequencing to classic assays such as chromatin immunoprecipitation and episomal reporter assays has allowed for high-throughput interrogation of enhancers in various cell types and tissues. This characterization of enhancers has revealed an enrichment of disease-associated variants within enhancers, as well as large scale turnover of enhancers over an evolutionary timescale. As these next generation sequencing-based assays improve, we expect that they will play an ever growing role in our ability to interpret genetic variation. We are excited to see how increased sensitivity and specificity of enhancer assays furthers our understanding of their roles in evolution and disease.

BIBLIOGRAPHY

Allawi, H. T.; Santalucia, J. Thermodynamics and NMR of Internal G , T Mismatches in DNA. *ACS Biochem.* **1997**, *2960* (96), 10581–10594.

Aparicio-Prat, E.; Arnan, C.; Sala, I.; Bosch, N.; Guigó, R.; Johnson, R. DECKO: Single-Oligo , Dual-CRISPR Deletion of Genomic Elements Including Long Non-Coding RNAs. *BMC Genomics* **2015**, *16* (846).

Aran, D.; Hellman, A. DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. *Cell* **2013**, *154*, 11–13.

Arnold, C. D.; Gerlach, D.; Stelzer, C.; Boryn, L. M.; Rath, M.; Stark, A.; Boryń, Ł. M.; Rath, M.; Stark, A. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq. *Science (80-.).* **2013**, *339* (6123), 1074–1077.

Arnold, C. D.; Gerlach, D.; Spies, D.; Matts, J. A.; Sytnikova, Y. A.; Pagani, M.; Lau, N. C.; Stark, A. Quantitative Genome-Wide Enhancer Activity Maps for Five Drosophila Species Show Functional Enhancer Conservation and Turnover during Cis-Regulatory Evolution. *Nat. Genet.* **2014**, *46* (7), 685–692.

Arnone, M. I.; Dmochowski, I. J.; Gache, C. Using Reporter Genes to Study Cis-Regulatory Elements. *Methods Cell Biol.* **2004**, *74*, 621–652.

Arnosti, D. N.; Kulkarni, M. M. Transcriptional Enhancers: Intelligent Enhanceosomes or Flexible Billboards? *J. Cell. Biochem.* **2005**, *94* (5), 890–898.

Ashkenazy, H.; Penn, O.; Doron-Faigenboim, A.; Cohen, O.; Cannarozzi, G.; Zomer, O.; Pupko, T. FastML: A Web Server for Probabilistic Reconstruction of Ancestral Sequences. *Nucleic Acids Res.* **2012**, *40* (W1), 580–584.

Atchinson, M. L. Enhancers: Mechanisms of Action and Cell Specificity. *Ann. Rev. Cell Biol.* **1988**, *4*, 127–153.

Banerji, J.; Rusconi, S.; Schaffner, W. Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences. *Cell* **1981**, *27* (2 PART 1), 299–308.

Banerji, J.; Olson, L.; Schaffner, W. A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes. *Cell* **1983**, *33*, 729–740.

Bang, D.; Church, G. M. Gene Synthesis by Circular Assembly Amplification. *Nat. Methods* **2008**, *5* (1), 37–39.

Beaucage, S. L.; Caruthers, M. H. Deoxynucleoside Phosphoramidites-a New Class of Key Intermediates for Deoxynucleotide Synthesis. *Tetrahedron Lett.* **1981**, *22*, 1859–1862.

Bellen, H. J.; O'Kane, C. J.; Wilson, C.; Grossniklaus, U.; Pearson, R. K.; Gehring, W. J. P-Element-Mediated Enhancer Detection: A Versatile Method to Study Development in Drosophila. *Genes Dev.* **1989**, *3*, 1288–1300.

Berasi, S. P.; Xiu, M.; Yee, A. S.; Paulson, K. E. HBP1 Repression of the p47phox Gene : Cell Cycle Regulation via the NADPH Oxidase. *Mol. Cell. Biol.* **2004**, *24* (7), 3011–3024.

Bersaglieri, T.; Sabeti, P. C.; Patterson, N.; Vanderploeg, T.; Schaffner, S. F.; Drake, J. A.; Rhodes, M.; Reich, D. E.; Hirschhorn, J. N. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *Am. J. Hum. Genet.* **2004**, *74*, 1111–1120.

Bier, E.; Vaessin, H.; Shepherd, S.; Lee, K.; McCall, K.; Barbel, S.; Ackerman, L.; Carretto, R.; Uemura, T.; Grell, E.; et al. Searching for Pattern and Mutation in the Drosophila Genome with a P-lacZ Vector. *Genes Dev.* **1989**, *3*, 1273–1287.

Binkowski, B. F.; Richmond, K. E.; Kaysen, J.; Sussman, M. R.; Belshaw, P. J. Correcting Errors in Synthetic DNA through Consensus Shuffling. *Nucleic Acids Res.* **2005**, *33* (6).

Blanchard, A. P.; Kaiser, R. J.; Hood, L. E. High-Density Oligonucleotide Arrays. *Biosens. Bioelectron.* **1996**, *11* (6), 687–690.

Borovkov, A. Y.; Loskutov, A. V; Robida, M. D.; Day, K. M.; Cano, J. A.; Olson, T. Le; Patel, H.; Brown, K.; Hunter, P. D.; Sykes, K. F. High-Quality Gene Assembly Directly from Unpurified Mixtures of Microarray-Synthesized Oligonucleotides. *Nucleic Acids Res.* **2010**, *38* (19).

Boyle, A. P.; Davis, S.; Shulha, H. P.; Meltzer, P.; Margulies, E. H.; Weng, Z.; Furey, T. S.; Crawford, G. E. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **2008**, *132* (2), 311–322.

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**, *46* (2), 111–138.

Buenrostro, J. D.; Giresi, P. G.; Zaba, L. C.; Chang, H. Y.; Greenleaf, W. J. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nat. Methods* **2013**, *10* (12), 1213–1218.

Canver, M. C.; Smith, E. C.; Sher, F.; Pinello, L.; Sanjana, N. E.; Shalem, O.; Chen, D. D.; Schupp, P. G.; Vinjamur, D. S.; Garcia, S. P.; et al. BCL11A Enhancer Dissection by Cas9-Mediated in Situ Saturating Mutagenesis. *Nature* **2015**, *527* (7577), 192–197.

Cao, J.; Packer, J. S.; Ramani, V.; Cusanovich, D. A.; Huynh, C.; Daza, R.; Qiu, X.; Lee, C.; Furlan, S. N.; Steemers, F. J.; et al. Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism. *Science (80-.).* **2017**, *357* (6352), 661–667.

Carr, P. A.; Park, J. S.; Lee, Y.; Yu, T.; Zhang, S.; Jacobson, J. M. Protein-Mediated Error Correction for de Novo DNA Synthesis. *Nucleic Acids Res.* **2004**, *32* (20), 1–9.

Chan, Y. F.; Marks, M. E.; Jones, F. C.; Jr, G. V.; Shapiro, M. D.; Brady, S. D.; Southwick, A. M.; Absher, D. M.; Grimwood, J.; Schmutz, J.; et al. Adaptive Evolution of Pelvic Reduction of a Pitx1 Enhancer. *Science (80-.).* **2010**, *327* (5963), 302–305.

Chong, J. X.; Buckingham, K. J.; Jhangiani, S. N.; Boehm, C.; Sobreira, N.; Smith, J. D.; Harrell, T. M.; McMillin, M. J.; Wiszniewski, W.; Gambin, T.; et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **2015**, *97* (2), 199–215.

Cong, L.; Ann, R. F.; Cox, D.; Lin, S.; Barretto, R.; Habib, N.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A.; et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **2013a**, *339* (6121), 819–823.

Cong, L.; Ran, F. A.; Cox, D.; Lin, S.; Barretto, R.; Hsu, P. D.; Wu, X.; Jiang, W.; Marraffini, L. A. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science (80-.).* **2013b**, *339* (6121), 819–823.

Cooper, D. N.; Mort, M.; Stenson, P. D.; Ball, E. V; Chuzhanova, N. A. Methylation-Mediated Deamination of 5-Methylcytosine Appears to Give Rise to Mutations Causing Human Inherited Disease in CpNpG Trinucleotides , as Well as in CpG Dinucleotides. *Hum. Genomics* **2010**, *4* (6), 406–410.

Cooper, G. M.; Stone, E. A.; Asimenos, G.; Green, E. D.; Batzoglou, S.; Sidow, A. Distribution and Intensity of Constraint in Mammalian Genomic Sequence. *Genome Res.* **2005**, *15* (7), 901–913.

Corradin, O.; Saiakhova, A.; Akhtar-Zaidi, B.; Myeroff, L.; Willis, J.; Cowper-Sallari, R.; Lupien, M.; Markowitz, S.; Scacheri, P. C. Combinatorial Effects of Multiple Enhancer Variants in Linkage Disequilibrium Dictate Levels of Gene Expression to Confer Susceptibility to Common Traits. *Genome Res.* **2014**, *24* (1), 1–13.

Cotney, J.; Leng, J.; Yin, J.; Reilly, S. K.; Demare, L. E.; Emera, D.; Ayoub, A. E.; Rakic, P.; Noonan, J. P. The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. *Cell* **2013**, *154* (1), 185–196.

Crawford, G. E.; Davis, S.; Scacheri, P. C.; Renaud, G.; Halawi, M. J.; Erdos, M. R.; Green, R.; Meltzer, P. S.; Wolfsberg, T. G.; Collins, F. S. DNase-Chip : A High-Resolution Method to Identify DNase I Hypersensitive Sites Using Tiled Microarrays. *Nat. Methods* **2006a**, *3* (7), 503–509.

Crawford, G. E.; Holt, I. E.; Whittle, J.; Webb, B. D.; Tai, D.; Davis, S.; Margulies, E. H.; Chen, Y.; Bernat, J. A.; Ginsburg, D.; et al. Genome-Wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS). *Genome Res.* **2006b**, *16*, 123–131.

Creyghton, M. P.; Cheng, A. W.; Welstead, G. G.; Kooistra, T.; Carey, B. W.; Steine, E. J.; Hanna, J.; Lodato, M. A.; Frampton, G. M.; Sharp, P. A.; et al. Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State. *PNAS* **2010**, *107* (50), 21931–21936.

Cusanovich, D. A.; Daza, R.; Adey, A.; Pliner, H. A.; Christiansen, L.; Gunderson, K. L.; Steemers, F. J.; Trapnell, C.; Shendure, J. Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing. *Science (80-.).* **2015**, *348* (6237), 910–914.

Degner, J. F.; Pai, A. A.; Pique-Regi, R.; Veyrieras, J.-B.; Gaffney, D. J.; Pickrell, J. K.; De Leon, S.; Michelini, K.; Lewellen, N.; Crawford, G. E.; et al. DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation. *Nature* **2012**, *482* (7385), 390–394.

Denechaud, P.; Lopez-mejia, I. C.; Giralt, A.; Lai, Q.; Blanchet, E.; Delacuisine, B.; Nicolay, B. N.; Dyson, N. J.; Bonner, C.; Pattou, F.; et al. E2F1 Mediates Sustained Lipogenesis and Contributes to Hepatic Steatosis. *J. Clin. Invest.* **2016**, *126* (1), 137–150.

Diao, Y.; Li, B.; Meng, Z.; Jung, I.; Lee, A. Y.; Dixon, J.; Maliskova, L.; Guan, K.; Shen, Y.; Ren, B. A New Class of Temporarily Phenotypic Enhancers Identified by CRISPR/Cas9-Mediated Genetic Screening. *Genome Res.* **2016**, *26* (3), 397–405.

Diao, Y.; Fang, R.; Li, B.; Meng, Z.; Yu, J.; Qiu, Y.; Lin, K. C.; Huang, H.; Liu, T.; Marina, R. J.; et al. A Tiling-Deletion-Based Genetic Screen for Cis-Regulatory Element Identification in Mammalian Cells. *Nat. Methods* **2017**, *14* (6), 629–635.

Dormitzer, P. R.; Suphaphiphat, P.; Gibson, D. G.; Wentworth, D. E.; Stockwell, T. B.; Algire, M. A.; Alperovich, N.; Barro, M.; Brown, D. M.; Craig, S.; et al. Synthetic Generation of Influenza Vaccine Viruses for Rapid Response to Pandemics. *Sci. Transl. Med.* **2013**, *5* (185), 1–14.

Dowell, R. D. Transcription Factor Binding Variation in the Evolution of Gene Regulation. *Trends Genet.* **2010**, *26* (11), 468–475.

Durham, T. J.; Libbrecht, M. W.; Howbert, J. J.; Bilmes, J.; Noble, W. S. PREDICTD PaRallel Epigenomics Data Imputation with Cloud-Based Tensor Decomposition. *Nat. Commun.* **2018**, *9*, 1402.

Eckner, R.; Ewen, M. E.; Newsome, D.; Gerdes, M.; DeCaprio, J. A.; Lawrence, J. B.; Livingston, D. M. Molecular Cloning and Functional Analysis of the Adenovirus E1A-Associated 300-kD Protein (p300) Reveals a Protein with Properties of a Transcriptional Adaptor. *Genes Dev.* **1994**, *8* (8), 869–884.

Edwards, S. L.; Beesley, J.; French, J. D.; Dunning, M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am. J. Hum. Genet.* **2013**, *93* (5), 779–797.

Ernst, J.; Kellis, M. ChromHMM: Automating Chromatin State Discovery and Characterization. *Nat. Methods* **2013**, *9* (3), 215–216.

Ernst, J.; Kellis, M. Large-Scale Imputation of Epigenomic Datasets for Systematic Annotation of Diverse Human Tissues. *Nat Biotechnol* **2015**, *33* (4), 364–376.

Ernst, J.; Kheradpour, P.; Mikkelsen, T. S.; Shoresh, N.; Ward, L. D.; Epstein, C. B.; Zhang, X.; Wang, L.; Issner, R.; Coyne, M.; et al. Mapping and Analysis of Chromatin State Dynamics in

Nine Human Cell Types. *Nature* **2011**, *473* (7345), 43–49.

Ernst, J.; Melnikov, A.; Zhang, X.; Wang, L.; Rogov, P.; Mikkelsen, T. S.; Kellis, M. Genome-Scale High-Resolution Mapping of Activating and Repressive Nucleotides in Regulatory Regions. *Nat. Biotechnol.* **2016**, *34* (11), 1180–1190.

Farley, E. K.; Olson, K. M.; Zhang, W.; Brandt, A. J.; Rokhsar, D. S.; Levine, M. S. Suboptimization of Developmental Enhancers. *Science (80-.).* **2015**, *350* (6258), 325–328.

Findlay, G. M.; Boyle, E. a.; Hause, R. J.; Klein, J. C.; Shendure, J. Saturation Editing of Genomic Regions by Multiplex Homology-Directed Repair. *Nature* **2014**, *513* (7516), 120–123.

Floor, S. N.; Doudna, J. A.; States, U.; Initiative, I. G. Tunable Protein Synthesis by Transcript Isoforms in Human Cells. *Elife* **2016**, *4*, 1–25.

Frankel, N.; Davis, G. K.; Vargas, D.; Wang, S.; Stern, D. L. Phenotypic Robustness Conferred by Apparently Redundant Transcriptional Enhancers. *Nature* **2010**, *466* (July), 490–493.

Fuhrmann, M.; Oertel, W.; Berthold, P.; Hegemann, P. Removal of Mismatched Bases from Synthetic Genes by Enzymatic Mismatch Cleavage. *Nucleic Acids Res.* **2005**, *33* (6), 6–13.

Fulco, C. P.; Munshauer, M.; Anyoha, R.; Munsom, G.; Grossman, S. R.; Perez, E. M.; Kane, M.; Cleary, B.; Lander, E. S.; Engreitz, J. M. Systematic Mapping of Functional Enhancer-Promoter Connections with CRISPR Interference. *Science (80-.).* **2016**, *354* (6313), 769–773.

Gao, X.; Tsang, J. C. H.; Gaba, F.; Wu, D.; Lu, L.; Liu, P. Comparison of TALE Designer Transcription Factors and the CRISPR/dCas9 in Regulation of Gene Expression by Targeting Enhancers. *Nucleic Acids Res.* **2014**, *42* (20).

Gasperini, M.; Findlay, G. M.; McKenna, A.; Milbank, J. H.; Lee, C.; Zhang, M. D.; Cusanovich, D. A.; Shendure, J. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.* **2017**, *101* (2), 192–205.

Ghandi, M.; Lee, D.; Mohammad-noori, M.; Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features. *PLoS Comput. Biol.* **2014**, *10* (7).

Ghindilis, A. L.; Smith, M. W.; Schwarzkopf, K. R.; Roth, K. M.; Peyvan, K.; Munro, S. B.; Lodes, M. J.; St, A. G.; Bernards, K.; Dill, K.; et al. CombiMatrix Oligonucleotide Arrays : Genotyping and Gene Expression Assays Employing Electrochemical Detection ☆. *Biosen* **2007**, *22*, 1853–1860.

Gilbert, L. A.; Larson, M. H.; Morsut, L.; Liu, Z.; Brar, G. A.; Torres, S. E.; Stern-ginossar, N.; Brandman, O.; Whitehead, E. H.; Doudna, J. A.; et al. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **2013**, *154* (2), 442–451.

Gillies, S. D.; Morrison, S. L.; Oi, V. T.; Tonegawa, S. A Tissue-Specific Transcription Enhancer Element Is Located in the Major Lntron of a Rearranged Lmmunoglobulin Heavy

Chain Gene. *Cell* **1983**, *33*, 717–728.

Gompel, N.; Prud, B.; Wittkopp, P. J.; Kassner, V. A.; Carroll, S. B. Chance Caught on the Wing : Cis -Regulatory Evolution and the Origin of Pigment Patterns in Drosophila. **2005**, 481–487.

Grant, C. E.; Bailey, T. L.; Noble, W. S. FIMO : Scanning for Occurrences of a given Motif. *Bioinformatics* **2011**, *27* (7), 1017–1018.

Groner, A. C.; Meylan, S.; Ciuffi, A.; Zangger, N.; Ambrosini, G. KRAB-Zinc Finger Proteins and KAP1 Can Mediate Long-Range Transcriptional Repression through Heterochromatin Spreading. *Plos Genet.* **2010**, *6* (3).

Gross, D. S.; Garrard, W. T. Nuclease Hypersensitive Sites in Chromatin. *Annu Rev Biochem* **1988**, *57*, 159–197.

Grossniklaus, U.; Bellen, H. J.; Wilson, C.; Gehring, W. J. P-Element-Mediated Enhancer Detection Applied to the Study of Oogenesis in Drosophila. *Development* **1989**, *107*, 189–200.

Hamada, H. Activation of an Enhancerless Gene by Chromosomal Integration. *Mol. Cell. Biol.* **1986a**, *6* (12), 4179–4184.

Hamada, H. Random Isolation of Gene Activator Elements from the Human Genome. *Mol. Cell. Biol.* **1986b**, *6* (12), 4185–4194.

Heintzman, N. D.; Stuart, R. K.; Hon, G.; Fu, Y.; Ching, C. W.; Hawkins, R. D.; Barrera, L. O.; Calcar, S. Van; Qu, C.; Ching, K. A.; et al. Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome. *Nat. Genet.* **2007**, *39* (3), 311–318.

Heintzman, N. D.; Hon, G. C.; Hawkins, R. D.; Kheradpour, P.; Stark, A.; Harp, L. F.; Ye, Z.; Lee, L. K.; Stuart, R. K.; Ching, C. W.; et al. Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression. *Nature* **2009**, *459* (7243), 108–112.

Hilton, I. B.; Ippolito, A. M. D.; Vockley, C. M.; Thakore, P. I.; Crawford, G. E.; Reddy, T. E.; Gersbach, C. A. Epigenome Editing by a CRISPR-Cas9-Based Acetyltransferase Activates Genes from Promoters and Enhancers. *Nat. Biotechnol.* **2015**, *33* (5), 510–517.

Hoffman, M. M.; Buske, O. J.; Wang, J.; Weng, Z.; Bilmes, J. A.; Noble, W. S. Unsupervised Pattern Discovery in Human Chromatin Structure through Genomic Segmentation. *Nat. Methods* **2012**, *9* (5), 473–476.

Hughes, T. R.; Mao, M.; Jones, A. R.; Burchard, J.; Marton, M. J.; Shannon, K. W.; Lefkowitz, S. M.; Ziman, M.; Schelter, J. M.; Meyer, M. R.; et al. Expression Profiling Using Microarrays Fabricated by an Ink-Jet Oligonucleotide Synthesizer. *Nat. Biotechnol.* **2001**, *19* (April).

Inoue, F.; Kircher, M.; Martin, B.; Cooper, G. M.; Witten, D. M.; Mcmanus, M. T.; Ahituv, N.; Shendure, J. A Systematic Comparison Reveals Substantial Differences in Chromosomal versus

Episomal Encoding of Enhancer Activity. *Genome Res.* **2016**, *27* (1), 38–52.

Jiang, M.; Anderson, J.; Gillespie, J.; Mayne, M. uShuffle : A Useful Tool for Shuffling Biological Sequences While Preserving the K-Let Counts. *BMC Bioinformatics* **2008**, *9* (192).

Jin, Q.; Yu, L. R.; Wang, L.; Zhang, Z.; Kasper, L. H.; Lee, J. E.; Wang, C.; Brindle, P. K.; Dent, S. Y.; Ge, K. Distinct Roles of GCN5/PCAF-Mediated H3K9ac and CBP/p300-Mediated H3K18/27ac in Nuclear Receptor Transactivation. *EMBO J* **2011**, *30*, 249–262.

Kearns, N. A.; Pham, H.; Tabak, B.; Genga, R. M.; Silverstein, N. J.; Garber, M.; Maehr, R. Functional Annotation of Native Enhancers with a Cas9-Histone Demethylase Fusion. *Nat. Methods* **2015**, *12* (5), 401–403.

Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M.; Haussler, D. The Human Genome Browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006.

Kheradpour, P.; Ernst, J.; Melnikov, A.; Rogov, P.; Wang, L.; Alston, J.; Mikkelsen, T. S.; Kellis, M. Systematic Dissection of Regulatory Motifs in 2000 Predicted Human Enhancers Using a Massively Parallel Reporter Assay. **2013**, 562.

Kim, H.; Han, H.; Ahn, J.; Lee, J.; Cho, N.; Jang, H.; Kim, H.; Kwon, S.; Bang, D. " Shotgun DNA Synthesis " for the High-Throughput Construction of Large DNA Molecules. *Nucleic Acids Res.* **2018**, *40* (18).

Kimura, K.; Wakamatsu, A.; Suzuki, Y.; Ota, T.; Nishikawa, T.; Yamashita, R.; Yamamoto, J.; Sekine, M.; Tsuritani, K.; Wakaguri, H.; et al. Diversification of Transcriptional Modulation : Large-Scale Identification and Characterization of Putative Alternative Promoters of Human Genes. *Genome Res.* **2006**, *16*, 55–65.

King, M.; Wilson, A. C. Evolution at Two Levels in Humans and Chimpanzees. *Science (80-.).* **1975**, *188* (4184), 107–116.

Kircher, M.; Witten, D. M.; Jain, P.; O'Roak, B. J.; Cooper, G. M.; Shendure, J. A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.* **2014**, *46* (3), 310–315.

Klann, T. S.; Black, J. B.; Chellappan, M.; Safi, A.; Song, L.; Hilton, I. B.; Crawford, G. E.; Reddy, T. E.; Gersbach, C. A. CRISPR-Cas9 Epigenome Editing Enables High-Throughput Screening for Functional Regulatory Elements in the Human Genome. *Nat. Biotechnol.* **2017**, *35* (6), 561–568.

Kong, D. S.; Carr, P. A.; Chen, L.; Zhang, S.; Jacobson, J. M. Parallel Gene Synthesis in a Microfluidic Device. *Nucleic Acids Res.* **2007**, *35* (8), 1–9.

Korkmaz, G.; Lopes, R.; Ugalde, A. P.; Nevedomskaya, E.; Han, R.; Myacheva, K.; Zwart, W.; Elkon, R.; Agami, R. Functional Genetic Screens for Enhancer Elements in the Human Genome Using CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34* (2), 192–198.

Kosuri, S.; Church, G. M. Large-Scale de Novo DNA Synthesis : Technologies and Applications. *Nat. Methods* **2014**, *11* (5), 499–507.

Kosuri, S.; Eroshenko, N.; Leproust, E. M.; Super, M.; Way, J.; Li, J. B.; Church, G. M. Scalable Gene Synthesis by Selective Amplification of DNA Pools from High-Fidelity Microchips. *Nat. Biotechnol.* **2010**, *28* (12).

Kristiansson, E.; Thorsen, M.; Tama, M. J.; Nerman, O. Evolutionary Forces Act on Promoter Length : Identification of Enriched Cis -Regulatory Elements. *Mol. Biol. Evol.* **2009**, *26* (6), 1299–1307.

Kulakovskiy, I. V; Medvedeva, Y. A.; Schaefer, U.; Kasianov, A. S.; Vorontsov, I. E.; Bajic, V. B.; Makeev, V. J. HOCOMOCO : A Comprehensive Collection of Human Transcription Factor Binding Sites Models. *Nucleic Acids Res.* **2012**, *41*, 195–202.

Kunarso, G.; Chia, N.; Jeyakani, J.; Hwang, C.; Lu, X.; Chan, Y.; Ng, H.; Bourque, G. Transposable Elements Have Rewired the Core Regulatory Network of Human Embryonic Stem Cells. *Nat. Genet.* **2010**, *42* (7), 631–634.

Kwasnieski, J. C.; Fiore, C.; Chaudhari, H. G.; Cohen, B. A. High-Throughput Functional Testing of ENCODE Segmentation Predictions. *Genome Res.* **2014**.

Lee, W.; Mitchell, P.; R, T. Purified Transcription Factor AP-1 Interacts with TPA-Inducible Enhancer Elements. *Cell* **1987**, *49*, 741–752.

Levine, M. Transcription Regulation and Animal Diversity. *Nature* **2003**, *424*, 147–150.

Levine, M. Transcriptional Enhancers in Animal Development and Evolution. *Curr. Biol.* **2010**, *20* (17), R754–R763.

Li, H. Aligning Sequence Reads , Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv* **2013**.

Linshiz, G.; Yehezkel, T. Ben; Kaplan, S.; Gronau, I.; Ravid, S.; Adar, R.; Shapiro, E. Recursive Construction of Perfect DNA Molecules from Imperfect Oligonucleotides. *Mol. Syst. Biol.* **2008**, *4*.

Liu, Y.; Chang, J.-C.; Hon, C.-C.; Fukui, N.; Tanaka, N.; Zhang, Z.; Lee, M. T. M.; Minoda, A. Chromatin Accessibility Landscape of Articular Knee Cartilage Reveals Aberrant Enhancer Regulation in Osteoarthritis. *Biorxiv* **2018**.

Long, H. K.; Prescott, S. L.; Wysocka, J. Review Ever-Changing Landscapes : Transcriptional Enhancers in Development and Evolution. *Cell* **2016**, *167* (5), 1170–1187.

Long, M. D.; Smiraglia, D. J.; Campbell, M. J. The Genomic Impact of DNA CpG Methylation on Gene Expression ; Relationships in Prostate Cancer. *Biomolecules* **2017**, *7* (15).

Luyten, F. P.; Tylzanowski, P.; Lories, R. J. Wnt Signaling and Osteoarthritis. *Bone* **2009**, *44* (4),

522–527.

Maathuis, M.; Colombo, D.; Kalisch, M.; Bühlmann, P. A Method and Server for Predicting Damaging Missense Mutations. *Ann. Stat. Cell Stat. Soc. Ser. B J. Roy. Stat. Soc. Ser. B Biol* **2000**, *37* (16), 3133–3164.

Macosko, E. Z.; Basu, A.; Satija, R.; Nemesh, J.; Shekhar, K.; Goldman, M.; Tirosh, I.; Bialas, A. R.; Kamitaki, N.; Martersteck, E. M.; et al. Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **2015**, *161* (5), 1202–1214.

Markham, N. R.; Zuker, M. UNAFold: Software for Nucleic Acid Folding and Hybridization. *Methods Mol. Biol.* **2008**, *453*.

Matzas, M.; Stähler, P. F.; Kefer, N.; Siebelt, N.; Boisguérin, V.; Leonard, J. T.; Keller, A.; Stähler, C. F.; Häberle, P.; Gharizadeh, B.; et al. Letters High-Fidelity Gene Synthesis by Retrieval of Sequence-Verified DNA Identified Using High-Throughput Pyrosequencing. *Nat. Biotechnol.* **2010**, *28* (12), 1291–1295.

Maurano, M. T.; Humbert, R.; Rynes, E.; Thurman, R. E.; Haugen, E.; Wang, H.; Reynolds, A. P.; Sandstrom, R.; Qu, H.; Brody, J.; et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **2012**, *337* (6099), 1190–1195.

McCarthy, Mark I; Abecasis, G. R.; Cardon, L. R.; Goldstein, D. B.; Little, J.; Ioannidis, J. P. A.; Hirschhorn, J. N. Genome-Wide Association Studies for Complex Traits : Consensus , Uncertainty and Challenges. *Nature* **2008**, *9* (May), 356–369.

McKenna, A.; Findlay, G. M.; Gagnon, J. A.; Horwitz, M. S.; Schier, A. F. Whole Organism Lineage Tracing by Combinatorial and Cumulative Genome Editing. *Science (80-.).* **2016**, *353* (6298).

Melnikov, A.; Murugan, A.; Zhang, X.; Tesileanu, T.; Wang, L.; Rogov, P.; Feizi, S.; Gnirke, A.; Callan, C. G.; Kinney, J. B.; et al. Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay. *Nat. Biotechnol.* **2012**, *30* (3), 271–277.

Merika, M.; Williams, A. J.; Chen, G.; Collins, T.; Thanos, D. Recruitment of CBP/p300 by the IFNb Enhanceosome Is Required for Synergistic Activation of Transcription. *Mol. Cell* **1998**, *1*, 277–287.

Mikkelsen, T. S.; Xu, Z.; Zhang, X.; Wang, L.; Gimble, J. M.; Lander, E. S.; Rosen, E. D. Comparative Epigenomic Analysis of Murine and Human Adipogenesis. *Cell* **2010**, *143*, 156–169.

Mitchell, P. J.; Wang, C.; Tjian, R. Positive and Negative Regulation of Transcription in Vitro: Enhancer-Binding Protein AP-2 Is Inhibited by SV40 T Antigen. *Cell* **1987**, *50*, 847–861.

Moreau, P.; Hen, R.; Wasylyk, B.; Everett, R.; Gaub, M. P.; Chambon, P. The SV40 72 Base Repair Repeat Has a Striking Effect on Gene Expression Both in SV40 and Other Chimeric

Recombinants. *Nucleic Acids Res.* **1981**, *9* (22), 6047–6068.

Mouse, E. C. An Encyclopedia of Mouse DNA Elements (Mouse ENCODE). *Genome Biol.* **2012**, *13* (8), 418.

Neuberger, M. S. Expression and Regulation of Immunoglobulin Heavy Chain Transfected into Lymphoid Cells. *EMBO* **1983**, *2* (8), 1373–1378.

Ng, P. C.; Henikoff, S. SIFT: Predicting Amino Acid Changes That Affect Protein Function. *Nucleic Acids Res.* **2003**, *31* (13), 3812–3814.

Nguyen, T. A.; Jones, R. D.; Snavely, A. R.; Pfenning, A. R.; Kirchner, R.; Hemberg, M.; Gray, J. M. High-Throughput Functional Comparison of Promoter and Enhancer Activities. *Genome Res.* **2016**, *26* (8), 1023–1033.

Nguyen-Dumont, T.; Pope, B. J.; Hammet, F.; Southey, M. C.; Park, D. J. A High-Plex PCR Approach for Massively Parallel Sequencing. *Biotechniques* **2013**, *55* (2).

Nitta, K. R.; Jolma, A.; Yin, Y.; Morgunova, E.; Kivioja, T.; Akhtar, J.; Hens, K.; Toivonen, J.; Polytechnique, F. Conservation of Transcription Factor Binding Specificities across 600 Million Years of Bilateria Evolution. *Elife* **2015**, 1–20.

Noonan, J. P.; Mccallion, A. S. Genomics of Long-Range Regulatory Elements. *Annu. Rev. Genomics Hum. Genet.* **2010**, *11*, 1–23.

Nowock, J.; Borgmeyer, U.; Pdischel, A. W.; Rupp, R. A. W.; Sippel, A. E.; Biologie, M.; Zmbh, H.; Feld, I. N. The TGGCA Protein Binds to the MMTV-LTR, the Adenovirus Origin of Replication, and the BK Virus Enhancer. *Nucleic Acids Res.* **1985**, *13* (6), 2045–2061.

O'Kane, C. J.; Gehring, W. J. Detection in Situ of Genomic Regulatory Elements in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 9123–9127.

Patwardhan, R. P.; Lee, C.; Litvin, O.; Young, D. L.; Pe, D.; Shendure, J. High-Resolution Analysis of DNA Regulatory Elements by Synthetic Saturation Mutagenesis. *Nat. Biotechnol.* **2009**, *27* (12), 1173–1175.

Patwardhan, R. P.; Hiatt, J. B.; Witten, D. M.; Kim, M. J.; Smith, R. P.; May, D.; Lee, C.; Andrie, J. M.; Lee, S.-I.; Cooper, G. M.; et al. Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo. *Nat. Biotechnol.* **2012**, *30* (3), 265–270.

Piette, J.; Kryszke, M.; Yaniv, M. Specific Interaction of Cellular Factors with the B Enhancer of Polyoma Virus. *EMBO J.* **1985**, *4* (10), 2675–2685.

Project Consortium, E. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (80-.).* **2004**, *306* (October), 636–641.

Project Consortium, E. Identification and Analysis of Functional Elements in 1% of the Human Genome by the ENCODE Pilot Project. *Nature* **2007**, *447* (7146), 799–816.

Project Consortium, E. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol* **2011**, *4*, e1001046.

Project Consortium, E. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **2012**, *489* (7414), 57–74.

Qi, L. S.; Larson, M. H.; Gilbert, L. A.; Doudna, J. A.; Weissman, J. S.; Arkin, A. P.; Lim, W. A. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **2013**, *152* (5), 1173–1183.

Quan, J.; Saaem, I.; Tang, N.; Ma, S.; Negre, N.; Gong, H.; White, K. P.; Tian, J. Parallel on-Chip Gene Synthesis and Application to Optimization of Protein Expression. *Nat. Biotechnol.* **2011**, *29* (5), 449–452.

Queen, C. Lmmunoglobulin Gene Transcription Is Activated by Downstream Sequence Elements. *Cell* **1983**, *33*, 741–748.

Quinlan, A. R.; Hall, I. M. BEDTools : A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26* (6), 841–842.

Raine, E. V. A.; Wreglesworth, N.; Dodd, A. W.; Reynard, L. N.; Loughlin, J. Gene Expression Analysis Reveals HBP1 as a Key Target for the Osteoarthritis Susceptibility Locus That Maps to Chromosome 7q22 Correspondence to. *Ann. Rheum. Dis.* **2012**, *71* (12), 2020–2027.

Rajagopal, N.; Srinivasan, S.; Kooshesh, K.; Guo, Y.; Edwards, M. D.; Banerjee, B.; Syed, T.; Emons, B. J. M.; Gifford, D. K.; Sherwood, R. I. High-Throughput Mapping of Regulatory DNA. *Nat. Biotechnol.* **2016**, *34* (2), 167–174.

Reynolds, N.; Dvinge, H.; Hynes-allen, A.; Balasooriya, G.; Leaford, D.; Behrens, A.; Bertone, P.; Hendrich, B. NuRD-Mediated Deacetylation of H3K27 Facilitates Recruitment of Polycomb Repressive Complex 2 to Direct Gene Repression. *EMBO J.* **2012**, *31* (3), 593–605.

Saaem, I.; Ma, K.; Marchi, A. N.; Labean, T. H.; Tian, J. In Situ Synthesis of DNA Microarray on. *ACS Appl. Mater. Interfaces* **2010**, *2* (2), 491–497.

Sambrook, J.; Fritsch, E. F.; Maniatas, T. Molecular Cloning: A Laboratory Manual. *Cold Spring Harb. NY Cold Spring Harb. Lab. Press* **1989**.

Sanjana, N. E.; Wright, J.; Zheng, K.; Shalem, O.; Fontanillas, P.; Joung, J.; Cheng, C.; Regev, A.; Zhang, F. High-Resolution Interrogation of Functional Elements in the Noncoding Genome. *Science (80-.).* **2016**, *353* (6307), 1545–1549.

Schlabach, M. R.; Hu, J. K.; Li, M.; Elledge, S. J. Synthetic Design of Strong Promoters. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (6), 1–6.

Schultz, D. C.; Ayyanathan, K.; Negorev, D.; Maul, G. G.; Rauscher, F. J. 3rd. SETDB1: A Novel KAP-1-Associated Histone H3, Lysine 9-Specific Methyltransferase That Contributes to HP1-Mediated Silencing of Euchromatic Genes by KRAB Zinc-Finger Proteins. *Genes Dev.*

2002, *16* (8), 919–932.

Schwartz, J. J.; Lee, C.; Shendure, J. Accurate Gene Synthesis with Tag-Directed Retrieval of Sequence-Verified DNA Molecules. *Nat. meth* **2012**, *9* (9), 913–915.

Scott, J. L.; Gabrielides, C.; Davidson, R. K.; Swingler, T. E.; Ian, M.; Wallis, G. A.; Boot-handford, R. P.; Kirkwood, T. B. L.; Robert, W.; Young, D. A. Superoxide Dismutase Downregulation in Osteoarthritis Progression and End-Stage Disease. *Ann Rheum Dis* **2010**, *69* (8), 1502–1510.

Sen, R.; Baltimore, D. Multiple Nuclear Factors Interact with the Immunoglobulin Enhancer Sequences. *Cell* **1986**, *46*, 705–716.

Serfling, E.; Jasin, M.; Schaffner, W. Enhancers and Eukaryotic Gene Transcription. *Trends Genet* **1985**, *1*, 224–230.

Shalem, O.; Sanjana, N. E.; Hartenian, E.; Shi, X.; Scott, D. A.; Mikkelsen, T. S.; Heckl, D.; Ebert, B. L.; Root, D. E.; Doench, J. G.; et al. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science (80-.).* **2014**, *343* (6166), 84–87.

Sharon, E.; Kalma, Y.; Sharp, A.; Raveh-Sadka, T.; Levo, M.; Zeevi, D.; Keren, L.; Yakhini, Z.; Weinberger, A.; Segal, E. Inferring Gene Regulatory Logic from High-Throughput Measurements of Thousands of Systematically Designed Promoters. *Nat. Biotechnol.* **2012**, *30* (6), 521–530.

Shlyueva, D.; Stampfel, G.; Stark, A. Transcriptional Enhancers: From Properties to Genome-Wide Predictions. *Nat. Rev. Genet.* **2014**, *15* (4), 272–286.

Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; Thompson, J. D.; Higgins, D. G.; Mcwilliam, H.; et al. Fast , Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7* (539).

Simeonov, D. R.; Gowen, B. G.; Boontanrart, M.; Roth, T. L.; Gagnon, J. D.; Mumbach, M. R.; Satpathy, A. T.; Lee, Y.; Bray, N. L.; Chan, A. Y.; et al. Discovery of Stimulation-Responsive Enhancers with CRISPR Activation. *Nature* **2017**.

Skarnes, W. C. Entrapment Vectors: A New Tool for Mammalian Genetics. *Nat. Biotechnol.* **1990**, *8* (9), 827–831.

Smith, J.; Modrich, P. Removal of Polymerase-Produced Mutant Sequences from PCR Products. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (June), 6847–6850.

Smith, R. P.; Taher, L.; Patwardhan, R. P.; Kim, M. J.; Inoue, F.; Shendure, J.; Ovcharenko, I.; Ahituv, N. Massively Parallel Decoding of Mammalian Regulatory Sequences Supports a Flexible Organizational Model. *Nat. Genet.* **2013**, *45* (9), 1021–1028.

Sripathy, S. P.; Stevens, J.; Schultz, D. C. The KAP1 Corepressor Functions to Coordinate the Assembly of de Novo HP1-Demarcated Microenvironments of Heterochromatin Required for

KRAB Zinc Finger Protein-Mediated Transcriptional Repression. *Mol. Cell. Biol.* **2006**, *26* (22), 8623–8638.

Taher, L.; Mcgaughey, D. M.; Maragh, S.; Aneas, I.; Bessling, S. L.; Miller, W.; Nobrega, M. A.; Mccallion, A. S.; Ovcharenko, I. Genome-Wide Identification of Conserved Regulatory Function in Diverged Sequences. *Genome Res.* **2011**, *21* (7), 1139–1149.

Tewhey, R.; Kotliar, D.; Park, D. S.; Liu, B.; Winnicki, S.; Steven, K. Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay. *Cell* **2016**, *165* (6), 1519–1529.

Thakore, P. I.; Ippolito, A. M. D.; Song, L.; Safi, A.; Shivakumar, N. K.; Kabadi, A. M.; Reddy, T. E.; Crawford, G. E.; Gersbach, C. A. Highly Specific Epigenome Editing by CRISPR-Cas9 Repressors for Silencing of Distal Regulatory Elements. *Nat. Methods* **2015**, *12* (12), 1143–1149.

Tian, J.; Gong, H.; Sheng, N.; Zhou, X.; Gulari, E.; Gao, X.; Church, G. Accurate Multiplex Gene Synthesis from Programmable DNA Microchips. *Nature* **2004**, *432* (23/30), 1050–1054.

Tie, F.; Banerjee, R.; Stratton, C. A.; Prasad-Sinha, J.; Stepanik, V.; Zlobin, A.; Diaz, M. O.; Scacheri, P. C.; Harte, P. J. CBP-Mediated Acetylation of Histone H3 Lysine 27 Antagonizes Drosophila Polycomb Silencing. *Development* **2009**, *136*, 3131–3141.

Tishkoff, S. A.; Reed, F. A.; Ranciaro, A.; Voight, B. F.; Babbitt, C. C.; Silverman, J. S.; Powell, K.; Mortensen, H. M.; Hirbo, J. B.; Osman, M.; et al. Convergent Adaptation of Human Lactase Persistence in Africa and Europe. *Nat. Genet.* **2007**, *39* (1), 31–40.

Trizzino, M.; Park, Y.; Holsbach-beltrame, M.; Aracena, K. Transposable Elements Are the Primary Source of Novelty in Primate Gene Regulation. *Genome Res.* **2017**, *10*.

True, J. R.; Carroll, S. B. Gene Co-Option in Physiological and Morphological Evolution. *Annu. Rev. Cell Dev. Biol.* **2002**, *18*, 53–80.

Tsai, S. Q.; Zheng, Z.; Nguyen, N. T.; Liebers, M.; Topkar, V. V; Thapar, V.; Wyvekens, N.; Khayter, C.; Iafrate, A. J.; Le, L. P.; et al. GUIDE-Seq Enables Genome-Wide Profiling of off-Target Cleavage by CRISPR-Cas Nucleases. *Nat. Biotechnol.* **2015**, *33* (2), 187–197.

Ulirsch, J. C.; Nandakumar, S. K.; Wang, L.; Giani, F. C.; Rogov, P.; Melnikov, A.; Mcdonel, P.; Do, R.; Tarjei, S. Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **2016**, *165* (6), 1530–1545.

Vidigal, J. A.; Ventura, A. Rapid and Efficient One-Step Generation of Paired gRNA CRISPR-Cas9 Libraries. *Nat. Commun.* **2015**, *6*.

Villar, D.; Berthelot, C.; Flicek, P.; Odom, D. T.; Villar, D.; Berthelot, C.; Aldridge, S.; Rayner, T. F.; Lukk, M.; Pignatelli, M. Enhancer Evolution across 20 Mammalian Species Article Enhancer Evolution across 20 Mammalian Species. *Cell 160* (3), 554–566.

Villar, D.; Berthelot, C.; Flicek, P.; Odom, D. T.; Villar, D.; Berthelot, C.; Aldridge, S.; Rayner, T. F.; Lukk, M.; Pignatelli, M. Enhancer Evolution across 20 Mammalian Species Article Enhancer Evolution across 20 Mammalian Species. *Cell* **2015**, *160* (3), 554–566.

Visel, A.; Blow, M. J.; Li, Z.; Zhang, T.; Akiyama, J. A.; Plajzer-frick, I.; Shoukry, M.; Wright, C.; Chen, F.; Afzal, V.; et al. ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers. *Nature* **2009**, *457* (7231), 854–858.

Vockley, C. M.; Guo, C.; Majoros, W. H.; Nodzenski, M.; Scholtens, D. M.; Hayes, M. G.; Lowe, W. L.; Reddy, T. E. Massively Parallel Quantification of the Regulatory Effects of Noncoding Genetic Variation in a Human Cohort. *Genome Res.* **2015**, *25* (8), 1206–1214.

Wan, W.; Li, L.; Xu, Q.; Wang, Z.; Yao, Y.; Wang, R.; Zhang, J.; Liu, H.; Gao, X.; Hong, J. Error Removal in Microchip-Synthesized DNA Using Immobilized MutS. *Nucleic Acids Res.* **2014**, *42* (12).

Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **2008**, *456*, 470–476.

Wang, X.; Hou, J.; Quedenau, C.; Chen, W. Pervasive Isoform-Specific Translational Regulation via Alternative Transcription Start Sites in Mammals. *Mol. Syst. Biol.* **2016**, *12* (875).

Welter, D.; MacArthur, J.; Morales, J.; Burdett, T.; Hall, P.; Junkins, H.; Klemm, A.; Flicek, P.; Manolio, T.; Hindorff, L.; et al. The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations. *Nucleic Acids Res.* **2014**, *42* (D1), 1001–1006.

Whyte, W. A.; Bilodeau, S.; Orlando, D. A.; Hoke, H. A.; Frampton, G. M.; Foster, C. T.; Cowley, S. M.; Young, R. A. Enhancer Decommissioning by LSD1 during Embryonic Stem Cell Differentiation. *Nature* **2012**, *482* (7384), 221–225.

Wilson, C.; Pearson, R. K.; Bellen, H. J.; O'Kane, C. J.; Grossniklaus, U.; Gehring, W. J. P-Element-Mediated Enhancer Detection: An Efficient Method for Isolating and Characterizing Developmentally Regulated Genes in Drosophila. *Genes Dev.* **1989**, *3*, 1301–1313.

Wittkopp, P. J.; Kalay, G. Cis -Regulatory Elements : Molecular Mechanisms and Evolutionary Processes Underlying Divergence. *Nat. Rev. Genet.* **2011**, *13* (1), 59–69.

Wittkopp, P. J.; True, J. R.; Carroll, S. B. Reciprocal Functions of the Drosophila Yellow and Ebony Proteins in the Development and Evolution of Pigment Patterns. *Development* **2002**, *1858* (129), 1849–1858.

Wray, G. A. The Evolutionary Significance of Cis -Regulatory Mutations. **2007**, *8* (March), 206–217.

Wu, C. The 5′ Ends of Drosophila Heat Shock Genes in Chromatin Are Hypersensitive to DNase I. *Nature* **1980**, *286*, 854–860.

Wu, C.; Wong, Y. C.; Elgin, S. C. The Chromatin Structure of Specific Genes: II. Disruption of Chromatin Structure during Gene Activity. *Cell* **1979**, *16*, 807–814.

Wuestefeld, T.; Pesic, M.; Rudalska, R.; Dauch, D.; Longerich, T.; Kang, T.; Yevsa, T.; Heinzmann, F.; Hoenicke, L.; Hohmeyer, A.; et al. A Direct In Vivo RNAi Screen Identifies MKK4 as a Key Regulator of Liver Regeneration. *Cell* **2013**, *153* (2), 389–401.

Xi, H.; Shulha, H. P.; Lin, J. M.; Vales, T. R.; Fu, Y.; Bodine, D. M.; McKay, R. D.; Chenoweth, J. G.; Tesar, P. J.; Furey, T. S.; et al. Identification and Characterization of Cell Type-Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome. *PLoS Genet.* **2007**, *3* (8), e136.

Xie, S.; Duan, J.; Li, B.; Xie, S.; Duan, J.; Li, B.; Zhou, P.; Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* **2017**, *66* (2), 285–299.

Xu, D.; Nussinov, R. Favorable Domain Size in Proteins. *Fold. Des.* **1998**, *3* (1), 11–17.

Yamashita, R.; Sathira, N. P.; Kanai, A.; Tanimoto, K.; Arauchi, T.; Tanaka, Y.; Hashimoto, S.; Sugano, S.; Nakai, K.; Suzuki, Y. Genome-Wide Characterization of Transcriptional Start Sites in Humans by Integrative Transcriptome Analysis. *Genome Res.* **2011**, *21*, 775–789.

Yang, H.; Wang, H.; Shivalila, C. S.; Cheng, A. W.; Shi, L.; Jaenisch, R. One-Step Generation of Mice Carrying Reporter and Conditional Alleles by CRISPR / Cas-Mediated Genome Engineering. *Cell* **2013**, *154* (6), 1370–1379.

Yao, T. P.; Oh, S. P.; Fuchs, M.; Zhou, N. D.; Ch'ng, L. E.; Newsome, D.; Bronson, R. T.; Li, E.; Livingston, D. M.; Eckner, R. Gene Dosage-Dependent Embryonic Development and Proliferation Defects in Mice Lacking the Transcriptional Integrator p300. *Cell* **1998**, *93* (3), 361–372.

Young, L.; Dong, Q. Two-Step Total Gene Synthesis Method. *Nucleic Acids Res.* **2004**, *32* (7).

Zemojtel, T.; Kielbasa, Szymin, M.; Arndt, P. F.; Behrens, S.; Bourque, G.; Vingron, M. CpG Deamination Creates Transcription Factor – Binding. *Genome Biol Evol* **2011**, *3*, 1304–1311.

Zhang, F.; Lupski, J. R. Non-Coding Genetic Variants in Human Disease. *Hum. Mol. Genet.* **2015**, *24* (R1), R102–R110.

Zhang, J.; Kobert, K.; Flouri, T.; Stamatakis, A. PEAR: A Fast and Accurate Illumina Paired-End Read Merger. *Bioinformatics2* **2014**, *30*, 614–620.

Zheng, W.; Gianoulis, T. A.; Karczewski, K. J.; Zhao, H.; Snyder, M. Regulatory Variation Within and Between Species. *Annu. Rev. Genomics Hum. Genet.* **2011**, *12*, 327–346.

Zhou, X.; Cai, S.; Hong, A.; You, Q.; Yu, P.; Sheng, N.; Srivannavit, O.; Muranjan, S.; Rouillard, J. M.; Xia, Y.; et al. Microfluidic PicoArray Synthesis of Oligodeoxynucleotides and Simultaneous Assembling of Multiple DNA Sequences. *Nucleic Acids Res.* **2004**, *32* (18), 5409–

5417.

Zhu, S.; Li, W.; Liu, J.; Chen, C.; Liao, Q.; Xu, P.; Xu, H.; Xiao, T. Genome-Scale Deletion Screening of Human Long Non-Coding RNAs Using a Paired-Guide RNA CRISPR-Cas9 Library. *Nat. Biotechnol.* **2016**, *34* (12), 1279–1286.

VITA

Jason C. Klein grew up in Cherry Hill, NJ, where he attended Cherry Hill East High School. After high school, he matriculated at Duke University in 2012, where he majored in Biology with a certificate in Genome Sciences. During the summer of 2009, he began working in the lab of David McClay as an Institute for Genome Sciences and Policy summer fellow. Jason continued working in the lab throughout his time at Duke, studying the role of alternative splicing of *Erg* on epithelial to mesenchymal transition. He also spent a summer as a Pasteur Foundation Fellow in the lab of Francois Schweisguth studying the role of an uncharacterized gene in gastrulation and development and spent two years as a student and teaching assistant with Eric Spana mapping and cloning Drosophila mutations. For his Genome Sciences capstone project, he worked with Maynard Olson on a case-study of the National Human Genome Research Institute's Ethical, Legal and Social Implications Research Program. In 2012, Jason graduated from Duke University and entered the Medical Scientist Training Program at the University of Washington. In 2013, he rotated with Jay Shendure, and began the research described in this dissertation. Outside of classes and research, Jason is a competitive triathlete and runner.