Massively parallel functional dissection of regulatory elements

Rupali P. Patwardhan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Jay Shendure, Chair

Raymond Monnat

Robert Waterston

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Massively parallel functional dissection of regulatory elements

Rupali P Patwardhan

Chair of the Supervisory Committee:
Associate Professor Jay Shendure
Department of Genome Sciences

Massively parallel sequencing has accelerated the cataloging of *cis*-regulatory elements in mammalian genomes. However, it remains challenging to estimate the functional effects of variation in *cis*-regulatory elements. The current methods to measure such effects are labor-intensive and involve testing each variant separately. This dissertation describes the development of methods to interrogate functional effects of *cis*-regulatory variants in a massively parallel fashion. First, I present a method that takes advantage of massively parallel DNA synthesis and massively parallel sequencing to test the functional effects of all possible single nucleotide variants of a given *cis*-regulatory element *en masse* in a single assay. As a proof of concept, this method was applied to perform saturation mutagenesis of three bacteriophage core promoters and three core promoters recognized by the mammalian Pol II transcriptional machinery. Microarray synthesized mutant promoters, each with a unique 20bp tag sequence downstream of the transcription start site were subjected to *in vitro* transcription and the resulting RNA-derived tags were sequenced. The relative abundance of each programmed tag provided a digital readout of the transcriptional efficiency of its *cis*-linked mutant promoter. Next, I describe a method to generate long, accurate reads from short, error-prone reads produced by the current massively parallel sequencing platforms. This strategy, referred to as “subassembly”, is of broad utility in a wide range of contexts including but not limited to metagenomics, *de novo* genome assembly and detection of rare variants in clinical samples. It also enables the interrogation of longer regulatory elements beyond the current read-lengths supported by massively parallel sequencing platforms. Finally, I present an improved version of the saturation mutagenesis method, including incorporation of the

“subassembly” technique and use it to dissect mammalian enhancers up to 620bp long in a massively parallel *in vivo* assay. Development of such methods for rapid functional analysis of regulatory elements will not only facilitate interpretation of variation and understanding of the architecture and grammar of these elements, but also enable design of novel synthetic regulatory elements.

ACKNOWLEDGEMENTS

Over the course of my graduate career I have been incredibly fortunate to have had the chance to meet and work with the most amazing group of people, and I would like to acknowledge their contributions.

First and foremost, I would like to thank my mentor Jay Shendure for giving me the opportunity to work with him, for his guidance and support throughout the last five years, and for providing me with the best possible environment that a graduate student could hope for. I honestly could not have asked for a better mentor.

I would also like to thank members of my dissertation committee Phil Green, Ray Monnat, Bob Waterston and Alan Weiner for their advice and guidance throughout the course of my graduate career here at UW.

I am immensely grateful to all the members of the Shendure lab for their help. There are several to whom I owe a special mention of gratitude. I would like to thank Joe Hiatt for being an amazing partner on several projects and for serving as a sounding board on a daily basis. I would like to thank Sarah Ng for being a great friend and counselor during all moments of indecision. I would like to thank Emily Turner for her pragmatic advice and being a wonderful mentor. I would like to thank Charlie Lee for his infinite patience, his willingness to help with any task and for teaching me the ropes of bench work. I would also like to thank Jacob Kitzman for patiently offering several crucial tips and insights, Ruolan Qiu for sharing her experimental expertise and Akash Kumar for his contagious good-spirited attitude.

For several of my projects, I got a chance to interact and collaborate with people outside the lab. I would like to acknowledge their contribution. In particular, Daniela Witten lent crucial statistical

expertise. Interactions with Nadav Ahituv and Len Pennacchio spawned several new projects. Discussions with Ben Hall and members of his lab have always been educational and inspiring.

My classmates from entering class of 2007 played a huge role in making graduate school enjoyable and memorable. I will always remember the several hours we spent each day in the graduate student lounge during our first year in the program, be it working on homework, preparing for puzzle competitions or enjoying pot-luck lunches.

I would also like to thank several people who mentored me during my days at Indiana University, Bloomington, in particular, Mehmet Dalkilic, Peter Cherbas, Lucy Cherbas, John Colbourne and Justen Andrews.

I was also incredibly fortunate to be taught by excellent teachers and mentors throughout middle school and high school years. In particular I would like to thank Dr. Nandini Deshmukh for kindling my interest in genetics.

Finally, the support and encouragement of my parents and my grandparents, and other members of my family, as well as my friends from back home in India has played a crucial part in keeping me motivated to successfully complete any of my endeavors. My husband Gaurang Zaveri has been a constant source of strength and inspiration. Without his support, this work would not have been possible.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1 Introduction

The first human genome was sequenced just over a decade ago [1, 2], at the cost of 300 million US dollars and required an effort spanning ten years involving hundreds of scientists. Since then, there have been rapid advances in DNA sequencing technologies [3] such that the human genome can now be routinely sequenced in just a couple of days for a few thousand dollars, and this trend is expected to continue as newer and faster sequencing technologies are being developed. The feasibility of sequencing the entire genome of a fetus using non-invasive methods has also been demonstrated [4, 5]. These advances have made personal genome sequencing a reality. A large number of human genomes have already been sequenced [6-12], and with the availability of commercial whole genome sequencing services at reasonable costs, it might not be long before genome sequences become a standard component of our medical records. These sequences can pinpoint the specific set of mutations we carry. However, simply having access to this list of mutations is not useful. The true potential of this information can only be realized if we have the ability to predict the functional effects of these variants.

Predicting these functional effects is a challenging problem [13]. Variation in the portion of the genome that encodes proteins (called coding regions) is easier to interpret due to our deeper understanding of gene structure, and the rules governing translation. As a result a majority of current studies looking for mutations underlying diseases tend to only focus on these variants and using this strategy, these studies have successfully identified the mutations responsible for several rare genetic disorders [14-20], as well as for individual cases of more common but complex syndromes such as autism [21], schizophrenia [22] and mental retardation [23].

However, coding regions account for only around 1% of the human genome. Non-coding portions of the genome also play an important role in genome biology. In particular, regions that control the expression of genes, thus called regulatory elements, form a critical component of

the intricate logic that governs how the instructions encoded in our genome are read and executed. Predicting functional effects of variation in these regions is more challenging, largely due to our lack of understanding of the architecture of these elements as well as the lack of experimental methods to test the functional effects of these variants in a high-throughput manner.

1.1 Organization

This dissertation describes the development of such massively parallel methods to interrogate the functional effects of mutations in regulatory elements. In Chapter 2, I provide an overview of the current methods to identify regulatory elements in the genome and describe existing methods for functional analysis of these elements. In Chapter 3, I describe a method to interrogate the effect of all possible single-nucleotide mutations in a core promoter in a single experiment, and demonstrate its use on three bacteriophage core promoters, as well as three core promoters recognized by the mammalian Pol-II transcription machinery. In Chapter 4, I describe a method called “subassembly” that can generate longer and more accurate effective reads from short sequencing reads. This technology enables the application of massively parallel functional dissection method to larger regulatory elements. In Chapter 5, I present an improved version of the massively parallel functional assay and demonstrate its use to dissect mammalian enhancers. I conclude with Chapter 6 in which I speculate about the possible directions in which these studies can be expanded in the future.

Chapter 2 Background

2.1 Regulatory elements

A vast majority of the functions required for keeping a cell alive are performed by proteins. The recipe for making each protein is encoded in the DNA sequence, which specifies the exact sequence of amino acids that will constitute the protein. However, not all proteins are required at all times, and in all cell types. In order to ensure proper functioning of the cell, when and how much of each protein gets made has to be very carefully orchestrated. For instance, certain proteins are only required at very specific time points during development. A classic example is the switch between fetal, embryonic and adult forms of hemoglobin (reviewed in [24]). Others need to be expressed only in response to some environmental trigger, such as presence of nutrients or toxins [25]. Moreover, proteins need to be expressed in very specific quantities in order to maintain the stoichiometric ratios and reaction kinetics with other interacting proteins. Disturbances in these ratios could result in improper functioning and disease. For example, an imbalance in the quantities of alpha and beta globin can lead to thalassemia [26]. The programming required to achieve this fine-scaled control is also encoded in the genome, and is implemented via the interaction of specific proteins with specific regions of the genome. These regions of the genome, and the specific nucleotide sequences they represent are broadly referred to as “regulatory regions” or “regulatory elements”, and are further classified into different classes such as promoters, enhancers and insulators, based on their function and position. These elements specifically regulate the transcription of genes into RNA, the first step in production of proteins. In addition to these transcriptional regulatory elements, there are several other kinds of regulatory elements that control a variety of downstream steps such as splicing of the nascent RNA transcript into messenger RNA (mRNA), modulating the stability of mRNA, and the efficiency of translation of mRNA into protein. In this dissertation, I will primarily

refer to promoters and enhancers, and will use the rest of this chapter to introduce these two elements and describe the current methods for their identification and analysis.

2.2 Promoters

A promoter is a segment of DNA that is capable of driving the transcription of its immediate downstream sequence. It accomplishes this due to the presence of specific sequence features that are recognized by the transcriptional machinery. The position from where transcription begins is called the transcription start site (TSS).

The precise sequence requirements for a functional promoter, especially in mammalian genomes, are still not completely understood. Initially, promoters were identified on the basis of the presence of a conserved A+T-rich sequence with the approximate consensus “TATAA” (and hence referred to as the “TATA-box”), approximately 30 bases upstream of the TSS [27, 28]. It was subsequently discovered that a conserved Initiator element (INR) centered over the TSS also played an important role in directing transcription [29], and was in fact capable of directing transcription in the absence of a TATA-box [30]. Studies in organisms like *Drosophila melanogaster* revealed another motif around 30 bases downstream of the TSS, named Downstream Promoter Element (DPE), capable of directing the transcriptional machinery to a precise TSS [31]. Availability of the complete human genome and the subsequent genome-wide computational analyses revealed that only a small percentage of human promoters contain any of these elements. The TATA-box for example is present in fewer than 20% of promoters and around 50% of the promoters lack any of these canonical elements [32, 33].

2.3 Enhancers

An enhancer is a segment of DNA that can activate and up-regulate the transcription of its target genes. Enhancers are typically located several kilobases away from their target genes and can act independently of their orientation [34]. Enhancers are believed to function via looping of the intervening DNA such that the enhancer and its target promoter are in physical proximity in three-dimensional space [35-37]. Enhancers are known to bear clusters of binding sites for several specific proteins called transcription factors, and are often responsible for the tissue-specific activation of the genes they regulate.

2.4 Discovery of regulatory elements

Identification of cis-regulatory elements, especially enhancers in mammalian genomes is a challenging problem because they can be scattered several kilobases (or more) away from their regulatory targets. However significant progress has been made in the last few years in the development of methods to discover them. These include both computational as well as experimental techniques.

2.4.1 Computational methods

A majority of the computational methods to discover regulatory elements rely on evolutionary conservation to identify non-coding sequences under functional constraint. This is a powerful technique, especially given that the genome sequences of a large number of organisms are now available. For example, several of the ultra-conserved non-coding regions identified by Bejerano et al. [38] on the basis of perfect conservation across human, mouse and rat genomes proved to be functional enhancers [39]. More recently Lindblad-Toh et al. generated a very large list of constrained elements and potentially functional regulatory elements using evolutionary conservation across 29 mammalian genomes [40]. However methods relying on sequence

conservation alone miss regulatory elements in rapidly evolving regions of the genome. Another category of methods overcomes this problem by relying on detecting an enrichment of known transcription factor binding sites (TFBSs) in localized regions [41-43]. Some methods use a combination of evolutionary conservation and TFBS enrichment by searching for conserved arrangements of TFBSs [44]. Computational methods can generate a large list of candidate regulatory elements. However, they still suffer from relatively high false positive rates and cannot at present replace experimental techniques for discovery as well as validation.

2.4.2 Experimental methods

Several effective experimental techniques have been recently developed and successfully applied to identify regulatory elements throughout the genome.

ChIP-Seq

ChIP-seq [45, 46] can be used to identify all locations in the genome bound by a protein of interest. Cells or tissues are first cross-linked with formaldehyde to preserve the DNA-protein associations. DNA is extracted, fragmented into smaller segments, and then immunoprecipitated using an antibody against the protein of interest. This enriches for DNA fragments bound by the protein. The cross-linking is then reversed and the identity of these DNA fragments is learned by sequencing (or by hybridization to microarrays in an earlier version of ChIP-seq called “ChIP-chip” [47]). As a part of the ENCODE project [48] as well as through numerous independent studies, this technique has been used to identify the binding sites of several transcription factors, as well as regions of the genome with specific chromatin modifications in a wide variety of tissues and cell types. The data from these methods have been used to identify locations of broader classes of regulatory elements. For example, ChIP-seq with the enhancer-associated protein p300 has been used to identify tissue-specific enhancers [49-51]. ChIP-chip and ChIP-seq have also been used to establish the chromatin

signatures associated with individual classes of regulatory elements. For example promoters are marked by trimethylation of lysine 4 residue of histone H3 (H3K4me3), whereas poised as well as active enhancers are marked by monomethylation of the same residue (H3K4me1) [52]. Active enhancers can further be distinguished from poised enhancers by acetylation of histone H3 at lysine 27 (H3K27ac) [53, 54], although it is now believed that H3K27ac more generally marks active elements, both enhancers as well as promoters. In another comprehensive study, Ernst et al. used ChIP-seq to map nine chromatin marks across nine cell types and used recurrent combinations of these marks to define distinct chromatin states corresponding to repressed, poised and active promoters, strong and weak enhancers, putative insulators, transcribed regions, and large-scale repressed and inactive domains [55].

DNase I hypersensitivity profiling

Functionally active regions of the genome need to be accessible to transcriptional factors and other proteins complexes. Hence they are more likely to have “open” chromatin, and thus likely to be “hypersensitive” to digestion by DNase I. DNase I hypersensitivity mapping can thus be used to identify potential regulatory regions in the genome [48, 56, 57]. Intact nuclei are extracted from lysed cells and digested with DNase I. The digested DNA, enriched in DNase I digested ends, is subjected to massively parallel sequencing (or in an older version, hybridized to tiling microarrays). Regions of the genome that are enriched for an increased number of read-starts along consecutive positions are designated as DNase I hypersensitive regions [58]. Even within these hypersensitive regions, there are often dips in the number of read-starts at certain positions. These are referred to as DNase footprints and represent regions that are protected likely because they are bound by proteins. A high resolution map of DNA-protein interactions, including potential transcription factor binding sites can thus be obtained [59].

DNA Calling Cards (Card-seq)

A technique called DNA calling cards is useful to identify all positions bound by a transcription factor of interest. The transcription factor is fused to a transposase of a transposon. When the transcription factor binds its targets in the genome, the transposase directs transposon insertions in the vicinity of the binding site. This transposon insertion serves as a persistent marker or a “calling card” recording the binding event. The locations of these insertions can be learned by sequencing (Card-seq) or by hybridization to an array. This method was initially developed in yeast making use of the *Ty5* transposon as the calling card [60, 61], but has since also been demonstrated in human cells using the *piggyBac* transposon [62]. Although Card-seq requires introduction of an expression construct into the cells and can thus not be readily applied to tissues, it could serve as an alternative to ChIP-seq, especially when no antibodies are available against the transcription factor of interest. Due to the permanent form in which an interaction is recoded, Card-seq could prove particularly useful in studying differentiation. For example, it confers the ability to correlate transcription-factor binding events in progenitor cells to the final fates of their progeny cells during development.

CAGE

Relative to other types of regulatory elements, identification of promoters might seem straightforward due to their constrained location relative to genes. However compiling a comprehensive list of all functional promoters in the genome is not a trivial task due to several reasons. First, there are a large number of unannotated non-coding genes in the genome (e.g. microRNAs, long non coding RNAs). Second, even annotated protein-coding genes often have several alternative transcripts, some of which are driven by alternative promoters. Third, for any downstream analysis, the precise locations of TSSs provide much greater power than approximate locations of the promoter regions. Cap Analysis of Gene Expression (CAGE) enables identification of promoters and TSSs by accurately capturing and sequencing the capped 5’ end of transcripts [32, 63, 64].

2.5 Functional analysis of regulatory elements

Techniques described in the previous section provide the coarse locations of regulatory elements. In this section I will describe current methods to functionally validate these candidate regulatory elements, discover their underlying architecture and to understand the effects of deviation from the wild-type sequences.

2.5.1 Measuring transcriptional activity

Testing whether a predicted transcriptional regulatory element is functional most commonly involves cloning the regulatory element to be tested upstream of a transcriptional cassette and subjecting it to *in vitro* or *in vivo* transcription. The next step is to detect and quantify RNA transcripts driven by that element. Different methods have been developed over the years for this purpose.

Nuclease protection assays

In this technique [65, 66], which was used in 1980s for many landmark studies [34, 67, 68], transcripts are hybridized to a molecular excess of radio-labeled complementary probe specific to the transcripts to be detected. The mixture is then exposed to a nuclease that digests single stranded RNA. Surviving RNA fragments correspond to the regions bound by the probe and thus a part of the transcripts of interest. These transcript fragments are run on an electrophoretic gel where they are detected by autoradiography and can be quantified by comparing the band intensity to a series of bands obtained by hybridization of the probe to different known quantities of the target sequence.

Reporter genes

The most common method to quantify activity of transcriptional regulatory elements since late 1980s right up to the present day is using reporter genes such as CAT [69], luciferase [70] or

fluorescent proteins [71]. The element to be tested is used to drive transcription of the reporter gene. The activity of the element is measured in terms of the activity of the protein encoded by the reporter gene. Reporter gene assays have been used in hundreds of studies to test individual regulatory elements. In recent years, the availability of the human genome sequence enabled characterization of much larger sets of elements using this technique. For example Trinklein et al. tested more than a hundred predicted human promoters for activity in four human cell lines using transient transfection luciferase assays [72]. Using the same approach, Cooper et al. expanded the study to almost 400 promoters across 16 different cell lines [73]. They also tested nested deletion fragments for 45 of these promoters. This led to interesting insights such as the presence of positive regulatory elements in the -350 to -40 region upstream of the TSS and repressive elements in the -500 to -1000 region.

2.5.2 Dissection of regulatory elements

Going a step further from validating candidate regulatory elements, functional dissection hopes to uncover the underlying architecture of a regulatory element by introducing variations in its wild-type sequence and measuring the activity of these variants. Knowing which changes lead to changes in activity of the element can provide clues to the functional parts of the element.

Variants can include nested deletions, scanning deletions, single nucleotide substitutions, scanning block substitutions or multiple substitutions randomly scattered across the element. When the activities of all possible substitutions in an element are systematically tested, the method is referred to as “saturation mutagenesis”. The substitution variants are traditionally generated using site-directed mutagenesis, error-prone PCR, or chemical treatment of DNA [74].

Saturation mutagenesis can yield incredibly rich functional data at single-nucleotide resolution. However, only a handful of regulatory elements have been analyzed in detail using saturation

mutagenesis because the process has traditionally been laborious and time consuming. Considerable effort is needed to generate and isolate constructs representing every possible single-nucleotide substitution, and then to measure the transcriptional activity each of these constructs individually.

One of the most well-known examples of saturation mutagenesis of a regulatory element is the functional dissection of the beta-globin promoter [67]. In this study, they used the *in vitro* mutagenesis method described in [74] to introduce 130 different random single base substitutions in the beta-globin promoter region. Briefly, this involved chemical treatment of single stranded DNA, synthesis of the complementary strand, separation of fragments bearing wild-type and mutant versions of the promoter by running them on a denaturing gel, cloning of these fragments into the appropriate expression vector, followed by sequencing to learn the identity of the mutations. Each mutant was then transiently transfected into HeLa cells and the transcripts harvested after 48 hours were analyzed using the S1 nuclease assay described in the previous section. Although laborious, this process not only allowed them to identify the three regions critical to the activity of the promoter (-87 to -95, -72 to -77 and -26 to -30), but also the precise effect of any specific substitution on promoter activity.

Although not as comprehensive as saturation mutagenesis, scanning mutagenesis in which several adjacent bases are mutated at once is also a powerful technique for functional dissection and has been used to dissect several regulatory elements over the last several decades, such as the beta-globin enhancer [75]. Fine-nested deletions have also been shown to be useful in mapping the critical functional regions, for example in the beta-interferon enhancer [76] as well as the adenovirus-2 major late and chicken conalbumin promoters [29].

More recently, attempts have also been made to learn the rules governing architecture and grammar of regulatory elements using synthetic regulatory elements. This is powerful because it

allows sampling of a much larger space than that offered by existing genomic elements. For example, Gertz et al. [77] constructed and tested synthetic promoter libraries consisting of random arrangement of binding sites for a handful of transcription factors. They used the results to construct thermodynamic models of gene regulation. In another study, basal promoters containing either strong, or weak or no TATA-boxes were cloned *in cis* with different combinations of TF binding sites as in Gertz et al. to generate three synthetic promoter libraries [78]. They were able to show that the TATA-box acts as a simple scaling factor such that gene expression scales with the strength of the TATA-box independent of the arrangement of transcriptional factor binding sites upstream of the TATA-box.

All of the studies described above used either nuclease protection assays or reporter genes for readout of transcriptional activity. While highly sensitive, these methods still suffer from a fundamental limitation, which is that the activity of each construct has to be measured separately. This makes testing of large numbers of such constructs infeasible. In the next chapter I will describe a method that allows quantitative readout of the transcriptional activity of thousands of constructs *en masse* using massively parallel sequencing.

Chapter 3 Functional dissection of core promoters

This chapter is based on the following published paper:

Rupali P Patwardhan, Choli Lee, Oren Litvin, David L Young, Dana Pe'er and Jay Shendure. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, 27, 1173 - 1175 (2009).

With one significant exception noted below, Choli (Charlie) Lee and I performed all of the experimental work. With one significant exception, I performed all of the data analysis, in consultation with Jay Shendure. Jay Shendure and I wrote the manuscript.

David Young performed the *in vivo* validation of individual synthetic T7 promoters described in section 3.4.3. Oren Litvin analyzed data from a pilot experiment to choose the optimal combinations of double mutants described in section 3.4.2.

3.1 Summary

This chapter describes a method that harnesses massively parallel DNA synthesis and massively parallel sequencing for the high-throughput functional analysis of core promoters at single-nucleotide resolution. As a proof of concept, the effects of all possible single-nucleotide mutations for three bacteriophage promoters and three mammalian core promoters were assayed in a single experiment per promoter. Mutant promoters were synthesized in parallel as DNA oligonucleotides (oligos) on a programmable microarray and then released into solution, resulting in a complex library including all mutant promoters. Each synthetic promoter in the library included a unique 20bp sequence tag downstream of the promoter's transcription start site (TSS). The synthetic promoters were subjected to *in vitro* transcription and the resulting transcripts were sequenced. The relative abundance of each programmed tag provided a digital readout of the transcriptional efficiency of its *cis*-linked mutant promoter. In addition to facilitating the functional analysis of core promoters already present in the genome, this method could also serve as a rapid screening tool for regulatory element engineering.

3.2 Introduction

In spite of the rapid advances in our understanding of genome biology, regulatory regions of the genome remain still poorly understood. One of the most systematic and high resolution methods to functionally characterize regulatory elements is to test the effect of every possible single nucleotide mutation on the function of that element. This is referred to as saturation mutagenesis. The traditional process of performing saturation mutagenesis involves constructing all the mutant versions of the element one at a time and testing each one separately in order to quantify the effect of that mutation on activity of the element. The mutant versions are generated using either site-directed or random mutagenesis methods. The effect of each mutation on activity of the element, at least in the context of transcriptional regulatory

elements, is quantified by having them drive the expression of a reporter gene such as luciferase or a fluorescent protein. To date, only a handful of regulatory elements have been analyzed in this manner [67, 79-82], largely due to the laborious and time-consuming nature of such assays.

Here we present a high-throughput method that allows us to overcome these bottlenecks and systematically analyze the effect of mutations at every position in a core promoter in a single experiment. To accomplish this, we took advantage of the advances in DNA synthesis as well as sequencing technologies.

3.3 Method overview

An overview of the method is presented in **Figure 3.1**. Mutant promoters are synthesized in parallel as DNA oligonucleotides on a programmable microarray and released into solution [83], resulting in a complex library. Each oligonucleotide in the library is designed to include a unique sequence tag downstream of the promoter's transcription start site (TSS). The oligonucleotides are transcribed *in vitro*, and the resulting transcripts are sequenced. The relative abundance of each programmed tag provides a digital readout of the transcriptional efficiency of its *cis*-linked mutant promoter.

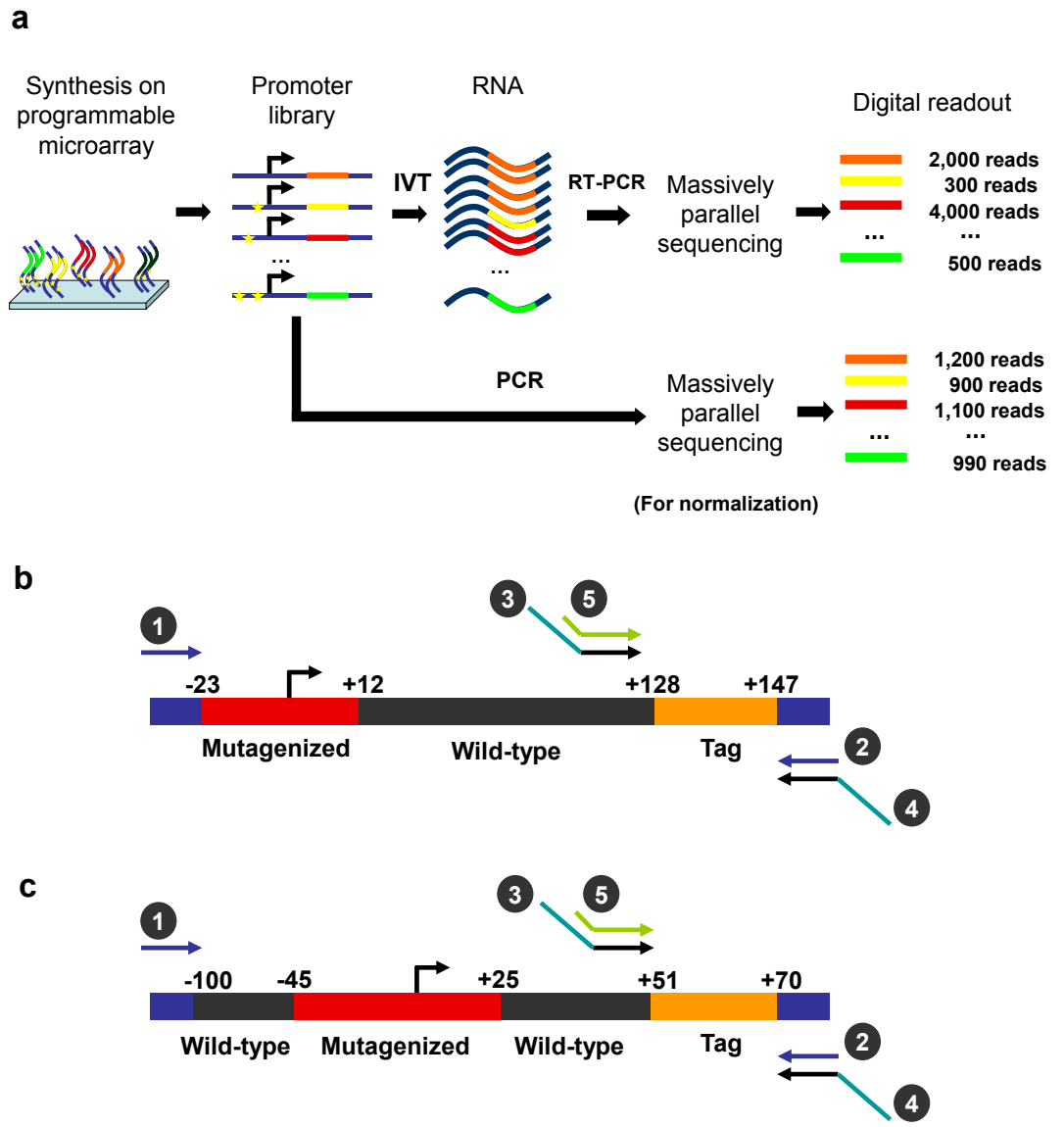


Figure 3.1 Overview of synthetic saturation mutagenesis.

(a) Promoter templates are synthesized on a programmable microarray, released into solution and amplified by PCR (primers 1 and 2). One fraction of the amplified promoter library is subjected to *in vitro* transcription followed by reverse-transcription PCR (primers 3 and 4). Another fraction is PCR amplified using the same primers. Tags within RNA- and DNA-derived amplicons are sequenced separately (sequencing primer 5). RNA-derived tag counts provide a digital readout of the transcriptional efficiency of associated promoters. DNA-derived tag counts are used to normalize for any non-uniformity in the initial oligonucleotide concentrations. **(b)** For bacteriophage promoters, each 200-nt oligonucleotide consists of the promoter (red), 115-nt of the wild-type downstream sequence (black), a variable 20-nt tag (orange) and 15-nt PCR primers (blue) on either side. **(c)** For mammalian Pol II promoters, each 200-nt oligonucleotide consists of the promoter region from −100 to +50 (black), including the region subjected to saturation mutagenesis (red), followed by a variable 20-nt tag (orange) and 15-nt PCR primers (blue) on either side.

3.4 Synthetic saturation mutagenesis of bacteriophage core promoters

As a proof of concept, this method was applied to three well-characterized bacteriophage promoters: T3 (class 3, phi13), T7 (class 3, phi10) and SP6 (SP6p32). We focused on a 35-nt region, spanning 23-nt upstream and 12-nt downstream of each promoter's TSS (**Figure 3.1b**). At each position, we mutated the wild-type nucleotide to every other nucleotide or introduced a single-nucleotide deletion. We also included several double mutation promoters, allowing us to compare the single mutants to their combination. Rather than including all possible pair-wise combinations, we only included promising candidate pairs chosen on the basis of single-base substitution data from a pilot experiment (See **Appendix A** for the criteria used). To guard against the potential influence of the tag itself on transcriptional activity, we represented each mutant variant of each native promoter by six distinct 20-nt tags (**Appendix A**). Wild-type promoters with no mutations were also included and were each represented by 270 different tags. These served as positive controls and provided a baseline against which to compare transcriptional efficiencies of mutant promoters. Templates with random sequence in place of the promoter were included as negative controls (**Table 3.1** and **Appendix A**).

Table 3.1 Synthetic promoter library constituents for bacteriophage promoters.

	Promoter variants			Tags per promoter variant
	T3	**T7**	**SP6**	
Single base substitutions	105	105	105	6
Single base deletions	35	35	35	6
Double base substitutions	553	453	464	6
Wild-type	1	1	1	270
Random	274	274	274	1

The promoter library was transcribed *in vitro* with one of three RNA polymerases (T7, T3 or SP6). The resulting RNA pools were reverse transcribed, PCR amplified and sequenced on an Illumina GAII system. Reads were then mapped back to the 20-nt tags that we had programmed in *cis* with each synthetic promoter. To control for potentially non-uniform representation of synthesized oligos (e.g., owing to differential synthesis efficiencies, systematic biases in PCR efficiency or biases inherent to the sequencer itself), we also PCR amplified the DNA library that served as input to the *in vitro* transcription reaction and sequenced it in a separate lane. A comparison between counts of DNA- and RNA-derived tags associated with each wild-type (unmutated) promoter found that although synthetic promoter concentrations varied, they maintained a linear relationship with transcription efficiency (**Appendix A and Figure A.1**). The RNA-based counts associated with each tag were therefore normalized by dividing by the corresponding DNA-based counts.

Counts of tags corresponding to the wild-type promoter established the baseline activity of the wild-type promoter and an empirical null distribution for assessing significance. The effect of each mutation was measured as a fold-change in transcription relative to the wild-type promoter. Based on the variation observed within each set of 270 tags associated with each wild-type promoter, we were able to call changes of twofold or greater as statistically significant ($P < 0.01$, after Bonferroni correction for multiple testing) (**Appendix A and Figure A.2**).

3.4.1 Analysis of single mutants

The observed transcriptional profiles clearly delineated a core 'footprint' for each promoter, within which substitutions and deletions caused a drastic drop in efficiency of transcription (**Figure 3.2**). We also observed a range of position and mutation-specific effects. For example,

the −10 position within the SP6 promoter core region could be substituted without decreasing activity. In fact, a T → A substitution at this position caused a significant increase in transcriptional efficiency, consistent with previous studies of this promoter [82]. At certain positions, substitution of the wild-type nucleotide by a specific nucleotide was tolerated whereas other nucleotides were not. For instance, the change from A → G at position −1 on the T3 promoter was deleterious, whereas changes A → C or A → T were benign. In general, the SP6 promoter was more efficient than T7 and T3, and correspondingly more sensitive to the disruptions we introduced. Data from the SP6 mutants was also used to compute an activity logo (**Figure A.3**) to enable direct comparison with results from a previous saturation mutagenesis study of this promoter [82].

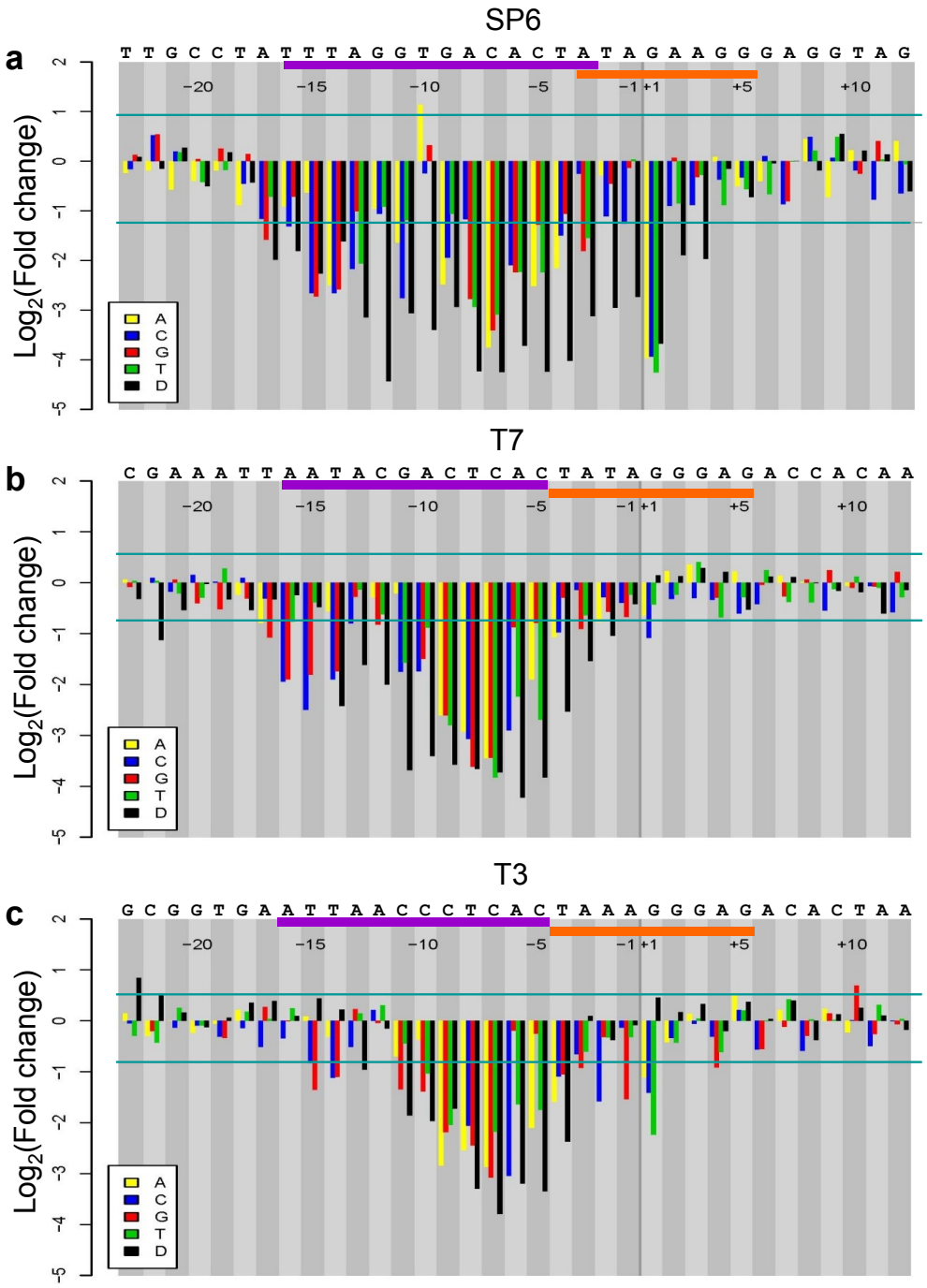


Figure 3.2 Synthetic saturation mutagenesis of bacteriophage core promoters.

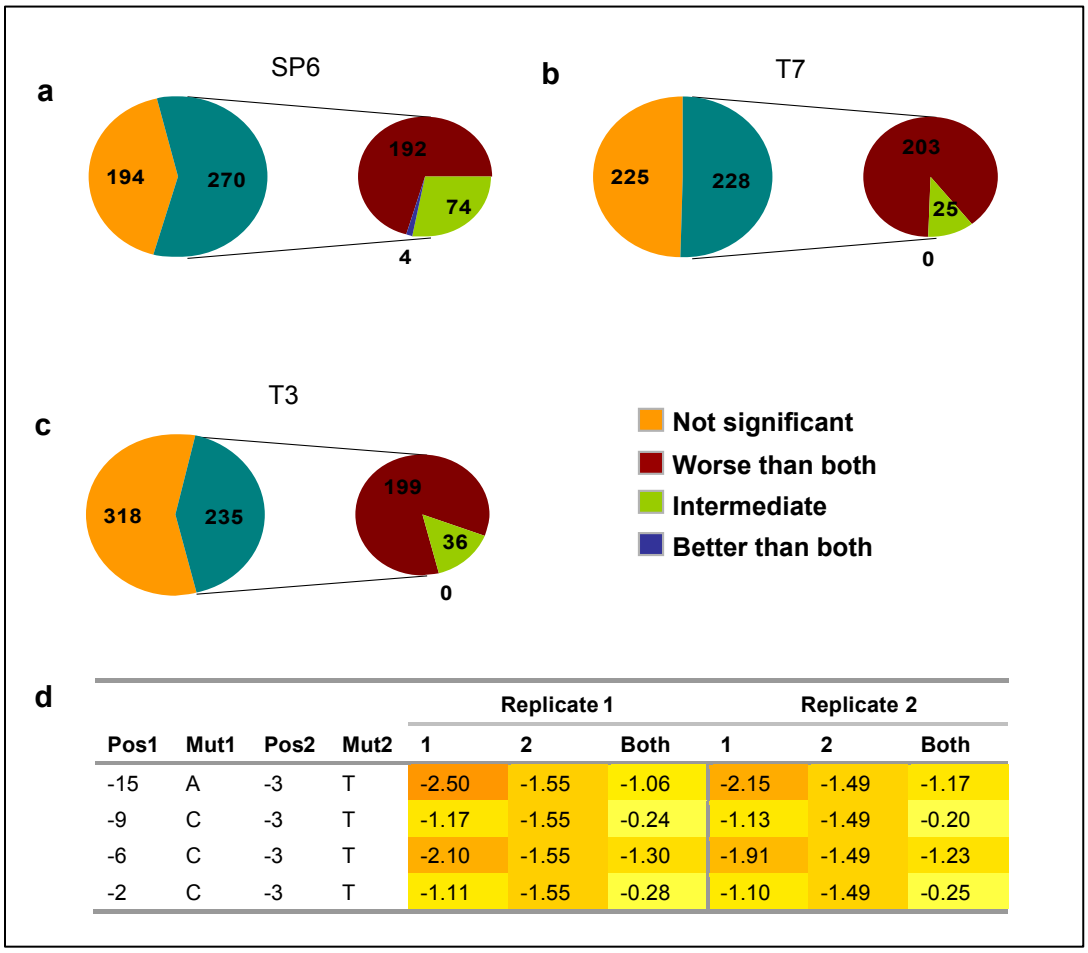
Changes in transcriptional efficiency (average of six tags) for each single-nucleotide substitution or deletion (D) relative to the wild-type promoter for bacteriophage promoters SP6 (a), T7 (b) and T3 (c) respectively. Horizontal lines mark significance cutoffs ($P < 0.01$). Horizontal axis denotes the position of the mutation relative to TSS, from −23 to +12, with wild-type nucleotides specified above. Polymerase binding (purple bar) and melting/initiation (orange bar) regions are also indicated above.

3.4.2 Analysis of double mutants and epistatic interactions

To explore whether we could detect synergistic or antagonistic associations between point mutations, our library of mutant promoters also included templates with substitutions at two positions within the promoter. Because it was not practical to test all possible permutations of double mutations, we used results of a pilot experiment consisting of only single mutants (data not shown) to choose a subset that provided a robust sampling of mutation position and severity (**Appendix A: Supplementary Methods**). We compared the double-mutant outcomes against predictions based on the corresponding single mutants, assuming a log-additive model. Although 65–70% of the double mutants matched predicted values, the rest showed deviations from this model, hinting at synergistic and compensatory interactions (**Figure A.4**). We filtered double mutants for the subset where at least one of either of the single mutants or the double mutant satisfied our significance threshold for fold-change relative to the native promoter (**Figure 3.3a-c**).

As expected, the effect of most double mutants was greater than either of the corresponding single mutants. However, there were also a number of cases where the effect of the combination of mutations was intermediate to the effects of the two corresponding single mutants, suggesting varying degrees of partial rescue. Finally, there were four SP6 double mutants that were less harmful than either of their corresponding single mutants. Notably, each of these four involved an A → T substitution at −3 as one of the mutations (**Figure 3.3d**). *In vitro* binding assays have shown that this mutation leads to a twofold increase in the strength of polymerase binding [82], which might explain the compensatory effect that we observe here. Although the single A → T mutation at −3 is associated with a decrease in transcriptional activity, we note that this is not necessarily inconsistent as we are measuring transcriptional activity rather than polymerase binding strength. For example, it may be that increased

polymerase binding directly underlies the observed decrease in transcriptional efficiency associated with the single A → T mutation at −3 (**Figure 3.2a**), whereas a second mutation occurring at any number of positions serves to reduce the strength of polymerase binding toward a more optimal level for transcription (**Figure 3.3d**).



d

				Replicate 1			Replicate 2		
Pos1	Mut1	Pos2	Mut2	1	2	Both	1	2	Both
-15	A	-3	T	-2.50	-1.55	-1.06	-2.15	-1.49	-1.17
-9	C	-3	T	-1.17	-1.55	-0.24	-1.13	-1.49	-0.20
-6	C	-3	T	-2.10	-1.55	-1.30	-1.91	-1.49	-1.23
-2	C	-3	T	-1.11	-1.55	-0.28	-1.10	-1.49	-0.25

Figure 3.3 Classification of double-mutant templates based on their effect on transcription.

(a-c) The templates where either the double mutant or at least one of the corresponding single mutants have a significant effect on transcription relative to the wild-type promoter are further classified based on the effect of the double mutant as compared to the two single mutants. (d) Details of the four SP6 double mutants whose transcriptional efficiency was higher than both the corresponding single mutants.

3.4.3 Validation of activity of synthetic promoters *in vivo* using luciferase assays

In synthetic biology, the multiplex *in vitro* evaluation of large numbers of synthetic promoters would represent an efficient empirical strategy for identifying variants that adjust the *in vivo* activity of a promoter with predictable magnitude. We sought to evaluate whether activities of individual synthetic promoters determined within our multiplex *in vitro* assay were recapitulated *in vivo*. Six T7 promoter variants were individually inserted upstream of a bacterial luciferase reporter in pCS26, a low-copy number plasmid [84], and the constructs were used to transform a T7 polymerase–expressing *Eshcerichia coli* strain. *In vivo* activities of the promoters as measured by luciferase luminescence correlated well with predictions based on the *in vitro* assay (r = 0.92) (**Figure 3.4** and **Appendix A**).

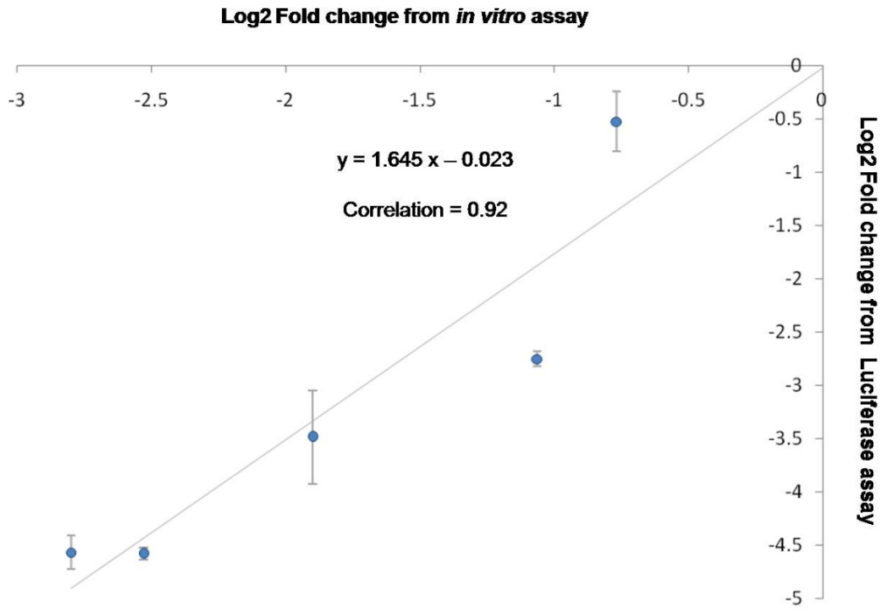


Figure 3.4 Comparison between activities of synthetic promoters predicted by massively parallel assay and individual luciferase assays

Correlation between the log2(fold-change) relative to the wild-type promoter for 5 T7 mutant promoters as seen in the tag-based *in vitro* assay versus individual *in vivo* analysis by the luciferase assay averaged across 9 replicates (three biological replicates, with three technical replicates each). The error bars indicate standard deviation. The luminescence was measured one hour after induction of the promoter. The differences in scaling between activities observed in the luciferase assay as compared to the *in vitro* values for low activity mutants could be due to differences in the nature of measurement as well as the challenges with measuring low activity levels accurately.

3.5 Synthetic saturation mutagenesis of mammalian core promoters

Next we evaluated whether this approach could be extended to promoters recognized by the mammalian transcriptional machinery. We assayed three core promoters: the immediate early promoter of the human cytomegalovirus (*CMV*), the promoter of the human beta globin gene (*HBB*) and the promoter of human S100 calcium binding protein A4 (*S100A4/PEL98*). The promoter region included on each oligonucleotide extended 100-nt upstream and 50-nt downstream of the TSS. For saturation mutagenesis, we focused on a 70-nt region spanning 45-nt upstream and 25-nt downstream of the TSS (**Figure 3.1c**). As previously described, we included six different tags per mutation. Wild-type promoters with no mutations were represented by 100 tags each (Table **3.2** and **Table A.2**).

Table 3.2 Synthetic promoter library constituents for Pol II promoters.

	Promoter variants			Tags per promoter variant
	CMV	HBB	S100A4	
Single Base Substitutions	210	210	210	6
Single Base Deletions	70	70	70	6
Wild-type	1	1	1	100
Random	60	60	60	1

In vitro transcription was performed using HeLa nuclear extracts. Libraries were separately generated from RNA and DNA and sequenced separately, and analysis was carried out as described for the bacteriophage promoters. In all three cases, we were able to detect changes in transcription that correlated with expectation (**Figure 3.5**).

For example, mutations disrupting the AT-rich groove that defines the TATA box of the CMV promoter (TATATA, −28 to −23) led to a clear drop in transcriptional efficiency. Substitutions of

C → A or C → T at −29 increased transcriptional efficiency, potentially secondary to the formation of a more optimal TATA box (−30 to −25) with respect to distance from the TSS (**Figure 3.5a**). Mutations disrupting the initiator element (TCAGATC, +1 to +7; **Appendix A: Supplementary Note**) also caused significant drops in transcription. Single-nucleotide deletions at any position between the TATA box and the initiator sharply reduced transcription, likely a result of violation of spacing constraints [85]. The results also suggested the presence of two additional elements, one near +16 and another near the −45 region.

The HBB promoter has a non-canonical TATA box (CATAAA, −32 to −27) [86], mutations in which have been documented in beta-thalassemia. As expected, our assay detected significant drops in transcription with changes to this motif (**Figure 3.5b**). Notably, a C → T substitution at −32 (creating a canonical TATA box, TATAAA) increased the strength of the promoter. However, we did not observe any significant effects of initiator or E-box mutations, in contrast with previous studies in a different cell type [87]. With the S100A4 core promoter, mutations disrupting both the canonical TATA box (TATAAA, −31 to −26) and the initiator element (CCATTCT, −2 to +5) led to drops in transcriptional efficiency (**Figure 3.5c**). Single-nucleotide deletions between the TATA box and the TSS did not show any significant effect on the HBB and S100A4 core promoters, in clear contrast with the CMV core promoter.

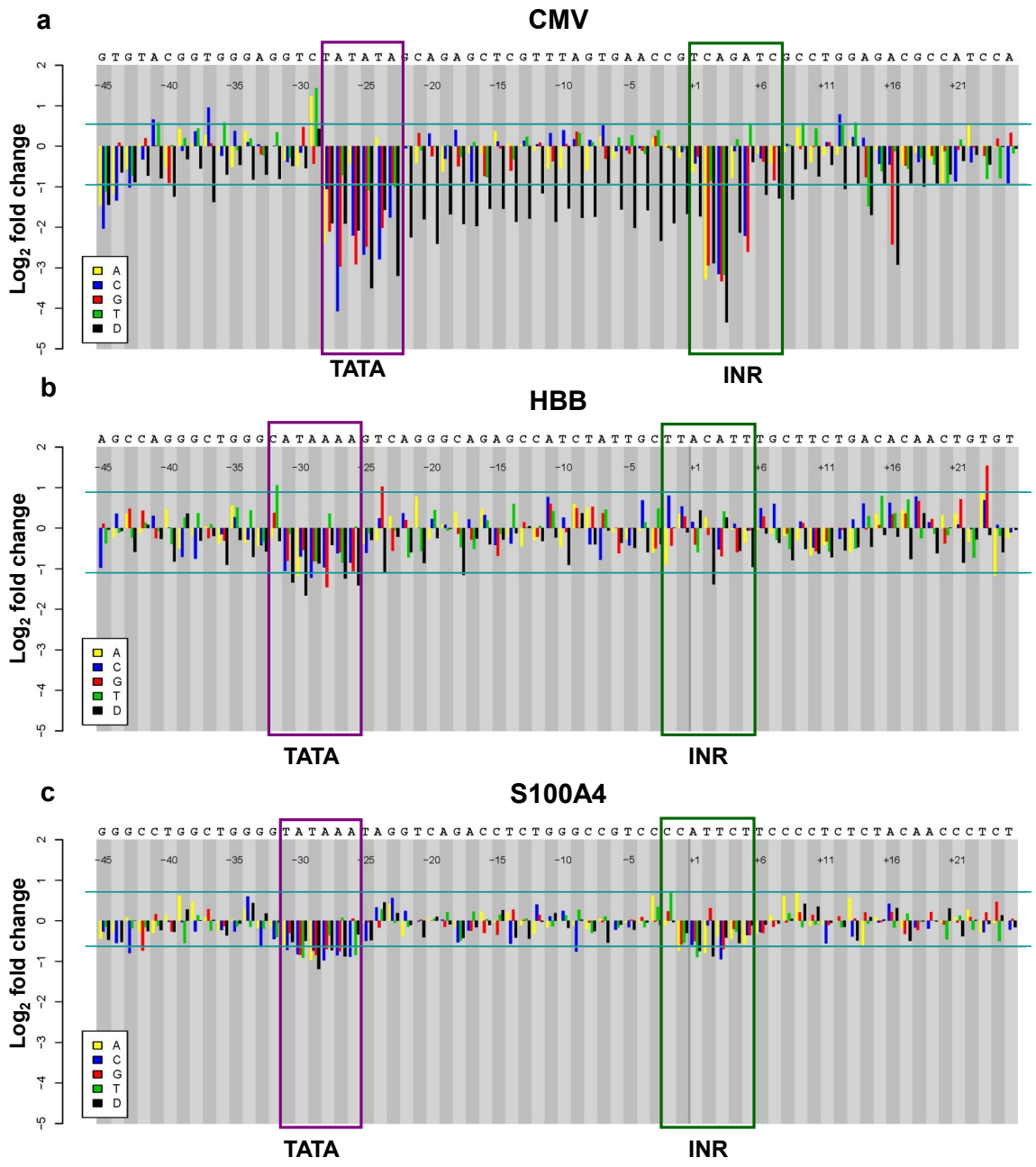


Figure 3.5 Mutagenesis of mammalian core promoters.

Transcriptional fold-change (average of six tags) for each single-nucleotide substitution or deletion (D) relative to the wild-type promoter for CMV (a), HBB (b) and S100A4 (c). Horizontal lines mark significance cutoffs ($P < 0.01$). Horizontal axis denotes the position of the mutation relative to TSS, from −45 to +25 with wild-type nucleotides specified above.

To evaluate reproducibility, we replicated the entire experiment for all six promoters. The distribution of observed fold-changes in transcriptional efficiency for each mutation as compared to the native promoter was reproducible, with correlation coefficients of 0.98, 0.97, 0.96, 0.99, 0.87 and 0.70 for the SP6, T7, T3, CMV, S100A4 and HBB core promoters respectively (**Figure A.5**). The lower reproducibility of S100A4 and HBB core promoters appears to be related to lower levels of transcriptional activity relative to the bacteriophage and CMV core promoters. The current experimental design required fitting the promoter, tag and other common sequences to the maximum available length of synthetic oligos (200 nt), whereas longer promoter fragments would have been likely to yield higher levels of activity [73].

3.6 Discussion

Synthetic saturation mutagenesis with quantitative readout by deep sequencing of *cis*-linked tags enables the measurement of the relative activities of thousands of core promoter variants in a single experiment. The use of programmable synthetic oligonucleotides also allows precise combinations of mutations to be studied in a directed fashion. Sequence tags eliminate the need for reporter genes or other cumbersome quantification techniques while allowing for a high level of multiplexing. Synthetic saturation mutagenesis may represent a useful and scalable tool for both regulatory element analysis and forward engineering of gene networks.

The method presented here is a good demonstration of how regulatory elements could be tested using a massively parallel assay. However, several limitations still remain before it could be directly applicable to more complex and larger elements such as distal promoters and enhancers. One of the most immediate concerns is the fact that the maximum length of array-synthesized oligonucleotides is currently 200–300 bp, whereas mammalian enhancers can be 1 kb or longer. Elements of that scale could potentially be constructed by polymerase chain assembly (PCA) of shorter overlapping oligonucleotides (e.g. 50-90bp), bearing either

programmed or random mutations. The resulting elements will thus have different combinations of mutations brought together, resulting in thousands of unique “haplotypes”. Further, these elements will have to be connected to tags at random, and the mapping, as well as the sequence of the element itself will have to be learned by sequencing the entire constructs. This poses a new challenge: the read-lengths on the current massively parallel sequencing platforms such as Illumina HiSeq and MiSeq are only 300bp, assuming 150-bp paired end reads. The next chapter describes a method called “subassembly”, which will allow us to solve this problem.

3.7 Notes

Data availability:

Raw Illumina sequencing reads have been submitted to the NCBI Short Read Archive under center name UWGS-JS.

Acknowledgements:

The authors would like to thank M.G. Surette (Univ. of Calgary) for the generous gift of the pCS26 plasmid used for the luciferase assays; E. LeProust and W. Woo (Agilent Technologies) for array-derived oligonucleotides libraries and E. Turner, J.B. Hiatt, S. Ng, J. Kitzman, R. Monnat, B. Stone, A. Dudley and N. Goddard for helpful discussions. D.P. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

Chapter 4 Subassembly: Parallel, tag-directed assembly of locally derived short sequence reads

This chapter is based on the following published paper:

Joseph B Hiatt, **Rupali P Patwardhan**, Emily H Turner, Choli Lee and Jay Shendure. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods*, 7, 119 - 122 (2010).

Bold face indicates equal contributors.

Emily Turner and Jay Shendure conceived the initial approach. Joseph Hiatt led the development of the subassembly method to its published form. All experimental work was performed by Joseph Hiatt and Emily Turner. I developed the computational framework to perform data analysis. Joseph Hiatt, Emily Turner and I performed data analysis. Choli Lee performed Illumina sequencing. Joseph Hiatt and Jay Shendure wrote the manuscript with contributions from me and Emily Turner.

4.1 Summary

"Subassembly" is an *in vitro* library construction method that extends the utility of short-read sequencing platforms to applications requiring long, accurate reads. A long DNA fragment library is converted to a population of nested sublibraries, and a tag sequence directs grouping of short reads derived from the same long fragment, enabling localized assembly of long fragment sequences. Subassembly can be applicable in a variety of contexts such as accurate de novo genome assembly, metagenome sequencing, rare variant detection and sequencing of long, randomly assembled synthetic DNA molecules.

4.2 Introduction

The cost and throughput advantages of massively parallel sequencing are offset by large tradeoffs with respect to read length and accuracy [3]. Although the availability of reference assemblies renders short reads sufficient for genomic re-sequencing and digital profiling [88, 89], other areas such as metagenomics [90], *de novo* assembly of complex genomes [91], immunoglobulin diversity profiling [92] and molecular haplotyping [93] are more challenging. In metagenomics, for example, sequences are derived from a population of related and unrelated genomes with highly varying abundances and a potentially enormous effective complexity. For identifying new open reading frames and for resolving related sequences within such a population, long reads remain indispensable [90].

As a means to deliver long reads using existing short-read massively parallel platforms, we developed a multiplex, *in vitro* strategy, termed subassembly, which is conceptually analogous to hierarchical shotgun genome assembly (**Figure 4.1**). In this approach, one of the two reads from a paired-end read serves as a sequence tag that identifies groups of short reads sharing a clonal origin, that is, deriving from the same longer DNA fragment (~500 bp). Each group of

short, locally derived reads is then collapsed to a long, subassembled (SA) read. To evaluate performance, we applied this method to two samples: genomic DNA from a (G+C)-rich organism, *Pseudomonas aeruginosa* strain PAO1, and a previously characterized metagenomic sample from lake sediment [94].

4.3 Method overview

For subassembly, we sheared DNA to relatively long lengths (for example, ~500 bp), ligated 'tag-adjacent' adaptors to the fragments and then diluted and PCR-amplified these fragments **(Figure 4.1** and **Appendix B**). The dilution step before PCR imposed a complexity bottleneck, such that a limited number (~10^5–10^7) of long fragments were amplified to high abundance (**Appendix B**). The PCR amplicons were concatemerized and then sonicated, and a single 'breakpoint-adjacent' adaptor was ligated to the sheared fragments. We performed a second round of PCR in which one primer corresponded to a tag-adjacent adaptor and the other primer corresponded to the breakpoint-adjacent adaptor. The resulting amplicons effectively comprise a population of nested sub-libraries derived from the original long-fragment library. The tag-adjacent adaptor provides access to genomic sequence that corresponds to the ends of the long fragments. As this end sequence will be consistent across amplicons derived from the same long fragment, it can serve as a tag to identify molecules that are clonally derived. After paired-end sequencing, the read primed by the tag-adjacent adaptor identifies the original long DNA fragment, and the read primed by the breakpoint-adjacent adaptor represents sequence from a shearing-determined breakpoint in that fragment. As a relatively short read could serve as a unique tag identifier, we obtained paired-end reads of unequal length (20-bp 'tag read' and 76-bp 'breakpoint read'). In the analysis, we used tag reads to group breakpoint reads and separately subjected each tag-defined read group (TDRG) to local assembly with phrap [95].

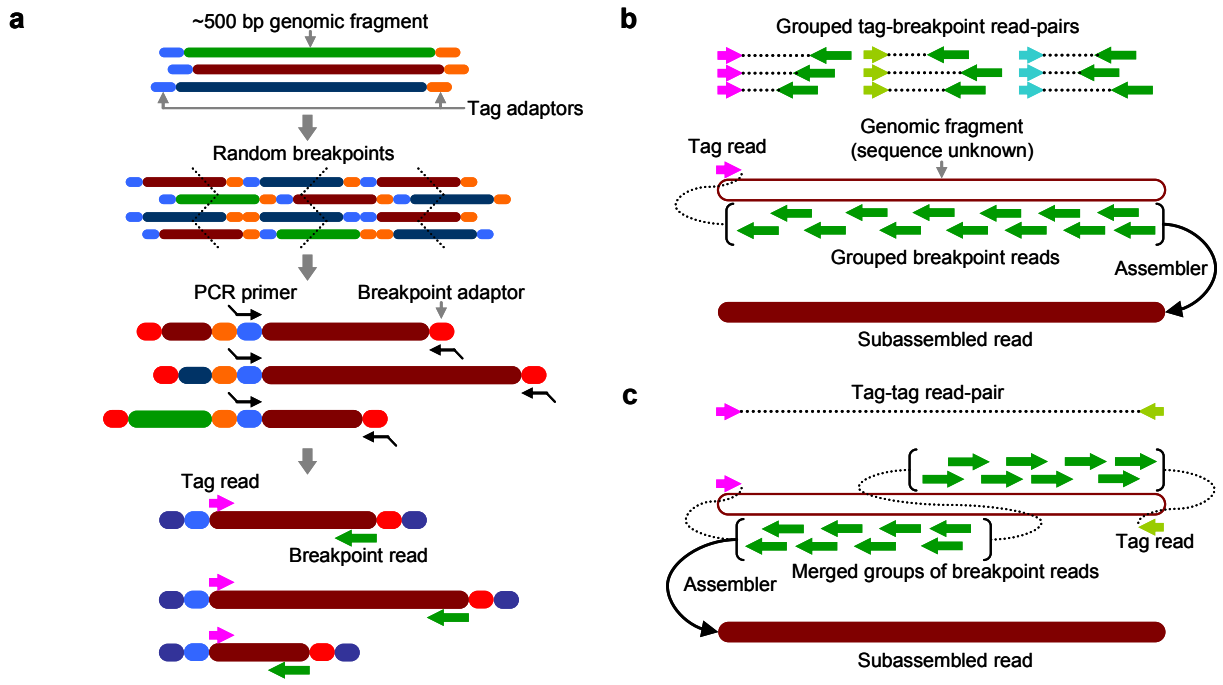


Figure 4.1 Schematic of subassembly process.

(a) Long DNA fragments are ligated to tag-adjacent adaptors, diluted and PCR-amplified. Dilution imposes a complexity bottleneck so that a limited number of long fragments are amplified. Concatemerized PCR products are then sheared by sonication and ligated to a breakpoint-adjacent adaptor. A second PCR amplification prepares amplicons for sequencing; one end of these amplicons corresponds to an end of a long fragment and the other end corresponds to a shearing breakpoint internal to that fragment. (b) Breakpoint reads are grouped *in silico* based on the sequence of the corresponding tag read. Breakpoint reads within a group, which derive from positions internal to the same parent long fragment, are subjected to local assembly to generate a subassembled read. (c) The metagenomic bottlenecked long-fragment library is subjected directly to paired-end Illumina sequencing to identify pairs of tag reads that were derived from opposite ends of the same original fragment. Two groups of breakpoint reads defined by distinct tag reads are merged and assembled together to generate one or more subassembled reads. In this study, this step was only applied to the metagenomic sample.

4.4 Application of subassembly to *P. aeruginosa* genome assembly

To rigorously assess performance, we applied subassembly to *P. aeruginosa* strain PAO1. After fragmenting genomic DNA, we size-selected it to ~550 bp (**Figure B.1a**) and processed the sample as illustrated in **Figure 4.1**. We used Illumina Genome Analyzer II (GA-II) to generate 56.8 million read pairs. We grouped the read pairs into TDRGs by the 20-bp tag (**Appendix B**) and separately subjected 76-bp breakpoint reads in each TDRG to local assembly with *phrap* to produce SA reads (**Table B.1**). We discarded SA reads not derived from identically oriented

breakpoint reads (1.2%) and those failing subassembly entirely (2.7%). For subsequent analyses, we considered only the longest SA read from TDRGs with ≥10 members.

This subset comprised 1.03 million SA reads with a median length of 338 bp (**Figure 4.2a** and **Table B.2**). The bimodal distribution may be due to uneven coverage of the original fragment secondary to imperfect size selection (**Figure B.2**). To assess quality, we mapped the SA reads to the *P. aeruginosa* strain PAO1 reference [96] and found that 99.82% had significant ($P < 10^{-6}$) alignments with basic local alignment search tool (BLAST) [97], with 98% of SA reads aligning along ≥95% of their full lengths. Although the contributing Illumina reads had an error rate of 2.4%, the substitution error rate of aligning SA reads was 0.25%. The longest correct SA read was 680 bp, likely an outlier from the gel-based size selection but nonetheless an indicator of the method's potential. We also estimated quality scores for bases in SA reads from the quality scores of contributing breakpoint reads (**Appendix B**). The 85% of bases in SA reads with the highest estimated quality scores were >99.99% accurate with respect to substitution errors when compared to the *P. aeruginosa* strain PAO1 reference (**Figure 4.2b**). Finally, we calculated the substitution error rate as a function of position along the SA read. The low overall error rate of one per 400 bp was maintained for hundreds of bases in the SA reads (**Figure 4.2c**).

.

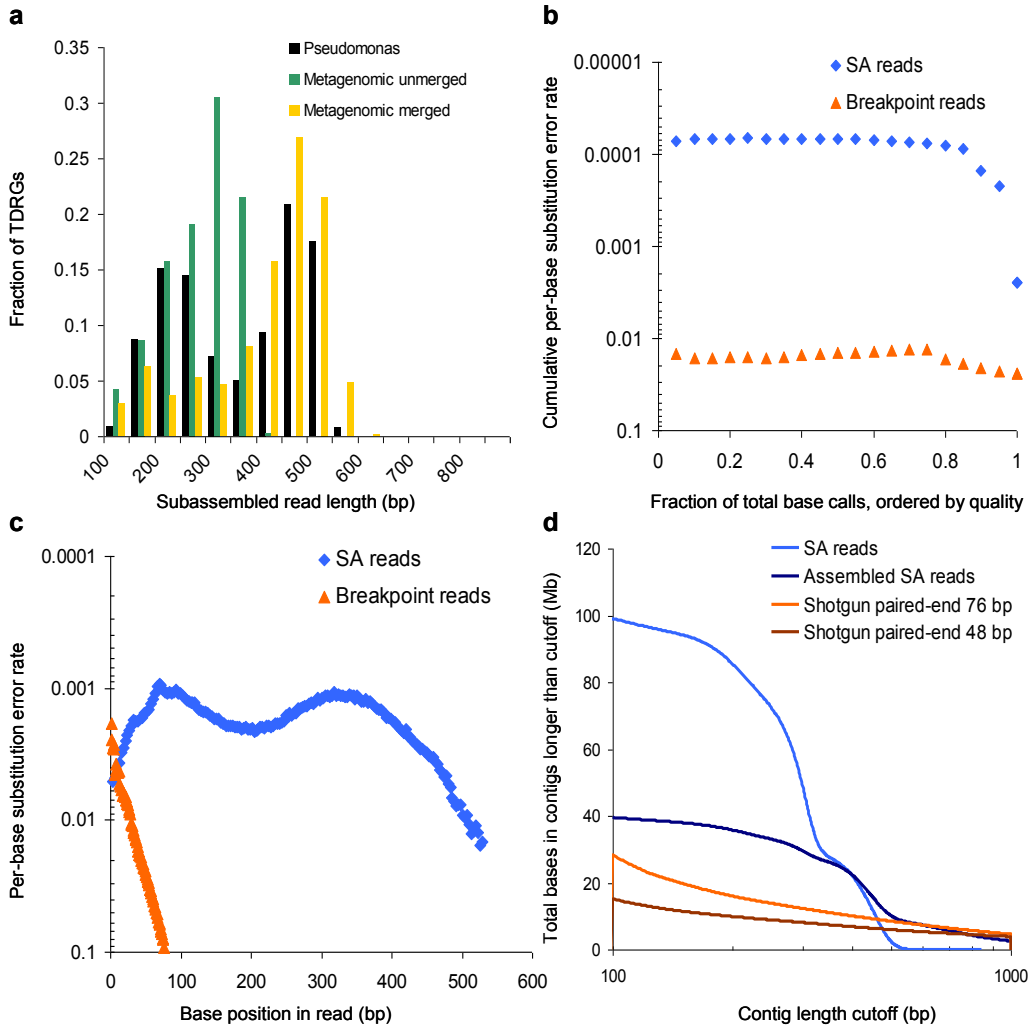


Figure 4.2 Evaluation of subassembly performance.

(a) Distribution of subassembled (SA) read length for P. aeruginosa sample and for methylamine metagenomic sample for unmerged and merged pairs of tag-defined read groups. (b) Cumulative per-base substitution error rate of base calls binned as a function of descending base quality in raw and SA reads, or the error rate of the x% of bases with the highest quality scores, after using BLAST to define the corresponding sequence in the reference. (c) Substitution error rate of base calls as a function of base position in raw and SA reads (binned every 3 bases). (d) Total length in sequences longer than a variable cutoff produced from SA reads compared to a standard shotgun library for the 100–1,000 bp range in which metagenomic analyses become possible. SA reads and assembled SA reads were compared to assembly of 48-bp or 76-bp paired-end reads from a standard Illumina shotgun library using Velvet with optimized parameters and an equivalent amount of raw sequence. Assembled SA reads refers to contigs produced by CABOG from SA reads.

Based on alignment with BLAST, SA reads covered 98.85% of the reference at a mean coverage of 63-fold. We observed bias against regions of extremely high G+C content (>70%)

relative to shotgun sequencing (**Figure B.3**), which could be mitigated by optimizing PCR conditions. We also observed slight systematic bias in the distribution of SA read quality scores across the reference that we conclude is unlikely to compromise accuracy at positions with adequate coverage (**Figure B.3**).

To explore the utility of subassembly for de novo genome assembly, we assembled all filtered SA reads using CABOG [98], resulting in 708 contigs ≥1 kilobase (kb) with an N50, or the length x such that 50% of the genomic length is in sequences at least x long, of 15 kb (**Table 4.1**). The substitution error rate was ~1/14,000, and there was a total of 65 bp of inserted or deleted sequence across 31 contigs. Contigs ≥20 kb, which comprised 2.3 Mb, were more accurate, with a substitution error rate of ~1/250,000 and 20 bp of insertion-deletions across eight contigs. BLAST alignment predicted 11 contigs ranging in size from 1 to 18 kb to contain local misassemblies, but four of these were related to differences between the strain used here and the reference (**Appendix B: Supplementary Note 2**), leaving only seven true misassemblies. Six of these were very local deletions or expansions of <400 bp (within contigs <20 kb long), and one 1,125 bp contig displayed a more complex BLAST alignment.

Table 4.1 De novo assembly of *P. aeruginosa* genome using subassembled (SA) reads

Input	Assembly strategy	# of contigs / scaffolds	Contig / scaffold N50	Longest contig / scaffold	Total sequence	Reference coverage
SA reads	Celera	708	15,070 bp	160,221 bp	6.07 Mb	96.2%
SA reads + PE fragment + jumping mate-pair	Celera + scaffolding	32	444,483 bp	915,353 bp	6.11 Mb	99.3%

Assembly of SA reads from *P. aeruginosa* using the Celera assembler produces long and accurate contigs and can be further extended by scaffolding contigs with short (~200 bp) and long (~2.5 kb) mate-pairing data. Listed is the data used for assembly, the assembly strategy (we used a custom scaffolding algorithm), the number of contigs (for SA reads, ≥ 1 kb) or scaffolds (for SA reads with shotgun data, ≥ 5 kb), the contig or scaffold N50, the longest contig or scaffold, and the coverage of the reference genome. Physical coverage (sequence covered by contigs and N's spanning contigs) is shown for the assembly derived from SA reads supplemented with paired-end and mate-pair data.

Shotgun assembly of SA reads therefore resulted in long and highly accurate sequences with contiguity likely limited by sequence content biases. To facilitate scaffolding, we included sequencing data from one lane of a paired-end fragment library (2 × 36 bp; insert size ~200 bp) and one lane of a mate-paired jumping library (2 × 36 bp; insert size ~2.5 kb). Using a custom iterative scaffolding algorithm (**Appendix B**), we generated 32 scaffolds ≥5 kb, with scaffold N50 of 445 kb, longest scaffold of 915 kb and 99.3% physical coverage of the reference (**Table 4.1**). Notably, scaffolding introduced only one misassembly, likely because of the presence of multiple nearly identical phage-like insertions (**Appendix B: Supplementary Note 2**). Our results, which were generated from a single platform, compare favorably to summary statistics of a published *de novo* assembly from a related organism that had been generated by combining long-read 454 and short-read Illumina data [99] (**Appendix B: Supplementary Note 3**).

To evaluate subassembly on a complex metagenomic sample, we used total DNA isolated from lake sediment and enriched for methylamine-fixing microbes [94]. We started with a slightly shorter long-fragment library (~450 bp; **Figures B.1b, B.4**) and imposed a more stringent complexity bottleneck by diluting the long-fragment library to ~105–106 molecules before PCR (**Appendix B**). We obtained 21.8 million read pairs, which resulted in 262,298 TDRGs, in which the median length of the longest SA read in filtered TDRGs was 256 bp (**Table B.2** and **Figure 4.2a**).

In addition to the nested breakpoint reads that we used to produce SA reads, we also obtained 1.8 million paired-end reads from the original long-fragment library (2 × 20 bp), allowing us to merge TDRGs whose tags were observed as a read pair (**Figure 4.1**). We merged ~68% of the metagenomic TDRGs in this fashion. Subjecting breakpoint reads from merged TDRGs to local assembly yielded SA reads with a median length of 408 bp (**Figure 4.2a** and **Table B.2**).

4.5 Application of subassembly to metagenomics

We hypothesized that localized, tag-directed assembly would be particularly useful in the context of metagenomics, for which the highly nonuniform representation of organisms complicates de novo assembly from short reads. To test this, we generated a standard Illumina shotgun paired-end library from the same metagenomic sample and assembled reads from this library with Velvet [100] using optimized parameters (**Table B.3** and **Figure B.5**). We evaluated shotgun assemblies from both paired-end 76-bp reads and paired-end 48-bp reads. For both assemblies, we used 2.2 Gb of raw sequence, which was equal to the amount of data used for subassembly.

CABOG assembly of SA reads yielded considerably more total sequence data in longer contigs than direct assembly of shotgun reads, generating greater than twice as much sequence in contigs ≥200 bp (**Figure 4.2d** and **Table B.3**). Unassembled SA reads comprised greater than five times as much sequence ≥200 bp. Notably, shotgun assemblies did achieve greater contiguity at the longest lengths (**Table B.3** and **Figure B.5**). These long contigs may be due to deep sampling of the most abundant genomes. However, many are likely to represent misassemblies, as we did not observe long BLAST alignments to the available Sanger sequence data [94] or to any sequence in the GenBank nt or env_nt databases.

To conservatively estimate each method's effective coverage, we compared assembled contigs to 37.2 Mb of Sanger sequence data recently reported for the same sample [94] (**Appendix B** and **Figure B.6**). Although the complexity of the metagenomic sample likely remains undersampled, subassembly covered at least 45% more of the Sanger sequence reference when compared to contigs assembled from the paired-end short-read library. In addition, subassembly generated a comparable amount of total sequence as compared to Sanger sequencing data (39.5 Mb versus 37.2 Mb) in somewhat shorter contigs (median of 390 bp

versus 835 bp) but with considerably less effort (three Illumina sequencing lanes versus hundreds of Sanger sequencing runs). In summary, subassembly produced substantially more sequence at lengths necessary for accurate phylogenetic classification [101] and gene discovery [102] than direct assembly from shotgun short reads and did so in better agreement with the available Sanger sequencing data, suggesting that the quality of assembled data may also be higher.

4.6 Discussion

Given that we observed accurate SA reads of nearly 700 bp, optimization of this method in concert with the tag-pairing approach (**Figure 4.1**) could potentially extend the effective length of SA reads to ~1 kb, that is, approaching the maximum length of Sanger sequencing data. One potential concern about the method as described is that tag sequences from different long DNA fragments can occasionally be identical by chance, especially if samples contain repetitive elements at high abundance. A simple modification would be to use a tag-adjacent adaptor containing an embedded degenerate sequence (for example, a randomized 20-bp segment), as this would completely decouple the tag sequence from the sample composition.

Finally, we note that subassembly offers a fundamental advantage in the way that a low error rate is achieved with a second-generation sequencing platform. Accurate assembly of short shotgun reads can be successful, provided that these reads are derived from relatively random sequence and that deep, uniform coverage can be obtained [100]. Platforms such as Roche 454 offer long reads at a cost that is likely similar to subassembly (**Appendix B: Supplementary Note 4**) but have error profiles comparable to those of other second-generation sequencing platforms. Therefore, achieving high consensus accuracy also depends on the assumptions of

uniform sampling and of a common origin for nearly identical reads. In contrast, because subassembly samples individual long DNA fragments and separately reconstructs a consensus sequence for each one, the production of long, accurate SA reads is insulated from nonuniform representation and sequence relatedness in the sample of interest.

4.7 Notes

Data availability:

Raw Illumina sequence reads have been deposited to the NCBI Short Read Archive under the accession number SRA010316.

Acknowledgements:

We thank L. Chistoserdova and M.G. Kalyuzhnaya (University of Washington) for the gift of the methylamine-enriched metagenomic DNA sample, C. Manoil (University of Washington) for the gift of *P. aeruginosa* strain PAO1 genomic DNA and P. Green for helpful discussions. J.B.H. is supported by US National Institutes of Health grant T32GM007266 and an Achievement Rewards for College Scientists fellowship.

Chapter 5 Functional dissection of enhancers

This chapter is based on the following published paper:

Rupali P Patwardhan, **Joseph B Hiatt**, Daniela M Witten, Mee J Kim, Robin P Smith, Dalit May, Choli Lee, Jennifer M Andrie, Su-In Lee, Gregory M Cooper, Nadav Ahituv, Len A Pennacchio and Jay Shendure. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology*, 30, 265–270 (2012).

Bold face indicates equal contributors.

I performed majority of the experimental work with significant exceptions noted below. Joseph Hiatt, Daniela Witten and I performed all of the data analysis, with one significant exception noted below. Joseph Hiatt, Jay Shendure and I wrote the manuscript.

Mee J Kim, Robin Smith and Dalit May performed tail vein injections and RNA extraction from mouse livers. Mee J Kim and Robin Smith performed luciferase assays for validation of six individual ALDOB mutants. Choli Lee performed Illumina sequencing. Gregory Cooper contributed the evolutionary conservation analysis.

5.1 Summary

The functional consequences of genetic variation in mammalian regulatory elements are poorly understood. Here we report the *in vivo* dissection of three mammalian enhancers at single-nucleotide resolution through a massively parallel reporter assay. For each enhancer, we synthesized a library of >100,000 mutant haplotypes with 2–3% divergence from the wild-type sequence. Each haplotype was linked to a unique sequence tag embedded within a transcriptional cassette. We introduced each enhancer library into mouse liver and measured the relative activities of individual haplotypes *en masse* by sequencing the transcribed tags. Linear regression analysis yielded highly reproducible estimates of the effect of every possible single-nucleotide change on enhancer activity. The functional consequence of most mutations was modest, with ~22% affecting activity by >1.2-fold and ~3% by >2-fold. Several, but not all positions with higher effects showed evidence for purifying selection, or co-localized with known liver-associated transcription factor binding sites, demonstrating the value of empirical high-resolution functional analysis.

5.2 Introduction

In Chapter 3, I described a method called 'synthetic saturation mutagenesis' in which programmable microarrays were used to synthesize variants of core promoters, each in *cis* with a downstream tag sequence. The population of core promoter variants was subjected to a cell-free *in vitro* assay, after which sequencing of the transcribed tags was performed to quantify the relative activity of specific core promoter variants. This method is very effective in the context of core promoters, and potentially other small elements. However several aspects limit its broader application and scalability: (i) when each regulatory element variant is synthesized as a separate array feature, the overall cost of synthesis remains high; (ii) the separate synthesis of individual variants also limits how many combinations of mutations can be simultaneously

programmed; (iii) the maximum length of array-synthesized oligonucleotides is currently 200–300 bp, whereas mammalian enhancers can be 1 kb or longer; (iv) access to array-derived oligonucleotide libraries remains restricted to a few groups; and (v) the cell-free, in vitro assay that we used poorly captures biological context.

To overcome these limitations and facilitate the high-resolution dissection of mammalian enhancers, we developed an improved method, termed massively parallel functional dissection (MPFD) (**Figure 5.1**). We then used MPFD to assess the extent to which all possible single-nucleotide variants (SNVs) affect the activity of three mammalian enhancers that are active in the liver, designated here ALDOB (hg19:chr9:104195570-104195828) [103-105], ECR11 (hg19:chr2:169939082-169939701) [106] and LTV1 (mm9:chr7:29161443-29161744).

5.3 Method overview

To apply the MPFD method (**Figure 5.1**) to the three enhancers of interest, each enhancer was synthetically constructed by polymerase cycling assembly using overlapping oligonucleotides (~90 bp) containing a programmed level of degeneracy. At each position, 97% of molecules were expected to be synthesized correctly with 1% doping of each possible single-nucleotide substitution (**Appendix C**). Therefore, each synthetic enhancer molecule contained, on average, three mutations per 100bp, randomly distributed along its length. The population of molecules was inherently complex, both with respect to representation of all possible SNVs of the wild-type enhancer as well as myriad unique combinations. Because nearly all synthetic enhancers contained multiple substitutions, they are referred to here as 'enhancer haplotypes'.

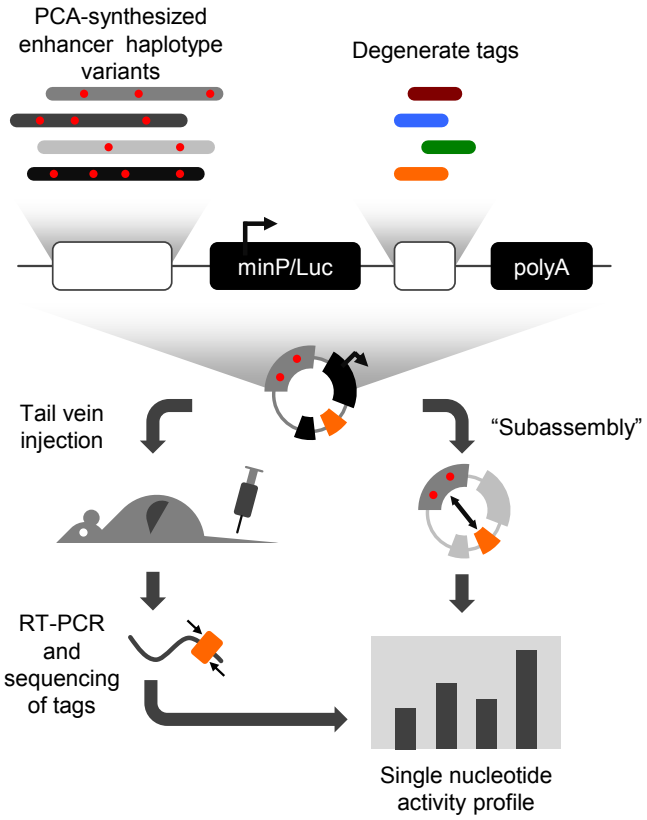


Figure 5.1 Overview of MPFD.

We used doped oligonucleotide synthesis and polymerase cycling assembly (PCA) to generate a highly complex library of enhancer haplotypes for each enhancer studied. On average, each enhancer haplotype diverged from wild type by ~2–3% (red circles represent mutations). These mutant enhancers, along with 20-bp degenerate tags, were cloned into an expression vector (pGL4.23) containing a minimal promoter driving transcription of luciferase (minP/Luc). We performed 'subassembly' on each library to determine the full sequence of each enhancer haplotype and to identify the 20-bp tag to which each haplotype was cloned in *cis*. Each library was then introduced into two mice through hydrodynamic tail vein injection, livers were harvested after 24 h and sequencing was performed to quantify abundance of transcribed 20-bp tags. These data were used to estimate the effect of each possible mutation on transcriptional activation.

Next, a library for assessing the activity of each enhancer haplotype was created by cloning the synthetic enhancers into a plasmid (Promega pGL4.23), which contains a minimal promoter upstream of the luciferase gene. In order to uniquely tag each enhancer haplotype, we cloned an oligonucleotide containing a 20-bp, fully degenerate subsequence to a separate site in the 3′ untranslated region (UTR) of the luciferase gene. The sequences of specific 20-bp tags cloned in *cis* with specific enhancer haplotypes were determined by massively parallel sequencing. As

the enhancer haplotypes were highly related sequences with lengths that exceeded the maximum read-length of the Illumina platform, we used tag-guided subassembly [107] to enable full-length, high-accuracy sequencing of individual enhancer haplotypes in association with their downstream tags. Each resulting library included >100,000 fully sequenced enhancer haplotypes, with nearly all containing multiple substitutions, and each associated with one or more unique tags.

The library was then subjected to what was effectively a massively parallel *in vivo* reporter assay. For the experiments described here, we used the hydrodynamic tail vein injection assay [106, 108] to assess *in vivo* enhancer activity in the mouse liver. Mice were euthanized 24 h after injection, at which time total RNA was extracted from each liver, followed by RT-PCR and massively parallel sequencing of cDNA from transcribed tags.

5.4 Results

We studied three mammalian enhancers identified by diverse methods (**Figure 5.2**). ALDOB (259 bp) is a human intronic enhancer of the aldose B gene [103-105]. ECR11 (620 bp) is a human enhancer located in an intron of dehydrogenase/reductase SDR family member 9 (DHRS9) [106]. LTV1 (302 bp) is a candidate mouse enhancer located on the 3′ side of zinc-finger protein 36 (Zfp36) (**Figure C.1a,b**). The activity of each wild-type enhancer was confirmed using a conventional hydrodynamic tail vein injection assay, in which luciferase activity in liver tissue was measured 24 h after injection (**Figure C.1c**).

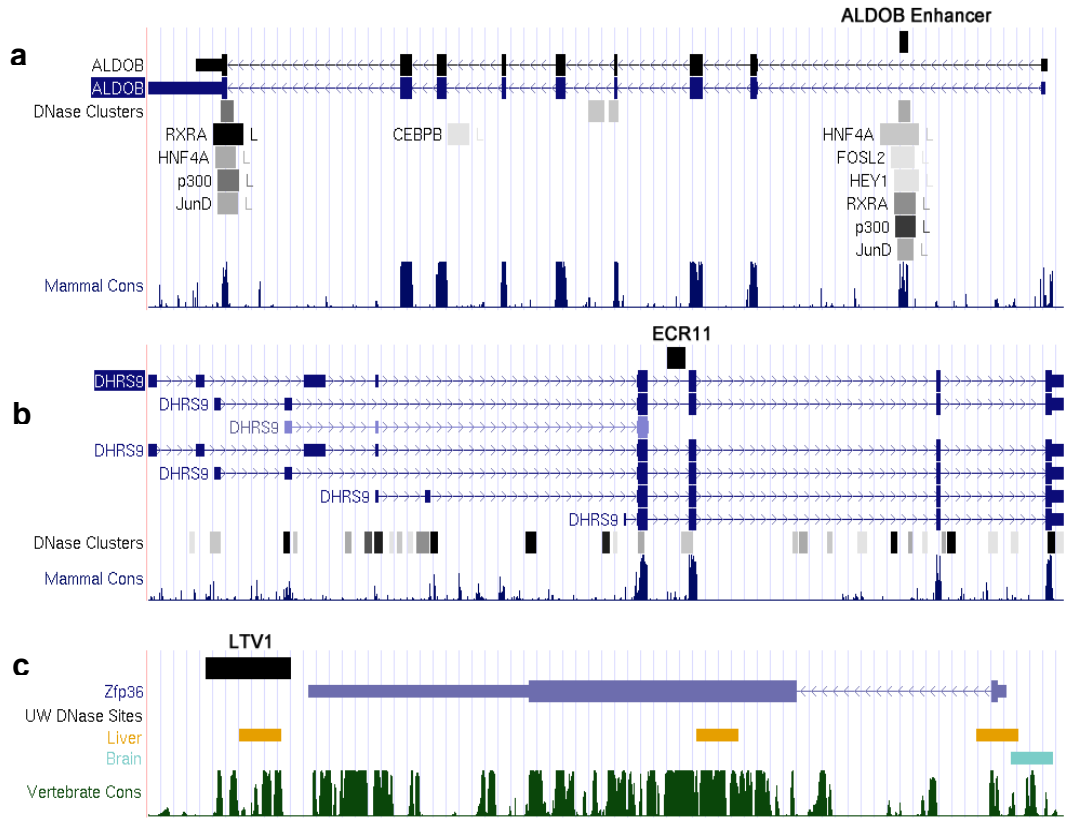


Figure 5.2 Schematics of candidate enhancer loci.

UCSC genome browser snapshots depicting the genomic locations of each of the three enhancers. Each enhancer was identified by diverse methods. (a) The human ALDOB enhancer is located in the first exon of ALDOB. It was identified and characterized by conventional transgenic assays [103-105], and overlaps extensively with heavily conserved clusters of ENCODE DNase hypersensitivity sites [48] and HepG2 ChIP-Seq peaks. (b) Human enhancer ECR11 is located in the fifth intron of DHRS9 in a region that overlaps with an ENCODE DNase hypersensitivity cluster on its 3` end and is conserved to mice. It was identified by comparative genomics and liver-specific transcription factor binding site analyses [106]. (c) Mouse enhancer LTV1 is located immediately downstream (3') of Zfp36 in a conserved region and overlaps with DNase hypersensitivity sites [48] from liver tissue but not brain. This enhancer was first identified by p300 ChIP-Seq on early adult mouse liver [L.A.P., unpublished data]. Using deletion experiments, we isolated a functionally equivalent 302bp core element that was used for mutagenesis (**Figure C. 1a,b**).

We applied MPFD to systematically dissect the functional consequences of all possible SNVs in these three enhancers. Sequencing with subassembly confirmed that the resulting libraries were complex, with a total of 641,135 distinct haplotypes associated with 1,186,696 tag sequences (**Table 5.1**). The observed number of mutations per haplotype approximated expectations, with ~2–3 substitutions per 100 bp (**Figure C.2**) and were well distributed (**Figure C.3**). All possible

substitution variants of each enhancer were represented in ≥42 uniquely tagged haplotypes. On average, each position was disrupted on ~4,000 distinct enhancer haplotypes. Furthermore, all possible pairs of positions were disrupted in ≥1 haplotype with the exception of a single pair of positions in LTV1.

Table 5.1 Enhancer haplotype library characteristics

Library	Number of haplotypes	Number of tags	% of possible substitutions in at least one haplotype	% of possible pairs of sites in at least one haplotype	Per-base mutation rate per haplotype (mean ± s.d.)
ALDOB	378,450	406,071	100% (777 of 777)	100% (33,411 of 33,411)	0.021 ± 0.010
ECR11	105,795	105,832	100% (1860 of 1860)	100% (191,890 of 191,890)	0.023 ± 0.006
LTV1 rep. 1	119,950	403,869	100% (906 of 906)	99.99% (45,449 of 45,451)	0.031 ± 0.010
LTV1 rep. 2	105,188	270,924	100% (906 of 906)	99.99% (45,449 of 45,451)	0.031 ± 0.010

For each library of enhancer haplotypes, we list the number of distinct haplotypes, the number of tags with which those distinct haplotypes are associated in *cis*, the percentage of possible single nucleotide substitutions that are present in at least one haplotype, the percentage of possible pairs of positions where both positions contain mutations together in at least one haplotype, and the per-base mutation rate in each library.

We introduced each library (one each for ALDOB and ECR11, and two independently constructed libraries for LTV1) into two mice by hydrodynamic tail vein injection (**Figure C.1d**). Total RNA from each mouse liver was split into several aliquots (ALDOB: N = 39; ECR11: N = 69; LTV1-1: N = 10; LTV1-2: N = 10), with each aliquot separately subjected to RT-PCR with primers flanking the 20-bp tag located in the 3′ UTR of the luciferase transcriptional cassette, and then to massively parallel sequencing on an Illumina GAIIx. Because target RNA was very scarce relative to cellular RNA, a modest number of target RNA molecules contributed to each

RT-PCR, leading to a complexity bottleneck. In other words, within each sequencing library, all reads corresponding to any single tag appeared to have been derived from amplification of a single RNA molecule. We therefore used the number of RNA aliquots in which a particular tag was observed, and not the total number of reads associated with a tag, as a measure of the relative transcriptional activity of its associated enhancer haplotype.

For each position in each enhancer, we constructed a linear model to assess the extent to which the presence of a mutation at that position is predictive of a change in the number of RNA aliquots in which an enhancer haplotype was observed. This is effectively a proxy for its effect on transcriptional activation, that is, 'effect size' (**Appendix C**). Specifically, we use the term 'effect size' to describe the log2-fold change in the predicted transcriptional activity, as measured by the number of RNA aliquots in which a tag-associated haplotype appeared, relative to the wild type. We first sought to assess reproducibility, so we calculated effect sizes separately for the two independently constructed LTV1 libraries (combining data from the two mice subjected to each of these libraries). For ALDOB and ECR11, we calculated effect sizes separately on the data from each mouse. For these two types of biological replicates, the effect sizes were highly correlated ($r = 0.96$ for LTV1, $r = 0.93$ for ALDOB, $r = 0.96$ for ECR11). Because reproducibility was high and to increase resolving power, we performed all subsequent analyses after combining data across mice for each enhancer haplotype library (data for one of the two LTV1 replicate libraries is shown in **Figure C.4**).

We next recalculated effect sizes in two ways (**Figure 5.3**, **Figure 5.4** and **Figure 5.5**). First, as for the reproducibility analysis, we constructed separate linear models for each position where mutational status was encoded as a single binary variable representing whether an enhancer haplotype was wild type or mutant at that position (**Figure 5.3a**, **Figure 5.4a** and **Figure 5.5a**). Second, we constructed separate multiple linear regression models for each position with three variables, each corresponding to a particular nucleotide substitution at that position (**Figure**

5.3b, **Figure 5.4b** and **Figure 5.5b**). For each enhancer, we also constructed a multiple linear regression model incorporating all positions. These models were also significantly predictive ($P < 0.01$) (**Appendix C: Supplementary Note** and **Table C.1**), and yielded effect-size profiles similar to models constructed independently for each position (**Figure. C.5**). As the coefficients from models constructed independently for each position are more naturally interpreted as position-specific effects, we used these models for subsequent analyses.

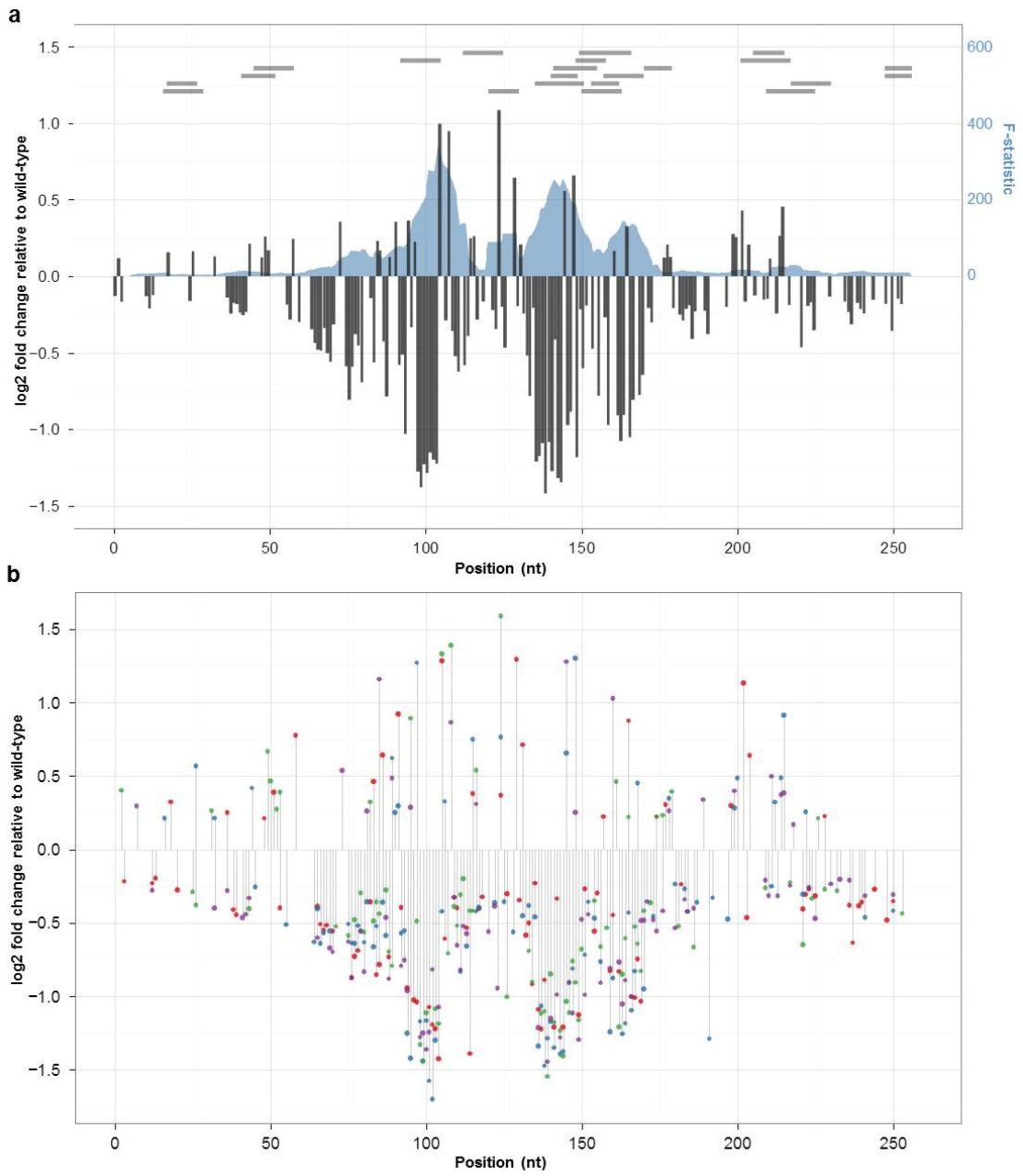


Figure 5.3 Effect size on transcriptional activity of all possible substitution mutations in ALODB enhancer.

Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for ALDOB (a and b, respectively), Effect sizes were estimated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: 39). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (Panel a, blue shadow, right axis). The locations of TFBS predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in a.

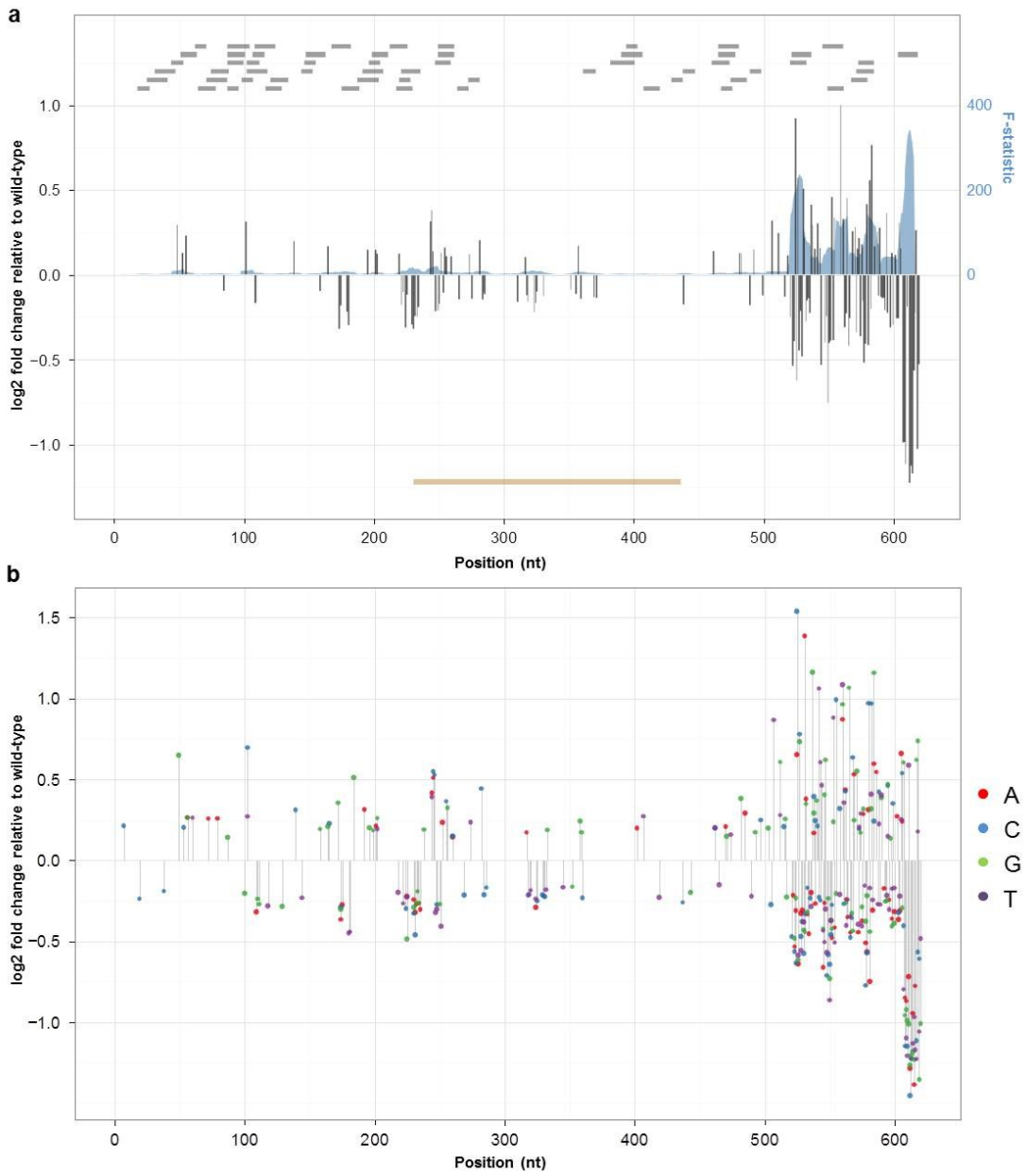


Figure 5.4 Effect size on transcriptional activity of all possible substitution mutations in ECR11 enhancer.

Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for ECR11 (a and b, respectively). Effect sizes were estimated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: 69). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (Panel a, blue shadow, right axis). The locations of TFBS predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in a. The location of a partial LINE element in ECR11 is shown as an orange bar at the bottom of a.

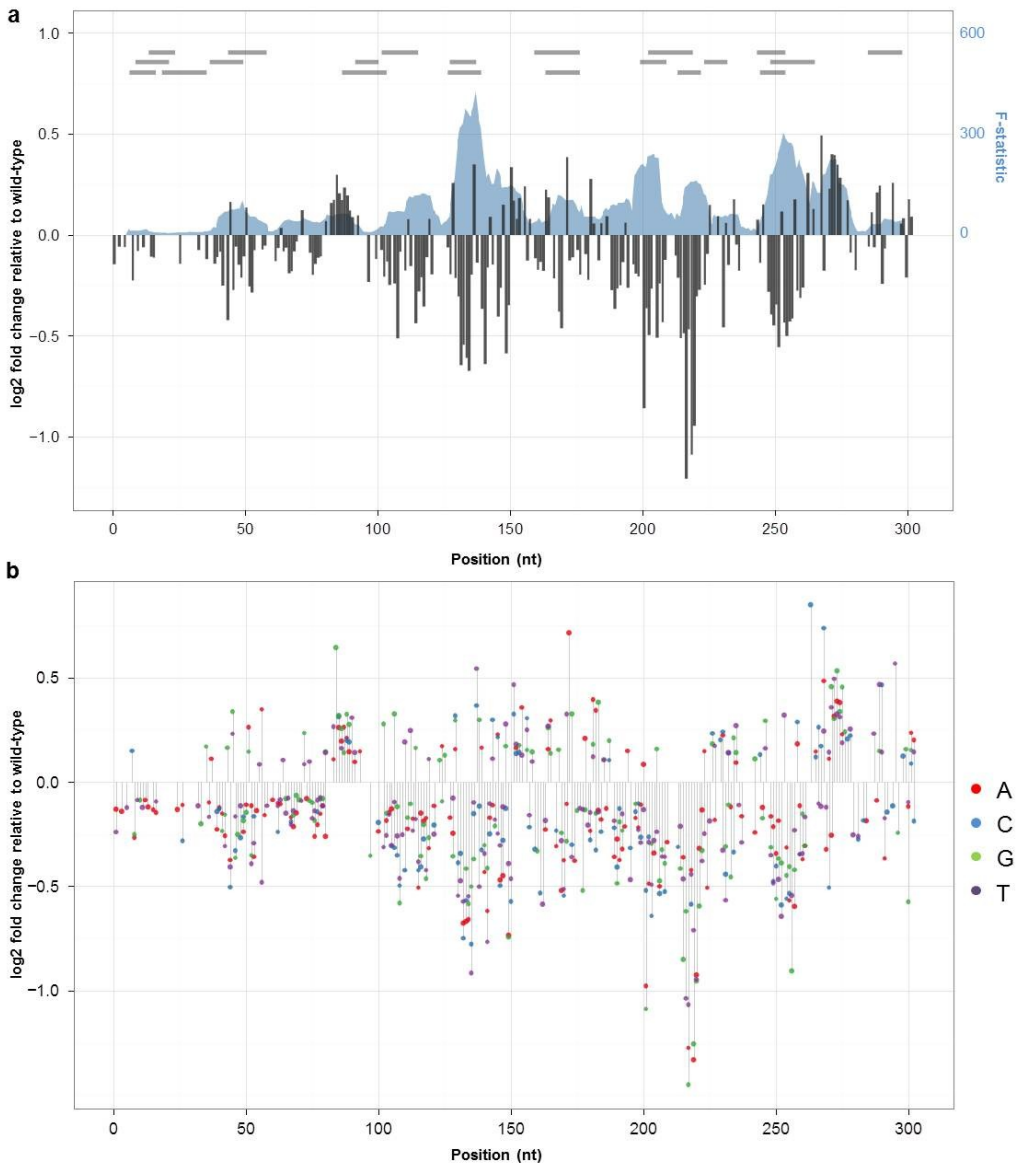


Figure 5.5 Effect size on transcriptional activity of all possible substitution mutations in LTV1 enhancer.

Estimated effect size of mutation at each position based on coefficients from univariate (gray columns, left axis) and trivariate (A:red, C:blue, G:green, T:purple) models are shown for LTV1 (a and b, respectively). Effect sizes were estimated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide (total number of aliquots sequenced per library: 10). Effect sizes are shown only for positions where model coefficients had associated P-values ≤ 0.01. We also used multiple linear regression with sets of ten adjacent positions as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (Panel a, blue shadow, right axis). The locations of TFBS predictions using the MATCH web server (with restriction to TFs present in liver) are shown as horizontal gray bars at the top of the plot in a.

To provide further validation, we also performed site-directed mutagenesis to individually introduce the six mutations in ALDOB that were predicted to have among the largest effect sizes (three increasing activity and three decreasing activity), and tested these individually using the hydrodynamic tail vein luciferase assay (**Figure 5.6**). Observed luciferase fold-changes were highly correlated with effect-size predictions from the models (R = 0.985).

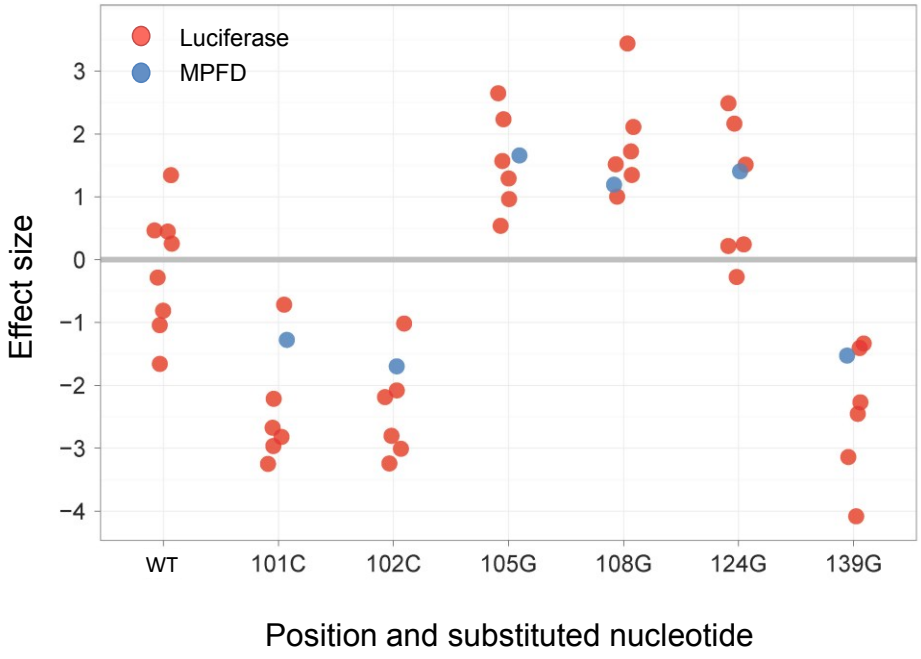


Figure 5.6 Validation of MPFD predictions using the hydrodynamic tail vein luciferase assay.

Shown are mutation effect sizes (log2 fold-change in expression of mutant compared to wild-type) for six single nucleotide ALDOB enhancer variants compared to the wild-type sequence. Each mutant was injected individually in at least six mice, and luciferase activity was measured at 24 hours post injection. Measurements from individual mice are shown. Effect sizes determined by the massively parallel reporter assay described here are shown for comparison. Effect sizes calculated via hydrodynamic tail vein luciferase assay were highly correlated with luciferase activity (R=0.985). On average, the observed fold-change in luciferase for individually tested mutations was ~25% greater in magnitude than the effect size predictions from our massively parallel reporter assay, although the predicted effect size based on the massively parallel reporter assay always fell within the range of effect sizes observed in the individual luciferase replicates. This may reflect differences between the assays or, alternatively, systematic but modest underestimation by our current methods.

5.4.1 Co-localization of high-impact positions and known TFBSs

Across each enhancer, the effect-size profiles exhibited spatial structure—that is, a clustering of positions with larger effect sizes. Positions separated by less than ~6 nucleotides had

significantly correlated effect sizes ($P < 0.01$) (**Figure C.6**). To further explore this, we performed multiple linear regression using mutational status at ten adjacent positions (that is, a binary variable for wild-type or mutant) at a time (**Appendix C**). These models remained predictive of transcriptional activity in a spatially resolved pattern (**Figure 5.3a**, **Figure 5.4a** and **Figure 5.5a**). We suspected that these clusters of correlated positions might represent transcription factor binding sites (TFBSs). Indeed, when we predict TFBSs [109] (**Figure 5.3a**, **Figure 5.4a**, **Figure 5.5a and Table C.2**), we observe striking overlap between predicted binding sites and clusters of highly predictive positions (**Figure 5.3a**, **Figure 5.4a** and **Figure 5.5a**). For example, a predicted binding site for HNF4 in the ALDOB enhancer (bases 94–105) coincides with a highly predictive localized model (**Figure 5.3a**). Furthermore, all mutations in this region had negative effects on activity, with the notable exception of mutations that increased identity with the consensus HNF4 binding site, which were activating (e.g., 95A→G and 105T→A) (**Figure 5.7a**). The same pattern was observed for other predicted sites as well, for example, a predicted HNF1 binding site at bases 135–148 in ALDOB (**Figure 5.7b**). Notably, independent experiments have established that these two transcription factors drive this element *in vivo* [105]. The spatial patterns may also reveal or refine broader features of activity—for example, the boundaries of functional elements. For example, in ECR11, computational prediction yielded a large number of predicted liver-specific TFBSs in the proximal 300 bases [106], but we observed that the highest impact SNVs were largely confined to the distal 160 bases (**Figure 5.4 and Figure C.7**).

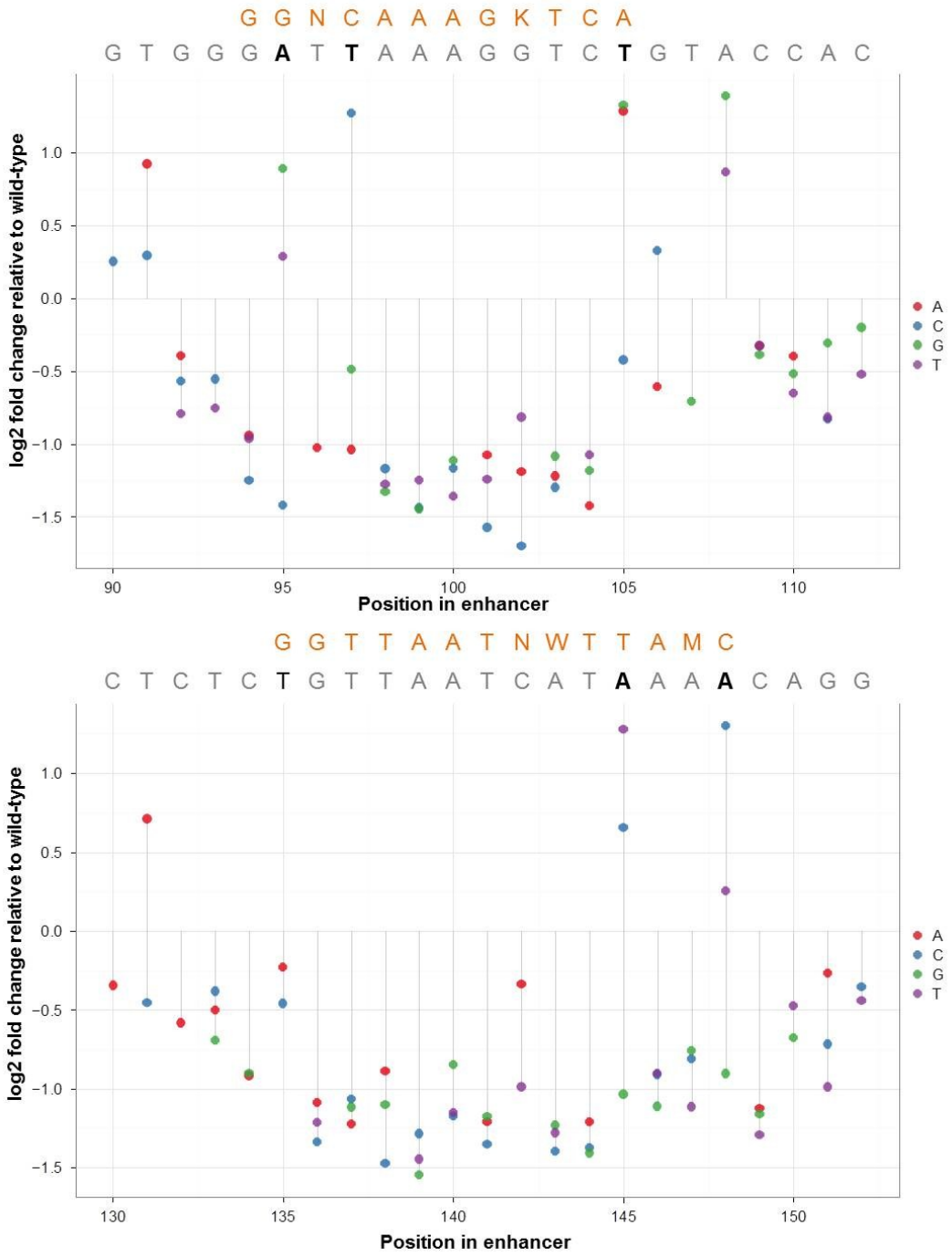


Figure 5.7 Profiles of mutation effect size in TFBSs.

For a predicted HNF4 site (positions 94–105) (a) and a predicted HNF1 site (positions 135–148) (b) in ALDOB, the effect size for each possible substitution, with the consensus TF binding sequence (orange) and the enhancer sequence (gray for consensus, black for nonconsensus) is plotted. Nonconsensus positions where rescue is observed after mutating to consensus are shown in boldface. HNF4 binding to the ALDOB enhancer region in human liver has been previously demonstrated [110], whereas *in vivo* occupancy data for HNF1 at this region is not yet available.

5.4.2 Relationship between evolutionary and functional constraint

Evolutionary constraint in noncoding, regulatory DNA has frequently served as a proxy for functional constraint [111-113]. However, recent studies have shown that many enhancers are evolving rapidly and that mammalian genomes contain large numbers of evolutionarily young, sometimes species-specific, enhancers [50, 110]. All three enhancers studied here are grossly conserved between human and mouse (**Figure 5.2**). We therefore investigated the relationship between functional constraint and evolutionary constraint at single-nucleotide resolution. For two of three enhancers, linear models, constructed to assess whether evolutionary constraint (that is, Genomic Evolutionary Rate Profiling (GERP) [114]) was predictive of functional constraint (that is, the absolute value of univariate model coefficients that we obtained), were significantly predictive with modest explanatory power (ALDOB: R^2 = 0.1232, P = 6.31e-9; LTV1: R^2 = 0.03911, P = 5.47e-4). For both enhancers, positions with the highest functional effect sizes were significantly associated with elevated evolutionary constraint scores ($P < 0.01$) (**Figure C.8**). However, not all positions with high GERP scores (≥4) had functional effect sizes in the top quartile for each enhancer (ALDOB: 33 of 61, 54%; ECR11: 5 of 25, 20%; LTV1: 0 positions with GERP≥4). These positions might have functions unrelated to the enhancer activity assayed here or might be of greater functional relevance in other contexts, for example, other tissues or developmental time points. On the other hand, a small set of highly functional positions, for example, most nucleotides within the distal-most C/EBP motif in ECR11, have lo GERP scores, consistent with lineage or species-specific activity.

5.4.3 Effect-size spectrum of single-nucleotide variants

A substantial proportion of polymorphisms and new mutations in mammalian genomes are single-nucleotide substitutions [11]. However, the functional dissection of regulatory elements has historically relied on introducing nested or scanning deletions, limiting the extent to which they inform the interpretation of naturally occurring variation. Our results provided an opportunity

to examine the distribution of effect sizes of SNVs in mammalian enhancers on the magnitude of transcriptional activation (**Figure 5.8**). Notably, we observed that the majority of SNVs result in only a modest change in transcription relative to the wild-type enhancer. Overall, <25% of the mutations alter transcriptional activity by >1.2-fold. Furthermore, only a few mutations, mostly in ALDOB, altered activity by a factor of >2. These results suggest that these enhancers are highly robust to the vast majority of potential SNVs. Further application of this method will be needed to assess whether this is a general property of mammalian enhancers.

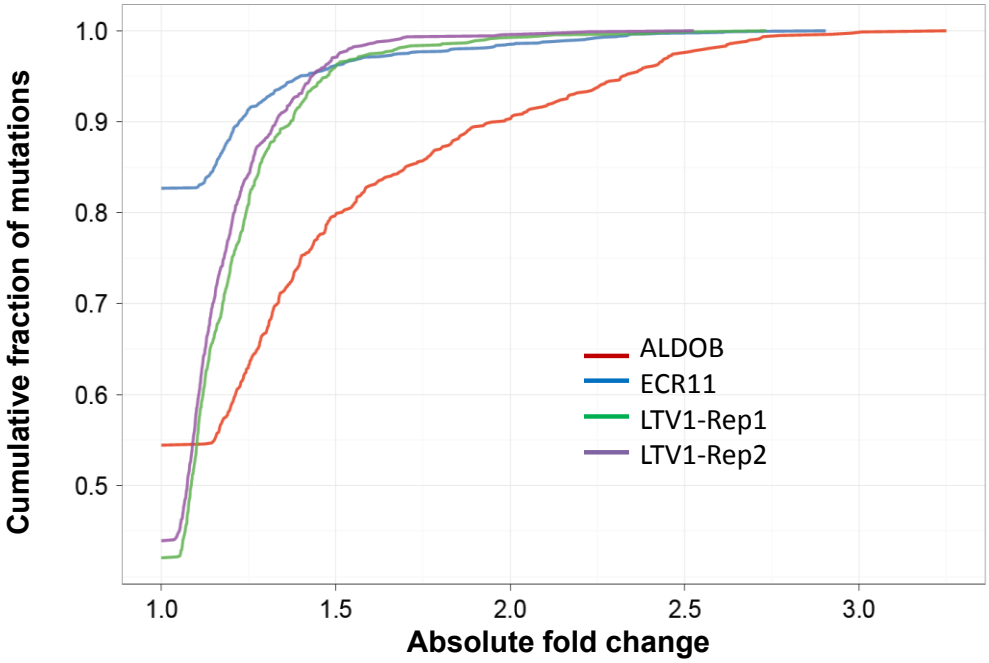


Figure 5.8 Distribution of effect sizes for all possible substitution mutations in three mammalian enhancers.

For the three enhancers studied (two replicate libraries for LTV1), the cumulative fraction of substitutions possessing a given effect size is expressed as the absolute value of the effect size of a given substitution. For example, across the three enhancers, between ~80% and ~95% of substitutions influence transcriptional activity by less than a factor of 1.5.

Perhaps as expected, the majority of functionally important mutations decreased activity (70% or 850/1,211). In general, only one substitution at a given position was activating, for example, substitutions that render a motif more like the consensus sequence (**Figure 5.7**). However, we observed some notable exceptions, including positions 83–93 and 272–278 in LTV1, where all or almost all substitutions were activating, consistent with binding of a repressive transcription factor. Positions 83–93 harbor a predicted binding site for NF-1, whereas there are no predicted sites in the immediate vicinity of positions 272–278, highlighting the value of experimental assessment of mutational impact.

5.4.4 Epistatic interactions

Finally, we sought to leverage the fact that our enhancer libraries contain multiple mutations on each haplotype to assess the degree of epistasis, or interaction, between positions in the enhancer. To obtain adequate power, we restricted our analysis to pairs of positions that were both mutated in at least 20 haplotypes. For each pair of positions that passed this cutoff, we built a multiple linear regression model consisting of three binary variables where the first two variables encoded mutation status (wild type or mutant) at each position independently and the third encoded whether both are mutant in a particular haplotype. With a false-discovery rate (FDR) cutoff of 0.05, we observed few pairs with a significant interaction term (ALDOB: 82 of 33,389, 0.25%; ECR11: 199 of 184,206, 0.10%; LTV1: 45 of 43,975, 0.10%), suggesting that the effects of multiple SNVs on the same haplotype are generally additive, or that our study lacked power to identify subtle interactions. Interacting pairs were significantly enriched for proximity (that is, pairs within 10 bp of each other versus pairs further apart, ALDOB: $P < 1e\text{-}4$; ECR11: $P < 1e\text{-}3$; LTV1: $P < 1e\text{-}4$), and we observed several different classes of interacting pairs with respect to the signs of the individual position effects and the sign of the interacting term (**Table C.3**).

5.5 Discussion

We developed a strategy to construct complex libraries of mammalian enhancers that contain all possible single-nucleotide substitutions and hundreds of thousands of distinct haplotypes. This method surpasses its predecessor [115] in terms of cost effectiveness, tunability, applicability to full-length regulatory elements and integration with an *in vivo* assay. We applied this method to empirically measure the distribution of effect sizes of all possible SNVs in three mammalian enhancers in an *in vivo* model. A key finding is that the vast majority of SNVs in these enhancers have highly reproducible yet remarkably modest effects on transcriptional activation. The distribution suggests that enhancers are highly robust to single-nucleotide changes. We also find that most combinations of single-nucleotide changes have additive effects on function. As expected, there is a clear relationship between the magnitude of functional impact and the location of predicted TFBSs, although not all predicted TFBSs are functional, and not all functional motifs are associated with predicted TFBSs. Similarly, evolutionary constraint, although clearly correlated with the magnitude of functional impact, does not predict it well on a nucleotide-by-nucleotide basis.

There remain some limitations of the method. First, although we exploited a mouse tail vein assay to assess function *in vivo*, the regulatory elements are episomal and therefore may not be subject to the same mechanisms governing elements residing on chromosomes. For example, because of the size of the synthetic construct, we were unable to assess the effects of mutations that may influence long-range interactions between regulatory elements. This might be addressed in part by transitioning to a lentiviral system, which would facilitate use in additional tissues and may also enable the application of other assays, for example, ChIP-Seq, to enhancer variant libraries. Furthermore, our results must also be considered specific to the minimal promoter used here until other promoter classes are tested. Second, we have assayed these enhancers in a single tissue and at a single time point. The activity profile of specific

positions could well be different in other tissues; this is the long-standing context problem [116]. Third, because of the scarcity of the target transcript relative to total RNA, we observed complexity bottlenecking, limiting the precision of our estimates of the effect size. This can be addressed by optimization of the RNA isolation step, for example, by hybridization-based enrichment. Fourth, we restricted our analysis to enhancer haplotypes containing only substitutions, as this was the dominant form of variation introduced during synthesis. To facilitate simultaneous dissection of the functional consequences of small insertions and deletions (indels), one could use reduced-fidelity oligonucleotide synthesis conditions, or polymerase cycling assembly with oligonucleotides containing programmed indels. Current efforts are directed at implementing these improvements, scaling this method to more enhancers and applying it to other classes of noncoding regulatory elements.

A fundamental goal of modern biology is to understand the human genome at single-nucleotide resolution. Single-nucleotide differences between genomes are causative for, or affect susceptibility to, a host of diseases, and single-nucleotide mutations are a primary source of raw material for evolution. We anticipate that the high-throughput, empirical measurement of the functional impact of single-nucleotide variants in enhancers will substantially facilitate the analysis of noncoding variants in genome-wide association study hits, the study of the mechanistic basis for enhancer activity and the engineering of enhancers with desired properties. Furthermore, with cost-effective, massively parallel methods for functional analysis, it may soon be realistic to empirically measure the functional effects of all possible single-nucleotide changes in all noncoding regulatory elements in the human genome.

5.6 Notes

Data availability:

Raw sequencing reads available in the NCBI Short Read Archive under accession number SRA049159. A full list of mutations interrogated for this work, along with the associated effect sizes and P values, are provided as Supplementary Data on the Nature Biotechnology website.

Acknowledgements:

We thank R. Qiu and J. Kitzman for advice on experimental strategies, and B. Cohen and D. Pe'er for helpful discussions. This work was supported in part by grants HG003988 from the National Human Genome Research Institute (L.A.P.), US National Institutes of Health (NIH) grant DP5OD009145 (D.M.W.), National Institute of General Medical Sciences (NIGMS) award number GM61390 (N.A.), National Institute of Child Health and Human Development (NICHD) grant number R01HD059862 (N.A.), the Pilot/Feasibility grant from the University of California, San Francisco Liver Center (P30 DK026743) (N.A.), AG039173 from the National Institute on Aging (J.B.H.) and a fellowship from the Achievement Rewards for College Scientists Foundation (J.B.H.). M.J.K. was supported in part by NIH Training grant T32 GM007175 and the Amgen Research Excellence in Bioengineering and Therapeutic Sciences Fellowship. R.P.S. is supported by a CIHR fellowship in the area of hepatology. Parts of the research were conducted at the E.O. Lawrence Berkeley National Laboratory and performed under Department of Energy Contract DE-AC02-05CH11231, University of California. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NICHD, NHGRI or the NIGMS.

Chapter 6 Future directions

In this dissertation, I presented massively parallel methods for *in vitro* as well as *in vivo* functional dissection of regulatory elements at single-nucleotide resolution. I demonstrated these methods on a small number of core promoters and distal enhancers, primarily as a proof-of-concept. Moving forward, these assays can now be applied in several different contexts and to larger sets of elements.

For example, the array-based method presented in Chapter 3 could be used to understand the architecture of different classes of promoters, such as the TATA-less promoters. The method could also be used to compare the relative strengths of all core promoters in the genome in a single experiment. Similarly, the functional effects of all known polymorphisms that lie in core promoter regions could be tested. Promoters of non-coding transcripts such as microRNAs, long non-coding RNAs, tRNAs and ribosomal RNAs could also be analyzed in this manner. Different distal and basal elements could be tested in pairs to understand the combinatorial regulatory logic behind regulation in a broader context.

In addition to promoters and enhancers, insulators form an important class of transcriptional regulatory elements. An enhancer can activate transcription from promoters located several kilobases away and in an orientation independent matter. However not all promoters within its vicinity might be the intended targets. Presence of an insulator between an enhancer and promoter sequence can protect promoters from being unintentional targets of a nearby enhancer, thus enforcing modularity of regulatory domains. Massively parallel functional dissection method presented in Chapter 5 could be used to systematically study the functional effects of variation in insulator sequence.

This method need not be limited to dissection of known regulatory elements. It can be used for massively parallel functional validation of candidate regulatory elements predicted by genome-wide discovery methods, both empirical, as well as computational.

Massively parallel strategies similar to the ones presented in this dissertation have recently been employed for dissection of transcriptional regulatory elements by several other groups [117-119].

In the first study, Kinney et al. targeted the -75 to -1 region of the *E. coli* lac promoter. A library with mutations in this region was generated using a single oligonucleotide synthesized with a programmed level of degeneracy. Plasmids containing these mutant *lac* promoters driving GFP expression were transformed into *E. coli*. Induced cells were partitioned into ten expression bins using Fluorescence Activated Cell Sorting (FACS). The mutant promoters in each FACS bin were sequenced using 454 pyrosequencing. The strength of each mutant promoter was quantified based on the expression bin in which it was preferentially observed.

The second group [117] used array-derived oligos, similar to the method presented in Chapter 3 [115] to generate a library of enhancer variants, each linked to a unique programmed tag. The library was cloned into a plasmid and transfected into human cells. Activity of each enhancer variant was quantified by sequencing the RNA-derived tags. They applied this technique to two inducible human promoters and observed distinct activity profiles for basal versus induced states.

The third group [118] also used array-derived oligos to obtain a library of yeast promoter variants, but instead of a tag-based readout, they used FACS coupled with massively parallel sequencing, similar to Kinney et al [119]. In addition to simple scanning mutations, their library included synthetic constructs to analyze various aspects of promoter grammar such as positioning of various TFBSs relative to the TSS.

While each of these studies including the ones presented in this dissertation have their own strengths and limitations, they serve to establish the feasibility of regulatory elements analysis using massively parallel methods and pave the way for future improvements towards more realistic assays. An obvious step forward would be to move away from assaying regulatory elements in an extra-chromosomal context (i.e. on transiently transfected plasmids) and instead assay them after stable integration at fixed or random locations in the genome. This will also allow interrogation of the effect of sequence variation on other processes, such as chromatin modifications.

In addition to dissection of transcriptional regulatory elements, massively parallel methods are also being developed and applied to study other functional elements in the genome. For example, Ke S et al. [120] tested all possible hexamers for their potential to function as exonic splicing elements. Pitt et. al. [121] used successive rounds of functional selection coupled with massively parallel sequencing to quantify the level of activity of each individual member from a mutagenized pool of an RNA ligase riboezyme. Fowler et. al. [122] developed a method called “deep mutational scanning” to evaluate the effect of every possible amino acid substitution in a protein domain on its ability to bind its cognate peptide. They applied this method to systematically quantify binding efficiencies of more than half a million sequence variants of the human WW protein domain and generated a high resolution map of its mutational preference. Recently, this method also was applied to optimize the affinity, specificity and function of two computationally designed inhibitors against H1N1 influenza hemagglutinin [123].

While the ability to read and interpret the genome is crucial, the next stage of the challenge is to be able to design novel functional elements from scratch. Such attempts have already been made [124-127]. Availability of massively parallel tools to rapidly screen large numbers of candidate synthetic elements will make such regulatory element engineering more efficient and feasible.

In conclusion, the studies described in this section together with the methods presented in this dissertation provide further validation of the potential of massively parallel methods for functional analysis. I anticipate that the success of these proof-concept experiments, together with concerted efforts such as the next phase of the ENCODE project will motivate accelerated development of such technologies and their application to several aspects of genome biology that have yet to be fully understood, moving us closer to the dream of a fully interpretable genome.

Appendix A. Supplementary material for Chapter 3

Supplementary Tables

Table A.1: Oligonucleotide sequences used for the bacteriophage promoters

PCR and Sequencing Primers		
Name		**Sequence**
BULK_AMP_FWD		TGCCTAGGACCGGATCAACT
BULK_AMP_REV		GAGCTTCGGTTCACGCAATG
RT_PCR_SP6_FWD		AATGATACGGCGACCACCGA TAGATAGTCTTCTCATTGA
RT_PCR_T3_FWD		AATGATACGGCGACCACCGA ATTCTGGAAGCTGAGGCATT
RT_PCR_T7_FWD		AATGATACGGCGACCACCGA TAAAGGTGTAGTTGCTGCTG
RT_PCR_COMM_REV		GCAGAAGACGGCATACGA GAGCTTCGGTTCACGCAATG
SEQ_SP6		ATACGGCGACCACCGAGTAGATAGTCTTCTCATTGA
SEQ_T3		GACCACCGAATTCTGGAAGCTGAGGCATT
SEQ_T7		ACCACCGATAAAGGTGTAGTTGCTGCTG
200-mer oligonucleotides		
Structure of the 200-nt oligo: PCR primer (15) + promoter (35) + spacer (115) + tag (20) + PCR primer (15) Variable bases are indicated by X's (promoter) or N's (tag)		
SP6	**Template**	AGGACCGGATCAACTXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX TCCTCCCTAACCTATCAACTTGATTTATAAGGAGATTATAATACATGTCT ACGCCGAACAACTTGACCAACGTTGCCGTTTCCGCTTCCGGGGAAGTAGA TAGTCTTCTCATTGANNNNNNNNNNNNNNNNNNNCATTGCGTGAACCGA
	WT promoter	TTGCCTATTTAGGTGACACTATAGAAGGGAGGTAG
	Tag seed	GAAGTTCAACGGTAAGGTCA
T3	**Template**	AGGACCGGATCAACTXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX TAGATACGAAGGGGGGGGGGGGGGTTAAAGCATTATGTATATTACAAAG TGTTTACAAAGCCACGCTGACAGCTTTAAGCCGTCCATAGAGGACATTCT GGAAGCTGAGGCATTNNNNNNNNNNNNNNNNNNNCATTGCGTGAACCGA
	WT promoter	GCGGTGAATTAACCCTCACTAAAGGGAGACACTAA
	Tag seed	GGGTGTCGAACCTAAAGTAA
T7	**Template**	AGGACCGGATCAACTXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX CGGTTTCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACA TATGGCTAGCATGACTGGTGGACAGCAAATGGGTACTAACCAAGGTAAAG GTGTAGTTGCTGCTGNNNNNNNNNNNNNNNNNNNCATTGCGTGAACCGA
	WT promoter	CGAAATTAATACGACTCACTATAGGGAGACCACAA
	Tag seed	GAGATAAACTGGCGTTGTTC

Table A.2: Oligonucleotide sequences used for the Pol II promoters

PCR and Sequencing Primers		
Name		**Sequence**
BULK_AMP_FWD		TGCCTAGGACCGGATCAACT
BULK_AMP_REV		GAGCTTCGGTTCACGCAATG
RT_PCR_CMV_FWD		AATGATACGGCGACCACCGA TTTTGACCTCCATAGAAGAC
RT_PCR_HBB_FWD		AATGATACGGCGACCACCGA AGCAACCTCAAACAGACACC
RT_PCR_S100A4_FWD		AATGATACGGCGACCACCGA CAGCGCTTCTTCTTTCTTGG
RT_PCR_COMM_REV		CAAGCAGAAGACGGCATACGA GAGCTTCGGTTCACGCAATG
SEQ_CMV		GGCGACCACCGA TTTTGACCTCCATAGAAGAC
SEQ_HBB		GACCACCGA AGCAACCTCAAACAGACACC
SEQ_S100A4		GACCACCGA CAGCGCTTCTTCTTTCTTGG
200-mer oligonucleotides		
Structure of the 200 base oligo: PCR primer (15) + promoter (150: -100 to +50) + barcode (20) + PCR primer (15) Variable bases are indicated by X's (promoter -45 to +25) or N's (tag)		
CMV	**Template**	AGGACCGGATCAACTGACTTTCCAAAATGTCGTAATAACCCCGCCCCGTT GACGCAAATGGGCGGTAGGCXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXCGCTGTTTTG ACCTCCATAGAAGACNNNNNNNNNNNNNNNNNNNCATTGCGTGAACCGA
	Native promoter	GTGTACGGTGGGAGGTCTATATAGCAGAGCTCGTTTAGTGAACCGTCAGA TCGCCTGGAGACGCCATCCA
	Barcode seed	ACCGGGACCGATCCAGCCTC
HBB	**Template**	AGGACCGGATCAACTTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTC CCAGGAGCAGGGAGGGCAGGXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTCACTAGCAA CCTCAAACAGACACCNNNNNNNNNNNNNNNNNNNCATTGCGTGAACCGA
	Native promoter	AGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATT TGCTTCTGACACAACTGTGT
	Barcode seed	ATGGTGCATCTGACTCCTGA
S100A4	**Template**	AGGACCGGATCAACTATCAGCCCACAGCAGGAAGGCAGTATCCGCTCTCC CCTGTCCCCTGCTATGGGCAXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXCTCCTCAGCG CTTCTTCTTTCTTGGNNNNNNNNNNNNNNNNNNNCATTGCGTGAACCGA
	Native promoter	GGGCCTGGCTGGGGTATAAATAGGTCAGACCTCTGGGCCGTCCCCATTCT TCCCCTCTCTACAACCCTCT
	Barcode seed	TTTGGTGAGTTGTGTTGGCC

Supplementary Figures

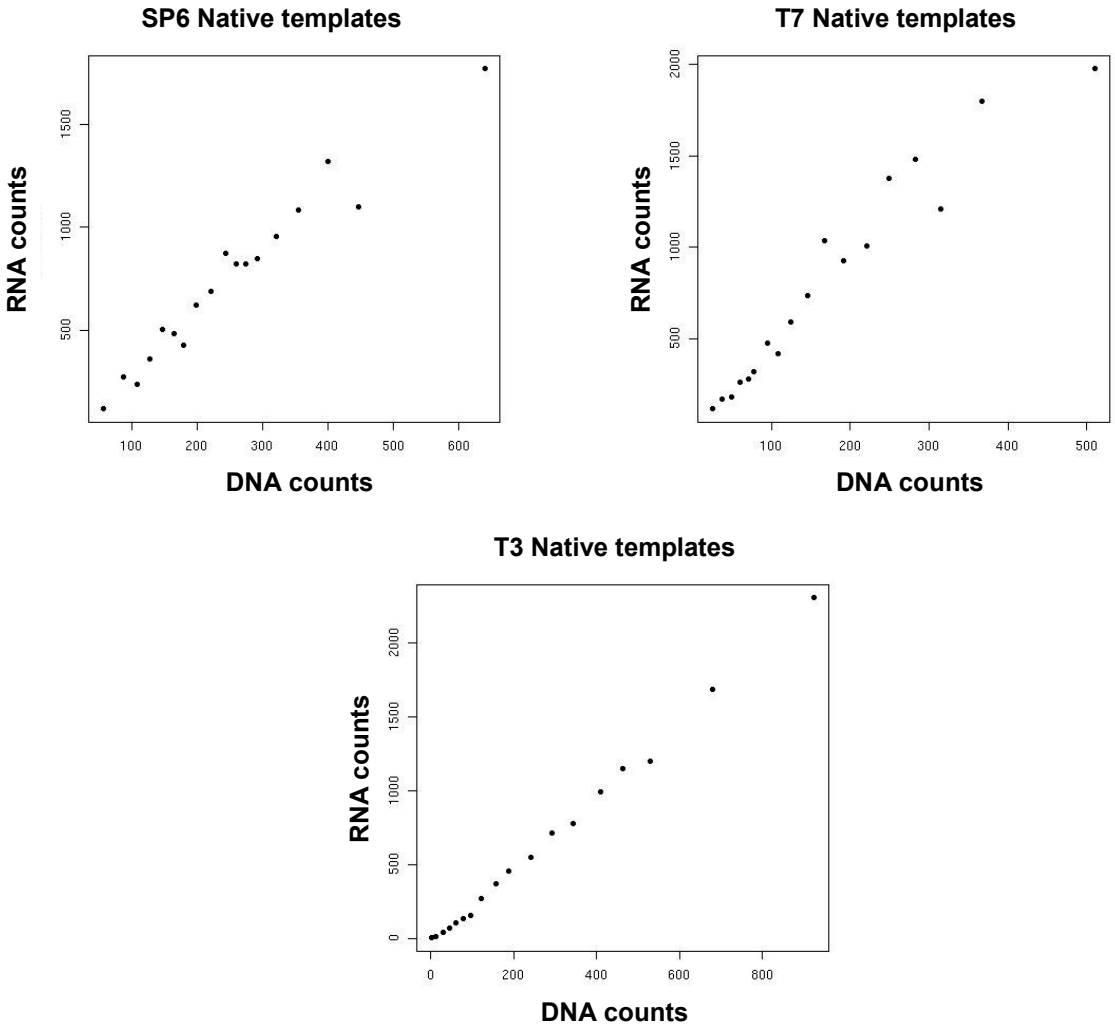


Figure A.1: Correlation between the number of times each tag was observed in the DNA and RNA-derived sequencing libraries for wild-type (native) promoter templates for SP6, T7 and T3.

For each of the three promoters, the native promoter templates (n=270) were rank-ordered and grouped into uniformly sized bins (bin size=15) based on their DNA counts. The mean DNA versus mean RNA counts were plotted for each bin. The relationship between the DNA and RNA tag counts appears reasonably linear across the full range, implying that individual concentrations of synthetic promoter templates are generally within the range of linear relationship with transcriptional efficiency.

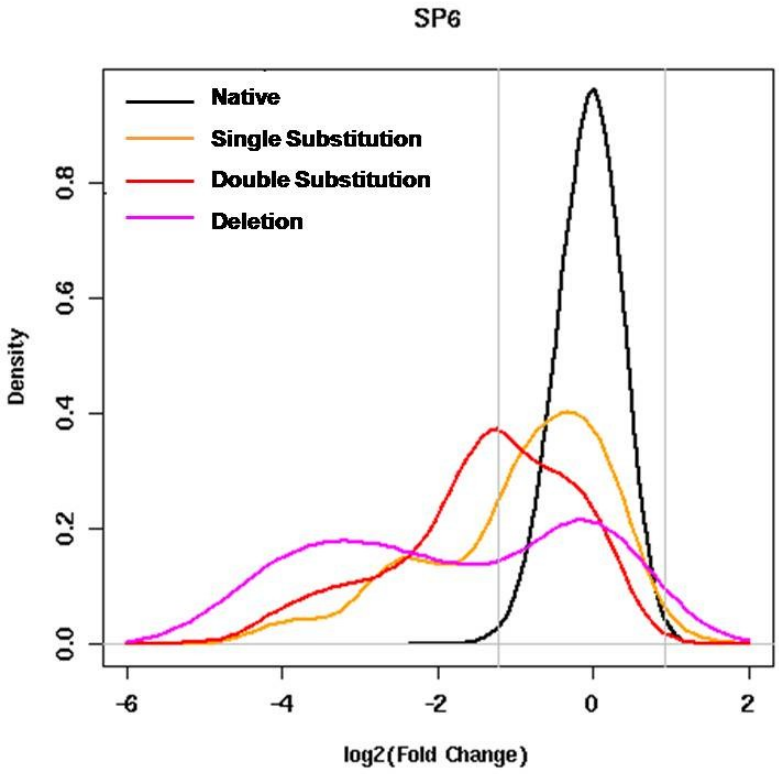


Figure A.2: Distribution of effect-sizes for various categories of mutations.

Fold change distributions for various populations of mutant promoter templates (single base substitutions in orange, double base substitutions in red and single base deletions in pink) contrasted against the empirical null distribution created from the native promoters (black). The grey lines represent significance cutoffs ($p<0.01$) determined on the basis of the empirical null. The empirical null distribution was generated by 100,000 samplings of 6 data-points selected randomly from the set of 270 barcodes associated with each native promoter sequence.

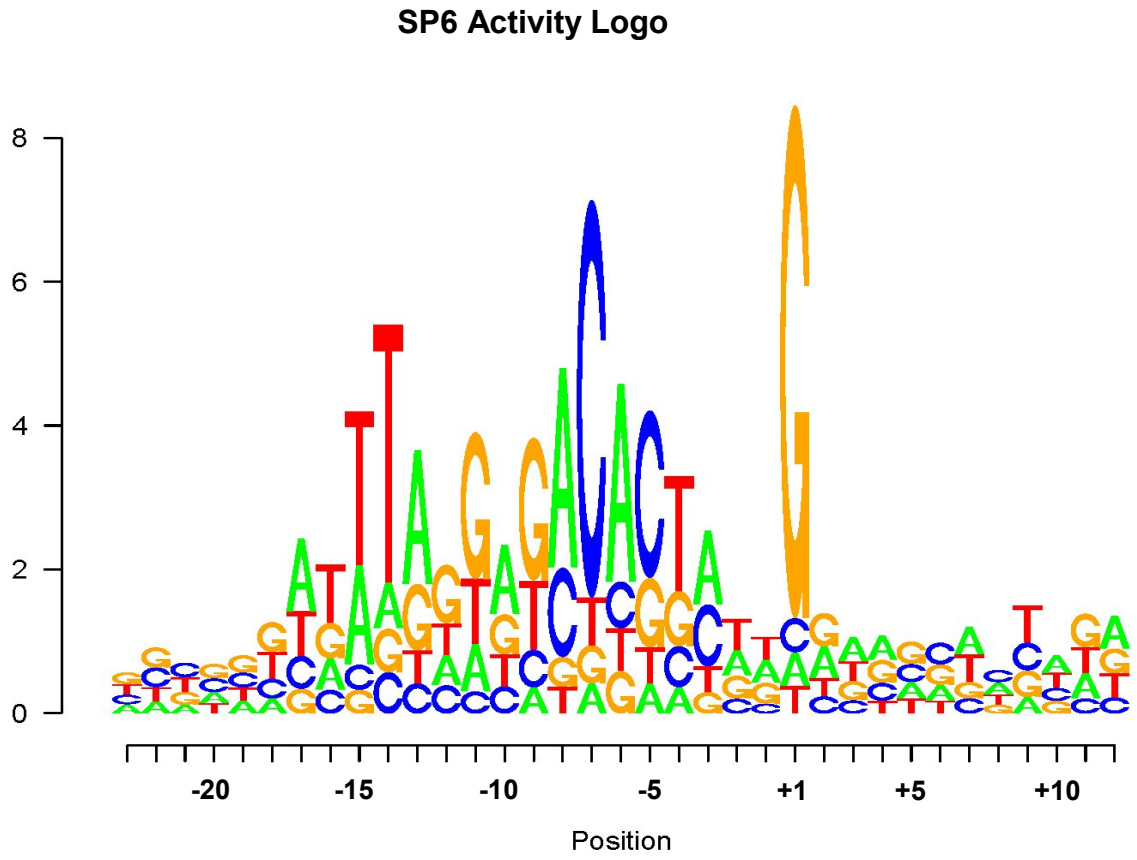


Figure A.3: Activity logo for SP6 promoter mutants, generated as per the method in Shin et al. [82].

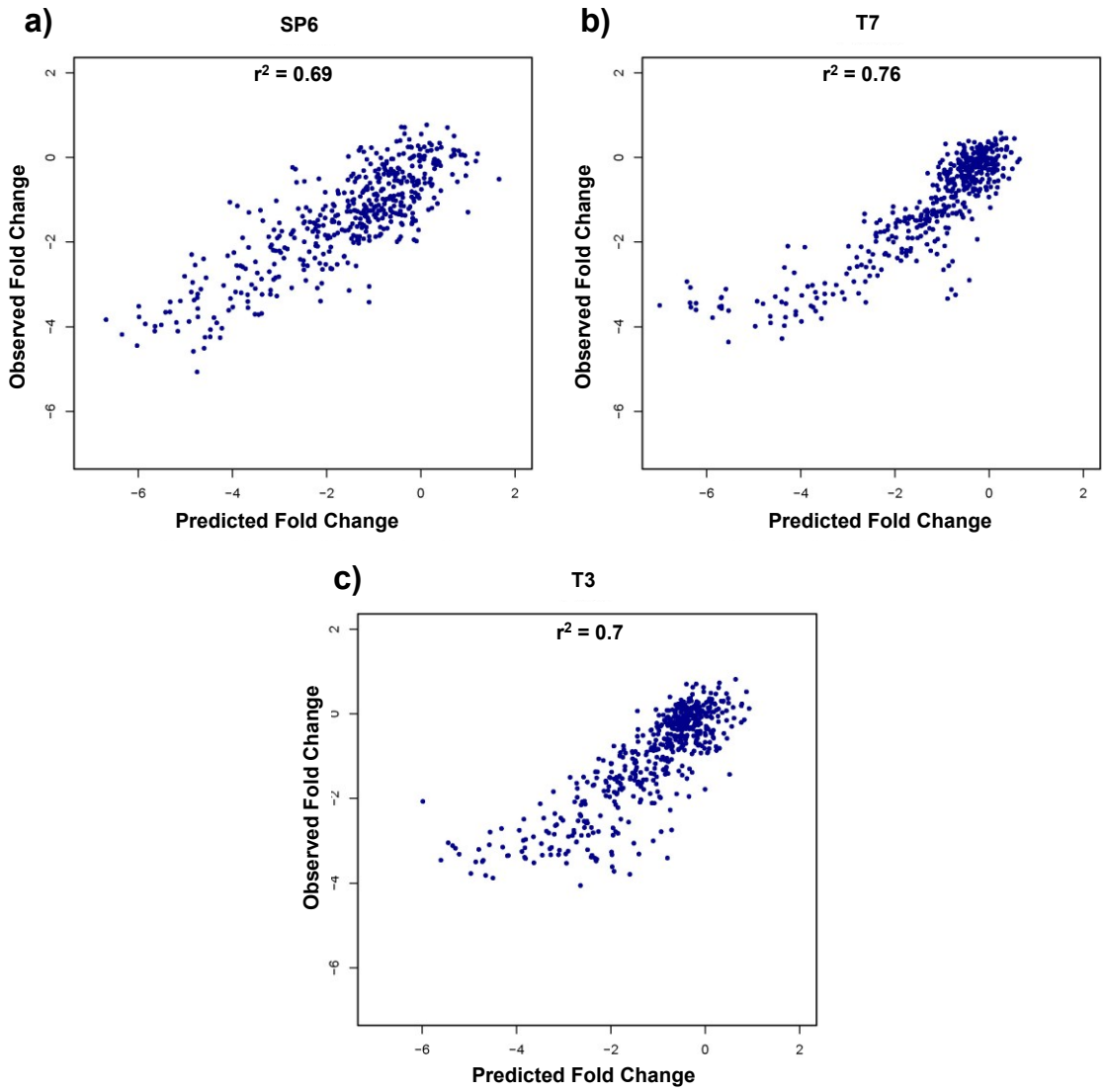


Figure A.4: Observed and predicted log2 (fold change) values for the SP6 (a), T7 (b) and T3 (c) double mutants. Predicted values were calculated as the sum of the log2(fold change) of the two single mutants.

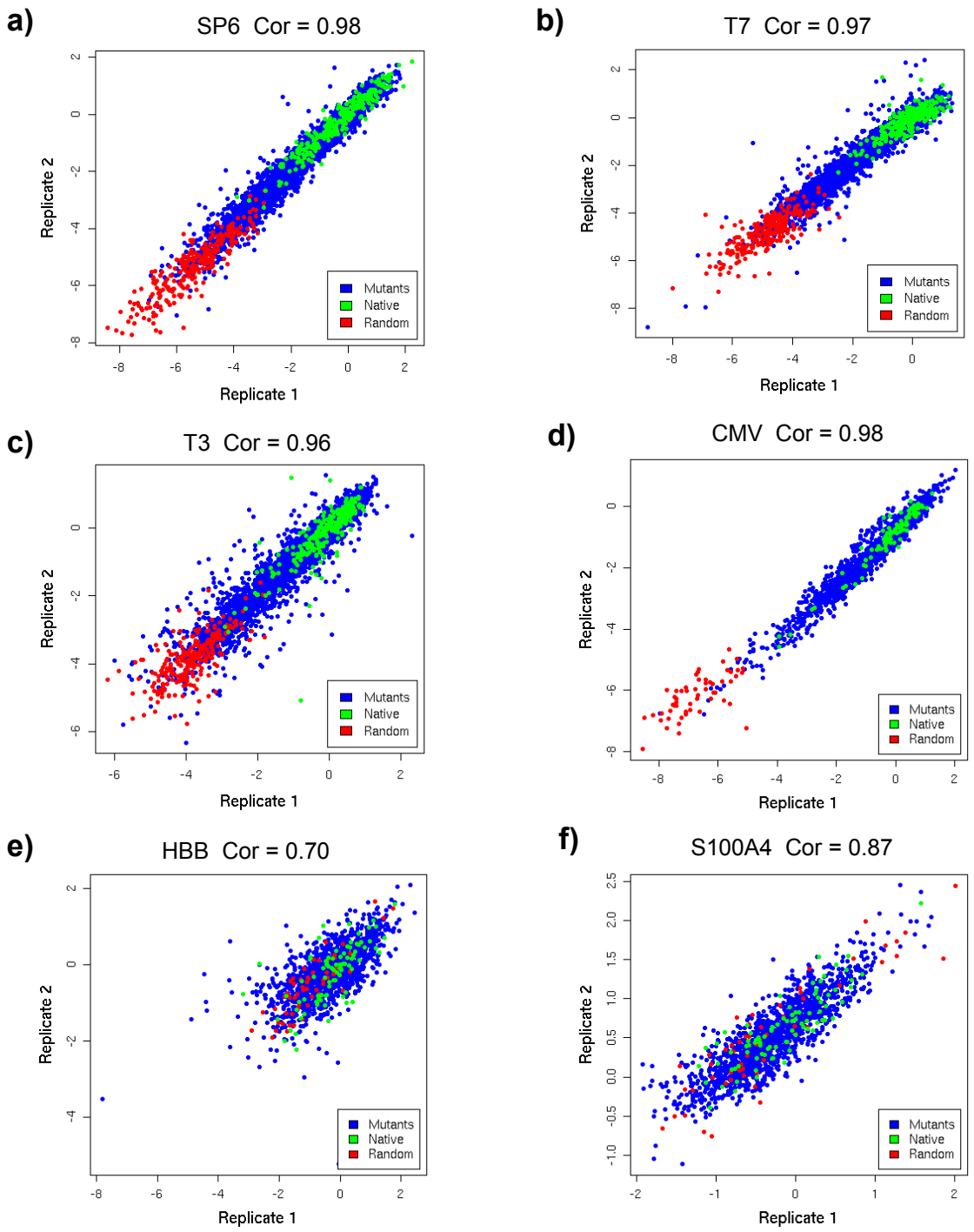


Figure A.5: Correlation between fold-change values (with respect to mean value of all native promoter variants) for all promoter templates in replicate experiments for SP6 (a), T7 (b), T3 (c), CMV (d), HBB (e), and S100A4 (f). The mutant templates are shown in blue, native templates are in green and the random promoter sequences are in red. All correlations were highly significant with a p-value < 2.2e-16.

Supplementary Note

In a typical initiator element (PyPyAN(A/T)PyPy), transcription is believed to initiate at the A nucleotide. As per this expectation, the +1 site for the CMV promoter is possibly mis-annotated and is likely to be at what we call the +3 site. The co-ordinates we use (in the text and **Figure 3.5**) correspond to the annotated start site as per the GenBank record for CMV IE promoter (Towne Strain). We are retaining these coordinates to be consistent with this source.

Supplementary Methods

Design of promoter templates

The overall layout of the bacteriophage phage promoter templates is shown in **Figure 3.1b**. The promoter and downstream flanking sequence (-23 to +147) for SP6, T3 and T7 were extracted from the whole genome sequences of these phages present in GenBank: SP6 - GI:31880044 (22,407-22,576), T3 - GI:17384270 (25,451-25,620), T7 - GI:431187 (22,881-23,050). Sequences are specified in **Table A.1**.

Choice of mutants:

Every single base substitution at all sites from -23 to +12 for each of SP6, T3 and T7 was included, in addition to every single-base deletion in that span. Double mutants were chosen on the basis of a pilot experiment consisting of only the single base substitution and deletion mutants. To ensure a good representation of different representative combinations in terms of both severity and position, we divided the promoter into sub-regions. Double mutants were designed to include pairs of point mutations with strong effect, one (or both) from each of the defined regions, selecting all possible combinations of regions. The number of double mutants chosen for each of SP6, T3 and T7 is specified in **Table 3.1**. Random promoters to be used as negative controls were created by emitting a string of nucleotides, all four nucleotides being equally likely.

Generation of tag sequences:

To minimize any effects that an artificial sequence might have on transcription, all tags were designed to be as similar as possible to the wild-type downstream sequence, but different enough from each other to be unambiguously identified after sequencing in spite of any sequencing errors. The tag was located 128 bases downstream of the transcription start site. We used the wild-type sequence in this position as the starting point to generate our set of tags. For each of the three promoters (SP6, T3 and T7), 20 bases of wild-type sequence (from +128 to +147) was used as the seed. Slight changes were made to this sequence using a custom Perl script to generate a set that a) differed from the original sequence at exactly three positions, and b) differed from every other barcode in the set at a minimum of two positions (and a maximum of six positions as a consequence of the first constraint). Since we planned to multiplex the synthesis of T3, T7 and SP6 templates, an additional check was performed to

make sure that combined set of 12,972 tags still satisfied the condition of differing from each other at two or more positions.

Synthesis of promoter templates

12,972 custom 200 base long oligonucleotides were ordered from Agilent Technologies. They were synthesized in parallel on a microarray slide using their in-situ ink-jet based protocol and then cleaved from the surface [83]. The resulting complex oligonucleotide library was shipped to us lypholized in a single tube and was subsequently re-suspended in EB (Qiagen) to a stock concentration of 100nM.

Amplification of promoter library

10% of the100nM library template was amplified in a 1ml bulk PCR reaction with 1X Phusion buffer (Finnzymes), 0.2mM dNTP mixture, 0.5uM each of primers BULK_AMP_FWD and BULK_AMP_REV, and 0.2X Sybr Green II. The reaction was assembled on ice. Thermal cycling was performed on Bio-Rad MiniOpticon Real-Time PCR system under the following conditions: 98°C for 30 seconds (for activation of the Hot Start Phusion enzyme), 20 cycles at 98°C for 10s, 54°C for 30s and 72°C for 15s. PCR products were purified using QIAquick PCR Purification Kit (Qiagen), and resuspended in EB.

***In vitro* transcription (IVT)**

A portion of the amplified promoter library was subjected to *in vitro* transcription using MAXIScript kit (Ambion). Separate reactions were performed for each of T3, T7 and SP6. The reactions were assembled as per manufacturer's protocol under strict RNase-free environment, using 500ng of template DNA instead of 1ug stated in the protocol and incubated at 37°C for one hour. After the incubation, each reaction was treated with 1uL of TURBO DNase I (Ambion) at 37°C for 1.5 hours to destroy template DNA. The mixture was then heated to 75°C for 10 minutes in the presence of 1uL of 0.5M EDTA to inactivate the DNase.

Gel Purification of RNA

The IVT product was mixed with Gel Loading Buffer II (Ambion) and denatured by heating to 95°C for 2 minutes before running it on a 6% polyacryamide urea gel at 150V for one hour. The band corresponding to the expected size of the transcript (162 bases) was excised, re-suspended in 200ul TE, and purified using 0.2um Nanosep column (VWR) followed by ethanol precipitation with 7.5M Ammonium Acetate and 100% ethanol. 0.05 mg/ml glycogen was added to the mixture and it was left overnight at -80°C. The pellets were collected the following day by

centrifuging at full speed for 30 minutes at 4°C. The pellets were washed twice with 75% ethanol and dried in Speed Vac at 30°C for around 5 minutes, and re-suspended in 10ul of EB.

RT-PCR

Gel-purified transcripts were reverse transcribed to cDNA and amplified by PCR using the Qiagen One-Step Kit. The reaction was assembled on ice in a 25uL total volume with the following reagents: 1X Qiagen One-Step RT-PCR buffer, 400uM of each dNTP, 0.6uM of relevant forward primer (RT_PCR_SP6_FWD, RT_PCR_T7_FWD or RT_PCR_T3_FWD), 0.6uM of reverse primer RT_PCR_COMM_REV, 0.2x Sybr Green II and 2 ul of RNA template (T3, T7 or SP6). Thermal cycling was performed on Bio-Rad MiniOpticon Real-Time PCR system with the following program: 50°C for 30m (reverse transcription), 95°C for 15m (inactivation of reverse transcriptase and heat-activation of the DNA polymerase), then 30 cycles of 94°C for 30s, 58°C for 30s and 72°C for 30s. Each sample was monitored and extracted from the PCR machine when the fluorescence began to plateau. The cDNA products were purified by QIAquick PCR Purification Kit (Qiagen) in 30ul EB. The primers used for the RT-PCR were hybrid Solexa adapters, thus the cDNA library obtained at the end of this step was sequencing-ready, eliminating the need for a separate sequencing library construction step.

Control reactions to detect DNA template contamination in the RNA sample:

The exact same RT-PCR reactions were run, but the RNA template was added to the tubes after the inactivation of reverse transcriptase and activation of DNA polymerase at the end of the RT-step (i.e. immediately after the 95°C for 15m step). No amplification was observed for this set of controls implying the absence of any significant amount of DNA template contamination.

PCR of DNA templates for normalization

A portion of the amplified promoter library was subjected to PCR using the same primers used for the RT-PCR reaction. The purpose of this set was to allow for the quantification of the bias in representation of different oligonucleotides in the library so that the counts of RNA transcripts obtained from each DNA template could be normalized to correct for this bias. The PCR was performed using the same master mixture used for the RT-PCR of RNA as described above, except for the use of 2ul of the DNA template instead of the RNA sample. Thermal cycling conditions were identical to the RT-PCR reaction except for the omission of the initial RT step of 50°C for 30m. PCR cycling was stopped right before the signal reached the plateau. The PCR products were purified by QIAquick PCR Purification Kit (Qiagen) in 30ul EB to yield a sequencing-ready library. Separate reactions were performed for each of T3, T7 and SP6 since they needed different forward primers complementary to their own spacer sequences.

Sequencing

T3, T7 and SP6 RT-PCR products were pooled and sequenced on a single lane on Illumina GA II using sequencing primers designed to read into the tag sequence. Pooling was possible because the tags were designed to be unique across the three promoter types as well. The PCR-amplified template DNA for T3, T7 and SP6 was pooled and sequenced on another lane. The RT-PCR products and the PCR-amplified template DNA from the replicate experiment were similarly sequenced on another two lanes of the flow-cell.

Analysis of sequenced reads

The first 20 bases of each read, representing the tag, were extracted and. the number of instances of each tag was counted, for both the DNA and RNA-derived lanes. To avoid spurious results due to small numbers, tags with less than five reads in the DNA lane (272 out of 12,972 in replicate 1, and 184 out of 12,972 in replicate 2) were discarded.

Normalization:

To ensure that all differences observed in the RNA-derived counts of different templates were a result of differences in their promoter activity and not due to non-uniformity in the abundances of templates in the original synthetic library itself, the count of each tag from the RNA-derived lane was normalized by dividing with the corresponding count from the DNA lane. This strategy was based on the assumption that the concentrations of all the individual DNA templates were within the range of linear relationship with transcriptional efficiency. To make sure that this assumption was valid, we took advantage of the 270 tag variants associated with each of SP6, T3 and T7 wild-type promoters. Since the promoter activity for all these tag variants should be the same, their RNA-derived counts should correlate linearly with abundance of their respective DNA templates. We verified this to be the case (**Figure A.1**).

Assessing significance:

Our null hypothesis was that there is no significant difference between the activities of a given mutant promoter and the wild-type (i.e. canonical or unmutated) promoter. Because the tag sequence itself had the potential to influence promoter activity, we sought to take an empirical approach to establishing significance. Specifically, we used the data from the wild-type promoter to establish an empirical null distribution and based our p-values for mutant promoters on this distribution. Each wild-type promoter was associated with 270 different tags, while each mutant promoter was associated with 6 tags. To arrive at an empirical distribution of expected values for the wild-type promoter, we picked 6 of the 270 wild-type promoter-associated tags at random, and calculated the mean of the activities estimated from them. We repeated this process 100,000 times. For each individual mutant promoter, we calculated the mean of the activities estimated from the six independent tags assigned to it. To translate this into a p-value, we compared this estimated activity against the distribution of 100,000 means that were based on the 6-value sampling of the wild-type promoter data.

Comparison of replicates

Fold changes in efficiency of transcription for each mutation as compared to the wild-type promoter were compared between the two replicates. 824 of 861 mutant promoters identified as causing a significant change in replicate 1 were also significant in replicate 2, where 878 mutant promoters achieved significance

Luciferase assay

Six T7 promoter variants (-4T→A, -14T→C, -9C→T, -2T→A, -4T→deletion, and wild-type) were assayed for *in vivo* activity and the results were compared with the results of our *in vitro* assay. Primers were used to add XhoI and BamHI restriction sites to the 35 bp region of the wild-type T7 promoter and to each T7 promoter mutant. These fragments were inserted into the cloning site 36bp upstream of a bacterial luciferase reporter (the luxCDABE operon) in pCS26, a low-copy number (pSC101) plasmid [84]. Plasmids were transformed into TOP10 electrocompetent cells (Invitrogen) and colonies were screened for inserts by colony PCR. Inserts and plasmids were verified by Sanger sequencing. Verified constructs were then transformed into E. Cloni EXPRESS electrocompetent BL21 (DE3) LysS cells (Lucigen). Promoter activity was assayed at several time-points over the course of 10 hours after IPTG induction with a Perkin Elmer Victor V 1420 Multilabel Counter. Data from the early time-points (up to 3h) correlated well with the *in vitro* values, beyond which the correlation decreased, likely due to accumulation of luciferase inside cells and the resultant toxicity. We chose to focus on the 1 hour time-point since the *in vitro* transcription reaction was also incubated for one hour.

Pol II Promoter Mutagenesis assay

The layout of the Pol II promoter templates is shown in **Figure 3.1**. For each promoter, the -100 to +70 region was extracted. CMV (Towne strain major immediate-early promoter): GI:330614 (390 to 559), HBB (hg18): chr11: 5204748 – 5204977, and S100A4 (hg18): chr1:151784807-151784976. The sequences are specified in **Table A.2**.

The -45 to +25 region for each of the three promoters was subjected to saturation mutagenesis. The +51 to +70 region was used as the tag seed. Other sites (-100 to -46, +26 to +50) remained unaltered. Constraints for generation of tag sequences were identical to those described for the bacteriophage experiment.

Promoter templates were ordered as 200-nt synthetic oligonucleotides from Agilent Technologies[2]. The library was re-suspended in EB to 100nM stock concentration. 10% of the library was amplified in a 1ml bulk PCR reaction with 1X iProof Master Mix (Bio-Rad), 0.5uM each of primers BULK_AMP_FWD and BULK_AMP_REV, and 0.2X Sybr Green II. The reaction was assembled on ice. Thermal cycling was performed on Bio-Rad MiniOpticon Real-Time PCR

system under the following conditions: 98°C for 30 seconds (for initial denaturation), 18 cycles at 98°C for 10s, 60°C for 30s and 72°C for 15s. PCR products were purified using QIAquick PCR Purification Kit (Qiagen) and eluted in EB.

A portion of the amplified promoter library was subjected to *in vitro* transcription using the HeLaScribe Nuclear Extract In vitro Transcription System (Promega). The reaction was assembled as per manufacturer's protocol under strict RNase-free environment, in a 25uL reaction volume with 8 units of HeLa nuclear extract, 0.4mM of each rNTP, ~100ng of template DNA and 7.5 uLof 1X Transcription Buffer. The reaction was incubated at 30°C for one hour. After the incubation, the reaction was treated with 1uL of DNase I (Fermentas) at 37°C for one hour to destroy template DNA. The reaction was cleaned up using the RNeasy kit (Qiagen) as per manufacturer's protocol; expect that the ethanol volume in the first step was increased to 700uL since our expected transcripts were smaller than 200bp. RNA was eluted in 50uL of RNase-free water and subjected to another round of DNase I treatment using 1uL of enzyme in the presence of 5uL of DNase I buffer (Fermentas). The reaction was incubated at 37°C for one hour and followed by another around of RNeasy cleanup. The final RNA was eluted in 35uL of RNase-free water.

The subsequent steps (RT-PCR, PCR of DNA templates for normalization, sequencing and analysis) were identical to those described for the bacteriophage library, except that a slightly higher annealing temperature of 61°C was used for the RT-PCR/PCR cycles, and in the analysis step, templates with less than 100 reads in the DNA-derived lane were discarded from the analysis. The primers used for RT-PCR and sequencing are specified in **Table A.2**.

Appendix B. Supplementary material for Chapter 4

Supplementary Tables

Table B.1. Phrap optimization

Min match	Min score	Force level	Index word size	# of TDRGs	Mean longest SA read	Median longest SA read	Fraction of non-BLASTing SA's	Fraction of SA's BLASTing <90% of length	Fraction of mismatches among BLASTing bases
12	12	1	10	2619	361.6	403	0.004964	0.02993	0.001513
10	12	1	10	2619	364.4	406	0.004964	0.0284	0.001543
10	12	1	8	2619	364.4	406	0.004964	0.0284	0.001543
10	10	1	8	2619	369.5	409	0.004964	0.04106	0.001551
8	10	1	8	2619	371.9	411	0.004964	0.04643	0.001579

A representative subset of 10,000 *Pseudomonas* TDRGs was randomly selected and subjected to phrap assembly using different parameters and the resulting lengths and qualities of the longest subassemblies from each TDRG were assessed. We determined that parameters of minmatch 10, minscore 12, force level 1, and index word size 8, achieved the optimal balance between assembly accuracy, measured as the fraction of subassembled reads BLASTing across at least 90% of their length in a single BLAST hit (and the fraction removed because of oppositely oriented reads, not shown), and subassembled read length.

Table B.2. Summary statistics for subassembled reads

Sample	Original fragment size	# of read-pairs	# of filtered TDRGs	Median length
P. aeruginosa	~550 bp	56.8M	1,031,537	338 bp
Metagenomic	~450 bp	21.8M	262,298	256 bp
Metagenomic (merged)	~450 bp	21.8M+1.8M	180,008 (90,004 pairs)	408 bp

For the two samples used and the two analyses performed of the methylamine-enriched metagenomic sample, listed is the approximate size of long fragments from which subassembly libraries were generated, the number of Illumina read-pairs that were used to generate subassembled (SA) reads (merged analysis also shows the number of reads used to pair tags), the number of TDRGs after filtering for successful assembly and properly oriented contributing reads, and the median length of the longest SA read from each filtered TDRG.

Table B.3. Summary statistics from assembly of metagenomic SA reads versus assembly of a standard shotgun library

Input	Assembly strategy	# of contigs	Median contig length	Sequence in contigs ≥ 200 bp	Longest contig
SA reads	Celera	86,418	390 bp	35.7 Mb	6,000 bp
Shotgun PE 48 bp	Velvet (exp_cov = 100)	17,618	332 bp	9.9 Mb	102,806 bp
Shotgun PE 76 bp	Velvet (exp_cov = 100)	33,374	315 bp	16.0 Mb	28,861 bp

Comparison of assembly of short reads from a standard Illumina shotgun library prepared from the metagenomic sample to Celera assembly of the full complement of SA reads from the same sample. Listed is the assembly input, the assembly strategy used, and, for contigs at least 200 bp long, the number of contigs produced, the median contig length, the total amount of sequence contained in such contigs, and the longest contig. 76 bp paired-end (PE) reads were collected from a standard shotgun library and were trimmed to 48 bp reads to match the amount of sequence collected per read-pair for subassembly (20+76). Velvet assembly was performed using both 48 bp and 76 bp paired-end reads, but the same total amount of raw sequence as collected for subassembly (2.2 Gb) was used in each shotgun assembly. Notably, while the shotgun assemblies achieve greater contiguity at the longest lengths, potentially due to deep sampling of abundant genomes or to misassemblies, subassembly produces at least twice as much sequence at the lengths necessary for phylogenetic analysis and gene prediction.

Table B.4. Oligo sequences

	Name	Sequence
Bottleneck adaptor oligos	Ad1	TCGCAATACAGAGTTTACCGCATT
	Ad1_rc	/5Phos/ATGCGGTAAACTCTGTATTGCGA
	Ad2	CTCTTCCGCATCTCACAACCTACT
	Ad2_rc	/5phos/GTAGGTTGTGAGATGCGGAAGAG
Illumina adaptor oligos	Illum_rev	CTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T
	Illum_rev_rc	/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Bottleneck PCR primers	Ad1_amp	/5phos/TCGCAATACAGAGTTTACCGCATT
	Ad2_amp	/5phos/CTCTTCCGCATCTCACAACCTACT
TDRG merging PCR primer	Illum_amp_r_Ad2	CAAGCAGAAGACGGCATACGAGATATCGAGAGCCTCTTCCGC ATCTCACAACCTACT
Sequencing PCR primers	Illum_amp_f_Ad1	AATGATACGGCGACCACCGAGATCTACACCAATGGAGCTCGC AATACAGAGTTTACCGCATT
	Illum_amp_f_Ad2	AATGATACGGCGACCACCGAGATCTACAC ATCGAGAGCCTCTTCCGCATCTCACAACCTACT
	Illum_amp_r	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGC TGAACCGCTCTTCCGATCT
Oligos used in sequencing	Ad1_seq	CAATGGAGCTCGCAATACAGAGTTTACCGCATT
	Ad2_seq	ATCGAGAGCCTCTTCCGCATCTCACAACCTACT
	Illum_seq_r	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Oligos were obtained from Integrated DNA Technologies. An asterisk indicates a phosphorothioate bond. /5Phos/ indicates a five-prime phosphate modification.

Supplementary Figures

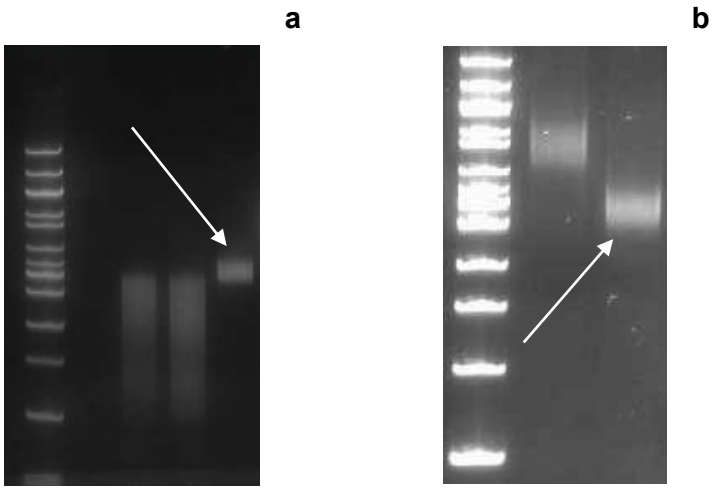


Figure B.1: Length of library fragments by PAGE

(a) PAGE of NEB 100 bp ladder and nebulized and size-selected ~550 bp *P. aeruginosa* fragments. (b) PAGE of NEB 100 bp ladder and Biorupted and size-selected Methylamine metagenomic fragments.

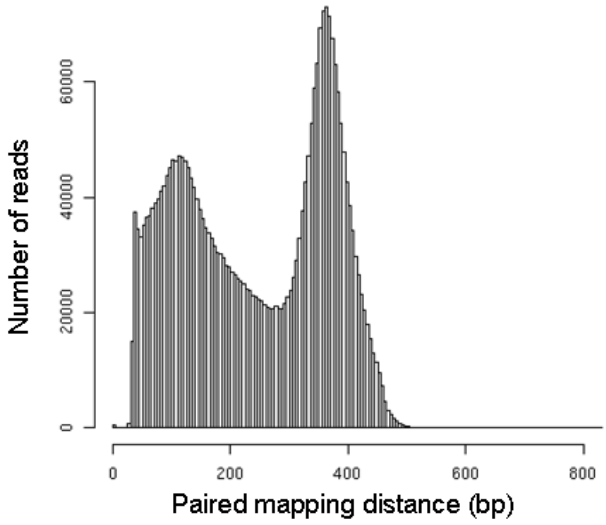


Figure B.2: Length distribution of subassembly fragments by paired-end sequencing

Histogram of mapping distance separating tag and breakpoint reads from the 450-600 bp size-selection performed at the end of the subassembly library construction protocol of a representative subset of the *Pseudomonas* data. Paired 20x76 bp reads were mapped to the PAO1 reference genome using *maq.* Shorter mapping distances are thought to arise from over-amplification during PCR, which causes shorter fragments to migrate with longer fragments during PAGE. Retained shorter fragments are then preferentially amplified and sequenced during the Illumina sequencing protocol. Careful PCR amplification is essential to prevent small fragments from completely dominating the sequencing reaction. The non-uniform nature of this distribution may contribute to the bimodal distribution of subassembled read length that we observed for this sample.

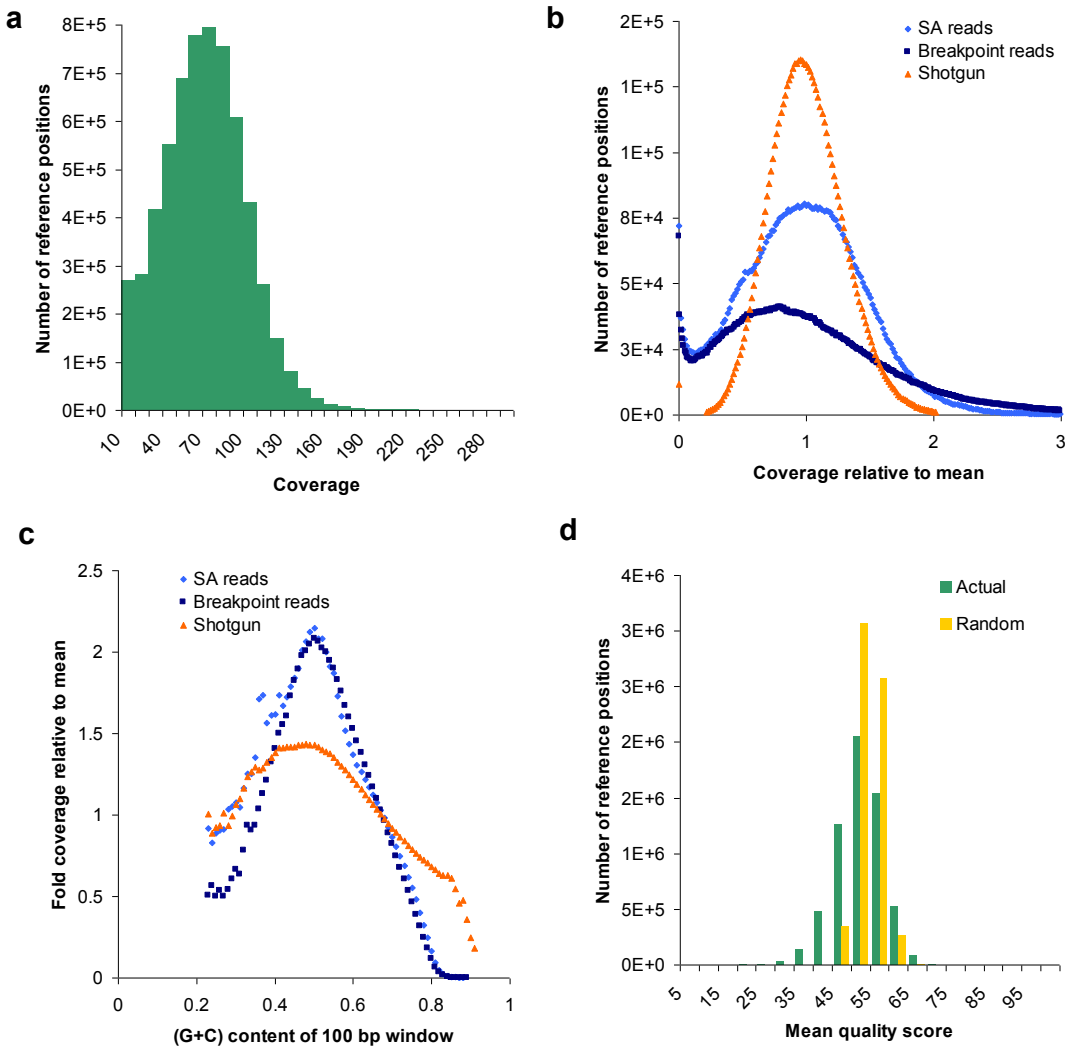


Figure B.3: Coverage of the PAO1 reference by SA reads

(a) Histogram of coverage of the PAO1 reference by SA reads as determined by BLAST alignment (bin size 10 bp). (b) Histogram of coverage of the PAO1 reference by SA reads, a standard Illumina paired-end 36 bp shotgun library, and the 76 bp breakpoint reads that contributed to SA reads. (c) Mean (G+C) content in the 100 bp window around reference positions with a given coverage on the x-axis by SA reads, a standard Illumina paired-end 36 bp shotgun library, and the 76 bp breakpoint reads that contributed to SA reads. A strong relationship between coverage and (G+C) content is observed. That is, reference bases in very high (G+C) content regions tend to have reduced coverage relative to the mean, and regions with intermediate (G+C) content are correspondingly overrepresented. This is likely due to (G+C) content biases present during the PCR steps of library construction, as a similar relationship is observed for the contributing 76 bp reads, and could likely be mitigated by PCR conditions designed to reduce (G+C) bias. (d) Distribution of mean quality score (and therefore predicted error rate) across the reference. The number of reference positions with a given mean quality score is plotted in green ("Actual"), while a simulated distribution was made by randomizing the full set of quality score assignments in SA reads and then recalculating mean quality scores for reference positions, and is plotted in yellow ("Random"). The standard deviation of the actual distribution was six compared to three for the random distribution, indicating a small systematic bias in quality score (and therefore error) distribution across the PAO1 genome.

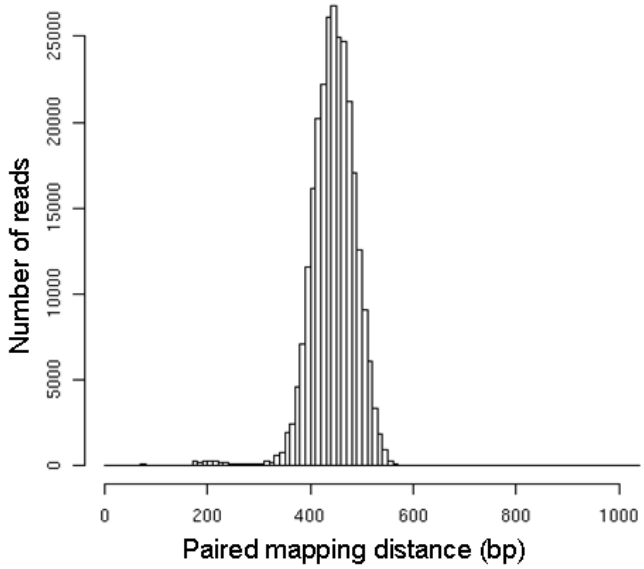


Figure B.4: Length distribution of metagenomic fragments

Histogram of the mapping distance separating paired tag reads from 36 bp paired-end sequencing data of the metagenomic library fragments (used to pair and merge TDRGs). Paired-end reads were mapped to the recently obtained Sanger data from the same sample using *maq.* Some selection for shorter molecules during the Illumina sequencing protocol may have taken place, shifting the peak of the distribution somewhat shorter than would be expected based on PAGE of the original fragments (Figure B.2).

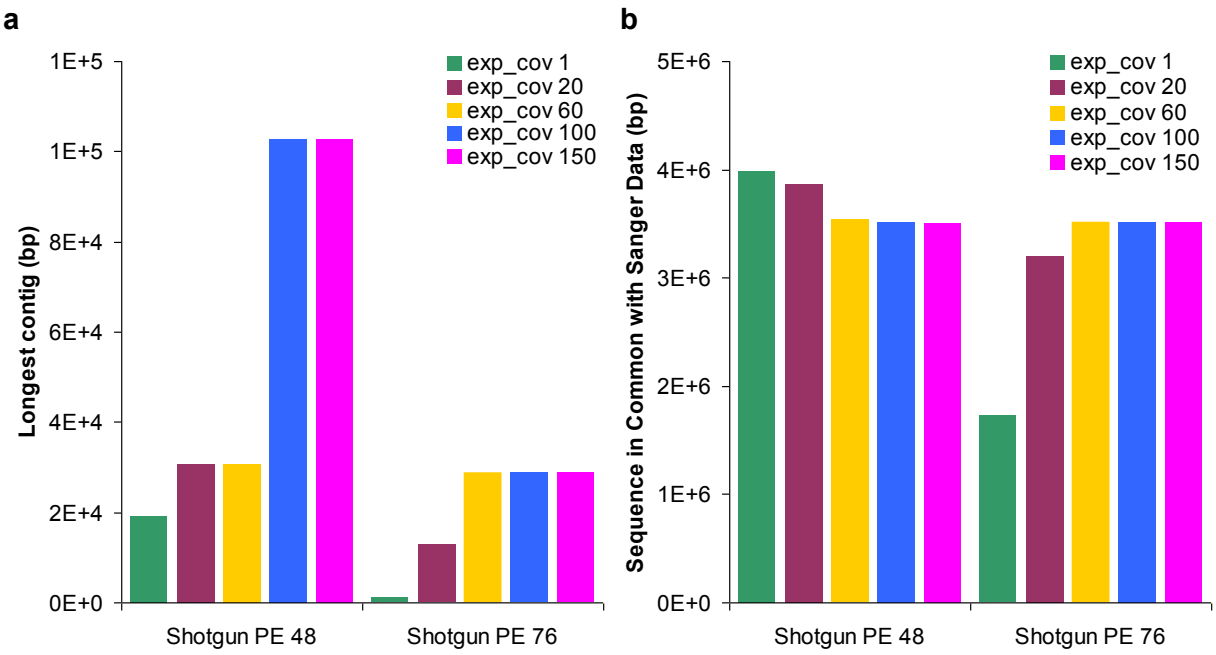


Figure B.5. Optimization of Velvet parameters for shotgun metagenomic assembly

We optimized Velvet parameters for shotgun metagenomic assembly with respect to contig length and sequence shared with available Sanger data. (a) Maximum contig length as a function of changing Velvet parameters for assembly of shotgun paired-end 48 bp and paired-end 76 bp reads. Contig length was found to be very sensitive to the exp_cov parameter. (b) Sequence in common with the available Sanger data from the same sample as a function of changing Velvet parameters as in (a). Shared sequence was found to be somewhat sensitive to the exp_cov parameter in an unpredictable fashion, with shared sequence decreasing with increased exp_cov for the 48 bp reads and increasing with increased exp_cov for the 76 bp reads. To optimize length and coverage, we chose to perform subsequent analyses with the exp_cov = 100.

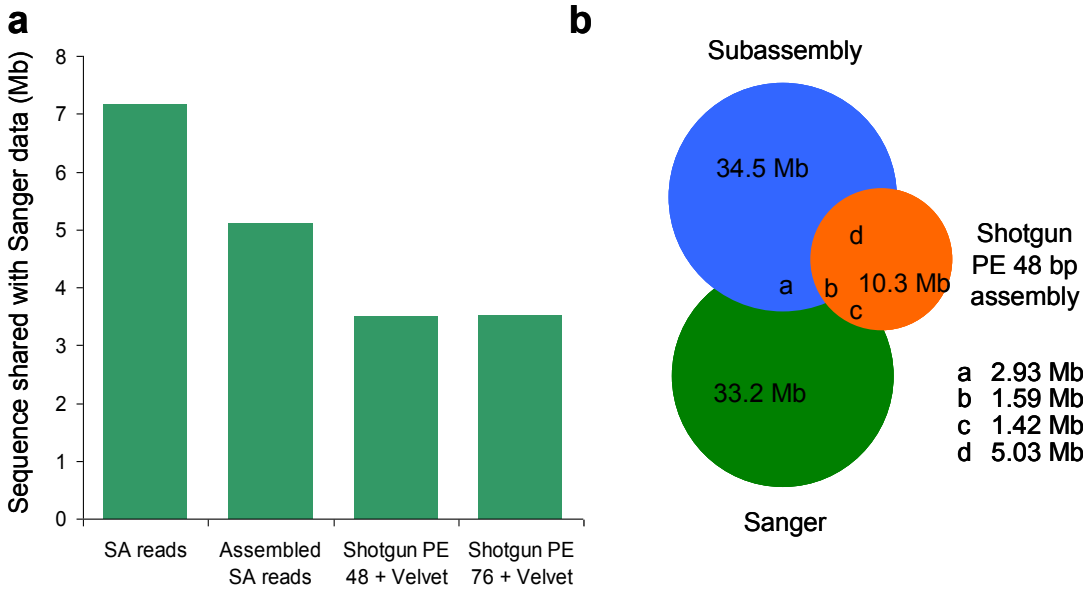


Figure B.6. Coverage overlap of metagenomic sample between sequencing methods

(a) Sequence shared with Sanger data for SA reads, assembled SA reads, and Velvet-assembled shotgun paired-end reads. Shared sequence was estimated by considering BLAST alignments with at least 98% identity across at least 100 bp. SA reads covered more than twice as much of the Sanger data as either shotgun assembly. (b) Venn diagram illustrating reciprocal coverage across data sets as determined by stringent BLAST analysis. Contigs produced by Celera assembly of SA reads, Velvet assembly of a 48 bp paired-end shotgun library with exp_cov=100, and the recently obtained Sanger sequencing data were compared to one another using BLAST. Coverage was defined as the best pair-wise match between bases as determined by the bit-score of the alignment as long as the alignment had at least 98% identity and was at least 100 bp long. The bases in common shown here are not in exact agreement with those presented in (a) because, for the purposes of constructing this diagram, each base was only allowed to align to one corresponding base in another dataset. Circles are drawn to scale; regions of overlap not to scale.

Supplementary Note 1. The importance of a complexity bottleneck

A complexity bottleneck is needed so that multiple overlapping, randomly-positioned breakpoint reads can be observed for each member of the long fragment library with a reasonable amount of sequencing. In other words, it is necessary to sample each nested sub-library in sufficient depth to reconstruct the sequence of the parent long molecule. For example, we obtained approximately 60 million read-pairs across six lanes of Illumina sequencing that enabled us to reconstruct part or all of the sequence of approximately one million long molecules. If we had used a library containing 100 million long molecules, we would only have observed, on average, less than one read-pair per long molecule, preventing any subassembly from taking place within most if not all sub-libraries. The only exception to this principle is in the case of a very small effective genome size and if the ends of molecules (and not degenerate synthetic adaptors) are used as tags. For example, in the case of a genome of only 500 kilobases, the maximum number of unique tag reads (assuming no repetitive sequence at the scale of the tag read) is one million, which we have shown to be a tractable library complexity. In such a situation, it might not be formally necessary to restrict library complexity.

Supplementary Note 2. Filtering of predicted misassemblies

Manual inspection of predicted misassemblies revealed four contigs that were incorrectly called misassemblies because of differences between the strain that we sequenced and the reference PAO1 strain. Three of these (2548, 2129 and 2115) exhibited extremely high sequence identity with a phage-like insertion in PAO1 that was recently added to GenBank (ID GQ141978.1) and that we have observed in independent shotgun sequencing data from our strain. Notably, the same phage-like insertion seems to have caused the lone misassembly in our scaffolds (Scaffold_LR7_3). The fourth contig (2622) spans a ~1 kb deletion in our strain that we have also observed in independent shotgun sequencing data (data not shown).

Supplementary Note 3. Comparison of *de novo* assembly to hybrid 454-Illumina approach

We compared the performance of our method to a recently published, high quality *de novo* assembly from a similar but significantly lower (G+C) content organism (66.6% versus 58.5%), which was generated by combining both long-read and long-range mate-paired 454 data with short distance paired-end Illumina data[99]. We find that our method compares very favorably to that approach with respect to N50 (445 kb versus 92 kb), longest scaffold (915 kb versus 389 kb), substitution error rate (~1/14,000 versus ~1/7,000), and number of rearrangements (one versus twenty). It should be noted that the authors of that study also performed sequencing and assembly of a related organism without a reference genome and achieved apparently better performance (N50 of 532 kb, longest contig of 794 kb), which they attempted to validate with limited Sanger sequencing. However, it is difficult to make a direct comparison with respect to accuracy in the absence of a reference genome. Our method also used significantly more raw data than that study, but only required a single sequencing platform, which may increase its general utility.

Supplementary Note 4. Estimated cost of subassembly protocol

Although it is difficult to draw firm conclusions in the face of rapidly changing costs associated with many second-generation sequencing platforms, it is clear that subassembly is significantly more expensive than standard shotgun Illumina sequencing if only the total amount of sequence produced is considered. However, as subassembly produces much longer reads at much higher per-base accuracy than the raw reads from the Illumina platform, such a comparison is not valid. Even the comparison to Roche/454 sequencing, which produced reads in the hundreds of base-pairs, is difficult because of the decreased accuracy of that method relative to the method we present here. Still, we estimate that our method is roughly cost-comparable to Roche/454 sequencing. For example, if a lane of sequencing is assumed to cost ~$2,000, from six lanes of sequencing we generated 405 Mb of long SA reads for the *P. aeruginosa* sample, which corresponds to a cost of ~$30/Mb, or about half that of recently published estimates of the cost of Roche/454 Sequencing [3]. However, the reduced error rate is a critical differentiator, making the cost comparison tenuous. A major advantage of subassembly is that extremely low error rates and long effective read length is maintained independent of sample complexity. In the case of short read sequencing (Illumina, AB SOLiD, Helicos), read length and error limitations can be overcome through the use of very high coverage. The ability to achieve high coverage depends implicity on sample complexity and can be complicated by relatedness of sequences therein. With Roche/454 sequencing, read lengths are longer, but once again, errors can only be overcome with high coverage, which again may be impossible in the case of either very high sample complexity or the presence of highly related sequences. We therefore conclude that subassembly produces equivalently long sequences at below or equal to the cost of Roche/454 sequencing with length and error performance that remains independent of sample complexity and sequence relatedness, a feature of no other currently available second-generation sequencing method.

Supplementary Methods

Subassembly library construction

Source DNA was fragmented by sonication, end-repaired and size-selected to ~550 bp (*P. aeruginosa*) or ~450 bp (metagenomic sample). Size-selected fragments were A-tailed and ligated to custom adaptors (Supplementary Table 4). Real-time PCR with phosphorylated primers was performed using serial dilutions of adaptor-ligated fragments to impose a complexity bottleneck and generate many copies of a limited number of long fragments. Complexity was estimated from the concentration of input material, the kinetics of PCR amplification and gel electrophoresis of the PCR product. After PCR, the product estimated to have resulted from ~105–107 long fragments was concatemerized to high molecular weight and then fragmented by sonication. Shearing products were end-repaired, A-tailed and ligated to the Illumina Read 2 adaptor. PCR amplification was then performed with one primer corresponding to the Read 2 adaptor and a second primer corresponding to one of the two original adaptors. Finally, the amplification products were size-selected to obtain a uniform distribution of shearing products across the original fragment (Supplementary Fig. 2). For the metagenomic effort, an aliquot of the bottleneck PCR was subjected to an additional round of PCR to prepare the long fragments for paired-end sequencing and subsequently used for tag-pairing and TDRG merging.

Shotgun library construction

P. aeruginosa short insert (~200 bp) and long insert (~2.5 kb), and metagenomic short insert shotgun libraries were constructed according to manufacturer's specifications, except that standard oligonucleotides were obtained from IDT. For the metagenomic library, to conserve source material, size selection to the desired fragment length was performed before A-tailing and adaptor ligation rather than afterward so that the longer size range could be used for subassembly.

Illumina sequencing

For subassembly libraries, an Illumina GA-II instrument was used to collect paired-end reads according to manufacturer's specifications, except that custom sequencing primers (Supplementary Table 4) were used, and asymmetric read lengths were collected (20-bp first read and 76-bp second read). For the tag-pairing metagenomic library, paired-end 36-bp reads were collected according to manufacturer's specifications with custom sequencing primers. For shotgun libraries, paired-end reads were collected according to manufacturer's specifications.

Organizing breakpoint short reads into TDRGs

For all experiments, breakpoint reads paired with identical or nearly identical tag sequences were grouped into TDRGs. As millions of tag reads were involved, an all-against-all comparison to cluster similar tags was not feasible. Instead, a two-step strategy was used to group tag sequences in each experiment. First, perfectly identical tags were collapsed using a simple hash to define a nonredundant set of clusters. From this set, clusters with four or more identical tags were identified as 'core' clusters and, in descending order by size, were compared to all other tags. Tags matching a given core cluster with up to one mismatch were grouped with that core cluster (and removed from further consideration if they themselves defined a smaller core cluster). TDRGs with more than 1,000 members were excluded from downstream analysis to limit analysis of adaptors or other low-complexity sequence.

Subassembly of TDRGs

Each TDRG was assembled separately using *phrap* with the following parameters: "-vector_bound 0 -forcelevel 1 -minscore 12 -minmatch 10 -indexwordsize 8". Pre-grouping reads into TDRGs allowed us to use less stringent parameters than the defaults used in traditional assemblies. Parameters were optimized to balance SA read length and accuracy (Supplementary Table 1). A short-read assembler, *Velvet*, was also tested but did not produce substantial gains in SA read length relative to *phrap* (data not shown).

Trimming and filtering of SA reads and assignment of consensus quality scores

SA reads were masked using the *cross_match* program provided as part of the *phrap* suite, using the following parameters: "-minmatch 5 -minscore 14 –screen". Determination of consensus quality scores and further trimming was performed as follows. Because it permits multiple alignments per read, the *Bowtie* short-read alignment tool [128] was used to map contributing 76-bp breakpoint reads to the SA reads to generate consensus quality scores for SA read base calls. Only alignments within TDRGs were allowed (that is, alignments of breakpoint reads to SA reads from another TDRG were ignored). *Bowtie* was also used to map the 20-bp tag reads back to the SA reads to facilitate end trimming where the SA read had extended into adaptor sequence. Next, SA reads were trimmed using both tag read mapping and adaptor masking information. SA reads were first trimmed from the 3′ end using the mapping location of the tag read; if bases remained that had been masked by *cross_match* because of the presence of adaptor, the masked bases were removed and the longest remaining continuous sequence was retained. Finally, any sequence containing a base call with quality below 10 within 5% of the 3′ end of the SA read was discarded.

In all subsequent analyses, only SA reads that were at least 77 bp long and were assembled from identically oriented short reads were considered. The read orientation filter was only applicable to SA reads from individual, unmerged TDRGs. In addition, for length and quality analyses, only the longest SA read from each TDRG was analyzed.

Quality assessment

The longest SA read (after trimming as described above) from each TDRG containing at least 10 member reads was aligned to the *P. aeruginosa* PAO1 reference genome using *BLAST* with the following parameters: "-p blastn -e 1e-6 -m 8 -F F -a 4".

Error rate as a function of quality score and position in the SA read was then determined as follows. BLAST alignments containing at least 95% of the length of the SA read query and without any gap openings were used to define the position in the reference of the SA read in question (the BLAST coordinates were extended to encompass the entire length of the SA read). Every base in an SA read whose alignment meets the above criteria was compared to the corresponding reference base. If less than 100% of the SA read aligned, the comparison was forced to extend to the ends of the SA read. From the base-by-base comparison, the error rate as a function of base call quality or position in the SA read was calculated.

We did not perform a base-by-base comparison for cases in which BLAST used a gap opening in making an alignment, which could potentially suppress our error rates if such SA reads were substantially more error-laden. Accuracy of such SA reads within aligned regions was slightly lower (99.56% accurate compared to 99.86% in SA reads without gaps), and such sequences only comprised less than 1% of the sequence being analyzed. We therefore concluded that errors in these sequences that fall outside of aligning regions are unlikely to substantially alter our estimates of error rate as a function of base quality. We performed a similar analysis for SA reads containing larger gaps with respect to the reference (those with a BLAST alignment less than 95% of their length), as we did not perform a base-by-base comparison for such SA reads either. Once again, the accuracy with aligned regions was somewhat lower (99.4% versus 99.86% in those with complete or nearly complete alignments). Such errors probably reflect larger-scale misassemblies owing to repetitive sequence in the true reference sequences. Notably, aggressive trimming substantially reduced the relative abundance of such sequences; only 1.5% of the total number of bases analyzed was contained in such sequences, and only 2.3% of BLAST alignments fell into this category. Once again, forcing the alignment to the very edges of such SA reads was not likely to substantially alter the relationship between error rate and base call quality score.

To analyze quality as a function of raw read base quality, *maq* was used to align contributing 76-bp breakpoint reads to the reference, Illumina base calls were compared to the reference and, for a randomly chosen subset of 1 million bases, the error rate as a function of Illumina base call quality was determined.

To analyze quality as a function of raw read position, a representative lane of contributing 76-bp breakpoint reads used for the subassembly process was aligned to the reference genome using maq, and the error rate at each position was determined by comparing read base calls to reference bases for each read.

Assembly of SA reads using the Celera assembler (*CABOG*)

For P. aeruginosa and metagenomic samples, all trimmed, orientation- and length-filtered SA reads (not only the longest per TDRG) were subjected to assembly using the Celera assembler. Assembly was guided by consensus quality scores generated as described above. The Celera assembler (CABOG) was run with default parameters and “unitigger=bog”.

Assessment of assembled SA read quality

Contigs produced by the Celera Assembler from SA reads were aligned to the reference using BLAST with the following parameters: “-p blastn -e 1e-6 -m 8 -F F”. Substitution error rate was measured as the number of mismatches within the best BLAST alignment for each contig. To account for a potentially higher error rate in misassembled contigs, if a contig aligned across less than 95% of its length, other BLAST alignments were also considered as long as they comprised at least 10% of the contig length.

Scaffolding of contigs for P. aeruginosa

For *de novo* assembly of the P. aeruginosa genome, we used independently produced shotgun sequencing libraries to scaffold the contigs produced from SA reads as follows. The resulting contigs were scaffolded using a custom script that used 36-bp shotgun paired-end Illumina reads from one lane each of short-insert (~200 bp) and long-insert (~2.5 kb) libraries. The gap between each pair of adjacent contigs in a scaffold was dynamically estimated based on the distance of the read pairs connecting the two contigs from the ends of the contigs and the expected insert size of the library from which they were derived. Scaffolds were then constructed by separating the contigs by a string of unknown nucleotides (Ns) as long as the estimated gap size. For cases where the expected gap size was close to zero or negative (indicating a possible overlap), the adjacent ends of the two contigs were subjected to a Smith-Waterman alignment and merged accordingly if a match was detected.

TDRG merging algorithm

Paired 36-bp reads were obtained from a sequencing library prepared from bottlenecked, adaptor-ligated metagenomic fragments, then trimmed computationally to 20 bp to correspond to the length of the tag reads that were obtained during sequencing of the subassembly libraries.

To prevent sequencing errors at the ends of the reads from creating spurious tags and tag pairs, we trimmed the reads further to the first 15 bp. If multiple TDRGs (defined by 20-bp tags) could correspond to a single 15-bp tag from a merging read pair, the TDRG with the most members was chosen. In descending order of tag-pair abundance, we defined TDRG pairs, removing tags that had been assigned to TDRG pairs as we proceeded.

Velvet assembly of shotgun metagenomic library

Paired-end shotgun reads constructed according to standard Illumina protocols were assembled using Velvet with the following parameters: "-cov_cutoff 2 -exp_cov [variable] -ins_length 250 -unused_reads yes".

If exp_cov was set to 1, cov_cutoff was set to 0. As *Velvet* (along with all other short-read assemblers) is not designed for assembly of metagenomic sequences, considerable effort was made to optimize its performance with respect to length of sequences produced and agreement with the available Sanger sequencing data to make the fairest comparison possible. We found that contig length was sensitive to the exp_cov parameter (**Figure B.5**). However, we observed unpredictable performance with respect to agreement with the Sanger sequencing data when altering this parameter, as agreement improved for the paired-end 76-bp reads but degraded for the paired-end 48-bp reads. We therefore chose an exp_cov value of 100 as the best compromise of sequence length and coverage for the comparator datasets.

Resulting scaffolds were then split into contigs that did not contain Ns, as we reasoned that key goals of metagenomic sequencing such as gene discovery and phylogenetic classification would depend solely on the length of contiguous regions of defined bases.

Comparison to Sanger sequencing data with BLAST

Contigs produced from SA reads with CABOG and contigs produced from shotgun short reads with Velvet were aligned to one another and to the recently collected Sanger sequencing data from the same sample (JGI IMG/M Taxon Object ID 2006207002, NCBI accession number ABSR01000000) using BLAST with the following parameters: "-p blastn -e 1e-6 -m 8 -F F". Two bases were considered to be a shared position between two datasets if they were contained in a BLAST alignment at least 100 bp long and with at least 98% identity. For the Venn diagram (Supplementary Fig. 6), an additional restriction was added so that mappings between the three datasets were not ambiguous: the two bases were required to be in the BLAST alignment with the highest bit score of all the BLAST alignments between the two datasets involving either base.

Appendix C. Supplementary material for Chapter 5

Supplementary Tables

Table C.1. Predictive power and significance of multiple linear regression models

Library	*n* term model		3*n* term model	
	R^2	p	R^2	p
ALDOB	0.03	< 0.005	0.05	< 0.01
ECR11	0.12	< 0.005	0.19	< 0.01
LTV1 rep. 1	0.21	< 0.005	0.29	< 0.01
LTV1 rep. 2	0.22	< 0.005	0.30	< 0.01

Multiple linear regression models taking into account all positions (*n* term model, where *n* is the length of the enhancer) or all mutations at all positions (3*n* term model), were constructed for each of the enhancers. Listed here are R^2 values and p-values (computed by constructing models for 200 or 100 random permutations of the outcome vector and comparing mean squared errors from the permuted data models to the actual data model).

Table C.2. Predicted transcription factor binding sites.

Enhancer	Start pos.	End pos.	Strand	Factor	Core Match	Matrix Match	Enhancer	Start pos.	End pos.	Strand	Factor	Core Match	Matrix Match
aldob	16	29	-	C/EBPbeta	0.833	0.799	ecr11	212	226	-	HNF-3beta	1	0.842
aldob	17	27	+	AP-1	1	0.975	ecr11	217	230	-	C/EBPbeta	0.829	0.786
aldob	41	52	+	Oct-1	1	0.91	ecr11	219	228	+	GATA-3	0.977	0.886
aldob	45	58	-	C/EBPbeta	0.833	0.776	ecr11	221	236	+	GR	1	0.848
aldob	92	105	-	HNF-4	0.825	0.811	ecr11	247	259	-	CHOP C/EBPalpha	-0.778	0.819
aldob	112	125	-	C/EBPbeta	0.821	0.762	ecr11	249	262	+	C/EBPbeta	0.833	0.863
aldob	120	130	-	AP-1	0.935	0.868	ecr11	249	262	-	C/EBPbeta	0.816	0.861
aldob	135	151	+	HNF-1	1	0.795	ecr11	264	273	+	USF	0.918	0.868
aldob	140	149	+	TATA	1	0.888	ecr11	272	281	+	USF	0.918	0.873
aldob	141	155	-	HNF-3beta	1	0.835	ecr11	361	371	+	AP-1	0.935	0.854
aldob	148	158	-	AP-1	0.935	0.866	ecr11	382	401	-	YY1	1	0.855
aldob	149	166	-	NF-1	1	0.983	ecr11	390	407	+	NF-1	0.911	0.878
aldob	150	163	-	HNF-4	0.796	0.802	ecr11	394	403	+	USF	0.905	0.88
aldob	153	162	+	USF	0.945	0.932	ecr11	407	420	-	C/EBPbeta	0.816	0.787
aldob	157	170	-	HNF-4	0.988	0.892	ecr11	429	438	+	USF	0.905	0.854
aldob	170	179	+	USF	0.931	0.852	ecr11	438	447	+	TATA	1	0.882
aldob	201	217	+	HNF-1	0.942	0.706	ecr11	460	474	-	HNF-3beta	1	0.894
aldob	205	215	+	AP-1	0.935	0.922	ecr11	465	478	-	C/EBPbeta	0.829	0.786
aldob	209	225	-	HNF-1	0.771	0.632	ecr11	465	481	-	HNF-1	0.829	0.712
aldob	217	230	+	C/EBPbeta	0.859	0.784	ecr11	467	476	+	GATA-3	0.977	0.886
aldob	247	256	+	GATA-3	1	0.935	ecr11	474	487	+	C/EBPbeta	0.883	0.815
aldob	247	256	-	GATA-3	1	0.946	ecr11	489	498	+	GATA-3	0.945	0.91
ecr11	18	27	+	GATA-3	0.968	0.912	ecr11	520	533	+	C/EBPbeta	0.865	0.78
ecr11	25	41	+	HNF-1	0.92	0.657	ecr11	521	537	-	HNF-1	0.874	0.71
ecr11	31	47	-	HNF-1	1	0.647	ecr11	545	561	-	HNF-1	0.835	0.732
ecr11	44	53	+	USF	0.918	0.865	ecr11	549	562	-	C/EBPbeta	0.883	0.829
ecr11	51	64	+	C/EBPbeta	0.854	0.811	ecr11	567	580	+	C/EBPbeta	0.833	0.788
ecr11	62	71	+	TATA	1	0.952	ecr11	571	585	+	HNF-3beta	1	0.907
ecr11	64	78	-	HNF-3beta	1	0.85	ecr11	573	585	-	CHOP C/EBPalpha	-0.882	0.81
ecr11	71	87	-	HNF-1	0.771	0.706	ecr11	603	619	+	HNF-1	0.796	0.657
ecr11	74	88	+	HNF-3beta	1	0.832	ltv1	6	16	+	AP-1	0.935	0.892
ecr11	87	100	-	C/EBPbeta	0.816	0.84	ltv1	8	21	+	HNF-4	1	0.873
ecr11	87	101	-	HNF-3beta	1	0.836	ltv1	13	23	+	AP-1	0.935	0.847
ecr11	87	104	-	NF-1	1	0.961	ltv1	18	35	+	NF-1	0.921	0.88
ecr11	87	96	-	TATA	1	0.9	ltv1	36	49	+	C/EBPbeta	0.848	0.789
ecr11	98	107	-	USF	0.945	0.931	ltv1	43	58	+	GR	0.978	0.85
ecr11	102	118	+	HNF-1	0.829	0.818	ltv1	86	103	+	NF-1	1	0.952
ecr11	102	112	+	AP-1	0.811	0.825	ltv1	91	100	-	USF	0.931	0.911
ecr11	106	116	+	AP-1	1	0.969	ltv1	101	115	-	HNF-3beta	1	0.879
ecr11	108	124	-	HNF-1	1	0.78	ltv1	126	139	-	C/EBPbeta	0.828	0.804

ecr11	116	129	-	C/EBPbeta	0.842	0.859	ltv1	127	137	-	AP-1	0.935	0.889
ecr11	120	134	-	HNF-3beta	0.93	0.868	ltv1	159	176	-	NF-1	0.921	0.88
ecr11	144	153	+	GATA-3	0.981	0.908	ltv1	163	176	+	C/EBPbeta	0.888	0.837
ecr11	146	155	-	GATA-3	0.981	0.908	ltv1	199	209	+	AP-1	0.935	0.887
ecr11	147	163	-	HNF-1	0.795	0.682	ltv1	202	219	-	NF-1	0.905	0.864
ecr11	167	182	-	GR	1	0.886	ltv1	213	222	+	USF	0.987	0.857
ecr11	175	189	-	HNF-3beta	1	0.926	ltv1	223	232	-	USF	0.926	0.883
ecr11	187	204	+	NF-1	0.921	0.875	ltv1	243	254	-	CREB	1	0.975
ecr11	191	207	-	HNF-1	0.794	0.672	ltv1	244	254	-	AP-1	0.935	0.868
ecr11	196	205	-	GATA-3	0.896	0.881	ltv1	248	265	-	NF-1	0.921	0.898
ecr11	198	211	-	C/EBPbeta	0.996	0.885	ltv1	285	298	-	C/EBPbeta	0.854	0.781

We used the MATCH web server[109] to predict transcription factor binding sites (TFBS) in the three enhancers under study using liver-specific profiles and cutoff selection set to minimize false negatives.

Table C.3: Characteristics of interacting positions from pairwise multiple regression models.

		ALDOB		**ECR11**		**LTV1**	
		<=10 nt	>10 nt	<=10 nt	>10 nt	<=10 nt	>10 nt
Not significant		2509	30798	5706	178301	2787	41143
Significant		22	60	17	182	28	17
p-value		< 1e-4		< 1e-3		< 1e-4	
		ALDOB		**ECR11**		**LTV1**	
Univar. model coeff. signs	Interaction term sign	Not significant	Significant	Not significant	Significant	Not significant	Significant
-/-	-	7378	0	4195	0	9690	0
-/-	+		2		4		18
+/-	-	4471	36	5387	20	8315	4
+/-	+		6		1		4
+/+	-	654	0	1682	1	1760	5
+/+	+		12		23		3

For pairs of positions that were mutated together in at least 20 haplotypes, we built multiple linear regression models with three binary variables to predict the number of RNA aliquots in which a haplotype was observed. Two variables encoded whether each position was mutant or wild-type in a given haplotype and the third encoded whether both were mutant together. We then compared whether pairs of positions with significant interaction terms (FDR<0.05) were enriched for nearby pairs (separated by ≤10 nt) compared to those with non-significant interaction terms (p-value obtained by comparing the number of nearby pairs with significant interaction terms to the null distribution of this quantity, obtained by randomly permuting the position vector 10,000 times and each time computing the number of nearby pairs with significant interaction terms). We also classified models on the basis of the sign (positive or negative) of the coefficient from the univariate position-by-position models ("Univar. model coeff. signs"), the interaction term sign, and whether or not the interaction term was significant in the pairwise model (note that non-significant interactions terms cannot be distinguished from zero and therefore do not have a sign).

Table C.4 Sequences of oligonucleotides and primers

Oligonucleotides used for PCA	
ALDOB_PCA_OLIGO1	AGGACCGGATCAACTTCTTCA
ALDOB_PCA_OLIGO2	TCCCTGTAAACAGTATTAGTTTGAATTATCATTTTCACTGTTATTCTGGTTGAGTCAGCATAC CCAGATTGAAGAAGTTGATCCGGTCCT
ALDOB_PCA_OLIGO3	ATAATTCAAACTAATACTGTTTACAGGGAGTTAAACTTCTACAGTGGGATTAAAGGTCTGTAC CACGTTAGCACAAATGTCACCTCTCTG
ALDOB_PCA_OLIGO4	CCATCCCAGGTTGTCCTCCTGTCTCCTTGTGGTGAACATTGGCCTGTGACCCTGTTTTATGA TTAACAGAGAGGTGACATTTGTGCTAAC
ALDOB_PCA_OLIGO5	GGAGGACAACCTGGGATGGGTAATGACAAAGAACGATTTCCGTACTCCTAAGCCTCTGCTC TCTCAGATCTCAAGCCATTGCGTGAACCG
ALDOB_PCA_OLIGO6	TCGGTTCACGCAATGGCTTG
ECR11_PCA_OLIGO1	AGGACCGGATCAACTCTCTGAAGCTCAAAAGCAATG
ECR11_PCA_OLIGO2	AAACATTTAGTATTTTTAAAGGTGTTGGAATTCAAGTGTTAAAAATCGAAGCCTTATCAAATCA TTGCTTTTGAGCTTCAGAGAGT
ECR11_PCA_OLIGO3	ATTCCAACACCTTTAAAAATACTAAATGTTTCCCATTTTAAACAAGCCAAGTGAATGACTGAA TTCTTAACCAAAAATAAATGTGA
ECR11_PCA_OLIGO4	GGCCAGAGAATATTTATATAATGTTCTGTATGGACAAAGAGTGATATCAATCTACTTCACATT TATTTTTGGTTAAGAATTCAGTC
ECR11_PCA_OLIGO5	ACAGAACATTATATAAATATTCTCTGGCCTTACTATCTAGCAAGGCAGGAAAAATAGATCAAT TTGTTCTCACTCATAGGTGGGAA
ECR11_PCA_OLIGO6	CCCCACAACAGGCCCCGATGTGTGATGTTCCCCTTCCTGTGTCCATGTGTTCTCATTGTTCA ATTCCCACCTATGAGTGAGAACAA
ECR11_PCA_OLIGO7	GGGGCCTGTTGTGGGGTGGGGGGAGGGGGGAGGGATAGCATTAGGAGATATATCTAACGT TAAATGACGTGTTAATGGGAGCAGCA
ECR11_PCA_OLIGO8	TAAGTTTTAGGGTACATGTGCACAACATGCAGTTTGTTACATATGTATACATGTGCCATGTTG GTGTGCTGCTCCCATTAACACGT
ECR11_PCA_OLIGO9	TTGTGCACATGTACCCTAAAACTTAAAGTATAATAAGAAAAATAGATCAATTTACTCTACATCT GAGATTAAAAAGCAGAAAGACT
ECR11_PCA_OLIGO10	TTCTCGCTGTTACTCTATTTCTGGTTCTGAATGTCAAATACTGAAACTCTGTGAGTGAGTCTT TCTGCTTTTTAATCTCAGATGTA
ECR11_PCA_OLIGO11	ACCAGAAATAGAGTAACAGCGAGAACTTGAACTATTTCAGTTTAGCCTCCCACCCTCTCTGC TATCACTTCCCAAAACATTGCGTG
ECR11_PCA_OLIGO12	TCGGTTCACGCAATGTTTTGGGAAGTG
LTV1_PCA1_OLIGO1	ATCACAAGTTTGTACAAAAAAGCAGGCTCCGCGGCCGCCCCCTTCACCTTTGGGTGACCCC TGACCCTGGCCGCCTGGGCTC
LTV1_PCA1_OLIGO2	ACAGGGCCAAGGAAGGAGGGCGGGGTGGGGCGGGGCGGCGAGGACGGAATGTGCGGGA AGGCGAGCCCAGGCGGCCAGGGTC
LTV1_PCA1_OLIGO3	CCCTCCTTCCTTGGCCCTGTGGGGACGGAAACATCCCGTTCCTGCCCAAGCTGGGTCAAGA GCCGGAGGGACAGGACCAGAG
LTV1_PCA1_OLIGO4	AGGCGTGGCGAGATGAGGTCACCCAGTAGGAACAAGGAGAGCTAGTTCTGGCGTAAGGGG TGCTCTGGTCCTGTCCCTCCGG
LTV1_PCA1_OLIGO5	GACCTCATCTCGCCACGCCTCCTCAGGTGAACACCCGGGCTGGTAACGTCACTTCCTGCCA

	GGTAAGCGCCCCCAGGCAGCA
LTV1_PCA1_OLIGO6	ATCACCACTTTGTACAAGAAAGCTGGGTCGGCGCGCCCACCCTTCAGACCTTTCCGTGAGC AGTGCTGCCTGGGGGCGCTTAC
LTV1_PCA2_OLIGO1	AGCAGGCTCCGCGGCCGCCCCCTTCACCTTTGGGTGACCCCTGACCCTGGCCGCCTGGGC TCGCCTTCCCGCACATTCCG
LTV1_PCA2_OLIGO2	GGATGTTTCCGTCCCCACAGGGCCAAGGAAGGAGGGCGGGGTGGGGCGGGGCGGCGAG GACGGAATGTGCGGGAAGGCGA
LTV1_PCA2_OLIGO3	CTGTGGGGACGGAAACATCCCGTTCCTGCCCAAGCTGGGTCAAGAGCCGGAGGGACAGGA CCAGAGCACCCCTTACGCCA
LTV1_PCA2_OLIGO4	GTTCACCTGAGGAGGCGTGGCGAGATGAGGTCACCCAGTAGGAACAAGGAGAGCTAGTTC TGGCGTAAGGGGTGCTCTGG
LTV1_PCA2_OLIGO5	CCACGCCTCCTCAGGTGAACACCCGGGCTGGTAACGTCACTTCCTGCCAGGTAAGCGCCC CCAGGCAGCACTGCTCACGG
LTV1_PCA2_OLIGO6	ATCACCACTTTGTACAAGAAAGCTGGGTCGGCGCGCCCACCCTTCAGACCTTTCCGTGAGC AGTGCTGCCTGG
LTV1_PCA1_P1	GCTAGCCTCGAGGATATCACAAGTTTGTACAAAAAAGCAGGCTCCG
LTV1_PCA1_P2	ACGGGCCAAGGAAGGAGGGC
LTV1_PCA1_P3	CCCTCCTTCCTTGGCCCTGTGG
LTV1_PCA1_P4	AGGCGTGGCGTGATGAGGTCAC
LTV1_PCA1_P5	GACCTCATCTCGCCACGCCTCC
LTV1_PCA1_P6	AGGCCAGATCTTGATATCACCACTTTGTACAAGAAAGCTGGGTCG
LTV1_PCA2_P1	GCTAGCCTCGAGGATATCACAAGTTTGTACAAAAAAGCAGGCTCCGCGGCC
LTV1_PCA2_P2	GGATGTTTCCGTCCCCACAGGG
LTV1_PCA2_P3	CTGTGGGGACGGAAACATCCCG
LTV1_PCA2_P4	GTTCACCTGAGGAGGCGTGGCG
LTV1_PCA2_P5	CCACGCCTCCTCAGGTGAACAC
LTV1_PCA2_P6	AGGCCAGATCTTGATATCACCACTTTGTACAAGAAAGCTGGGTC
LTV1_OUTER_F	GCTAGCCTCGAGGAT
LTV1_OUTER_R	AGGCCAGATCTTGAT
Oligonucleotides used for cloning, tagging and sequencing	
VH_F	GCTAGCCTCGAGGATTGCCTAGGACCGGATCAACT
VH_R	AGGCCAGATCTTGATGAGCTTCGGTTCACGCAATG
ENHANCER_FWD	AATGATACGGCGACCACCGAGATCTACACTGCCTAGGACCGGATCAACT
LTV1_F	GGGAGGTATTGGACAGGCCGC
Nextera Adapter1	AATGATACGGCGACCACCGAGATCTACACGCCTCCCTCGCGCCATCAG
Nextera BP1	AATGATACGGCGACCACCGA
BARCODE_PE_F	AATGATACGGCGACCACCGAGATCTACACAGTCGCCTATACGGTGATGG
BARCODE_PE_R	CAAGCAGAAGACGGCATACGAGATATGGGATTAAACGGGGAGAC
BARCODE_PE_R_ILMNINDX1	CAAGCAGAAGACGGCATACGAGATCGTGATTCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_PE_R_ILMNINDX2	CAAGCAGAAGACGGCATACGAGATGCCTAATCGACTCTAGATGGGATTAAACGGGGA

	GAC
BARCODE_PE_R_ILMNINDX3	CAAGCAGAAGACGGCATACGAGATCACTGTTCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_PE_R_ILMNINDX4	CAAGCAGAAGACGGCATACGAGATATTGGCTCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_PE_R_ILMNINDX5	CAAGCAGAAGACGGCATACGAGATTCAAGTTCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_PE_R_ILMNINDX6	CAAGCAGAAGACGGCATACGAGATCTGATCTCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_PE_R_ILMNINDX7	CAAGCAGAAGACGGCATACGAGATAAGCTATCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_PE_R_ILMNINDX8	CAAGCAGAAGACGGCATACGAGATGTAGCCTCGACTCTAGATGGGATTAAACGGGGA GAC
BARCODE_SEQ_F	GACCACCGAGATCTACACAGTCGCCTATACGGTGATGG
BARCODE_SEQ_INDEX	GTCTCCCCGTTTAATCCCATCTAGAGTCGA
TAG_OLIGO	GTGTAATAATTCTAGAAGCTTAGTCGCCTATACGGTGATGGNNNNNNNNNNNNNNNN NNNNGTCTCCCCGTTTAATCCCATCTAGAGTCGGGGCGG
TAG_EXTEND	CCGCCCCGACTCTAGATG

Supplementary Figures

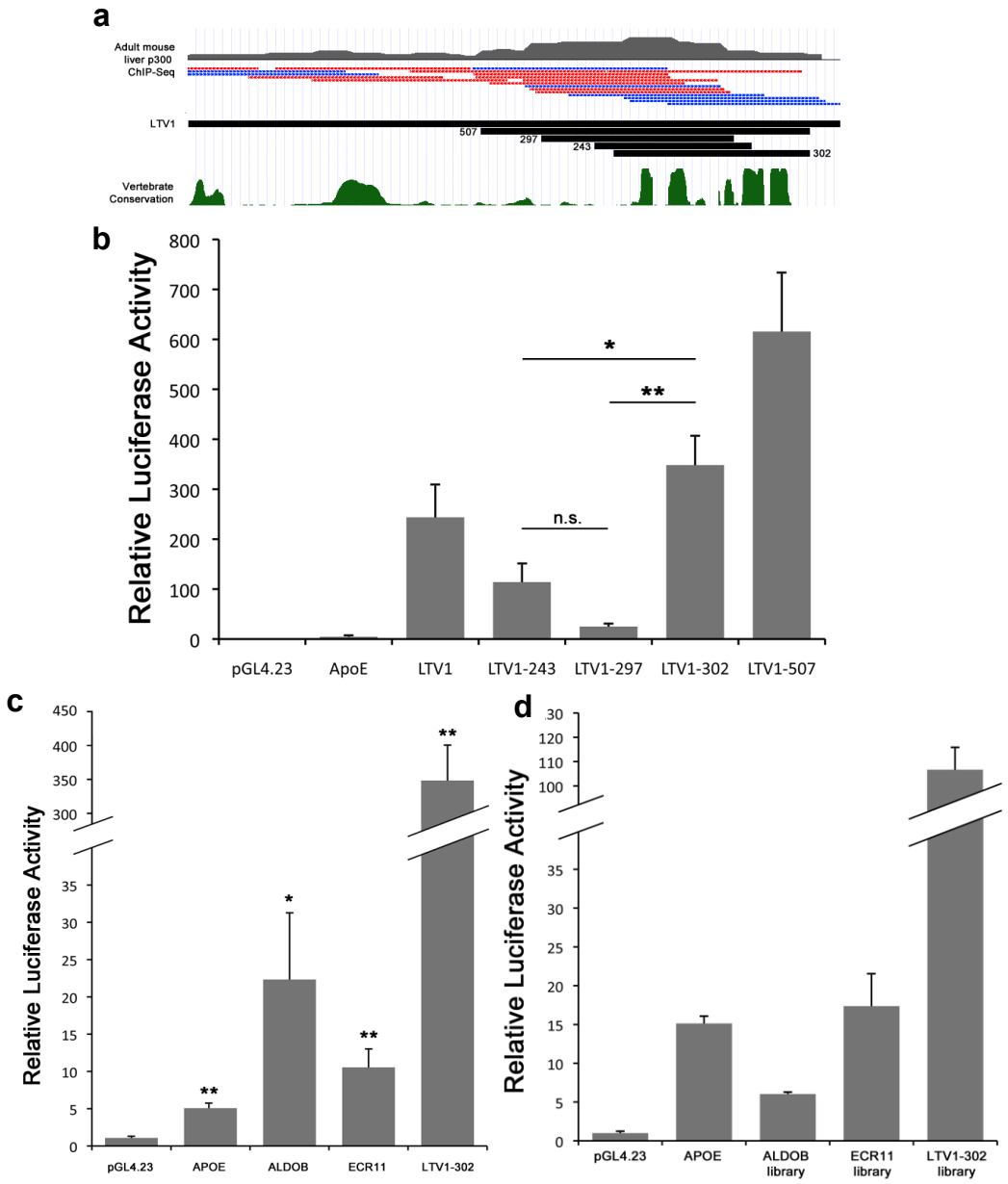


Figure C.1. Activity of wild-type enhancers and variant pools by tail vein luciferase assay.

(a) Identification of LTV1 based on p300 ChIP-Seq from early adult mouse liver and position of deletion fragments constructed to refine enhancer position. The labels indicate the name and size (bp) of each fragment (b) Relative luciferase activity driven by the various LTV1 fragments compared to the APOE liver enhancer and minimal promoter only (pGL4.23). *:p<0.05, **:p<0.01, One way analysis of variance (ANOVA) with Tukey post-hoc test to compare groups. (c) Relative luciferase activity driven by the three wild-type enhancers used in this study compared to the APOE liver enhancer and minimal promoter only (pGL4.23). *:p<0.05, **:p<0.01, Student's unpaired two tailed t-test (d) Aggregate relative luciferase activity driven by a pool of all the enhancer haplotypes for each of the three enhancers under study and compared to the ApoE liver enhancer and minimal promoter only (pGL4.23).

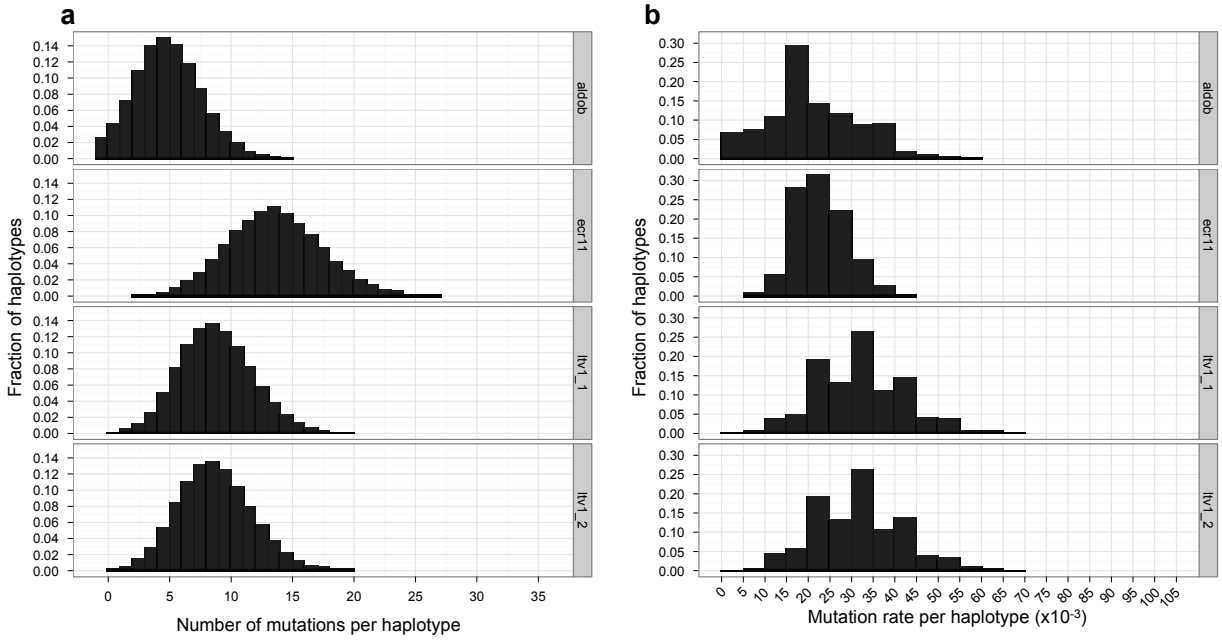


Figure C.2. Distribution of mutations per enhancer haplotype.

The fraction of enhancer haplotypes containing a given number of mutations (a) and the fraction of enhancer haplotypes with a given per-base mutation rate (b).

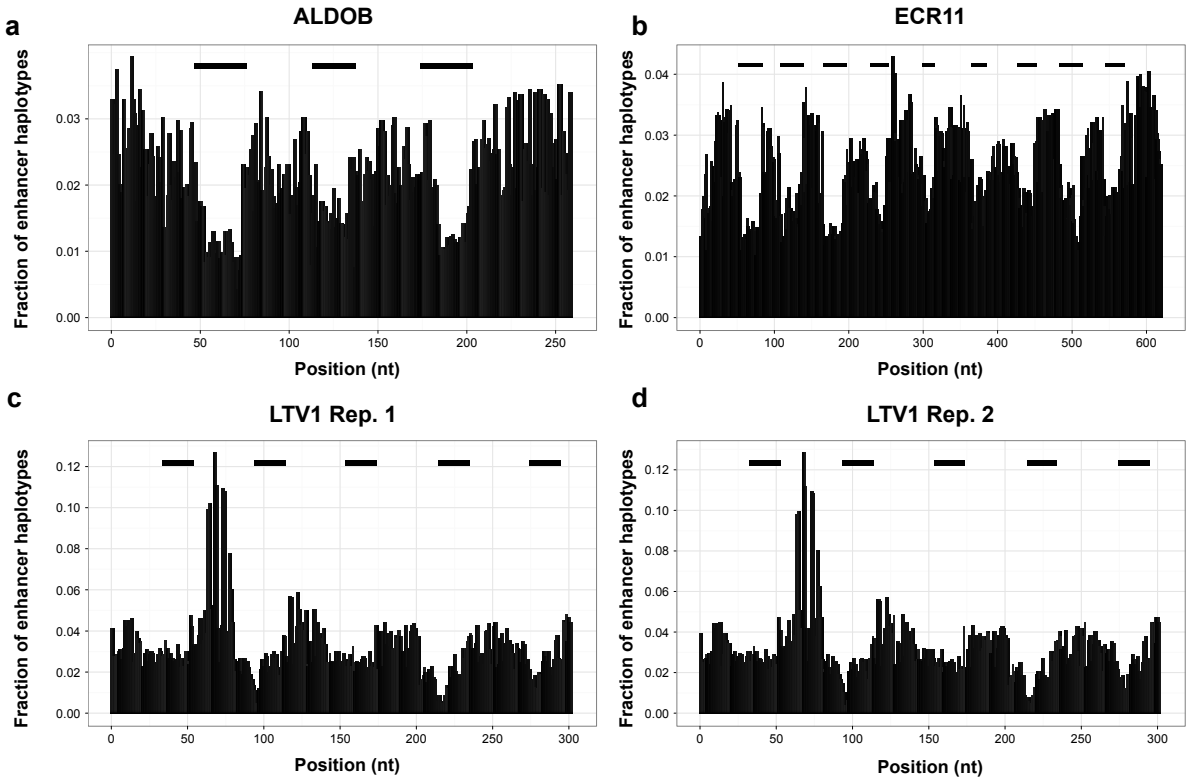


Figure C.3. Distribution of mutations by position in enhancer haplotypes.

Per-base mutation rate as a function of position in the enhancer for ALDOB (a), ECR11 (b), and the two replicates of LTV1 (c, d). As would be expected, dips in mutation rate correspond to overlap regions during the PCA process (horizontal gray bars). Nonetheless, all possible substitution mutations were observed in at least 42 distinct enhancer haplotypes and all pairs of positions were disrupted together in at least one haplotype with the exception of a single pair of positions in LTV1.

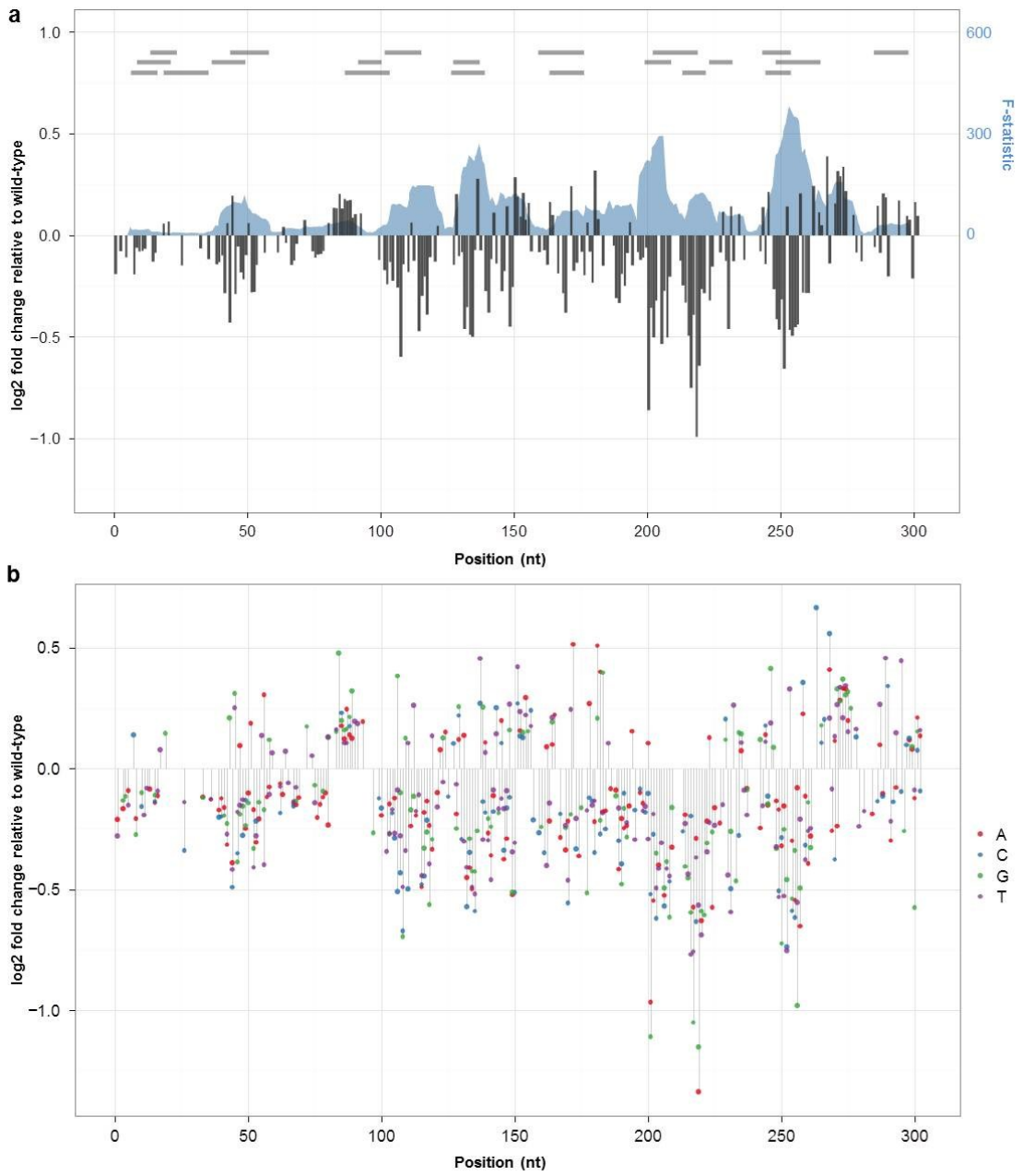


Figure C.4. Mutation effect size in the second replicate of LTV1.

Position-specific mutation effect sizes based on coefficients from univariate (grey columns, left axis) (a) and trivariate models (A:red, C:blue, G:green, T:purple) (b) are plotted here. Effect sizes were calculated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide. Effect sizes are only shown for positions where model coefficients had associated p-values less than or equal to 0.01. Multiple linear regression was used to predict the number of aliquots in which a given enhancer haplotype was observed, using sets of 10 adjacent positions (coded as binary vectors based on whether a mutation was present in each enhancer haplotype) as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (blue shadow, right axis) (a). The locations of TFBS predictions using the MATCH web server are shown as horizontal grey bars at the top of the plot.

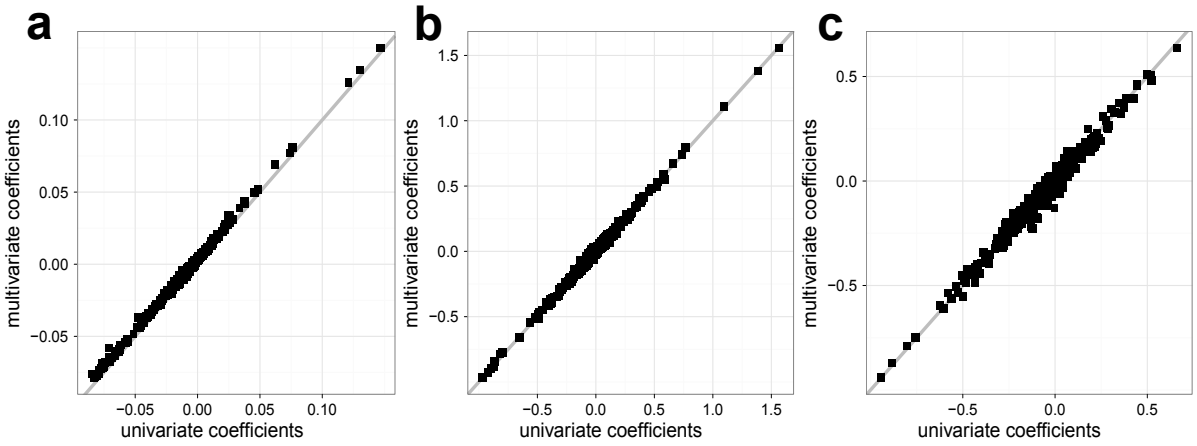


Figure C.5. Comparison between univariate and multivariate linear regression coefficients.

Coefficients calculated via univariate linear regression (i.e. only considering mutational status at a single position) are plotted against coefficients calculated via multivariate linear regression (simultaneously considering mutational status at all sites in the enhancer) for ALODB (a), ECR11 (b), and LTV1 (c). The line y=x is shown in gray in all three plots.

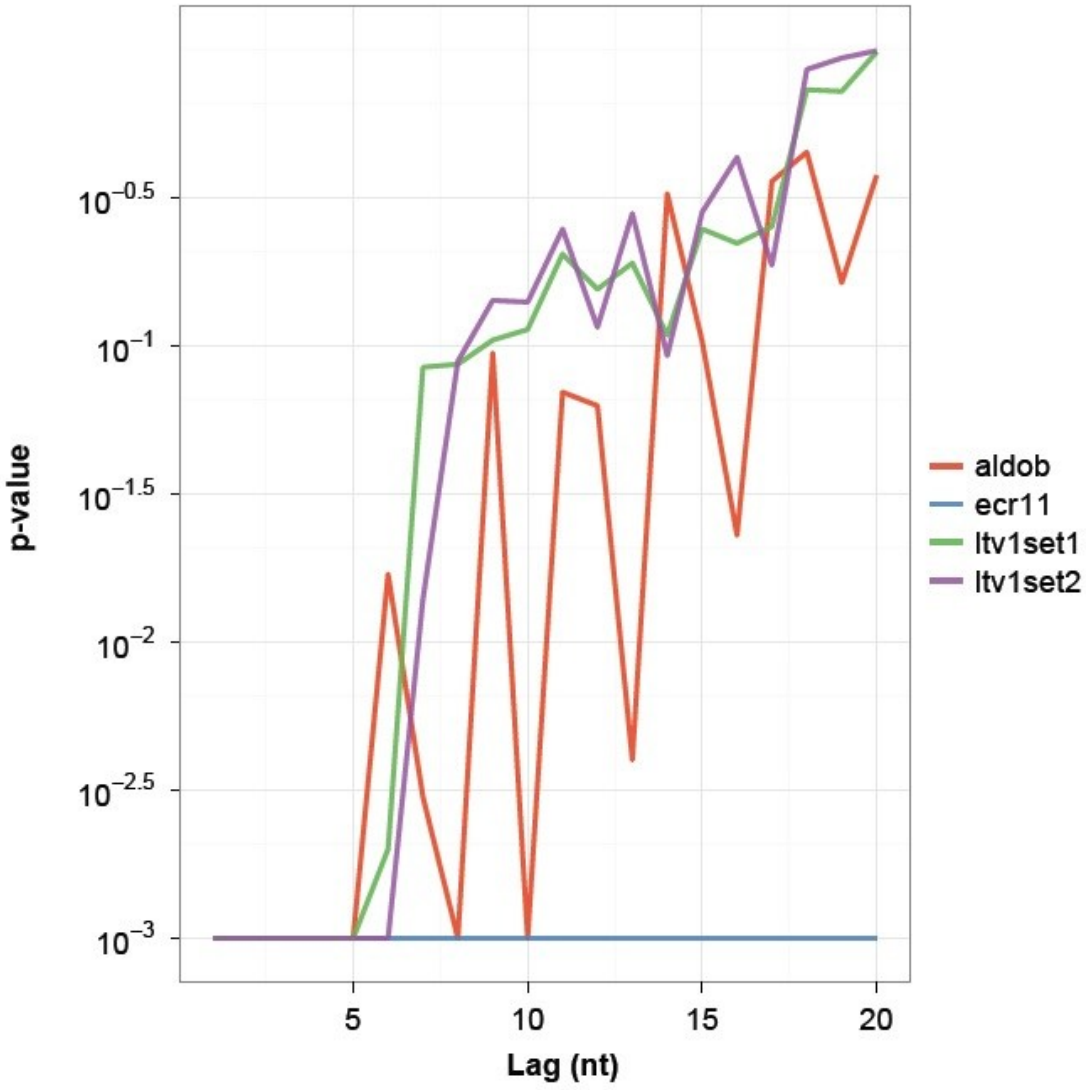


Figure C.6. P-value for the similarity of effect sizes of nearby positions.

To assess the similarity of the effect sizes of mutations at nearby positions in each enhancer, we summed the absolute difference between effect sizes at all positions separated by a fixed "lag" distance. We then recalculated this quantity 1000 times after randomly permuting the effect sizes. We obtained a p-value by calculating the fraction of times that the quantity computed on the permuted effect sizes was at least as small as the quantity computed on the real data. This was repeated for a range of values of the lag distance. The p-value is plotted here as a function of the lag distance. Positions separated by ~5 nucleotides or fewer show substantially similar effect sizes ($p<0.01$) across all three enhancers assayed.

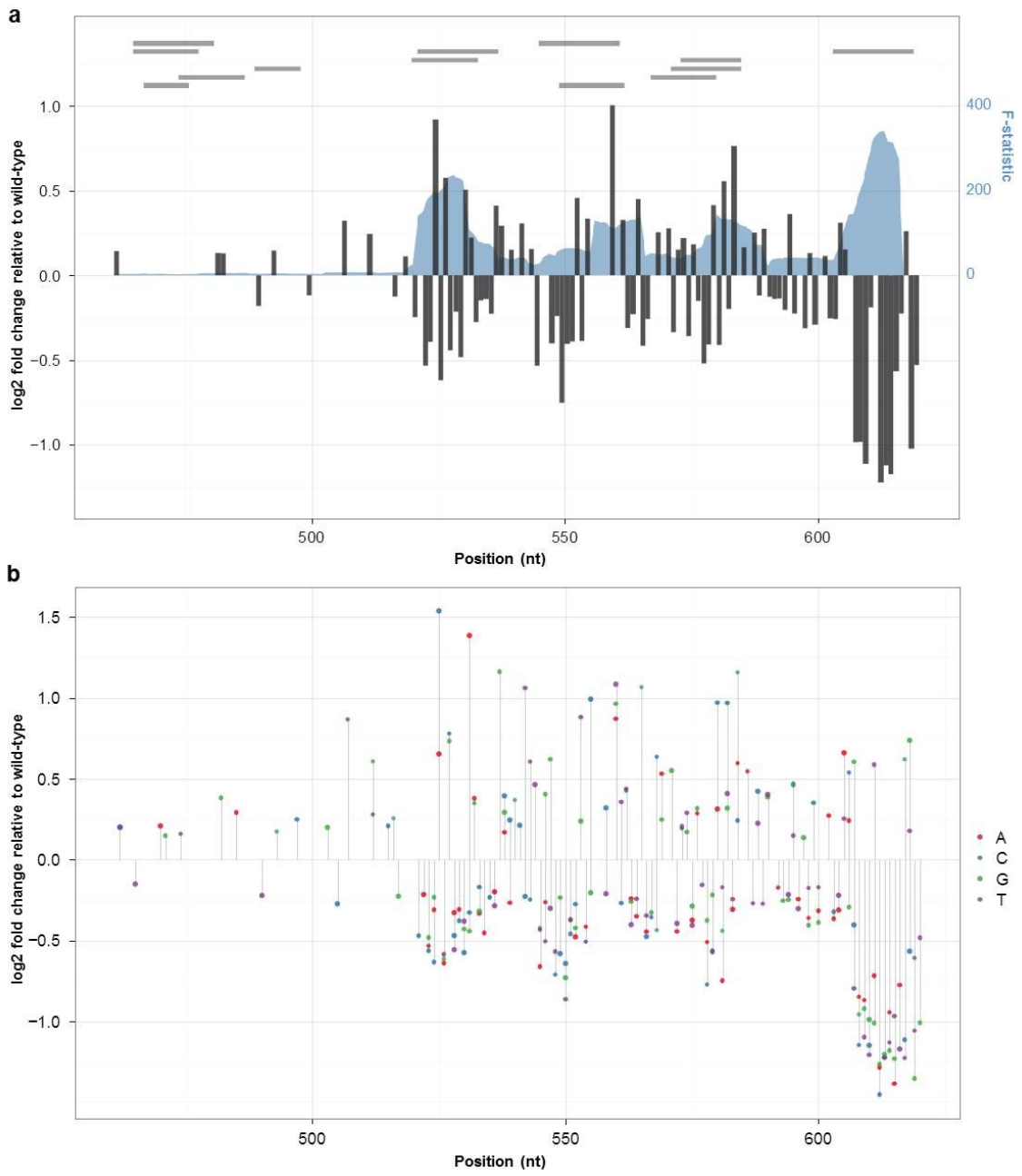


Figure C.7. Mutation effect size in the distal 160 nt of ECR11.

Position-specific mutation effect sizes based on coefficients from univariate (grey columns, left axis) (a) and trivariate models (A:red, C:blue, G:green, T:purple) (b) are plotted here. Effect sizes were calculated by taking the log2 of the ratio of the number of aliquots predicted by the model with a mutation to the number of aliquots predicted for the wild-type nucleotide. Effect sizes are only shown for positions where model coefficients had associated p-values less than or equal to 0.01. Multiple linear regression was used to predict the number of aliquots in which a given enhancer haplotype was observed, using sets of 10 adjacent positions (coded as binary vectors based on whether a mutation was present in each enhancer haplotype) as predictors. The F-statistic of these models, representing the extent to which the model is predictive of the outcome, is plotted (blue shadow, right axis) (a). The locations of TFBS predictions using the MATCH web server are shown as horizontal grey bars at the top of the plot.

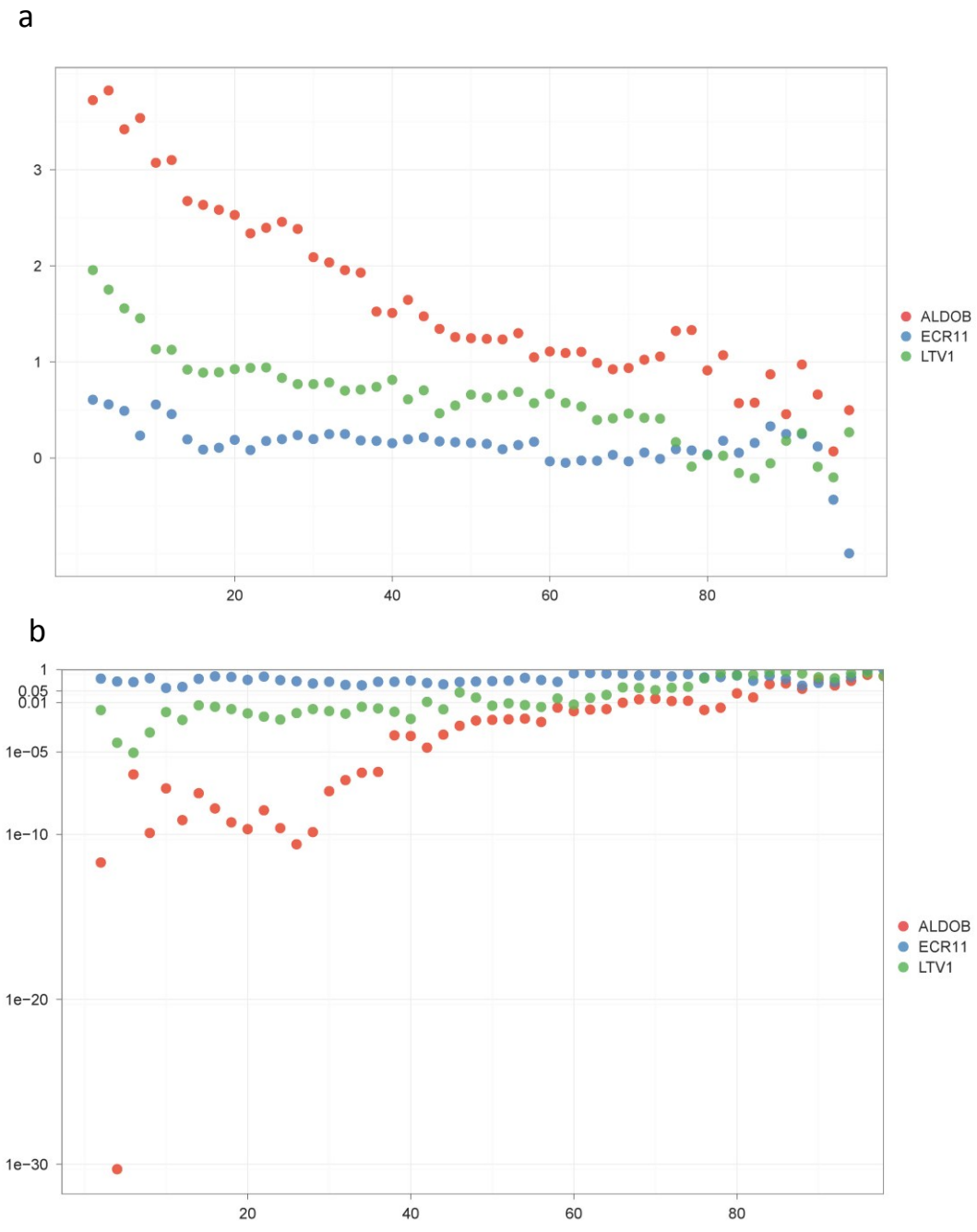


Figure C.8. Single nucleotide relationships between evolutionary and functional constraint.

Positions were rank-ordered based on the absolute vale of their effect size (from position-based, i.e. univariate, linear models) and the difference in the mean conservation score for the top *x* percent of positions versus the mean conservation score for the bottom 100-*x* percent of positions is shown (a). For example, a cutoff at the tenth percentile separates the highest impact ten percent of positions from the lowest impact ninety percent. A *t*-test was then performed to compare the means of the two distributions of conservation scores for a given impact threshold cutoff and the p-values associated with each test are shown in (b). The highest impact mutations tend to be significantly more conserved than the remainder of positions for ALDOB and ECR11.

Supplementary Note 1: Multiple linear regression on entire haplotypes

While linear models constructed on a position-by-position basis best represent the effect size of individual mutations, they may not perform optimally as predictors of the transcriptional activity of entire haplotypes, which contain many such mutations. To assess the ability of models constructed from our data to predict overall haplotype activity, we built two multiple linear regression models for each enhancer. The first model was composed of *n* binary variables (where *n* is the length of the enhancer) for whether or not a position was wild-type in an enhancer haplotype, and the second model was composed of 3*n* binary variables for whether a position was a particular mutant nucleotide in an enhancer haplotype (**Table C.1**). While all the models were significant as measured by comparison of mean squared error calculated from actual versus data versus data with the outcome vector permuted ($p<0.01$), the explanatory power of these models (R^2) ranged from 0.03 to 0.3, suggesting that complexity bottlenecking has limited the ability of our models to explain large fractions of the observed variation for entire haplotypes. Specifically, the relatively few numbers of tags with which individual haplotypes are associated, and the relatively few aliquots in which individual tags are observed, adds considerable stochastic noise to the system.

Supplementary Methods

Construction of enhancer haplotypes from short, doped oligonucleotides using PCA

Sets of overlapping oligonucleotides for each enhancer were designed either by manual inspection (LTV1) or using the program DNAWorks (ALDOB and ECR11). Common flanking sequences were included on either side to allow for amplification of the full-length enhancer haplotypes during PCA. For LTV1, two versions of overlapping oligonucleotides were designed, such that the overlap region in each was different. Oligonucleotides were synthesized by Integrated DNA Technologies (IDT). All positions corresponding to the enhancer region were synthesized using a hand-mix doped at a ratio of 97:1:1:1 (that is, designated base at a frequency of 97%, and every other base at a frequency of 1%). Sequences of all oligonucleotides are listed in **Table C.5**.

For ALDOB as well as ECR11, the full-length haplotypes were assembled in a single step. We used 50 fmol of each oligonucleotide (ALDOB_PCA_OLIGO[1...6] or ECR11_PCA_OLIGO[1...12]) in a 25 µl PCR reaction volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), and 0.5× SYBR Green II, with the following cycling conditions: 95 °C for 3 min; followed by 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Four such reactions were carried out in parallel and then pooled together for each enhancer. The PCR product representing a complex pool of enhancer haplotypes was purified using QIAquick columns (Qiagen). The assembled enhancer haplotypes were then subjected to an additional around of PCR to add 15 bp of vector homology on either side to render them competent for cloning using InFusion (Clontech). We used 20 ng of template in a 25 µl PCR reaction volume with 1× KapaHiFi Hot Start Ready Mix, 0.5× SYBR Green II, and each primer (VH_F and VH_R) at 0.3 µM final concentration. Thermal cycling was done with the following program: 95 °C for 3 min; followed by 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. Sixteen such reactions were carried out in parallel and then pooled together for each enhancer. The PCR product was purified using QIAquick columns (Qiagen).

The two LTV1 designs were assembled separately. For each design, pairs of oligonucleotides, that is, oligonucleotides 1 and 2, oligonucleotides 3 and 4, and oligonucleotides 5 and 6, were each assembled in parallel and the products of the three reactions were then assembled together into the final product in a single reaction as follows:

	Templates	**Primers**
Step 1, Reaction 1	LTV1_PCA[1/2]_OLIGO1, LTV1_PCA[1/2]_OLIGO2	LTV1_PCA[1/2]_P1, LTV1_PCA[1/2]_P2
Step 1, Reaction 2	LTV1_PCA[1/2]_OLIGO3, LTV1_PCA[1/2]_OLIGO4	LTV1_PCA[1/2]_P3, LTV1_PCA[1/2]_P4
Step 1, Reaction 3	LTV1_PCA[1/2]_OLIGO5, LTV1_PCA[1/2]_OLIGO6	LTV1_PCA[1/2]_P5, LTV1_PCA[1/2]_P6
Step 2	Products of reactions 1, 2, and 3	LTV1_OUTER_F, LTV1_OUTER_R

Each 50 µl PCR reaction was prepared on ice with 1× iProof Ready Mix (Bio-Rad), 0.5× SYBR Green II, forward and reverse primers each at 0.5 µM final concentration and 50 fmol of each template oligo. Thermal cycling was done in a MiniOpticon Real-time PCR system (Bio-Rad) with the following program: 98 °C for 30 s, followed by 30 cycles of 98 °C for 10 s, 62 °C for 30 s and 72 °C for 15 s. Each sample was monitored and extracted from the PCR machine when fluorescence began to plateau. PCR products were purified on a QIAquick column (Qiagen). The haplotypes obtained from each of the two LTV1 designs were pooled after the PCA step. Two aliquots were drawn from this pool, and then carried through subsequent steps as two independent samples and were associated with entirely different sets of tags.

Cloning of enhancer haplotypes and the degenerate tag into pGL4.23 plasmid

For ALDOB and ECR11, we first cloned in the degenerate tag to create a complex library of tagged pGL4.23 plasmids. We then cloned in the enhancer haplotypes into these tagged pGL4.23 plasmids. For LTV1, we first cloned in the enhancer haplotypes and then cloned in the degenerate tag. Details of each cloning step remained the same, irrespective of the order in which they were carried out, and are described below.

Cloning of degenerate tag into pGL4.23 plasmid

The tag oligonucleotide (TAG_OLIGO) was made double-stranded using primer extension in a 50 µl reaction volume with 1× iProof Master Mix, 0.5 µg single-stranded tag oligo, 0.5 µg reverse primer (TAG_EXTEND). The reaction was incubated at 95 °C for 3 min, 61 °C for 10 min and then 72 °C for 5 min. The product was purified using a QIAquick column and eluted in 50 µl EB. It was further subjected to ExoI treatment in 40 µl reaction volume for 1 h at 37 °C to degrade any remaining single-stranded DNA, and purified again using QIAquick columns. The resulting double-stranded tag oligo was then cloned into pGL4.23 at the XbaI site (at 1,799 bp) using standard InFusion (Clontech) protocol. The InFusion reaction was diluted to 100 µl using TE8. We used 1.5 µl of this diluted cloning reaction to transform 50 µl of chemically competent FusionBlue cells (Clontech) using the standard protocol. When the tag was being cloned in first, 16 such transformation reactions were pooled and grown overnight in four 50-ml liquid cultures at 37 °C in a shaking incubator. DNA was extracted using the Invitrogen Charge Switch Mini Prep Kit for ALDOB and ECR11, and the Invitrogen Charge Switch Midi Prep Kit for LTV1.

Cloning enhancer haplotypes into pGL4.23 vector

The enhancer haplotypes were cloned into the EcoRV site (at 42 bp) of the pGL4.23 plasmid, using standard InFusion protocol. We used 1.5 µl of the cloning reaction to transform 50 µl of chemically competent FusionBlue cells using standard protocol. Five transformations reactions were pooled and grown overnight in 50 ml liquid cultures at 37 °C in a shaking incubator. DNA was extracted using the Invitrogen Charge Switch Mini Prep Kit for ALDOB and ECR11, and the Invitrogen Charge Switch Midi Prep Kit for LTV1.

Tail vein injections

Enhancers were injected using methods as previously described [106]. Briefly, each library was injected into mice using the TransIT EE Hydrodynamic Gene Delivery System (Mirus Bio) following the manufacturer's protocol. We injected 10 μg of each library, alongside 2 μg of pGL4.74[hRluc/TK] vector to correct for injection efficiency, into the tail vein of CD1 mice (Charles River). After 24 h, mice were euthanized and livers were harvested.

Measurement of luciferase activity

Firefly and renilla luciferase activity were measured on a Synergy 2 Microplate Reader (BioTek Instruments) for each liver using the Dual Luciferase Reporter Assay System (Promega). The firefly luciferase to renilla luciferase ratios were determined and expressed as relative luciferase activity. All mouse work was approved by the UCSF Institutional Animal Care and Use Committee.

Isolation of RNA from mouse livers

Fresh liver tissue was immediately stabilized in RNAlater solution (Ambion). Samples were homogenized in TRIzol reagent (Invitrogen) and RNA was isolated from the samples according to the manufacturer's instructions.

DNase treatment of RNA

To remove any DNA contamination in the RNA extracted from mouse livers, it was subjected to DNaseI treatment using DNA-free (Ambion). Each reaction was prepared with 1× DNA-free buffer, 1 μl of rDNaseI enzyme, 10 μg of RNA and RNase-free water to 50 μl. The reactions were incubated at 37 °C for 1 h, with an additional 1 μl of enzyme added mid-way through the incubation. The reaction was stopped by adding 7 μl of the inactivation reagent and incubating for 2 min at 25 °C with frequent shaking. The reaction was centrifuged in a microcentrifuge at 10,000g for 1.5 min, and the supernatant containing RNA was carefully transferred to a fresh tube.

RT-PCR

Aliquots of RNA obtained after DNase treatment were reverse transcribed to cDNA and amplified by PCR using the Qiagen One-Step Kit. The PCR sought to amplify the 20-bp degenerate tag encoded at the 3′ end of the luciferase transcript. The reactions were assembled on ice in a 25 μl total volume with the following reagents: 1× Qiagen One-Step RT-PCR buffer, 400 μM of each dNTP, 0.6 μM of forward primer (BARCODE_PE_F), 0.6 μM of relevant reverse

primer (BARCODE_PE_R_ILMN_INDEX[1-8]), 0.5× SYBR Green II and 5 µl (~1 µg) of RNA template. Thermal cycling was done on a Bio-Rad MiniOpticon Real-Time PCR system with the following program: 50 °C for 30 min (reverse transcription), 95 °C for 15 min (inactivation of reverse transcriptase and heat-activation of the DNA polymerase), then 30 cycles of 94 °C for 30 s, 65 °C for 30 s and 72 °C for 30 s. Each reaction was monitored and extracted from the PCR machine when the fluorescence began to plateau. The cDNA products were purified using the QIAquick PCR Purification Kit (Qiagen) and eluted in 35 µl EB. The primers used for the RT-PCR contained the necessary sequences for compatibility with the Illumina flow-cell. Thus, the cDNA library obtained at the end of this step was ready for sequencing, eliminating the need for a separate sequencing-library construction step. The reverse primer additionally included 6 bp barcodes allowing for several RT-PCR reactions to be pooled into a single lane for sequencing.

Sequencing of RNA-derived tags

The pooled RT-PCR reaction products were sequenced on an Illumina GAIIx using a sequencing primer (BARCODE_SEQ_F) designed to read into the tag sequence. Each run was 36 cycles with an additional 6 cycles to read the indexing barcode using the index sequencing primer (BARCODE_SEQ_INDEX).

For each aliquot, reads were filtered based on the quality scores for the first 20 bases, which correspond to the degenerate tag. The numbers of occurrences of each tag were counted and tags that were supported by at least ten reads were classified as being 'present' in that aliquot.

Associating tags with enhancer haplotypes

The enhancer haplotypes and tags were situated more than 1,000 bp away from each other on the pGL4.23 plasmid. To bring them adjacent and facilitate the subassembly method, we digested the pGL4.23 plasmids using HindIII, which had two cut sites, one just 3′ of the enhancer, and one just 5′ of the tag, thus resulting in excision of the intervening region. Cut site 1 was already a part of the pGL4.23 backbone. Cut site 2 was engineered in as a part of the tag oligo. The digest was carried out in a 50 µl volume with 1× NEB Buffer 2, 1 µg of plasmid and 1 µl of HindIII Enzyme (New England BioLabs) and incubated at 37 °C for 3 h. The digested plasmid was purified using a QIAquick column.

The digested plasmids were then recircularized using intramolecular ligation, resulting in the tag becoming adjacent to the 3′ end of the enhancer. Ligation was performed using T4 DNA ligase (New England BioLabs) in a 20 µl reaction with 15 ng of template per reaction. The reaction was incubated for 15 min at 25 °C, followed 20 min at 65 °C to inactivate the ligase.

The enhancer and tag region were amplified from recircularized plasmids using PCR with the forward primer targeting the region immediately 5′ of the enhancer (ENHANCER_F for ALDOB and ECR11, and LTV1_F for LTV1) and the reverse primer targeting the region immediately 3′ of the tag (BARCODE_PE_R). The reaction was carried out in a 25 µl volume with 1× KapaHiFi

Hot Start Ready Mix (Kapa BioSystems), 0.5× SYBR Green II, 5 µl of the ligation reaction, and each primer at 0.3 µM final concentration. Thermal cycling was done using Bio-Rad MiniOpticon Real-Time PCR system using the following program: 95 °C for 3 min; and then 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each reaction was monitored and removed from the PCR machine when the fluorescence began to plateau. The reactions were then pooled and purified using QIAquick columns.

The amplicons were then subjected to the subassembly protocol as conceptually described in [107] with some modifications as follows. The random fragmentation step was carried out using the Nextera Tn5 transposase (EpiCentre) instead of mechanical shearing. The Nextera reaction was purified using MinElute column (Qiagen) and size-selected by PAGE (LTV1: 100+; ECR11:100-300,300+; ALDOB: no size-selection performed). The size-selected fragments were subjected to PCR in a 25 µl reaction volume with 1× KapaHiFi Hot Start Ready Mix (Kapa BioSystems), 0.5× SYBR Green II, 5 µl of the ligation reaction, Nextera Adaptor 1 at 10 nM final concentration, and primers Nextera BP1 and BARCODE_PE_R at 0.3 µM final concentration each. Thermal cycling was carried out using BioRad Mini Opticon System using the following program: 95 °C for 3 min; and then 30 cycles of 98 °C for 20 s, 65 °C for 15 s, 72 °C for 15 s. Each reaction was monitored and removed from the PCR machine when the fluorescence began to plateau. The PCR products were purified using a QIAquick column and then sequenced on either an Illumina GAIIx or a Hi-Seq 2000. Read1 collected 76 bp/101 bp of the enhancer sequence staring at random breakpoints along the enhancer. Read 2 collected the 20-bp tag sequence.

The reads were then grouped by tag. Reads belonging to each group were then aligned to the wild-type enhancer sequence to identify the mutations on the haplotype associated with that tag using a custom analysis framework.

Estimation of effect size of mutation at each position along the enhancer (univariate model)

All linear regression analyses were done using the lm() or lsfit() functions available in the R Statistical Package. To quantity the effect of mutation at any given position on the number of aliquots in which an enhancer haplotype was observed, we built a separate linear regression model at every position along the enhancer, with a single predictor representing whether the given position was wild type or mutant. The predictor was thus a binary variable representing presence (1) or absence (0) of a mutation at that position.

$$y_i = \beta_{0j} + \beta_{1j} X_{ij}$$

where,

y_i = number of aliquots in which the *i*th haplotype was observed (referred to as aliquot counts), and

X_{ij} = 1 if position *j* was mutant and 0 if position *j* was wild type in the *i*th haplotype.

To facilitate comparison between positions and between enhancers, we calculated the effect size of mutation at a position *j* as

$$\log_2\left(\frac{\beta_{0j}+\beta_{1j}}{\beta_{0j}}\right)$$

The P-value reported by the model for β_{1j} was used to judge whether the effect size was significant.

For LTV1, as a single haplotype was typically associated with multiple tags, we normalized the aliquot counts for a given haplotype by dividing by the number of tags associated with that haplotype. In the case of ALDOB and ECR11, as the enhancer haplotypes were cloned in second, almost all haplotypes were associated with single tags, and thus the aliquot counts for tags were used directly as the aliquot counts of their linked haplotypes.

Estimation of effect size of each specific nucleotide change at each position along the enhancer (trivariate model)

To explore whether the estimated effect sizes for each position were being driven by specific nucleotide substitutions, we modified the model just described to include three predictors, each representing one of the three possible nucleotide substitutions at that position. The factors were set up as binary variables representing the presence (1) or absence (0) of the particular change at that position.

$$y_i = \beta_{0j} + \beta_{1j}X_{ij_1} + \beta_{2j}X_{ij_2} + \beta_{3j}X_{ij_3}$$

Effect sizes were then calculated from the coefficients produced by the models as follows (for *k* = 1,2,3):

$$\log_2\left(\frac{\beta_{0j}+\beta_{kj}}{\beta_{0j}}\right)$$

The P-value reported by the model for β_{kj} was used to judge whether the effect of a given nucleotide substitution at a given position was significant.

Spatial structure

To quantify whether nearby positions tend to have similar effect sizes, we calculated the sum of the absolute values of the differences in effect sizes between positions located at a given distance (lag) from each other. In other words, we calculated

$$S(k) = \sum_{j=k+1}^{N} |r_j - r_{j-k}| ,$$

where k = 1, 2, ..., 20 denotes the lag, N denotes the length of the enhancer, and r_i is the effect size of position *i*.

For each value of the lag k, we also calculated $S_{1^*}(k),...,S_{1000^*}(k)$, each of which measures the sum of the absolute values of the differences in effect sizes between positions at a distance k from each other, after permuting the effect sizes ($r_1,...,r_N$). We then calculated a P-value associated with each value of the lag k as the fraction of the $S_{1^*}(k),...,S_{1000^*}(k)$ that was as small or smaller than S(k).

Models to estimate combined predictive power of blocks of adjacent positions

To further characterize the nature of the spatial structure of the effect sizes and to explore whether certain regions along the enhancer were enriched for positions with larger effect sizes, we focused on blocks of adjacent positions in a 10-bp sliding window along the length of the enhancer. For each window, we built a multiple linear regression model with one predictor for each position within the window. Each predictor was set up as a binary variable denoting the presence (1) or absence (0) of mutation at that position. The response variable y was the number of aliquots in which a given haplotype was seen.

$$y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{i(j+1)} + \ldots + \beta_{10} X_{i(j+9)}$$

The F-statistic from each model was used as a measure of the collective predictive power of positions within each window.

Multiple linear regression models based on the entire haplotype

The multiple linear regression model included one predictor for each position along the enhancer, encoded as a 1 or 0 to indicate presence or absence of a mutation at that position on a given haplotype, and the response variable y represented the number of aliquots in which the haplotype was observed. Here N is the number of positions within a given enhancer.

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_N X_{iN}$$

A P-value for the model was calculated by comparing the mean squared error (MSE) of the model to MSEs of 200 models built using randomly shuffled versions of the response variable. A P-value for the model was estimated by calculating the fraction of times that the MSE for models built using a shuffled response vector was at least as small as the MSE computed using real data.

We then expanded the model, such that each position was represented by three predictors to indicate which of the three possible nucleotide substitutions was observed at that position.

$$y_i = \beta_0 + \beta_{1j}X_{i1_1} + \beta_{2j}X_{i1_2} + \beta_{3j}X_{i1_3} + \cdots + \beta_{1j}X_{iN_1} + \beta_{2j}X_{iN_2} + \beta_{3j}X_{iN_3}$$

A P-value for the model was calculated by repeatedly permuting the outcome vector as described immediately above; however, only 100 permutations were used, due to the high computational burden of constructing this model.

Identification of epistatic interactions (that is, nonadditive effects) among pairs of mutations

For each pair of positions, we built a linear multiple regression model with three predictors: one predictor each to indicate the presence (1) or absence (0) of a mutation at each of the two positions and a third (referred to as the “interaction term”) whose value was set to 1 if both positions were mutant on the given haplotype and 0 otherwise. Only pairs of positions that were both mutant on at least twenty haplotypes were considered.

$$y_i = \beta_{0jk} + \beta_{1jk}X_{ij} + \beta_{2jk}X_{ik} + \beta_{3jk}X_{ij}X_{ik}$$

We used the P-values for the interaction terms for the resulting models to calculate a FDR for each interaction term (using the p.adjust() function in R, with method = “BH”). Interaction terms with FDR < 0.05 were considered significant and used for downstream analyses of epistatic interactions.

Bibliography

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

2. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

3. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.

4. Kitzman, J.O., et al., *Noninvasive whole-genome sequencing of a human fetus.* Sci Transl Med, 2012. **4**(137): p. 137ra76.

5. Fan, H.C., et al., *Non-invasive prenatal measurement of the fetal genome.* Nature, 2012.

6. Levy, S., et al., *The diploid genome sequence of an individual human.* PLoS Biol, 2007. **5**(10): p. e254.

7. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.

8. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

9. Wang, J., et al., *The diploid genome sequence of an Asian individual.* Nature, 2008. **456**(7218): p. 60-5.

10. Pushkarev, D., N.F. Neff, and S.R. Quake, *Single-molecule sequencing of an individual human genome.* Nat Biotechnol, 2009. **27**(9): p. 847-50.

11. The 1000 Genomes Project Consortium, *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

12. Kitzman, J.O., et al., *Haplotype-resolved genome sequencing of a Gujarati Indian individual.* Nat Biotechnol, 2011. **29**(1): p. 59-63.

13. Cooper, G.M. and J. Shendure, *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* Nat Rev Genet, 2011. **12**(9): p. 628-40.

14. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.

15. Ng, S.B., et al., *Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome.* Nat Genet, 2010. **42**(9): p. 790-3.

16. Hoischen, A., et al., *De novo mutations of SETBP1 cause Schinzel-Giedion syndrome.* Nat Genet, 2010. **42**(6): p. 483-5.

17. Gilissen, C., et al., *Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome.* Am J Hum Genet, 2010. **87**(3): p. 418-23.

18. Lupski, J.R., et al., *Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.* N Engl J Med, 2010. **362**(13): p. 1181-91.

19. Harakalova, M., et al., *Dominant missense mutations in ABCC9 cause Cantu syndrome.* Nat Genet, 2012. **44**(7): p. 793-6.

20. van Bon, B.W., et al., *Cantu Syndrome Is Caused by Mutations in ABCC9.* Am J Hum Genet, 2012. **90**(6): p. 1094-101.

21. O'Roak, B.J., et al., *Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations.* Nat Genet, 2011. **43**(6): p. 585-9.

22. Xu, B., et al., *Exome sequencing supports a de novo mutational paradigm for schizophrenia.* Nat Genet, 2011. **43**(9): p. 864-8.

23. Vissers, L.E., et al., *A de novo paradigm for mental retardation.* Nat Genet, 2010. **42**(12): p. 1109-12.

24. Stamatoyannopoulos, G., *Control of globin gene expression during development and erythroid differentiation.* Exp Hematol, 2005. **33**(3): p. 259-71.

25. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.

26. Chalevelakis, G., J.B. Clegg, and D.J. Weatherall, *Imbalanced globin chain synthesis in heterozygous beta-thalassemic bone marrow.* Proc Natl Acad Sci U S A, 1975. **72**(10): p. 3853-7.

27. Goldberg, M., *Sequence analysis of Drosophila histone genes*, 1979, Stanford University.

28. Gannon, F., et al., *Organisation and sequences at the 5' end of a cloned complete ovalbumin gene.* Nature, 1979. **278**(5703): p. 428-34.

29. Corden, J., et al., *Promoter sequences of eukaryotic protein-coding genes.* Science, 1980. **209**(4463): p. 1406-14.

30. Smale, S.T. and D. Baltimore, *The "initiator" as a transcription control element.* Cell, 1989. **57**(1): p. 103-13.

31. Burke, T.W. and J.T. Kadonaga, *Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters.* Genes Dev, 1996. **10**(6): p. 711-24.

32. Carninci, P., et al., *Genome-wide analysis of mammalian promoter architecture and evolution.* Nat Genet, 2006. **38**(6): p. 626-35.

33. FitzGerald, P.C., et al., *Clustering of DNA sequences in human promoters.* Genome Res, 2004. **14**(8): p. 1562-74.

34. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2 Pt 1): p. 299-308.

35. Dynan, W.S. and R. Tjian, *Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins.* Nature, 1985. **316**(6031): p. 774-8.

36. Ptashne, M., *Gene regulation by proteins acting nearby and at a distance.* Nature, 1986. **322**(6081): p. 697-701.

37. Su, W., et al., *DNA-looping and enhancer activity: association between DNA-bound NtrC activator and RNA polymerase at the bacterial glnA promoter.* Proc Natl Acad Sci U S A, 1990. **87**(14): p. 5504-8.

38. Bejerano, G., et al., *Ultraconserved elements in the human genome.* Science, 2004. **304**(5675): p. 1321-5.

39. Pennacchio, L.A., et al., *In vivo enhancer analysis of human conserved non-coding sequences.* Nature, 2006. **444**(7118): p. 499-502.

40. Lindblad-Toh, K., et al., *A high-resolution map of human evolutionary constraint using 29 mammals.* Nature, 2011. **478**(7370): p. 476-82.

41. Zhou, Q. and W.H. Wong, *CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.* Proc Natl Acad Sci U S A, 2004. **101**(33): p. 12114-9.

42. Burzynski, G., et al., *Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control.* Genome Res, 2012.

43. Narlikar, L., et al., *Genome-wide discovery of human heart enhancers.* Genome Res, 2010. **20**(3): p. 381-92.

44. Taher, L., et al., *Genome-wide identification of conserved regulatory function in diverged sequences.* Genome Res, 2011. **21**(7): p. 1139-49.

45. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

46. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.

47. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.

48. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.

49. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.

50. Blow, M.J., et al., *ChIP-Seq identification of weakly conserved heart enhancers.* Nat Genet, 2010. **42**(9): p. 806-10.

51. May, D., et al., *Large-scale discovery of enhancers from human heart tissue.* Nat Genet, 2012. **44**(1): p. 89-93.

52. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.

53. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-83.

54. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.

55. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.

56. Crawford, G.E., et al., *Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).* Genome Res, 2006. **16**(1): p. 123-31.

57. Sabo, P.J., et al., *Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays.* Nat Methods, 2006. **3**(7): p. 511-8.

58. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.

59. Hesselberth, J.R., et al., *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nat Methods, 2009. **6**(4): p. 283-9.

60. Wang, H., M. Johnston, and R.D. Mitra, *Calling cards for DNA-binding proteins.* Genome Res, 2007. **17**(8): p. 1202-9.

61. Wang, H., et al., *Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins.* Genome Res, 2011. **21**(5): p. 748-55.

62. Wang, H., et al., *"Calling cards" for DNA-binding proteins in mammalian cells.* Genetics, 2012. **190**(3): p. 941-9.

63. Valen, E., et al., *Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE.* Genome Res, 2009. **19**(2): p. 255-65.

64. Kodzius, R., et al., *CAGE: cap analysis of gene expression.* Nat Methods, 2006. **3**(3): p. 211-22.

65. Berk, A.J. and P.A. Sharp, *Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids.* Cell, 1977. **12**(3): p. 721-32.

66. Sambrook, J. and D.W. Russell, *Molecular cloning : a laboratory manual*. 3rd ed2001, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

67. Myers, R.M., K. Tilly, and T. Maniatis, *Fine structure genetic analysis of a beta-globin promoter.* Science, 1986. **232**(4750): p. 613-8.

68. Zinn, K., D. DiMaio, and T. Maniatis, *Identification of two distinct regulatory regions adjacent to the human beta-interferon gene.* Cell, 1983. **34**(3): p. 865-79.

69. Gorman, C.M., L.F. Moffat, and B.H. Howard, *Recombinant genomes which express chloramphenicol acetyltransferase in mammalian cells.* Mol Cell Biol, 1982. **2**(9): p. 1044-51.

70. de Wet, J.R., et al., *Firefly luciferase gene: structure and expression in mammalian cells.* Mol Cell Biol, 1987. **7**(2): p. 725-37.

71. Chalfie, M., et al., *Green fluorescent protein as a marker for gene expression.* Science, 1994. **263**(5148): p. 802-5.

72. Trinklein, N.D., et al., *Identification and functional analysis of human transcriptional promoters.* Genome Res, 2003. **13**(2): p. 308-12.

73. Cooper, S.J., et al., *Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.* Genome Res, 2006. **16**(1): p. 1-10.

74. Myers, R.M., L.S. Lerman, and T. Maniatis, *A general method for saturation mutagenesis of cloned DNA fragments.* Science, 1985. **229**(4710): p. 242-7.

75. Reitman, M. and G. Felsenfeld, *Mutational analysis of the chicken beta-globin enhancer reveals two positive-acting domains.* Proc Natl Acad Sci U S A, 1988. **85**(17): p. 6267-71.

76. Goodbourn, S., H. Burstein, and T. Maniatis, *The human beta-interferon gene enhancer is under negative control.* Cell, 1986. **45**(4): p. 601-10.

77. Gertz, J., E.D. Siggia, and B.A. Cohen, *Analysis of combinatorial cis-regulation in synthetic and genomic promoters.* Nature, 2009. **457**(7226): p. 215-8.

78. Mogno, I., et al., *TATA is a modular component of synthetic promoters.* Genome Res, 2010. **20**(10): p. 1391-7.

79. Gaal, T., et al., *Saturation mutagenesis of an Escherichia coli rRNA promoter and initial characterization of promoter variants.* J Bacteriol, 1989. **171**(9): p. 4852-61.

80. Singh, K., et al., *Saturation mutagenesis of the octopine synthase enhancer: correlation of mutant phenotypes with binding of a nuclear protein factor.* Proc Natl Acad Sci U S A, 1989. **86**(10): p. 3733-7.

81. Baliga, N.S. and S. DasSarma, *Saturation mutagenesis of the haloarchaeal bop gene promoter: identification of DNA supercoiling sensitivity sites and absence of TFB recognition element and UAS enhancer activity.* Mol Microbiol, 2000. **36**(5): p. 1175-83.

82. Shin, I., et al., *Effects of saturation mutagenesis of the phage SP6 promoter on transcription activity, presented by activity logos.* Proc Natl Acad Sci U S A, 2000. **97**(8): p. 3890-5.

83. Cleary, M.A., et al., *Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis.* Nat Methods, 2004. **1**(3): p. 241-8.

84. Goh, E.B., et al., *Transcriptional modulation of bacterial gene expression by subinhibitory concentrations of antibiotics.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 17025-30.

85. Ponjavic, J., et al., *Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters.* Genome Biol, 2006. **7**(8): p. R78.

86. Wobbe, C.R. and K. Struhl, *Yeast and human TATA-binding proteins have nearly identical DNA sequence requirements for transcription in vitro.* Mol Cell Biol, 1990. **10**(8): p. 3859-67.

87. Leach, K.M., et al., *Characterization of the human beta-globin downstream promoter region.* Nucleic Acids Res, 2003. **31**(4): p. 1292-301.

88. Hillier, L.W., et al., *Whole-genome sequencing and variant discovery in C. elegans.* Nat Methods, 2008. **5**(2): p. 183-8.

89. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.

90. Hamady, M. and R. Knight, *Microbial community profiling for human microbiome projects: Tools, techniques, and challenges.* Genome Res, 2009. **19**(7): p. 1141-52.

91. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.

92. Weinstein, J.A., et al., *High-throughput sequencing of the zebrafish antibody repertoire.* Science, 2009. **324**(5928): p. 807-10.

93. Bentley, G., et al., *High-resolution, high-throughput HLA genotyping by next-generation sequencing.* Tissue Antigens, 2009. **74**(5): p. 393-403.

94. Kalyuzhnaya, M.G., et al., *High-resolution metagenomics targets specific functional types in complex microbial communities.* Nat Biotechnol, 2008. **26**(9): p. 1029-34.

95. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.

96. Stover, C.K., et al., *Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen.* Nature, 2000. **406**(6799): p. 959-64.

97. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

98. Myers, E.W., et al., *A whole-genome assembly of Drosophila.* Science, 2000. **287**(5461): p. 2196-204.

99. Reinhardt, J.A., et al., *De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae.* Genome Res, 2009. **19**(2): p. 294-305.

100. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.

101. Brady, A. and S.L. Salzberg, *Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.* Nat Methods, 2009. **6**(9): p. 673-6.

102. Delcher, A.L., et al., *Alignment of whole genomes.* Nucleic Acids Res, 1999. **27**(11): p. 2369-76.

103. Sabourin, J.C., et al., *An intronic enhancer essential for tissue-specific expression of the aldolase B transgenes.* J Biol Chem, 1996. **271**(7): p. 3469-73.

104. Gregori, C., et al., *Expression of the rat aldolase B gene: a liver-specific proximal promoter and an intronic activator.* Biochem Biophys Res Commun, 1991. **176**(2): p. 722-9.

105. Gregori, C., et al., *In vivo functional characterization of the aldolase B gene enhancer.* J Biol Chem, 2002. **277**(32): p. 28618-23.

106. Kim, M.J., et al., *Functional characterization of liver enhancers that regulate drug-associated transporters.* Clin Pharmacol Ther, 2011. **89**(4): p. 571-8.

107. Hiatt, J.B., et al., *Parallel, tag-directed assembly of locally derived short sequence reads.* Nat Methods, 2010. **7**(2): p. 119-22.

108. Zhang, G., V. Budker, and J.A. Wolff, *High levels of foreign gene expression in hepatocytes after tail vein injections of naked plasmid DNA.* Hum Gene Ther, 1999. **10**(10): p. 1735-7.

109. Kel, A.E., et al., *MATCH: A tool for searching transcription factor binding sites in DNA sequences.* Nucleic Acids Res, 2003. **31**(13): p. 3576-9.

110. Schmidt, D., et al., *Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.* Science, 2010. **328**(5981): p. 1036-40.

111. Loots, G.G., et al., *Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.* Science, 2000. **288**(5463): p. 136-40.

112. Margulies, E.H., et al., *Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome.* Genome Res, 2007. **17**(6): p. 760-74.

113. Visel, A., et al., *Ultraconservation identifies a small subset of extremely constrained developmental enhancers.* Nat Genet, 2008. **40**(2): p. 158-60.

114. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.

115. Patwardhan, R.P., et al., *High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis.* Nat Biotechnol, 2009. **27**(12): p. 1173-5.

116. Botstein, D. and D. Shortle, *Strategies and applications of in vitro mutagenesis.* Science, 1985. **229**(4719): p. 1193-201.

117. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nat Biotechnol, 2012. **30**(3): p. 271-7.

118. Sharon, E., et al., *Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters.* Nat Biotechnol, 2012. **30**(6): p. 521-30.

119. Kinney, J.B., et al., *Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9158-63.

120. Ke, S., et al., *Quantitative evaluation of all hexamers as exonic splicing elements.* Genome Res, 2011. **21**(8): p. 1360-74.

121. Pitt, J.N. and A.R. Ferre-D'Amare, *Rapid construction of empirical RNA fitness landscapes.* Science, 2010. **330**(6002): p. 376-9.

122. Fowler, D.M., et al., *High-resolution mapping of protein sequence-function relationships.* Nat Methods, 2010. **7**(9): p. 741-6.

123. Whitehead, T.A., et al., *Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing.* Nat Biotechnol, 2012. **30**(6): p. 543-8.

124. Searle, P.F., G.W. Stuart, and R.D. Palmiter, *Building a metal-responsive promoter with synthetic regulatory elements.* Mol Cell Biol, 1985. **5**(6): p. 1480-9.

125. Amit, R., et al., *Building enhancers from the ground up: a synthetic biology approach.* Cell, 2011. **146**(1): p. 105-18.

126. Schlabach, M.R., et al., *Synthetic design of strong promoters.* Proc Natl Acad Sci U S A, 2010. **107**(6): p. 2538-43.

127. Liu, W., et al., *Rapid in vivo analysis of synthetic promoters for plant pathogen phytosensing.* BMC Biotechnol, 2011. **11**: p. 108.

128. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

VITA

Rupali Patwardhan was born and raised in Mumbai (Bombay), India, where she graduated with a Bachelor of Engineering degree in Information Technology at the University of Mumbai in 2003. She then moved to Bloomington, Indiana, USA where she earned a Master of Science degree in Bioinformatics from Indiana University in 2005. From 2005 to 2007, she worked as a research scientist at the Center for Genomics and Bioinformatics affiliated with Indiana University, Bloomington, before moving to the Seattle to continue her graduate studies at the University of Washington. Rupali studied under the mentorship of Dr. Jay Shendure, and in 2012 she graduated with a Doctor of Philosophy in Genome Sciences.