DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Assaying 3-Prime Untranslated Region Variants in a Massively Parallel Fashion

Citation

Griesemer, Dustin S. 2017. Assaying 3-Prime Untranslated Region Variants in a Massively Parallel Fashion. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:42061507

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Assaying 3-Prime Untranslated Region Variants in a Massively Parallel Fashion

A dissertation presented

by

Dustin S. Griesemer

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biomedical Informatics

Harvard University

Cambridge, Massachusetts

June 2017

© 2017 Dustin S. Griesemer
All rights reserved.

Dissertation Advisors: Pardis Sabeti and Carl Novina Dustin S. Griesemer

Assaying 3-Prime Untranslated Region Variants in a Massively Parallel Fashion

Abstract

Rapid advances in human genome sequencing technologies over the past decade have unraveled thousands of loci in the genome containing variants that cause inherited disease or have been placed under positive selective pressure to protect against acquired diseases such as infectious agents, climate change, and diet restrictions. Although genic regions, including 3'-Untranslated Regions (3'-UTRs), are highly enriched in disease and selection-associated loci, linkage disequilibrium and the lengthy timescale of low-throughput validation hinder functional follow-up of each locus. To address this challenge, we developed MPRAu, a tool for identifying functional 3'-UTR variants by combining the Massively Parallel Reporter Assay with Polysome Profiling. Interestingly, we find that under standard growth conditions in the HEK293 cell line, short (101bp) 3'-UTR segments have surprisingly little impact on translation. Furthermore, out of 9,754 variants tested (one-seventh of 3'-UTR variants with a minor allele frequency greater than 5%), we identify a shortlist of 115 variants with significant allelic differences in steady-state RNA expression, including some implicated in sodium transport, leptin signaling, and creatinine synthesis. Though functional variation identified by MPRA is concordant with known regulatory mechanisms,

the vast majority of identified variants fall outside of predicted motifs. This unbiased approach provides concrete leads as well as a means to quickly identify a wealth of causal variants within untranslated regions of the genome.

Table of Contents

Abstract ..... **iii**

Table of Contents ..... **v**

List of Figures ..... **ix**

List of Tables ..... **xi**

Acknowledgements ..... **xiii**

Chapter 1 : Introduction **1**

1.1 Millions of Variants, but What Do They Do? 2

1.2 Building the Haystack: GWAS and Positive Selection Studies 4

1.3 Finding the Needle: Causal Variant Identification 7

1.4 Beyond the Genome: Post-transcriptional Regulation 12

1.5 Summary of Thesis 15

Chapter 2 : Development of a Novel Massively Parallel Reporter Assay **17**

2.1 Abstract 18

2.2 Introduction 18

2.3 Results 21

2.3.1 Repurposing the MPRA for the 3'-UTR 21

2.3.2 Maintaining a Diverse Library 24

2.3.3 Technical Challenges 30

2.4 Discussion 35

2.5 Experimental Protocols 38

2.5.1 Oligo Synthesis and Amplification 38

2.5.2 MPRAu Vector Assembly 40

2.5.3 Validation of Outlier Barcodes 40
2.5.4 MPRAu Experiments 41
2.5.5 RNA Extraction and cDNA Synthesis 42
2.5.6 qPCR and Library Construction 43
2.5.7 Oligo Read Alignment 44
2.5.8 Modeling Oligo Counts in the Plasmid Pool 44
2.6 Contributions 45
2.7 Acknowledgements 45
Chapter 3 : A Resource of Common Variation in Untranslated Regions for Selection and Disease 48
3.1 Abstract 49
3.2 Introduction 49
3.3 Results 52
3.3.1 MPRAu: a Massively Parallel Screen of Causal UTR Variants 52
3.3.2 Short 3'-UTR segments Overwhelmingly Function at the RNA Level 55
3.3.3 Sensitive Detection of Functional 3'-UTR Variants 60
3.3.4 Allelic Effects are Congruent with Complementary Approaches 62
3.3.5 A Prioritized Set of Functional 3'-UTR Variants for Selection and Disease 68
3.4 Discussion 70
3.5 Experimental Procedures 75
3.5.1 Variant Selection 75
3.5.2 Oligo Synthesis and Amplification 75

3.5.3 MPRAu Vector Assembly 76
3.5.4 MPRAu Experiments 77
3.5.5 RNA Extraction and cDNA Synthesis 78
3.5.6 qPCR and Library Construction 79
3.5.7 Luciferase Assays 80
3.5.8 Oligo Read Alignment 81
3.5.9 Expression and Allelic Skew Quantification 82
3.5.10 Annotations Used for Enrichment Analysis 83
3.5.11 Conservation Analysis 83
3.5.12 Concordance of Allelic Skews with GTEx Data 84
3.5.13 Selection and GWAS Associations 85
3.6 Contributions 86
3.7 Acknowledgements 86
Chapter 4 : Perspectives 87
4.1: Utility of MPRAu 88
4.1.1 Massively Parallel Validation 88
4.1.2 Sensitivity to Subtle Allelic Differences 88
4.1.2 Unbiased Testing 89
4.2 Limitations of MPRAu 93
4.2.1 Technical Limitations 93
4.2.2 Conceptual Limitations 94
4.3 Future Directions of MPRAu 96
4.3.1 Application to Rare Variants 96

4.3.2 Application to 5’-UTR Variants 97

4.3.3 Application to Specific Post-transcriptional Mechanisms 99

4.3.4 Application to Relevant Physiological Contexts 100

List of Figures

Figure 1.1: Formulae used by Mendel to derive the mechanism of inheritance. .. 3

Figure 1.2: GWAS study design (case-control). 5

Figure 1.3: A selective sweep. 7

Figure 1.4: Direct identification of promoter/enhancer variants. 11

Figure 1.5: MS risk and selection mediated by 3'-UTR shortening. 13

Figure 2.1: Repurposing the MPRA for the 3'-UTR. 22

Figure 2.2: Technical challenges during oligo amplification. 25

Figure 2.3: Monitoring library diversity. 27

Figure 2.4: Composition of oligos in the plasmid pool. 29

Figure 2.5: SV40 promoter contamination results in "outlier" barcode expression.
..... 32

Figure 2.6: RT primer placement creates a translational "buffering" artifact. 34

Figure 3.1: Adapting MPRA for 3'-UTR post-transcriptional variation. 52

Figure 3.2: Experimental reproducibility. 54

Figure 3.3: Isolating regulatory effects of 3'-UTR segments. 57

Figure 3.4: Unbiased detection of functional 3'-UTR variants. 61

Figure 3.5: Enrichment of known regulatory elements. 64

Figure 3.6: Low-throughput validation of functional variants. 66

Figure 4.1: Mutagenesis of MPRAu-identified functional variants. 90

Figure 4.2: Tiling deletions of MPRAu-identified functional variants. 92

Figure 4.3: 5'-UTR Regulatory Mechanisms. 99

Figure A.1: Random hexamer effects on total RNA expression. 106

Figure A.2: Allelic skew p-value distribution. 107

Figure A.3: Plasmid vs RNA dispersions. 108

Figure A.4: Sucrose gradient qPCR. 109

Figure A.5: eQTL directionality. 110

List of Tables

Table 3.1: MPRAu identifies promising candidates for follow-up. 69

Table A.1: Differentially expressed variants identified by MPRAu. 104

This thesis is dedicated to my parents,
David and Seemean Griesemer,
and to my wife, Rani Bhatia.
Thank you for giving me perspective
and for showing me where true strength lies.

Acknowledgements

In so many ways, I have been incredibly fortunate to spend the last five years in the Sabeti and Novina labs at Harvard. One of the biggest lessons I will take away is that life should not be planned, because some of the best things that happen, even for one's career, are completely unplanned. I first met Pardis Sabeti on a train to New Haven. I can still remember how excited I was about the prospect of working with her, because I could immediately sense both her passion for chasing the difficult problems in science and the genuine pleasure she took in working with students. Indeed, few mentors would have allowed me to lead an interdisciplinary tech-dev project at this early stage in my training, or reached out to so many incredible investigators to help guide me. I have been humbled to train alongside such a fearless researcher with such a wonderful perspective on life.

I am also fortunate to have been given the opportunity to take on an interdisciplinary project, because this allowed me to be co-advised by Carl Novina. Few people can match Carl's infectious enthusiasm for science. No stone went unturned during our discussions, and I always left more excited about what I was doing then when I had arrived. Carl's enthusiasm helped me remain motivated through the difficult times, of which there were many, and reminded me of how lucky I was to be doing this work. I am grateful to have been a part of the Novina lab family, especially Cailin Joyce who provided valuable scientific insight and helped me to master the polysome profiling technique, and Maja Janas who trained me in luciferase assays and first introduced me to the

polysome profiling methodology.

Chance would also have it that I started graduate school at the same time that Ryan Tewhey, who first adapted the Massively Parallel Reporter Assay to variant testing, started his postdoctoral work. Ryan helped me conceive of this project, and has been an amazingly patient and generous mentor. Even greater, Ryan has served as a role model for me during this phase of my career; though there was a steep technical and conceptual learning curve involved, working alongside Ryan continually reminded me that mastering it was possible. I am struck by how powerful of a force this has been during my graduate training, and can only hope that I continue to have such wonderful role models to learn alongside and to inspire me in the future.

My interdisciplinary team extended beyond the Sabeti and Novina labs. Early on in my training, Pardis reached out to her undergraduate advisor, Dave Bartel, to assist me. Having developed an interesting in microRNA biology before starting graduate school, I was grateful simply to have an audience with Dave, and could not have anticipated how deeply involved he would become in my work. I am always impressed by how Dave seeks to understand microRNA biology (and related fields) from every possible angle, and by how this ideal is propagated to those working in his lab. Dave was always the first to ask the most difficult questions, and taught me the importance of addressing the difficult questions as quickly as possible. I am also grateful to Daniel MacArthur, whose insight was critical in applying the tool I was developing to better understand human genetic variation, as well as John Rinn, who I always enjoyed sharing war

stories with related to the complementary and inspiring work he was performing in his own lab.

I would also like to thank those who were involved in my graduate (and medical) training outside of the laboratory. I would like to thank Katherine Flannery, Catherine Haskell, Isaac Kohane, and Peter Park for helping me build a rigorous foundation in the multifaceted field of genomics. I would like to thank Amy Cohen, Loren Walensky, Dennis Kim, Patricia Cunningham, and Rick Mitchell for helping me to remember the ultimate goal of my graduate work – to apply this training to help patients in some form one day. I would like to thank the students living on my wing ("The Shire") for keeping me young, my friends for keeping me fun, and my family for keeping me sane.

And finally, I would like to acknowledge my own undergraduate advisors, Dr. Phil Thomas and Dr. Andrey Karamyshev, who taught me to seek out environments of curiosity and camaraderie, and who ultimately inspired me to pursue a scientific career.

Chapter 1: Introduction

1.1 Millions of Variants, but What Do They Do?

In 1866, an Augustinian friar by the name of Gregor Mendel published the results of experiments he had performed on the pea plants growing in the garden plot of a monastery[1]. By breeding thousands of plants, Mendel discovered two important laws in how traits were inherited. The Law of Segregation stated that every organism contains two factors for each trait, one inherited from each of their parents (Figure 1.1). The Law of Independent Assortment stated that factors for distinct traits are inherited independently of one another. These laws provided a concrete mechanism to an age-old observation that traits are passed on from parents to offspring, and sparked the modern field of genetics.

While Mendel provided the long-sought explanation to how traits are inherited from parent to offspring, many questions still remained. For example, if each trait is represented by two factors, how does one explain continuous traits such as human height? We now know that what Mendel described as factors are actually long stretches of DNA called chromosomes. Between any two people, 99.9% of the chromosome sequence is identical. Places where the sequence may differ, called genetic variants (each version of which is referred to as an “allele”), contribute to inherited traits. Continuous traits are encoded by multiple genetic variants, which may be inherited independently if they are far enough apart or on different chromosomes.

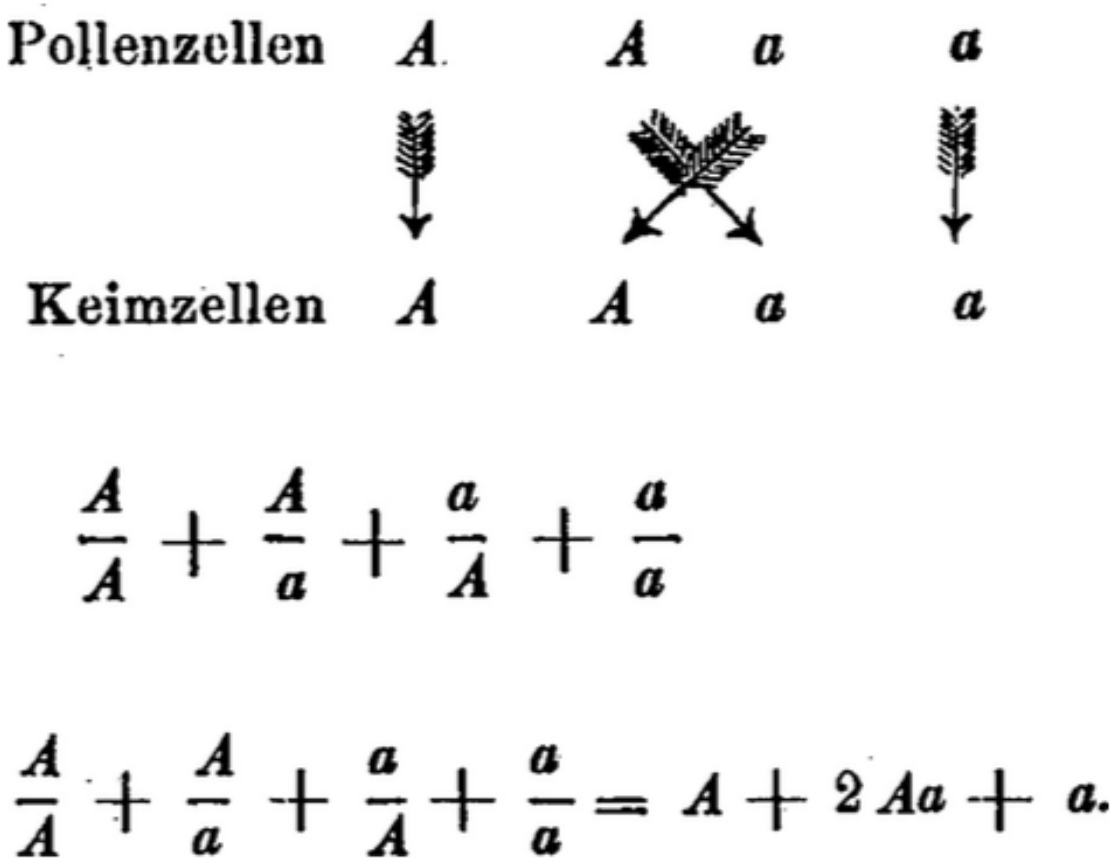


Figure 1.1: Formulae used by Mendel to derive the mechanism of inheritance. In the top diagram, four possibilities are shown for a hybrid (Aa) pollen cell, “Pollenzelen”, fertilizing a hybrid (Aa) ovarian cell, “Keimzellen.” In the middle equation, the genotypes of the offspring are shown. The pollen cell (male gamete) and the ovarian cell (female gamete) each contribute one factor (A or a) to the offspring. The bottom equation depicts how inheritance of factors explains the ratio of offspring Mendel observed in his experiments: 1 pure-breeding type-A to 2 mixed-breeding type-A (Aa) to 1 pure-breeding type-a (aa). Factor (allele) A is dominant[1].

But this sparked another question: what are the variants in the human genome, and which variants are important for each trait? The improvement and reduction of costs of DNA sequencing technology allowed geneticists to address this question in 2008 with the launching of the 1000 Genomes Project[2] (1kG).

1kG ultimately involved sequencing 2,504 whole genomes of individuals from 26 human populations, and led to the characterization of over 88 million variants[3]. 1kG provided a base set of variants to work with, but the latter question proves challenging to this day: what do these variants *do*?

1.2 Building the Haystack: GWAS and Positive Selection Studies

Genome-Wide Association Studies (GWAS) provide a framework for identifying which variants may affect a particular trait. According to the GWAS study design (Figure 1.2), individuals with a given trait (such as heart disease) are genotyped alongside individuals without the trait (i.e., no heart disease). Allele frequencies for each variant can then be compared with the trait of interest in order to identify which variants are statistically associated with trait status. With the availability of genotyping methods and characterization of millions of genetic variants, researchers and clinicians in nearly every field of medicine have jumped at the opportunity to perform a GWAS study for the diseases or traits that they study. In collaboration with the European Bioinformatics Institute (EBI), the National Human Genome Research Institute (NHGRI) curated a structured catalog of GWAS studies which currently contains more than 24,218 associations from 2,516 publications in 337 different journals[4].

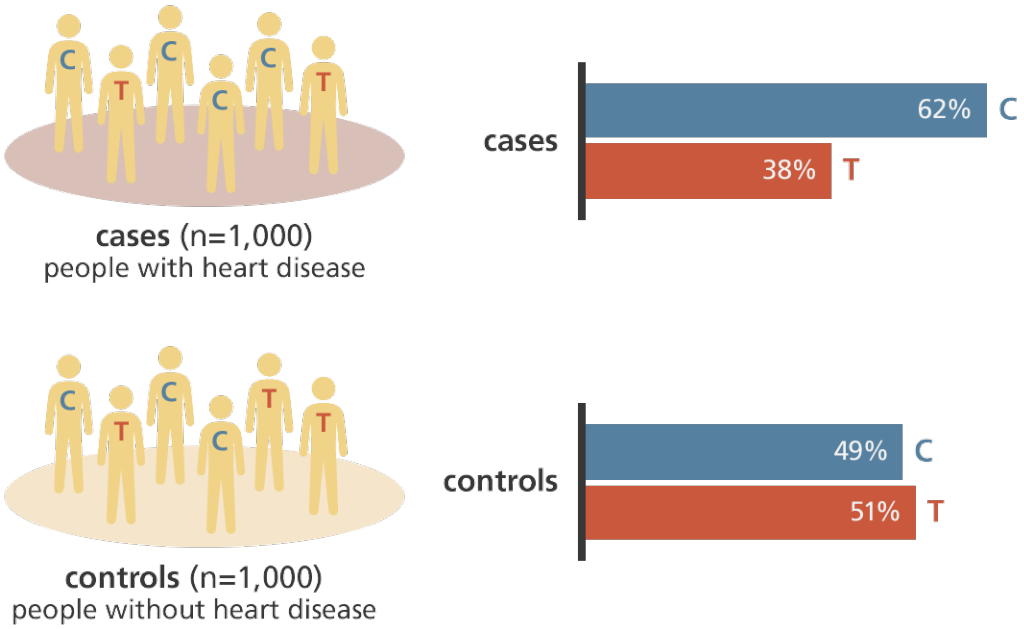


Figure 1.2: GWAS study design (case-control). Subjects with (cases) and without (controls) a disease (such as heart disease) are recruited. Each subject is genotyped. For each variant, allele frequencies are compared between cases and controls. Note that not all GWAS studies follow the case-control design setup. For continuous traits, such as height, subjects with a wide range in the trait of interest are recruited and each variant's genotype is correlated with the trait[5].

Yet despite this massive collection of variant-trait associations, only a small handful of the implicated variants have been proven to impact their associated trait. This paradox is largely due to an exception to Mendel's Law of Independent Assortment. Variants on the same chromosome undergo recombination, such that with enough recombination they are inherited independently on a population-level. However, variants that are nearby one another may not undergo enough recombination to be inherited independently. Such variants are "linked" with one another, or in "linkage disequilibrium" at a

population level. Because of linkage disequilibrium, each causal variant-trait association in a GWAS study carries with it dozens to hundreds of neutral (non-causal) variants. Thus, GWAS associations are not single variants, but rather loci containing many variants with the potential to impact the trait of interest. The causal variant is a needle in the haystack of the neutral variants it is linked with.

While most GWAS studies have targeted variants that underlie disease, the concept of linkage disequilibrium can itself be utilized to identify positively selected variants, or those that have had adaptive effects in evolutionary history. When a new mutation has a strong adaptive effect, such as survival during childbirth, the adaptive allele rises in frequency more quickly than recombination can act to break its linkage with nearby variants. This effect, termed a selective sweep (Figure 1.3), creates a longer than usual region of linkage disequilibrium, or haplotype[6]. In order to identify positively selected regions in the human genome, our group developed the Composite of Multiple Signals test (CMS), which combines multiple signals of positive selection including long haplotype, population differentiation, and derived allele frequency[7]. CMS identified 412 regions of positive selection in the human genome within 3 distinct populations[8]. As a result of the extraordinary linkage disequilibrium that often characterizes positively selected regions, each of these regions can be likened to a (typically) even larger haystack containing hundreds of implicated variants.

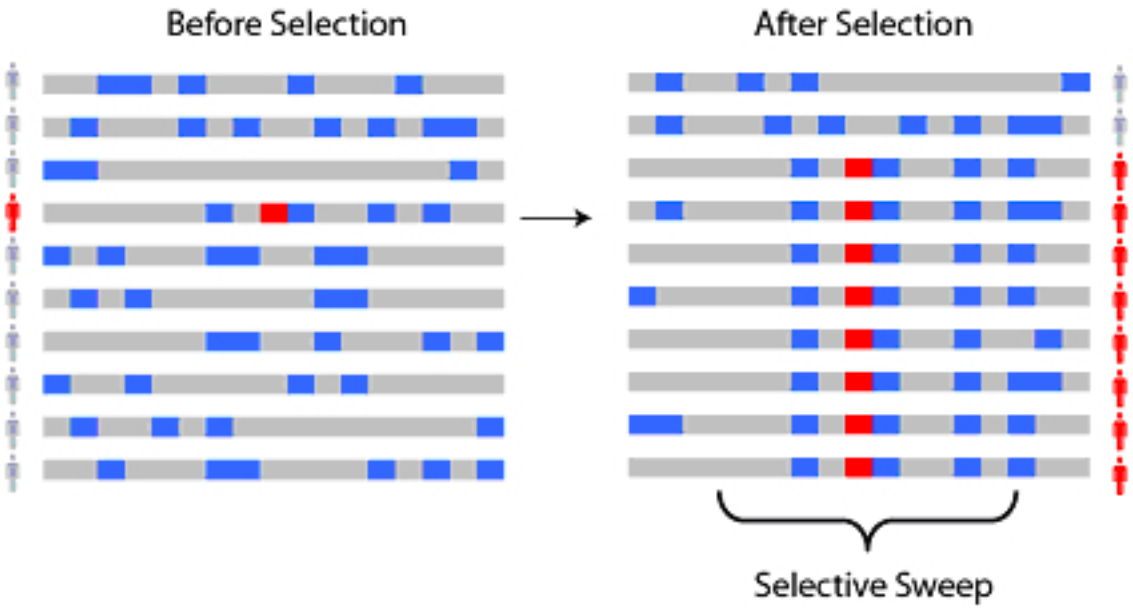


Figure 1.3: A selective sweep. A polymorphism under positive selection is shown in red, alongside neutral polymorphisms in blue, before and after a selective sweep. In a selective sweep, when an adaptive allele rapidly rises in frequency, there is not enough time for recombination to break its linkage with nearby variants, creating a long haplotype[9].

1.3 Finding the Needle: Causal Variant Identification

In the decade leading to the publication of the first human genome sequence[10] the causal gene (and oftentimes the causal variant) for hundreds of rare, severe disorders had been identified[11]. In contrast to most traits, the majority of these disorders were inherited in a simple Mendelian fashion, in which one high-impact variant was responsible for the disorder. One observation that allowed for the identification of causal variants in Mendelian disorders was that the causal variant typically resided within the protein-coding sequences that make up <1% of the genome. Many geneticists hypothesized then that the majority of causal variants for complex, multigenic traits would also reside within

protein-coding sequences.

However, the hundreds of GWAS studies of complex traits and common disorders that followed led to a surprising contradiction: most associations lay far from any known gene. Accounting for shared variance due to linkage disequilibrium, coding variants explained <10% of heritability[12]. Though this result indicated remarkable enrichment of coding regions, it suggested that most causal variants for common, multigenic disorders are regulatory in nature. Once the role of gene regulation in common disorders was discovered, the genomics community rallied to the challenging problem of identifying variants with regulatory roles.

However, unlike protein-coding regions, for which the genetic code may be used to decipher how variants in DNA sequence affect protein sequence and thereby function, regulatory sequences are subject to complex and multilayered mechanisms. For example, gene expression is in part affected by the binding of diverse transcription factors with distinct specificities and tissue expression profiles, and whose access to chromatin may in itself be regulated by CpG methylation and/or histone modification[13]. The availability of whole-genome sequencing led to the development of techniques capable of identifying targets of a given regulatory mechanism on a genome-wide scale. The ENCODE consortium formed to compile the massive list of regulatory annotations each of these technologies provided, including sites targeted by transcription factor binding (ChIP-seq), chromatin accessibility (DNase-seq, FAIRE-seq), CpG methylation (Bisulfite sequencing), and histone modifications (ChIP-seq) across

147 different cell types[14]. While each of these regulatory annotations is informative to regulatory variant interpretation, they vary between providing nonspecific information about a broad chromatin state (such as chromatin accessibility) and providing highly specific information (such as the binding of a specific transcription factor) that highlights a small fraction of regulatory elements.

Another means of identifying regulatory variants is to correlate each variant's genotype with mRNA expression, similar to a GWAS study in which the trait of interest is mRNA expression of a particular gene in a particular tissue. This is termed expression quantitative trait locus (eQTL) analysis. Several consortiums formed to robustly identify eQTLs. By sequencing lymphoblastoid cell lines catalogued as part of the Thousand Genomes Project, the Geuvadis Consortium identified eQTLs in European and African (Yoruban) populations[15]. The Genotype-Tissue Expression Consortium (GTEx) formed to identify eQTLs in many distinct tissues through the recruitment of non-living tissue donors[16], and currently involves over 500 donors in over 50 different tissues. While identifying eQTLs within a GWAS locus can help to identify candidate causal genes (i.e. the eQTL target gene), as well as aid in the prioritization of causal variants, due to their GWAS-like study design eQTL studies are themselves limited in their resolution by linkage disequilibrium.

As a result, causal variant identification remains a laborious endeavor, and even the best associations can be extraordinarily challenging to pair with a functional mechanism. For example, the best association for body-mass index,

an important clinical trait that has a high heritability (40% to 80%), was found in introns 1 and 2 of the *FTO* gene as early as 2007[17]. The causal variant was finally identified in 2015 after a heroic effort that involved chromatin state annotation across 126 cell types, eQTL analysis within adipocytes, haplotype-specific enhancer assays, and finally targeted genomic editing[18].

In order to address this problem, our group recently developed a genome-wide screen for regulatory function, based on the Massively Parallel Reporter Assay[19] (MPRA), to directly test variants for transcriptional activity[20] (Figure 1.4). To isolate general transcriptional effects observed in Geuvadis eQTLs to individual variants, an 180bp oligonucleotide sequence was designed for each allele of 32,373 variants from 3,642 quantitative trait loci in Geuvadis and synthesized on a microarray. These oligonucleotides were cloned upstream of a GFP reporter and RNA sequencing was performed to test for differences in expression between alleles. The ability to individually test thousands of variants in a high-throughput manner (conferred by the MPRA concept) was incredibly powerful. 842 variants showed differential expression between alleles, each of which had the potential to be the causal variant of the eQTL locus from which it was drawn, and 53 of which had the potential to drive disease and trait associations in the NHGRI-EBI GWAS Catalog.

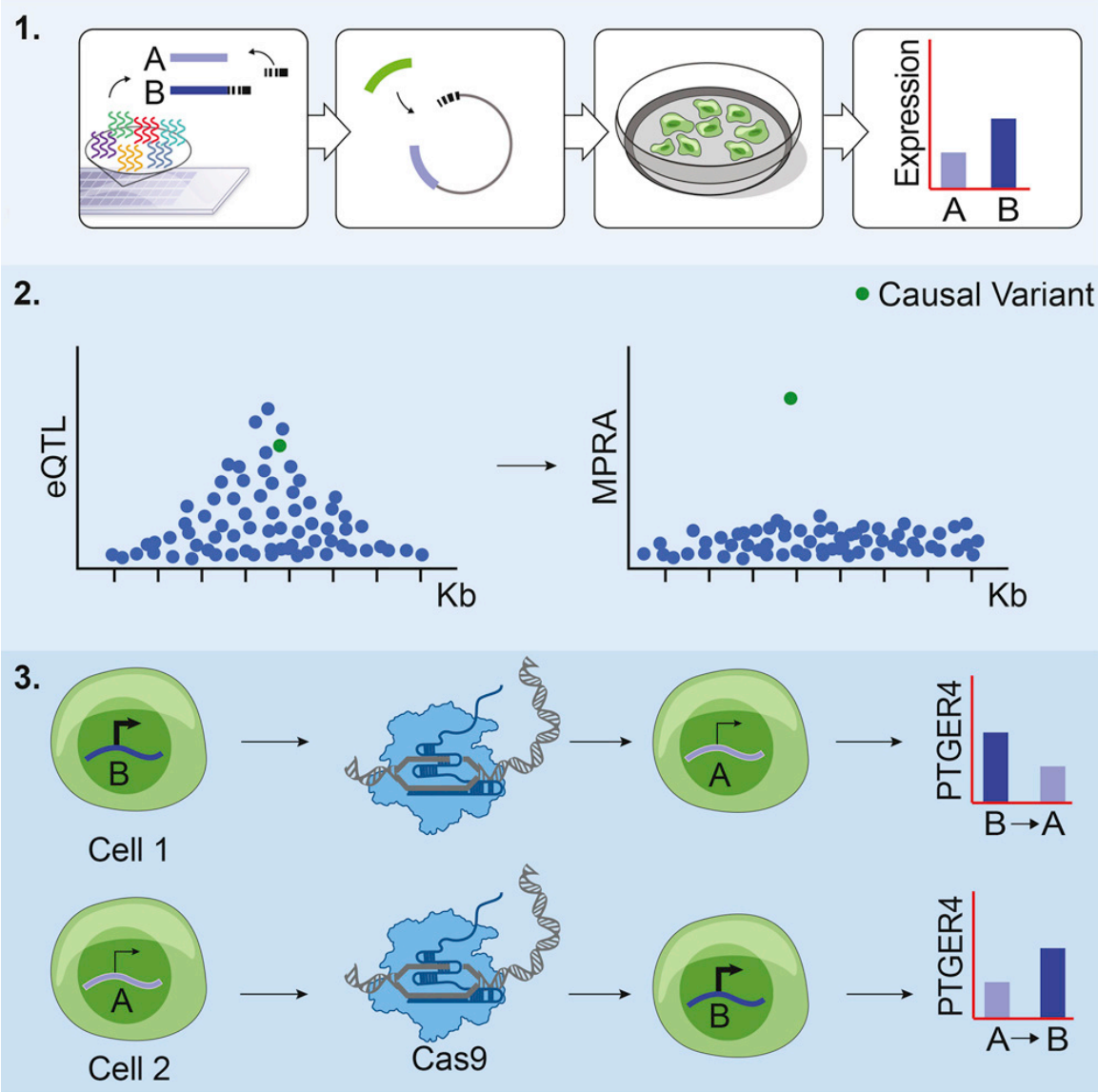


Figure 1.4: Direct identification of promoter/enhancer variants. Panel 1: Oligonucleotides were synthesized on an array for the reference and alternate alleles of 32,373 eQTL variants, inserted upstream of GFP, and transfected into lymphoblastoid cells to measure allelic differences in expression. Panel 2: Whereas eQTL causal variants are obscured by linkage with nearby variants, this MPRA-based approach allows the direct testing of each individual variant for regulatory activity. Panel 3: Endogenous allelic exchange was used to validate a variant driving allele-specific expression of the prostaglandin EP_4 receptor associated with ankylosing spondylitis[20].

1.4 Beyond the Genome: Post-transcriptional Regulation

The majority of GWAS causal variants as well as adaptive causal variants are predicted to reside within intergenic regions[8,12]. As intergenic variants are not carried forward into mRNA transcripts, they would be most likely to impact protein function by acting on transcription of other regions. Consequently, the annotation of regulatory variants has historically focused on control of transcription by a variety of mechanisms, such as those assayed by ENCODE (transcription factors, chromatin accessibility, CpG methylation, and histone modifications).

Regulation of protein expression occurs at several levels, beginning with the transcription of mRNA initiated by RNA polymerase II. However, there are further steps in protein production where regulation may take place. After transcription, a host of small RNAs[21] and RNA-binding proteins[22] interact primarily with the 5' and 3' untranslated regions (UTRs) of the transcript to determine its stability and translation efficiency. Thus, much of regulation occurs beyond interactions with the genome; this is only the first step of the central dogma, and the beginning of the road.

Although the majority of causal variants may reside outside transcripts, variants within transcripts – e.g. protein-coding regions and UTRs – make up a comparatively small proportion of common variation (1-2%). When this is accounted for, protein coding and UTR variants show significantly greater enrichments for causal variants than intergenic and intronic variants[12,23]. While protein-coding variants are often prioritized by their impact on protein sequence,

interpretation of UTR variants is less straightforward. Thus, a treasure trove of causal UTR variants are likely going unexplored.

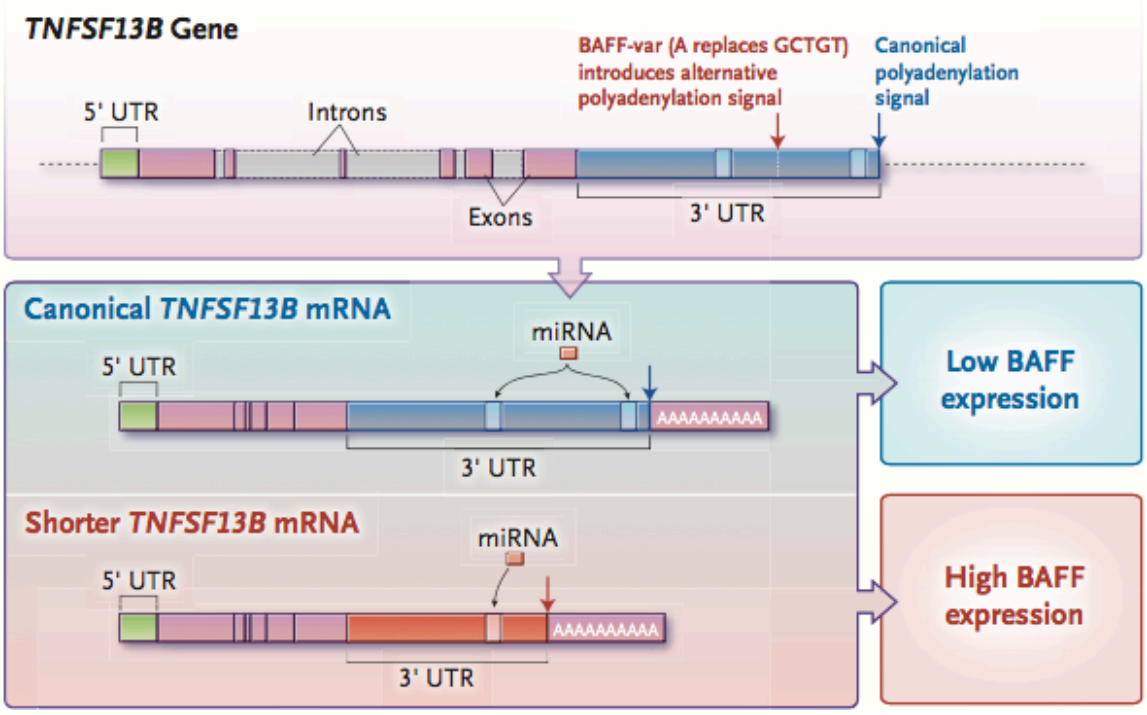


Figure 1.5: MS risk and selection mediated by 3'-UTR shortening. A deletion in *TNFSF13B* ("BAFF-var") introduces an alternative polyadenylation signal, producing a shortened UTR. The shortened UTR loses downstream microRNA binding sites, resulting in increased production of B-cell activating factor (BAFF). This confers an increased risk to multiple sclerosis (MS). In addition, BAFF-var has characteristics of positive selection, possibly due to protection against malaria[24].

Indeed, prominent examples of causal variants exist for both disease and adaptation in UTR regions. For example, Kulkarni and colleagues[25] identified a variant in the 3'-UTR of the *HLA-C* locus that alters binding of a microRNA, affecting the extent to which HLA-C is downregulated and consequently, HIV control. In addition, Steri and colleagues[24] identified a deletion in the 3'-UTR of

TNSF13B that introduces an upstream polyadenylation signal (PAS), leading to early truncation of the transcript and mRNA stabilization by removal of downstream microRNA binding sites (Figure 1.5). This resulted in increased production of B-cell activating factor (BAFF), explaining an association with multiple sclerosis as well as a positive selection signal in the Sardinia population (which the authors posited was provided protection against malaria).

UTRs may be an especially important source of common regulatory variation. Quantitative trait locus studies have found that RNA-decay[26] (rdQTLs) and polyadenylation[27] (paQTLs) are substantially affected by genetic variation, and hence post-transcriptional mechanisms such as those uncovered by Kulkarni et al. and Steri et al. may be more prevalent than previously thought. Furthermore, a recent analysis by Farh and colleagues[28] of expression Quantitative Trait Locus (eQTL) causal variants estimated that 36% of sites fall within transcripts, and only 23% are attributed to known promoter/enhancer elements. This suggests that a wealth of insight into the functional consequences of human genetic variation may be gained by a better understanding of post-transcriptional processes.

Finally, several lines of evidence suggest that translation similarly plays a significant role in health and adaptation. Genome-wide studies conducted on model organisms have shown that much of interspecies variation is due to differential regulation of translation[29,30]. This suggests that some phenotypes may be primarily caused by regulation at the translational level. Indeed, translation is known to be an important layer of gene expression regulation during

erythropoiesis as well as the cellular response to hypoxia[32]. Furthermore, twice as many genes involved in colorectal cancer progression were shown to be regulated at the translational level as at the transcriptional level.

Clearly, post-transcriptional regulation plays an important role in human physiology. Nonetheless, the contribution of post-transcriptional regulation to complex disease etiology and human population adaptation has largely been ignored in favor of cis-acting determinants of transcription factor binding and chromatin architecture, and comprehensive assays for variants that impact post-transcriptional variation are lacking.

1.5 Summary of Thesis

To aid in the identification of 3'-UTR variants contributing to health and adaptation, we developed MPRAu, a massively parallel reporter assay to test 3'-UTR variants for regulatory function. Chapter 2 is centered on the development of MPRAu. We discuss the utility of engineered oligonucleotide libraries. To assist groups working to incorporate oligonucleotide libraries into their work, we describe strategic choices we made in the MPRAu workflow as well as strategies we employed for identifying and resolving technical difficulties inherent to these libraries. Chapter 3 discusses the application of MPRAu to test 9,754 common 3'-UTR variants for regulatory function. We highlight interesting functional insights as well as potential causal variants. A list of variants with significant allelic differences can be found in Table A.1. Finally, in Chapter 4 we provide extended interpretations of our findings in the context of complementary studies and

discuss future applications of MPRAu.

Chapter 2: Development of a Novel Massively Parallel Reporter Assay

2.1 Abstract

In recent years, advances in large-scale DNA synthesis technologies have paved the way for advanced screening tools, such as massively parallel reporter assays. While this has exponentially expanded our functional genomic toolset, developing a screening tool that relies on synthesized oligonucleotide pools is not a straightforward task, as it requires optimizing each experimental step in order to avoid introducing bias between individual sequences. Here we describe the development of a massively parallel screen of 3'-UTR variants for regulatory function. We discuss strategic choices that we made while designing the assay workflow as well as technical difficulties inherent in working with large-scale DNA synthesis libraries.

2.2 Introduction

After the historic transition from Sanger sequencing to the high-throughput pyrosequencing method, a genomics revolution was born. Parallelization allowed a transition from routine sequencing of small bacteriophage genomes (on the order of thousands of nucleotides) to the routine sequencing of large mammalian genomes (on the order of billions of nucleotides). The demand of high-throughput sequencers fueled their advancement; in the years following their advent, their capabilities advanced at a rate approximately four times faster than that seen in the computing revolution described by Moore's law. As they became cheaper and more efficient, high-throughput sequencing technologies spurred efforts such as the Thousand Genomes project[2] (1kG), which sought to identify all common

genetic variants (with an allele frequency of 1% or greater) within 26 human populations by sequencing 2,504 whole genomes, and the Genotype-Tissue Expression project (GTEx), which has sequenced the transcriptomes of over 500 non-living tissue donors in over 50 tissues with the goal of linking human genetic variation with changes in gene expression.

While the ability to *read* a sequence has incredible potential in hypothesis generation, the ability to *write* a sequence has equal potential in hypothesis testing. Traditional DNA synthesis relies on microliter-scale reactions in glass columns. Analogously to pyrosequencing, the development of microarrays parallelized this process, allowing thousands of miniaturized reactions to be performed side-by-side over a slide[35]. Several modalities for array-based synthesis are currently available, all of which rely upon the phosphoramidite four-step process developed in the 1980s, but vary in how they employ deprotection (such as lithographic or electrochemical methods) and coupling (such as inkjet methods) in order to control whether a phosphoramidite monomer will be added to a particular spot on the array[36]. The major limitation to array-based synthesis is currently high stepwise failure rates in deprotection (up to 0.5%) and coupling (0.5-1.5%), which can lead to errors in synthesized oligos or incompletely synthesized (truncated) oligos[37]. This limits the length of synthesized sequences to approximately 200 bases. Nonetheless, the quality and length of DNA produced by array-based techniques has advanced remarkably in recent years[38].

The improvement of array-based synthesis and the availability of engineered oligonucleotide libraries have already sparked groundbreaking work

to better understand gene regulation. One technique that arose was the Massively Parallel Reporter Assay (MPRA), in which oligonucleotides are cloned into a reporter vector and tested for regulatory activity in a high-throughput manner. MPRA has been used to better understand known regulatory elements by testing thousands of synthetic variants of a single regulatory sequence[19], as well as to identify novel regulatory elements by testing thousands of sequences for regulatory function[39]. Recently, our own group utilized an MPRA-based approach to directly test thousands of common variants within expression quantitative trait locus (eQTL) regions for regulatory activity[20]. The power of this approach to identify variation modulating transcriptional activity motivated us to adapt the assay to allow the identification of post-transcriptional activity, particularly for variants in untranslated regions. We term this adapted approach MPRAu.

The power afforded by engineered oligonucleotide libraries is only beginning to come to light. Despite current limitations in length and high error rates, the ability to obtain thousands of sequences of interest dramatically expands the range of possible experimental testing approaches available to dissect genome-encoded functions. However, each new approach must be carefully optimized to prevent sequence biases from producing artifacts, and as of yet, few groups have expertise working with engineered oligonucleotide libraries. Here, we describe the development of MPRAu, a novel experimental testing approach that utilizes engineered oligonucleotide libraries. We discuss strategic choices that we made while designing the assay workflow as well as

technical difficulties inherent in working with large-scale DNA synthesis libraries.

2.3 Results

2.3.1 Repurposing the MPRA for the 3'-UTR

To develop a novel screen for 3'-UTR variation (MPRAu), we built on our experiences adapting MPRA for transcriptional variation (Figure 2.1). In our previous work dissecting causal expression quantitative trait loci[20], we synthesized oligonucleotides centered on either the reference or alternate allele of each variant of interest, surrounded by 150bp of sequence context. These oligos were amplified by emulsion PCR to minimize sequence bias and append a unique 20bp barcode tag to each oligo. After sequencing oligo-barcode pairs to catalog the unique barcodes tagging each oligonucleotide, the GFP coding sequence was inserted between the oligo and the barcode, such that the oligo was upstream of a minimal promoter and the barcode was expressed within the GFP transcript 3'-UTR. The resulting plasmid pool was transfected into lymphoblastoid (LCL) and hepatocarcinoma (HepG2) cells lines and mRNA sequencing of barcodes was performed to identify oligonucleotides with promoter/enhancer elements driving expression beyond the low baseline of the minimal promoter. Furthermore, by quantifying expression differences between barcodes tagging the reference allele oligos and barcodes tagging the alternate allele oligos, variants with significant allelic differences in mRNA expression could be identified.

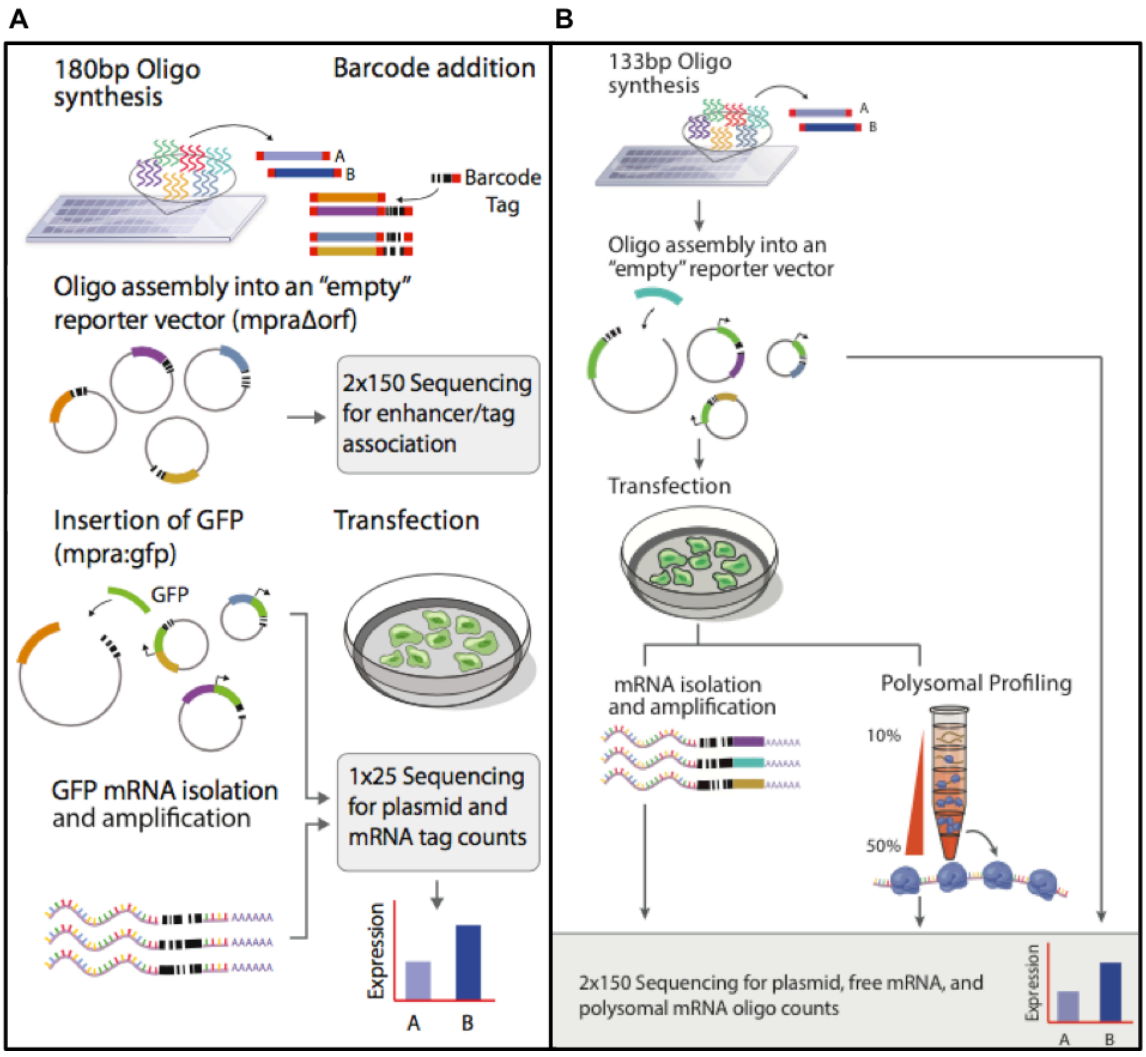


Figure 2.1: Repurposing the MPRA for the 3'-UTR. (A) MPRA workflow to test promoter/enhancer regions, taken from Tewhey et al (cite). After synthesis on an array, oligos were amplified to add a random 20mer barcode and cloned into a reporterless vector. The reporter gene, along with a minimal promoter, was cloned in between the oligo and the barcode, allowing oligo expression to be indirectly ascertained by mRNA-sequencing of the barcode site in the 3'-UTR. **(B)** MPRAu workflow to test 3'-UTR regions. After synthesis on an array, oligos were amplified to add a random hexamer barcode and cloned into the 3'-UTR of a reporter vector. Total and polysomal mRNA-sequencing of the oligo and the barcode within the same read allows steady-state mRNA expression and translation to be ascertained while monitoring library complexity.

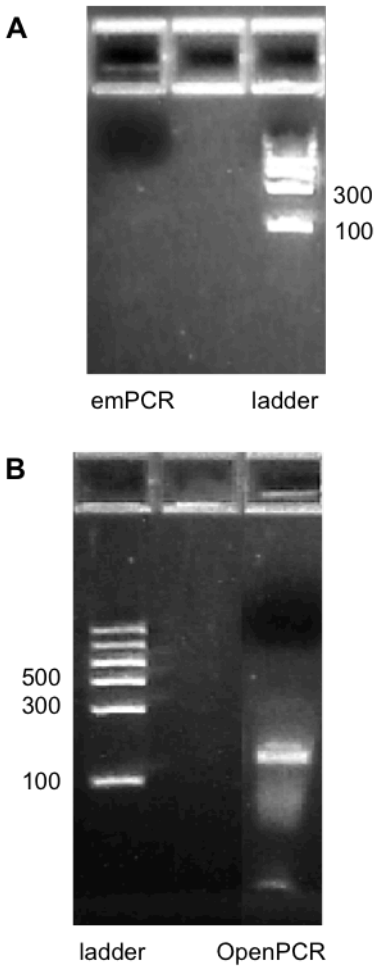
Adapting the MPRA approach to 3'-UTR variation simplified this approach in some ways, and complicated it in others. The majority of sequences that we

tested in our promoter/enhancer MPRA were intergenic or intronic sequences, and thus would be expected to impact regulation at the level of mRNA production and/or splicing. However, UTR sequences are carried into the mature mRNA sequence and have the potential to impact downstream regulatory processes, such as translation. Therefore, we incorporated polysome profiling, a classic experimental approach that involves isolating actively translating polyribosome-mRNA complexes on a sucrose gradient. By simultaneously quantifying allelic expression differences in total mRNA ("RNAskew") and allelic expression differences in polysomal mRNA ("POLYskew"), MPRAu has the potential to identify variants with a significant impact on translation initiation (see *Chapter 3, Sensitive Detection of Functional 3'-UTR Variants*). We note that while a great deal of work has involved shearing polyribosome-RNA complexes to determine ribosome occupancy within each transcript[40], this was not possible for testing 3'-UTR sequences which would not be expected to be occupied by the ribosome.

Finally, because 3'-UTR sequences are present in the expressed mRNA, barcodes were not necessary to tag oligonucleotides. Nonetheless, we decided to include random hexamer barcodes 15bp upstream of our oligonucleotides in the 3'-UTR. Rather than serving to tag our oligos, these barcodes could be sequenced in the same read as the oligo and thus served as unique molecular identifiers (UMIs).

2.3.2 Maintaining a Diverse Library

We synthesized a 3'-UTR oligonucleotide pool consisting of 25,408 unique oligonucleotides. We first attempted to amplify this oligo pool by emulsion PCR to minimize amplification bias (Figure 2.2A). However, we found that emulsion PCR was unsuccessful for our 3'-UTR pool. On the other hand, open PCR performed using similar reaction conditions was capable of amplifying our oligos (Figure 2.2B). We hypothesized that the poor yield of emulsion PCR was due to truncated oligos generated during the synthesis stage. Indeed, while a large yield of single-stranded DNA (57.1 ng/μl) was detected by fluorometry (Figure 2.2C), specific amplification by qPCR identified a very low yield of amplifiable template (10 pg/μl).



C

Method	Concentration
Fluorometry (Qubit ssDNA)	57.1 ng/µl
qPCR	10 pg/µl

Figure 2.2: Technical challenges during oligo amplification. (A) Emulsion PCR product visualized on a 2.2% Lonza FlashGel^{TM} (after SPRI purification). Expected product is 133bp. **(B)** As in (A), using open PCR (without SPRI purification). **(C)** Oligo input quantification by fluorometry and quantitative PCR.

Successful amplification by open PCR allowed us to append random hexamer barcodes as well as to generate enough yield of our oligo pool to clone the 3'-UTR oligos downstream of GFP in a reporter vector. We assessed library diversity before and after transformation by monitoring the oligo read count distribution (Figure 2.3A). Despite the use of open PCR, we obtained a highly diverse oligo library, capturing 97.6% of 25,408 oligos tested at a coverage of 20

reads or greater. We maintained this diversity throughout transformation by high-efficiency electroporation and outgrowth in large culture volumes (5L).

We could not take a similar approach to monitor library diversity throughout transfection, polysome isolation, and library preparation, because of the potential of the oligo read count distribution to be affected by regulatory effects observed in the assay. Instead, we monitored diversity in total and polysomal mRNA by counting unique oligo-barcode pairs (Figure 2.3B). Library diversity was maintained through high-efficiency transfections into approximately 18 million HEK293 cells and by using large reaction volumes during library preparation to accommodate as much input as possible (see 2.5.5 RNA Extraction and cDNA Synthesis).

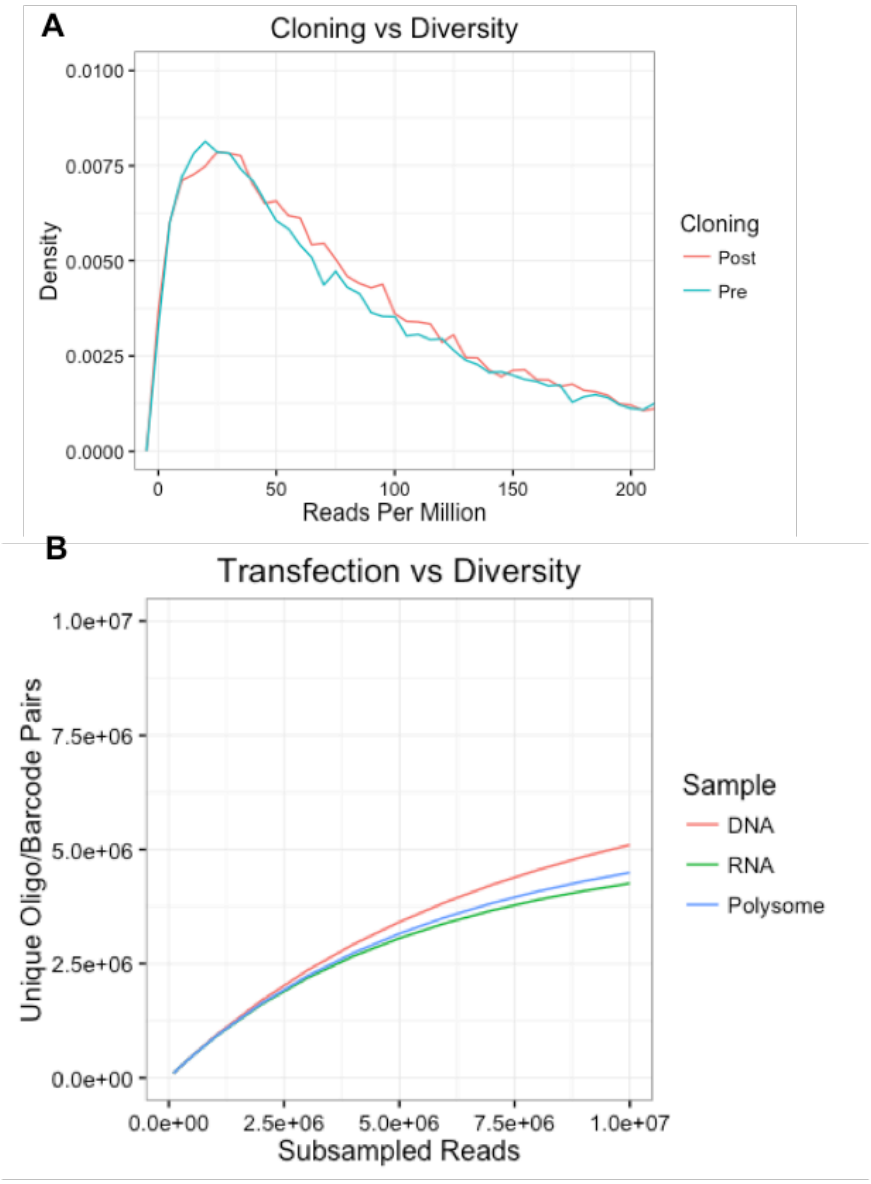


Figure 2.3: Monitoring library diversity. (A) Distributions of oligo counts in the open PCR product (pre-cloning, blue) and in the plasmid pool (post-cloning, red). **(B)** Library diversity before and after transfection was ascertained by counting unique oligo-barcode pairs in plasmid DNA, total RNA, and polysomal RNA reads sampled without replacement.

Although we obtained a diverse plasmid pool, we wondered if biases during synthesis and PCR affected the distribution of oligos in the plasmid pool. We first examined the number of mismatches in oligo sequencing reads.

Approximately 40% of reads were perfect (Figure 2.4A). From this, we estimate a binomial error rate of 0.7%, which includes sequencing errors in addition to errors during oligo synthesis. Mismatches were well distributed throughout the oligo pool, with close to 80% of oligos having a median of 1 mismatch (Figure 2.4B). Next, we examined the contributions of oligo sequence parameters to coverage in the plasmid pool (Figure 2.4C). Interestingly, the percent of thymidine bases was a better predictor of coverage than either GC-content or secondary structure predicted by RNAfold[41]. On further inspection, sequences with a high percent thymidine were overrepresented in the plasmid pool (Figure 2.4D). To determine if this thymidine bias occurred at the oligo synthesis or PCR amplification stage, we examined other oligonucleotide pools (unpublished) for a thymidine bias. Interestingly, arrays synthesized by electrochemical methods, such as our 3'-UTR oligo pool, contained a thymidine bias, while arrays synthesized by inkjet (piezoelectric) methods did not. This suggests that distinct synthesis modalities are predisposed to distinct synthesis biases, including mononucleotide content. In addition to oligos with high thymidine content, we also noticed that short oligos were overrepresented (Figure 2.4E). This may be expected to result from either oligo synthesis, in which shorter sequences would be less predisposed to early truncation, or during PCR amplification, in which shorter sequences would complete extension more efficiently.

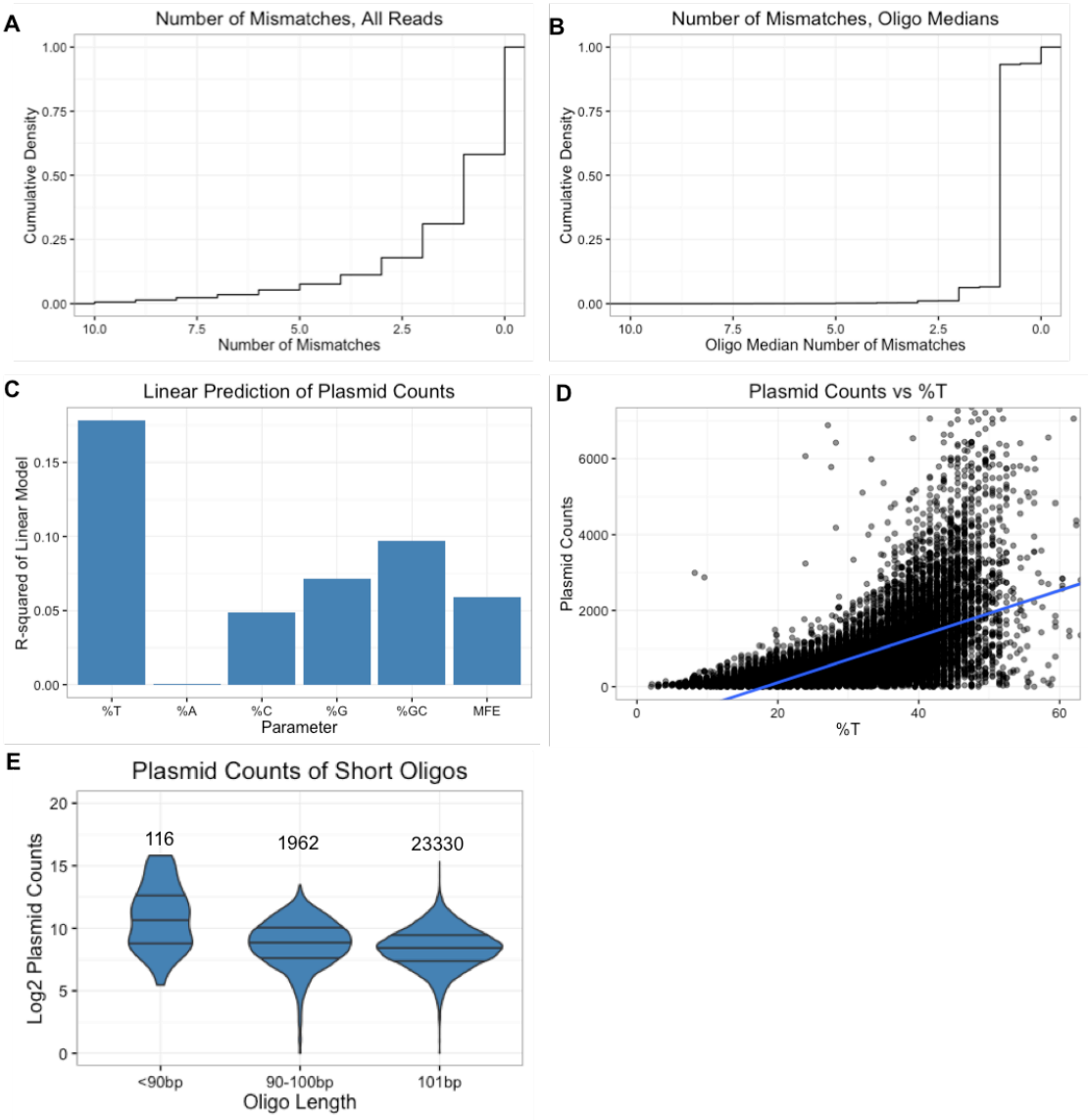


Figure 2.4: Composition of oligos in the plasmid pool. (A) Cumulative distribution of oligo sequence mismatches in the plasmid pool overall (prior to applying a 95% homology filter). **(B)** Cumulative distribution of each oligo's median number of sequence mismatches. **(C)** Univariate linear prediction of plasmid counts. MFE: minimum free energy, as determined by RNAfold. **(D)** Correlation of plasmid counts with the percent of T's in the oligo's sequence. **(E)** Plasmid counts broken down by oligo length. The number of oligos depicted within each group is shown.

2.3.3 Technical Challenges

During the development of our novel screening method, we encountered unique technical challenges inherent to working with thousands of engineered oligonucleotide sequences. The incorporation of random hexamer barcodes as unique molecular identifiers proved critical to identifying the source of technical issues.

During a pilot MPRAu experiment, we noticed that while each oligo's reads was evenly distributed across random hexamer barcodes in the plasmid pool, the majority of total and polysomal mRNA reads for a given oligo arose from just a few barcodes (Figure 2.5A). We termed these barcodes "outlier barcodes," so termed because they contained more reads than 95% of all barcodes paired with a given oligo. Although outlier barcodes were easily identifiable, their expression was so extreme that in total they comprised approximately three-quarters of reads.

We hypothesized that outlier barcodes arose as a result of an artifact, rather than bona fide regulatory interactions. However, to distinguish between these possibilities, we individually cloned and Sanger-validated a given oligo, rs398500_ref_2, adjacent to 12 barcodes with which it was observed to have high, medium, and low expression in the pilot MPRAu experiment. These oligos were combined into a plasmid pool and transfected alongside the original plasmid library in a 12-well plate. Whereas the original plasmid library resulted in high, medium, and low observed expression of the oligo depending on which barcode it was paired with, the Sanger-validated plasmid pool resulted in equal

expression for the oligo regardless of which barcode it was paired with (Figure 2.5B,C). This pointed to a flaw in the plasmid backbone that was being “tagged” by outlier barcodes. Sanger sequencing of specific oligo-barcode pairs revealed that the SV40 promoter, rather than the expected PGK promoter, drove outlier barcodes. We had previously inserted various promoters, including SV40, into the plasmid backbone to assess baseline expression. Furthermore inverse PCR had been used to linearize the PGK-driven plasmid backbone. Therefore, we postulated that our SV40 plasmid had contaminated the inverse PCR reaction used to linearize our PGK plasmid.

We regenerated our plasmid pool, substituting our plasmid’s ampicillin resistance gene for a kanamycin resistance gene and performing linearization by restriction digest in order to prevent contamination. We verified that our plasmid pool varied only in the barcode and oligo sequences by Nextera sequencing. Repeating MPRAu with our optimized plasmid pool demonstrated absence of outlier barcodes (Figure 2.5D).

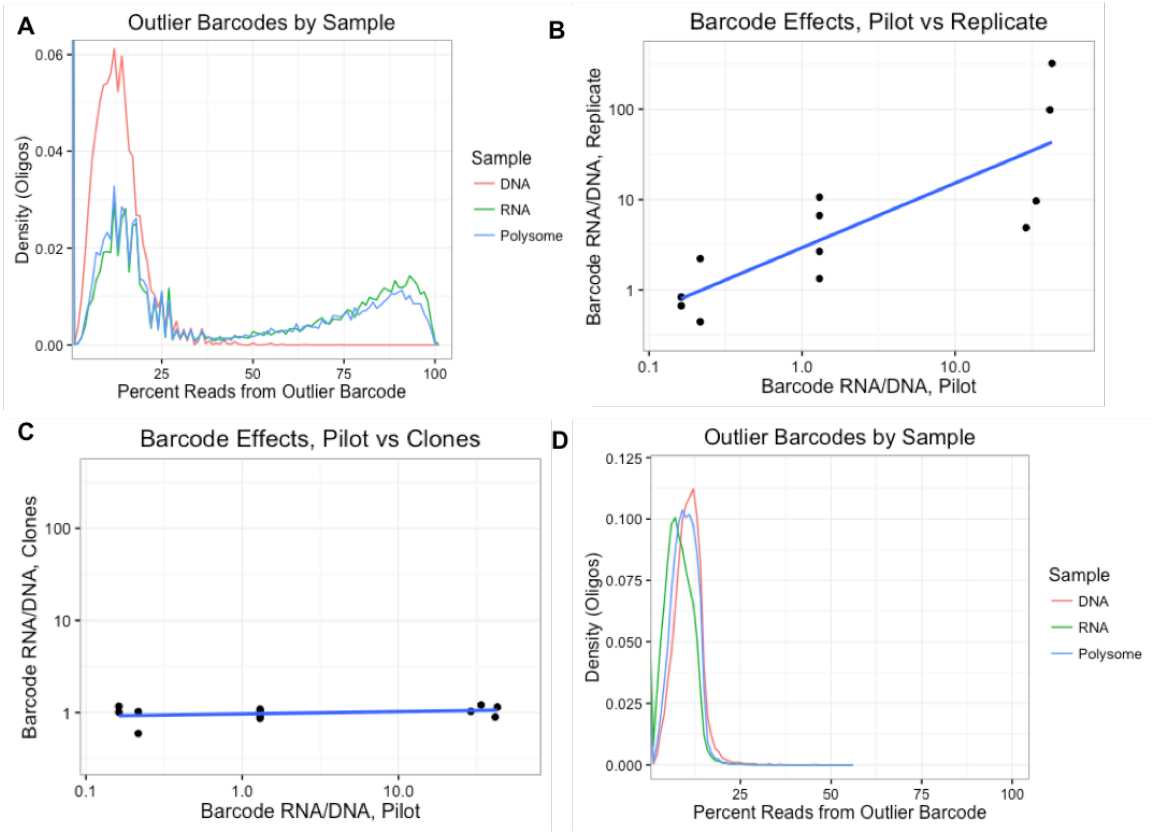


Figure 2.5: SV40 promoter contamination results in "outlier" barcode expression. (A) Distributions of oligo reads arising from outlier barcodes during a pilot assay in which the plasmid backbone contained SV40 promoter contamination. For each oligo, outlier barcodes are defined as barcodes with more reads than 95% of all barcodes paired with that oligo. **(B)** rs398500_ref_2 barcode counts in transfections of the SV40-contaminated plasmid pool performed in a 15cm plate ("Pilot") or a 2.2cm plate ("Replicate"), as well as in a transfection of individually cloned and sequence-verified plasmids performed in a 2.2cm plate ("Clones") **(C)**. For ease of comparison linear trendlines are shown. **(D)** As in (A) after removal of SV40 promoter contamination.

Incorporating polysome profiling into the MPRA also presented unexpected challenges. Comparing total RNA and polysome allelic skews amongst variants, we initially observed two distinct subsets of variants: variants whose allelic skews in total RNA were also present at the polysome level and variants whose allelic skews in total RNA were "buffered," or attenuated at the

polysome level (Figure 2.6A). Buffered variants did not differ from unbuffered variants in oligo coverage, suggesting that the buffering effect did not result from noise.

Interestingly, the large degree of buffering that we observed was in line with previous work that identified translational buffering in yeast[29,30]. However, buffered variants were not enriched in eQTLs found in GTEx to the extent that unbuffered variants were, which suggested that the total RNA expression differences that we observed in buffered variants were an artifact of the assay. Indeed, pooling all oligos across random hexamer barcodes, we noticed that certain random hexamer barcodes also appeared to undergo buffering, and buffered hexamers were predicted to form secondary structures overlapping the annealing site of the gene-specific primer (GSP) that we used during reverse transcription. As we had initially placed the GSP adjacent to the oligos, we reasoned that some oligos could form secondary structures that interfered with GSP binding. Thus, we moved the GSP 162bp downstream of the oligo in order to minimize secondary structure interactions. Using a downstream GSP, we no longer observed a subset of variants with buffered allelic expression differences (Figure 2.6B).

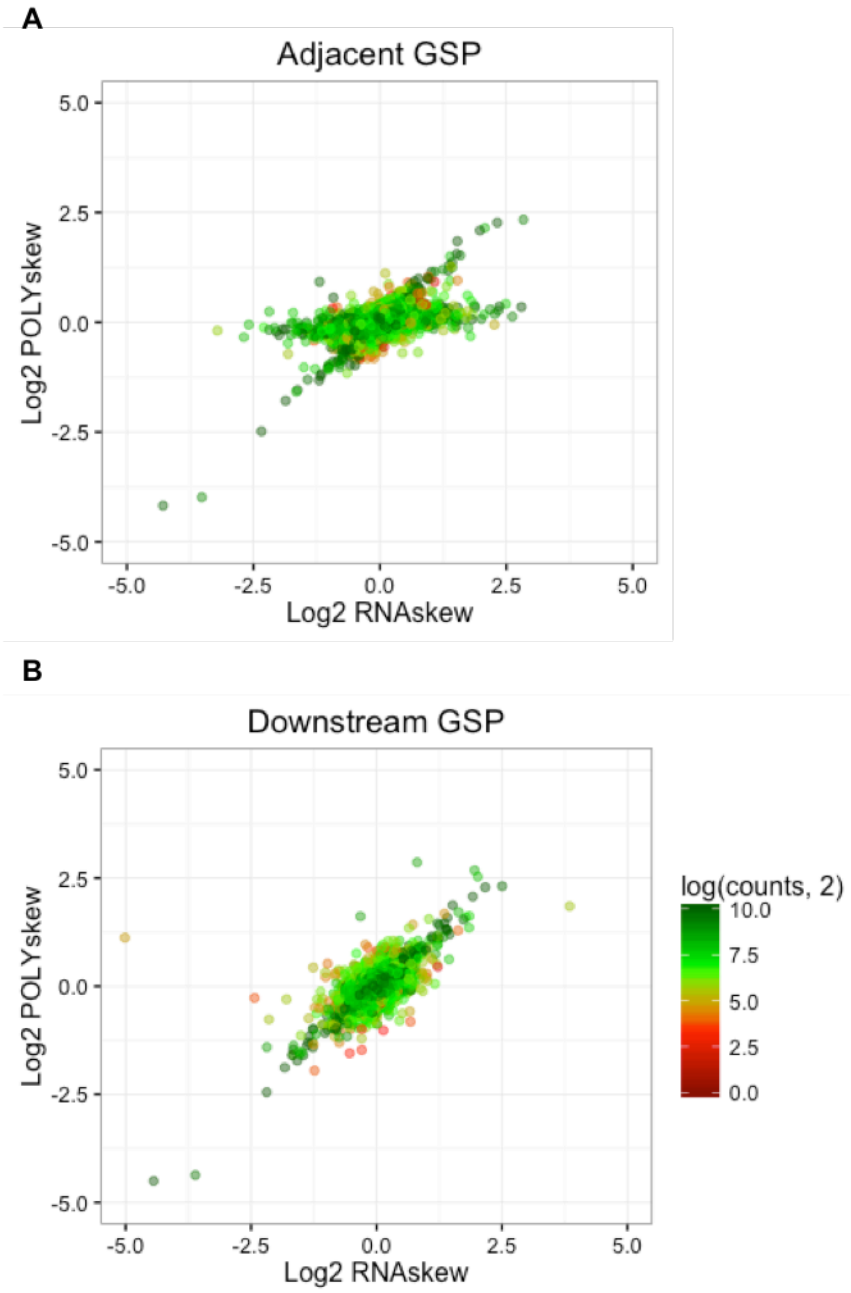


Figure 2.6: RT primer placement creates a translational "buffering" artifact. **(A)** Comparison of allelic skews in total RNA ("RNAskew") and polysomal RNA ("POLYskew") for each variant when the gene-specific primer (GSP) used during reverse transcription is placed directly adjacent to the oligo and **(B)** when the GSP is placed 162bp downstream of the oligo. Points are colored to represent read coverage in the plasmid pool.

2.4 Discussion

We developed MPRAu, a novel experimental approach that utilizes engineered oligonucleotide libraries to test regulatory variation in 3'-UTR regions. This involved significant adaptations to the MPRA approach, such as incorporating engineered oligonucleotides into the reporter (GFP) transcript and adding a polysome isolation step (Figure 2.1). We outlined important considerations to adapting standard molecular biology methods, such as PCR, molecular cloning, and transfection, to oligonucleotide libraries. In addition, we described unique technical challenges that we faced during the development of MPRAu as well as how we resolved them.

The design of MPRAu required careful consideration of the assay workflow, such as the incorporation of a random hexamer barcode 15bp upstream of the oligonucleotides in the 3'-UTR as unique molecular identifiers. Although barcodes were not necessary to identify oligos, as they were expressed in mRNA, their role as unique molecular identifiers proved critical to monitoring library diversity and troubleshooting technical issues. MPRAu differed not only in the orientation of the oligos within the reporter plasmid, but also in the oligonucleotides themselves. On average, 3'-UTRs have lower GC contents than the rest of the genome[42], and we noted that our oligonucleotides contained an average GC content of approximately 40%. Anticipating difficulties at the synthesis stage due to low sequence complexity, we opted to reduce the amount of sequence context from 150bp to 100bp. Indeed, differences in Qubit and qPCR yields suggested a high concentration of truncated synthesis products in

our library. These truncated products were likely responsible for the poor yields that we observed during emulsion PCR (Figure 2.2). This is because in emulsion PCR, PCR reagents are divided into micelles to avoid amplification bias. However, this may be particularly poorly suited to the scenario in which a large degree of truncated products are present, and PCR reagents must be allowed to diffuse freely to the small fraction of amplifiable products.

Despite how carefully an oligonucleotide-based assay is designed, plans should be made to monitor bias at each stage. We monitored library complexity after transformation as well as after transfection (Figure 2.3). In addition, the use of GFP as a reporter allows transfection efficiency to be monitored in real time by fluorescent microscopy. As array-based synthesis is prone to errors, we characterized the distribution of errors in the plasmid pool that we used for MPRAu experiments, confirming that most oligos had few errors (Figure 2.4). Finally, we identified biases that contributed to representation in the plasmid pool. Some of these biases, such as overrepresentation of short sequences, were expected. Others, such as overrepresentation of thymidine-rich sequences in oligonucleotides synthesized by electrochemical methods, were surprising. This emphasizes the importance of actively monitoring bias in each individual experiment, as differences may be observed depending on the sequence composition of the library and synthesis modality used.

While some forms of technical bias can go undetected without careful monitoring, other forms manifest in dramatic ways, such as the “jackpotting” of specific oligo-barcodes combinations that we observed after contamination with

the SV40 promoter. One possible interpretation of this result was a biological interaction between random hexamers and oligos. While this was unlikely, we formally ruled out this possibility through low-throughput validation (Figure 2.5). Notably, if we had not included random hexamers as UMIs, the SV40 contamination would have simply manifested as overexpression of particular oligonucleotides, a result that could plausibly be explained by the presence of enhancer elements in our 3'-UTR oligos. Thus, the inclusion of unique molecular identifiers was critical to quickly identifying contamination in the assay.

Indeed another ostensible manifestation of bias, the total RNA expression differences we observed as a result of secondary structure interference with the reverse transcription GSP annealing site (Figure 2.6), initially appeared to reproduce previous work describing translational buffering in yeast[29,30]. As low-throughput validation would have required a more involved approach, we overlapped our "buffered" variants with GTEx data and observed that expression differences of buffered variants were inherent to the assay. Our random hexamer barcodes were again critical in determining the source of the artifact, this time due to their degeneracy rather than their role as unique molecular identifiers.

Technologies that make use of DNA writing are still in their infancy. While working with engineered oligonucleotide libraries currently demands a specialized skillset, as their potential for hypothesis testing begins to be realized, basic steps such as amplification and cloning will soon become standardized. Still, it will always be important to carefully monitor bias at each step, because engineered libraries may be more sensitive to technical factors that favor

particular oligos than standard molecular biology approaches. Streamlining approaches to identifying and removing these technical factors, such as the use of unique molecular identifiers, simple low-throughput validation approaches, and the early use of existing functional data, will be critical as the next phase of the genomic revolution begins to take hold.

2.5 Experimental Protocols

2.5.1 Oligo Synthesis and Amplification

Oligos were synthesized by CustomArray as 133 bp sequences containing 100 bp of UTR context and adapters on either end (5'CGAGCTCGCTAGCCT [101 bp oligo] AGATCGGAAGAGCGTCG3'). For short UTRs for which 100bp of sequence context was not possible, we added additional sequence evenly outside of each adapter to allow the synthesis of 133bp oligos. Oligo concentration was quantified by Qubit ssDNA (ThermoFisher Q10212) and by quantitative PCR using PowerSYBR master mix (ThermoFisher, 4367659) and rs398500_ref_2 (synthesized as a gBlock, IDT) as a normalization control. Post synthesis, oligos were amplified by emulsion PCR or open PCR, using 5.7ng oligo (as determined by Qubit), 25 µL of Q5 NEBNext MasterMix (NEB, M0541S), 1 unit Q5 HotStart polymerase (NEB, M0493S), 0.5 µM oligo_pmir_F_min (CGAGCTCGCTAGCCT) and oligo_pmir_R_min (CGACGCTCTTCCGATCT) primers and 20 µg BSA (NEB, B9000).

For emulsion PCR (Figure 2.2A), PCR master mix was emulsified by vortexing with 220 µL Tegosoft (Evonik), 60 µL ABIL WE (Evonik) and 20 µL

Mineral Oli (Sigma, M5904) per 50 µL PCR reaction at 4°C for 5 min. 100 µL of Emulsion mixture was plated per well across a 96 well plate and cycled. Amplified emulsion mixture was broken and purified by adding 1 mL of 2-butanol (VWR, AA43315-AK), 50 µL of AMPure XP SPRI (Beckman Coulter, A63881) and 80 µL of binding buffer (2.5M NaCl, 20% PEG-8000) per 350 µL of Emulsion mix and vigorously vortexing followed by incubation for 10 minutes at room temperature. Broken emulsion/butanol mixture was spun at 2900 rcf for 5 min and the butanol phase was discarded. The aqueous phase was placed on a magnetic rack for 20 minutes prior to aspiration. Remaining beads were washed once with 2- butanol, three times with 80% EtOH and eluted in EB (Qiagen, 19086).

After we determined optimal amplification conditions, 6bp random hexamer barcodes and additional adapter sequences were added by performing 20 open PCR reactions each 50µl in volume containing 5.7ng of oligo, 25 µL of Q5 NEBNext MasterMix (NEB, M0541S), 1 unit Q5 HotStart polymerase (NEB, M0493S), 0.5 µM oligo_BAR_Bmt_F (CTAGTTGTTTAAAGCCCAACGCTAGTCNNNNNNCGAGCTCGCTAGCCT) and oligo_pmir_R_min (CGACGCTCTTCCGATCT) primers and 20 µg BSA (NEB, B9000). PCR reaction was purified by performing a 2.5X SPRI purification using Agencourt AMPure XP SPRI (Beckman Coulter, A63881) beads according to manufacturer instructions.

2.5.2 MPRAu Vector Assembly

Barcoded oligos were first inserted into pmirGLO:Δluc::gfp plasmid linearized by 30 cycles of inverse PCR using 10pg of template. However, after detecting SV40 contamination in this initial library, barcoded oligos were inserted into BmtI/XbaI digested pmirGLO:Δluc::gfp $\Delta Amp^R::Kan^R$ by gibson assembly (NEB, E2611). A 40µl reaction containing 1µg of digested vector and a 10:1 molar ratio (415ng) of amplified oligos was incubated for 60 min at 50°C followed by SPRI purification and eluted in 25µl EB, then concentrated to 10µl by vacuum centrifugation. The ligated vector was then split into 5µl aliquots, each of which was transformed into 100 µL of 10-beta E.coli (NEB, C3020K) by electroporation (2kV, 200 ohm, 25 µF). Electroporated bacteria were immediately split into five 1mL aliquots of SOC (NEB, B9020S) and recovered for 1 hour at 37°C then independently expanded in 1L of LB supplemented with 50µg/mL of kanamycin (Teknova, K2125) on a floor shaker at 37°C for 12 hours. After outgrowth aliquots were pooled prior to plasmid purification (Qiagen, 12991). For each of the aliquots we plated serial dilutions after SOC recovery and estimated a library size of $>10^8$ CFUs.

2.5.3 Validation of Outlier Barcodes

To validate barcode outlier effects, we individually cloned twelve barcodes upstream of rs398500_ref_2 into pmirGLO:Δluc::gfp by inverse PCR and Gibson assembly: four corresponding to low-range expression in the MPRAu pilot, four corresponding to medium-range expression, and four corresponding to high-

range (outlier) expression. We then confirmed the full length of each plasmid sequence by Sanger sequencing and combined the plasmids in equal concentration. 500ng of the plasmid clones (Figure 2.5C) was transfected into HEK293FT cells alongside 500ng of the plasmid pool used in the MPRAu pilot (Figure 2.5B) in separate wells of a 12-well plate by combining each sample with 2µl of Lipofectamine 2000 (ThermoFisher, 11668027). Cells were incubated with transfection reagents for 24 hours, monitored by fluorescent microscopy, and then split 1:4 into a new 12-well plate. After 24 hours (48 hours post-transfection), RNA extraction was performed with TRIzol LS (ThermoFisher, 10296). MPRAu total RNA libraries were prepared as described below.

2.5.4 MPRAu Experiments

HEK293FT cells were grown in DMEM (ThermoFisher, 10564) supplemented with 10% FBS (ThermoFisher, 26140). For all 5 transfections (Figure 2.1C), cells were plated in a 15cm plate and grown to a density of ~80-90%. Cells were transfected with 80µl Lipofectamine 2000 (ThermoFisher, 11668027) and 20µg DNA. Cells were incubated with transfection reagents for 24 hours, monitored by fluorescent microscopy, and then split 1:3 into 2 new 15cm plates. After 24 hours (48 hours post-transfection), media from each plate was replaced with 30mL DMEM + 100µg/mL cycloheximide (CHX) (Sigma-Aldrich, C4859) and incubated for 5 minutes at 37°C. Both plates were then washed with 10mL cold PBS (ThermoFisher, 14040) + 100µg/mL CHX, scraped in 1mL cold PBS (ThermoFisher, 14090) + 100µg/mL CHX, pooled and centrifuged for 5

minutes at 500xg, and finally resuspended in 350µl lysis buffer composed of 5mM Tris pH 7.5 (ThermoFisher, 15567), 2.5mM Magnesium Chloride (ThermoFisher, AM9530G), 1.5mM Potassium Chloride (ThermoFisher, AM9640G), 2µM DTT (VWR, 97061), 100µg/mL CHX, 5 mg/mL Triton-X (Sigma-Aldrich, T8787), 5 mg/mL Sodium Deoxycholate (Sigma-Aldrich, 30970). Cell lysate was then centrifuged for 5 minutes at 12,000xg and supernatant was flash-frozen in vapor phase nitrogen. For polysome profiling, approximately 200µl lysate was loaded onto 10-50% sucrose gradients in 15mM Tris pH 7.5, 15mM Magnesium Chloride, 150mM Sodium Chloride (ThermoFisher, AM9760G), and 100µg/mL CHX. Gradients were centrifuged in SW41Ti rotor at 35,000 rpm for 2.5 hours at 4°C and 0.5mL fractions were collected.

2.5.5 RNA Extraction and cDNA Synthesis

RNA was extracted with TRIzol LS (ThermoFisher, 10296) according to manufacturer instructions. Total RNA was purified from lysate. Polysomal RNA was purified from sucrose fractions, pooling fractions corresponding to three or more ribosomes. 7.5µg GlycoBlue (ThermoFisher, AM9515) was added each sample in order to visualize the pellet. mRNA was purified from total RNA using oligo $d(T)_{25}$ magnetic beads (NEB, S1419S) according to manufacturer's instructions, and eluted at 80°C. Purified total and polysomal mRNA were then subjected to Turbo DNase treatment (ThermoFisher, AM2239). The reaction was terminated in 1 mg/mL SDS (ThermoFisher, AM9822) and purified by performing a 2X SPRI purification using Agencourt RNAClean XP SPRI (Beckman Coulter,

A63987) beads according to manufacturer instructions. DNase-treated mRNA for each of the five replicates was diluted to the concentration of the lowest concentration sample, and first-strand cDNA was synthesized from concentration-normalized mRNA with SuperScript III (ThermoFisher, 18080) and a gene-specific primer 162bp downstream of the oligo (oligo_RT_R, GCATTCTAGTTGTGGTTTGTCCA) using the manufacturer's recommended protocol, modifying the total reaction volume to 40 µL and performing the elongation step at 55°C for 80 minutes. Single-stranded cDNA was purified by performing a 2X SPRI purification using Agencourt RNAClean XP SPRI beads.

The translational "buffering" artifact that we described (Figure 2.6A) was observed using a gene-specific primer directly downstream to the oligo (oligo_pmir_R_min, CGACGCTCTTCCGATCT).

2.5.6 qPCR and Library Construction

To minimize amplification bias during the creation of cDNA sequencing libraries, samples were amplified by qPCR to estimate relative concentrations of GFP. The highest Ct across all MPRAu cDNA samples was 20. Therefore, to add Illumina sequencing adapters, 10µl of cDNA from each replicate was amplified by 15 cycles of PCR followed by 2.5X AMPure SPRI purification and 5 cycles of PCR using a TruSeq index primer. For approximation of plasmid pool counts, the plasmid pool was diluted 1:30,000 and amplified as above in five independent PCR reactions (technical replicates). Indexed libraries were SPRI purified and pooled according to molar estimates from Agilent TapeStation quantifications.

Samples were sequenced using 2x150 bp chemistry on an Illumina HiSeq through the Broad Institute's walk-up sequencing service.

2.5.7 Oligo Read Alignment

Paired-end 150 bp reads were merged into single amplicons using Flash[43] v1.2.11 (flags: -M 150, -O). Amplicon sequences were kept if the sequence surrounding the barcode met the following conditions: (1) a perfect match was found to the 10bp sequence on either the left or right side of the barcode, (2) the 10bp sequences on both the left and right sides of the barcode matched with a Levenshtein distance of 3 or less, and (3) the 2bp immediately surrounding each side of the barcode matched perfectly. Oligo sequences from the passing reads were then mapped back to the expected oligo sequences using BWA-MEM[44] version 0.7.12 (flags: -M). Alignment scores were calculated as matching bases divided by the expected oligo size (up to 123bp maximum overlap between paired reads). For allelic skew calculation, reads with alignment scores of less than 0.95 were discarded.

2.5.8 Modeling Oligo Counts in the Plasmid Pool

To determine the contributions of oligo sequence parameters, such as mononucleotide content, GC content, and secondary structure, to coverage in the plasmid pool (Figure 2.4C), we constructed univariate linear models for each parameter and calculated a coefficient of determination (R^2). We observed

similar results using generalized additive models.

2.6 Contributions

Dustin S. Griesemer performed all experiments and wrote the chapter. Dustin S. Griesemer and James Xue performed data analysis. Dr. Ryan S. Tewhey and Dr. Cailin Joyce provided scientific and technical guidance. Drs. Carl D. Novina and Pardis C. Sabeti provided scientific guidance.

2.7 Acknowledgements

We would like to acknowledge Dr. Tarjei Mikkelsen for coordinating oligonucleotide synthesis and providing technical guidance with oligo amplification. This work was supported by the National Institutes of Health under Ruth L. Kirschstein National Research Service Award 5F30GM114940 from the National Institute of General Medical Sciences, as well as by the National Science Foundation under Doctoral Dissertation Improvement Grant DEB-1401237 from the Division of Environmental Biology.

Chapter 3: A Resource of Common Variation in Untranslated Regions for Selection and Disease

3.1 Abstract

Although genic regions, including 3'-UTRs, are highly enriched in disease and selection-associated loci, linkage disequilibrium and the lengthy timescale of low-throughput validation hinder functional follow-up of each locus. To address this challenge, we developed a tool for identifying functional 3'-UTR variants by combining the Massively Parallel Reporter Assay with Polysome Profiling. We find that under standard growth conditions in the HEK293 cell line, short (101bp) 3'-UTR segments have surprisingly small, if any, impact on translation. Furthermore, out of 9,754 tested variants (approximately one-seventh of the common 3'-UTR variants in the genome), we identify a shortlist of 115 variants with significant allelic differences in steady-state RNA expression which are implicated in physiologically important processes such as sodium transport, leptin signaling, and creatinine synthesis. Though functional variation identified by MPRA is strongly correlated with known 3'-UTR regulatory mechanisms, the vast majority of identified variants fall outside of predicted motifs. This unbiased approach provides concrete leads as well as a means to quickly identify a wealth of causal variants within untranslated regions of the genome.

3.2 Introduction

In the transformative years since the completion of the Human Genome Project, rapid advances in genotyping and sequencing technologies have made feasible large genome wide association studies (GWAS), which have uncovered thousands of loci linked to inherited disease[4]. At the same time, an influx of

genotype data from diverse ethnic groups has bolstered the power of unbiased scans for positive selection, such as the Composite of Multiple Signals (CMS) test[7], which identify loci underlying adaptation to historic environmental threats, such as infectious agents, that have plagued mankind for millennia. These developments set the stage for genotype-based diagnostics and promise to provide mechanistic insights that may be exploited to develop novel therapeutics.

Despite increased power to identify loci associated with disease and adaptation, it is still difficult to identify the causal variant(s) driving each association due to their tight correlation with dozens to hundreds of neutral polymorphisms nearby, i.e. linkage disequilibrium (LD). One solution is to identify polymorphisms with an observable molecular phenotype. However, quantitative trait locus (QTL) based approaches that associate genotype with molecular phenotypes similarly suffer from a large amount of LD[45,46]. Hence, identifying causal variants is currently a colossal task involving low-throughput functional dissection of all the candidates within a given locus[18,24]. As a result, even the best associations remain unexplained.

To address this, we previously adapted the massively parallel reporter assay (MPRA) to directly test thousands of candidate variants for transcriptional activity[20]. This allowed us to uncover hundreds of putative causal alleles for variation in gene expression within a single assay. However, we noted that a significant limitation to the sensitivity of our approach was its inability to detect post-transcriptional effects. While the majority of disease[13,47–49] and selection[8] associations are either intergenic or intronic, and thus likely to act on

transcription (and/or splicing), protein-coding as well as UTR variants are highly enriched within associations and may be more likely to be functional[12,23]. Furthermore, a recent analysis by Farh and colleagues[28] of expression Quantitative Trait Locus (eQTL) causal variants estimated that 36% of sites fall within transcripts, and only 23% are attributed to known promoter/enhancer elements, suggesting a substantial role for post-transcriptional activities in human regulatory variation. Far fewer studies have targeted regulatory variation in post-transcriptional activities, including mRNA decay, polyadenylation, and translation[26].

Here, we adapted the MPRA as a tool to quantify allelic steady-state expression differences for thousands of variants that lie within 3' untranslated regions (3'-UTRs). To additionally detect allelic differences in translation initiation, we extended the assay by incorporating polysome profiling, a classical method used to isolate polyribosomal-RNA complexes. Our method, *MPRAu*, is capable of detecting functional 3'-UTR variants in a highly sensitive and massively parallel manner. Here, we utilize MPRAu to characterize the regulatory landscape of common 3'-UTR variation as well as identify a shortlist of 115 variants with potential physiological roles in positive selection and disease.

3.3 Results

3.3.1 MPRAu: a Massively Parallel Screen of Causal UTR Variants

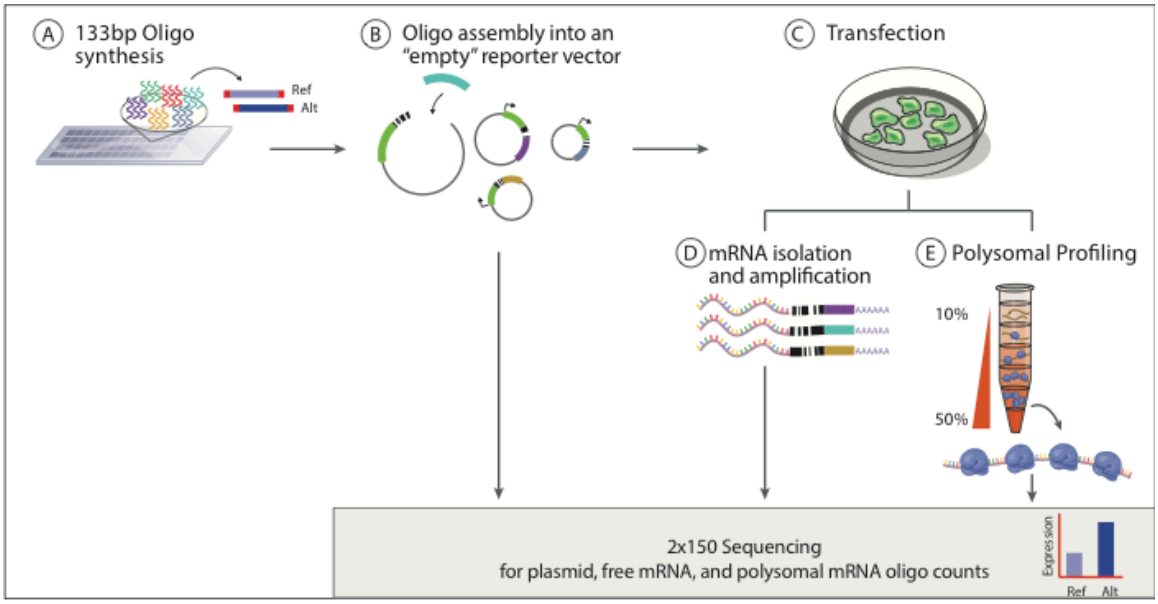


Figure 3.1: Adapting MPRA for 3'-UTR post-transcriptional variation. (A) Array based synthesis of thousands of oligonucleotides, each representing one allele of a variant of interest. **(B)** Oligos are PCR-amplified and inserted into a vector backbone downstream of GFP and adjacent to a random hexamer barcode. **(C)** The vector pool is transfected into HEK293 cells, split after 24h, and after 48h harvested for **(D)** mRNA sequencing and **(E)** polysome isolation and sequencing. Oligo counts from mRNA sequencing (D) and Polysome Profiling (E) are compared to baseline counts acquired from the sequencing of pre-transfected vector (C) in order to determine the relative expression and translation of ancestral and derived alleles.

To develop a functional screen for causal 3'-UTR variants, we repurposed MPRA, an approach we previously demonstrated to be effective in identifying hundreds of putative causal noncoding variants[20]. Unlike intergenic regions, untranslated regions have the potential to impact translation, particularly

translation initiation[50]. Therefore, we extended MPRA by incorporating polysome profiling (Figure 3.1; Experimental Protocols). By isolating the subset of actively translated (polysomal) RNA, differences the overall proportion of translated mRNA that result from regulation at the level of translation initiation may also be quantitated. We call this UTR-tailored approach *MPRAu.*

To select variants for testing by MPRAu, we selected common (minor allele frequency, MAF > 5%) Single Nucleotide Polymorphisms (SNPs) and indels within 412 positively selected 1Mb genomic tiles previously identified by CMS from Thousand Genomes pilot data. Notably, signals of positive selection within each of these tiles could be further localized to smaller regions of 27kb median size. However, we tested all 3'-UTR variants within each 1Mb tile in order to capture all 3'-UTR variants potentially contributing to positive selection in the pilot populations (CEU, CHB/JPT, YRI). This set of 9,754 variants comprises approximately one-seventh of all common 3'-UTR variants in the genome.

An oligonucleotide library capturing both the reference and alternate alleles of all 9,754 3'-UTR variants, including 100bp of 3'-UTR context, was amplified and cloned into a plasmid downstream of GFP. Following transfection of the plasmid pool, mRNA was purified from total (steady-state) and polysomal (actively translated) RNA of actively dividing cells. This was repeated for 5 biological replicates, performed on separate days. We observed high coverage, capturing 97.6% of 25,408 oligos tested at a coverage of 20 reads or greater. Reproducibility was excellent, with normalized read count correlations between experimental replicates of 0.99 for both total and polysomal RNA (Figure 3.2A,B).

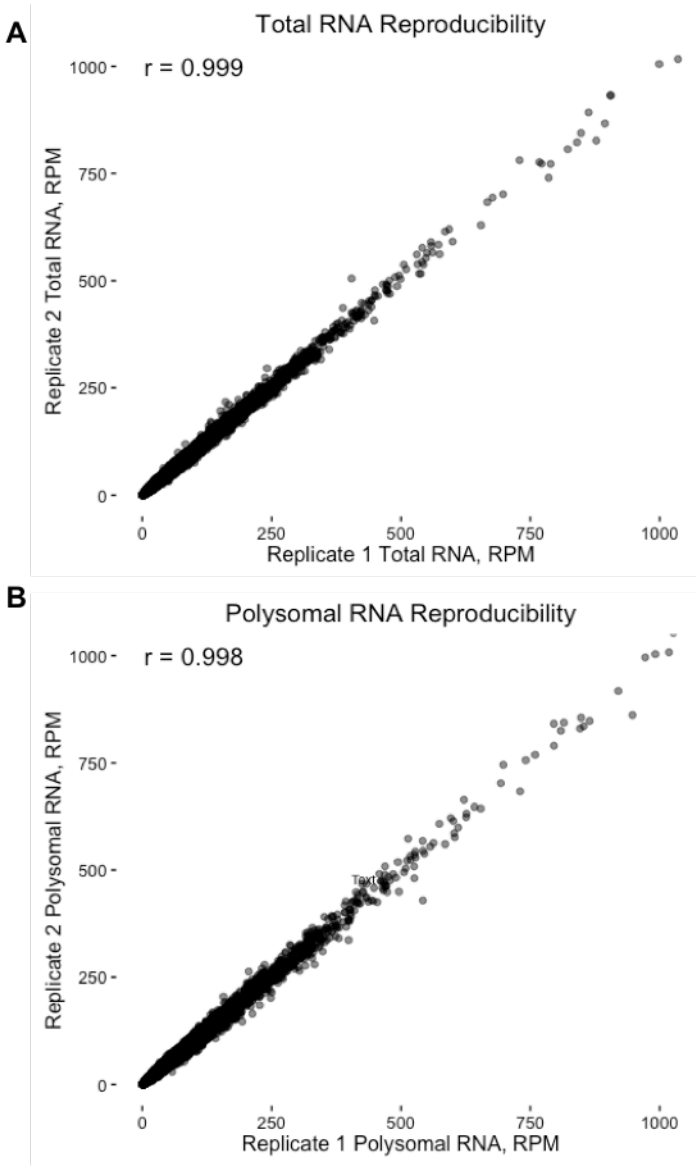


Figure 3.2: Experimental reproducibility. (A) Correlation of total RNA and **(B)** polysomal RNA between replicates. Correlation is provided as Pearson's R. **(C)** Library diversity was ascertained by counting unique oligo-barcode pairs in plasmid DNA, total RNA, and polysomal RNA reads sampled without replacement.

Oligos were directly sequenced from GFP mRNA as opposed to using barcode proxies. However, we carried over the use of barcodes as unique

molecular identifiers. Random hexamer barcodes were appended 15bp upstream of the oligo during amplification. This allowed us to confirm that library diversity was maintained throughout the protocol (Figure 3.2C) and helped us to troubleshoot technical issues that arose during the development of the screen (Figure 2.6).

3.3.2 Short 3'-UTR segments Overwhelmingly Function at the RNA Level

One utility of MPRAu is its ability to test thousands of short UTR segments in isolation from the rest of the transcript, such that each segment's regulatory impact may be attributed directly to the sequence of interest rather than nearby elements within the 3'-UTR, 5'-UTR, or even intergenic regions. Therefore, before investigating allelic effects, we sought to characterize the regulatory landscape of 3'-UTR sequences surrounding common variants (MAF>5%).

We used DESeq2 to estimate oligo fold-changes between plasmid (DNA), total RNA, and polysomal RNA using a paired negative binomial model. In order to normalize expression by initial representation in the plasmid pool, we performed independent amplification reactions of the plasmid pool to obtain 5 technical replicates for sequencing. Plasmid read counts correlated very highly between technical replicates (rho=0.999) and followed a Poisson distribution (Figure A.3). We therefore excluded plasmid replicates from the model's dispersion calculation in order to avoid underestimating noise.

Oligonucleotides were cloned 15bp downstream of a random hexamer

barcode, described above. As a result of the extremely high diversity of our library, each oligo was paired with over 200 barcodes on average. Similarly, each barcode was paired with over 1,500 oligos. This provided the unique opportunity to quantitate the regulatory impact of random hexamer motifs in the 3'-UTR somewhat independently of context. The regulatory effects of random hexamers on total RNA (Figure A.1) and polysomal RNA (Figure 3.3A) were normally distributed and small in magnitude. However one hexamer, AAUAAA, was an outlier for both steady-state and polysomal RNA. This hexamer constitutes the canonical polyadenylation signal sequence (PAS) and was associated with an apparent reduction in expression. This suggests that MPRAu is capable of detecting early termination events, likely due to loss of adapter sites required for reverse transcription and library amplification.

Our high library diversity as well as the modest regulatory effects of the barcodes further aided us in quantifying the regulatory effects of our 100bp 3'-UTR segments. The 3'-UTR is thought to play a large role in post-transcriptional regulation including mRNA stability, polyadenylation, and translation. Total RNA expression varied over a 30-fold range, with 11 oligos expressed greater than two-fold above the median and 658 oligos (2.6%) expressed greater than 2-fold below the median (Figure 3.3B). Hence, the largest regulatory effects resulted in a decrease in steady-state expression.

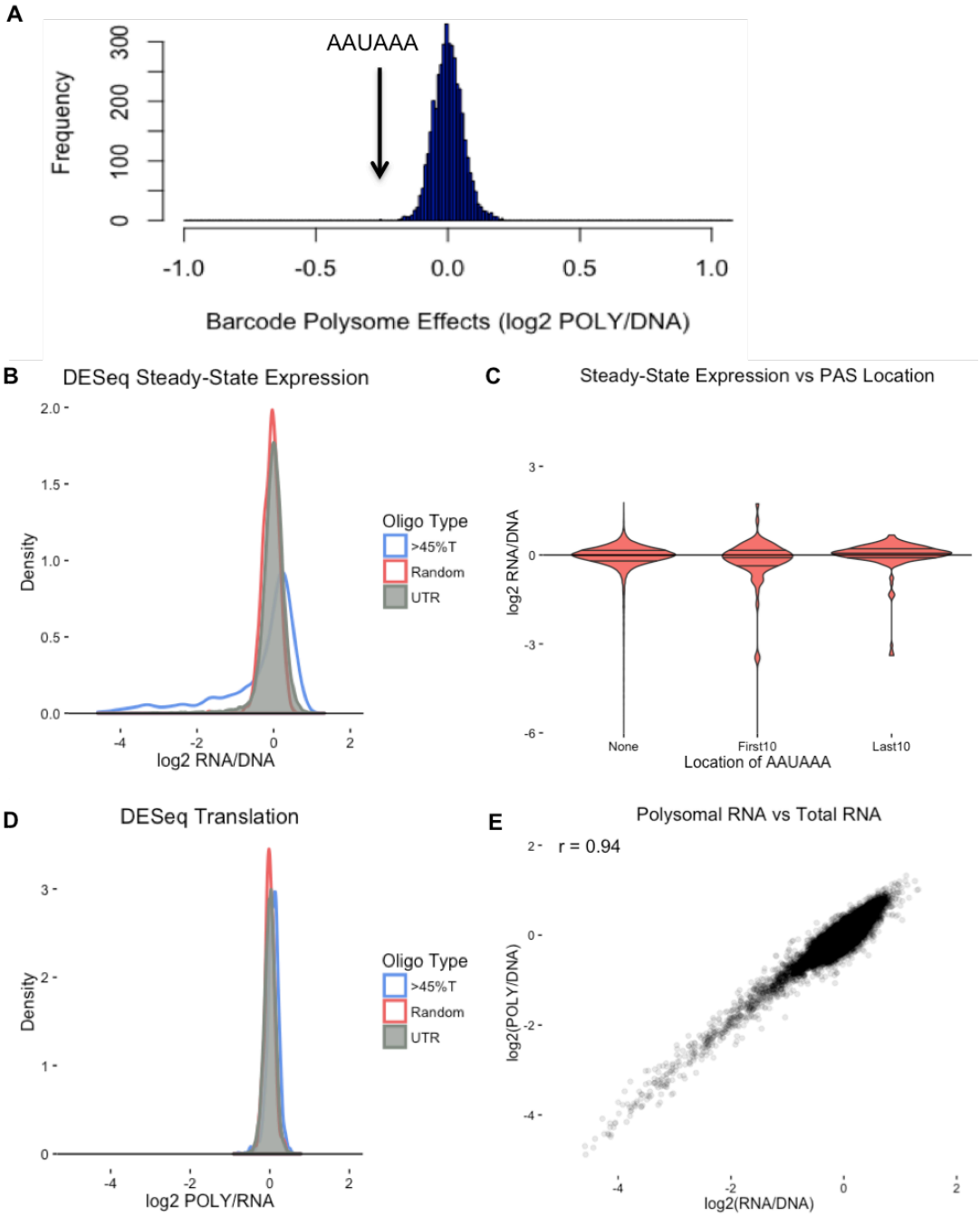


Figure 3.3: Isolating regulatory effects of 3'-UTR segments. (A) Histogram of random hexamer effects on plasmid-normalized polysomal RNA expression. The canonical polyadenylation signal (PAS), AAUAAA, is labeled. **(B)** Density plot of normalized total RNA expression, comparing oligos derived from 101bp of UTR (gray), oligos derived from random sequence (red), and oligos with greater than 45% uridine (blue). **(C)** Violin plots of steady-state RNA expression, comparing the presence and position of the PAS. **(D)** Density plot of polysomal RNA expression normalized by total RNA to quantify oligo effects on translation. **(E)** Correlation between total RNA and polysomal RNA (expressed as Pearson's R).

We compared the expression of oligos derived from genomic context surrounding 3'-UTR variants to randomly generated sequences of similar nucleotide content. Randomly generated oligos were more tightly clustered around the median expression, suggesting that they lacked evolved regulatory elements found in oligos derived from genomic 3'-UTRs. Of oligos derived from genomic 3'-UTRs, we observed that shorter oligos (derived from 3'-UTRs less than 100bp in length) were expressed more highly. This is in line with the general observation that transcript isoforms with shorter 3'-UTRs have on average greater expression[51,52].

After observing that the regulatory effect of the PAS was detectable in MPRAu, we searched for sequences that might be associated with early termination. Global mapping of transcript ends has identified enrichment for a PAS 20-30bp upstream of the polyA site, as well as U-rich sequence immediately downstream of the polyA site[53]. We examined the effects of PAS and U-rich sequences within our oligos. We found that the 4% of oligos composed of greater than 45% uridine were significantly enriched among oligos that reduced expression greater than 2-fold (199/658 oligos). Interestingly, PAS effects were position-dependent, with a PAS within the first 10bp of the oligo having an attenuating effect on expression and a PAS within the last 10bp of the oligo having an amplifying effect on expression (Figure 3.3C). This is consistent with our hypothesis that the reduction in expression observed when the PAS was in the barcode position (15bp upstream of the oligo) was due to loss of adapter sites. In contrast to a PAS located within the barcode position or the first 10bp of

the oligo, a PAS within the last 10bp might be expected to terminate downstream of the reverse transcription and sequencing adapters, resulting in a short but amplifiable transcript.

The 3'-UTR has also been shown to play a role in mediating translation initiation, mainly through re-initiation through the poly(A)-loop[54]. We sought to detect this layer of post-transcriptional regulation by quantifying the effects of our 3'-UTR oligos on polysomal (actively translated) RNA levels. Normalizing polysomal RNA expression by total RNA expression allowed us to obtain a measure differences in the fraction of translated mRNA, and thus translation initiation. Translational effects were smaller than effects on steady state expression, varying over a two-fold range (Figure 3.3D). Polysomal RNA was highly correlated with total RNA expression (rho=0.936, Figure 3.3E) even though only 36% of GFP total RNA resided in the polysomal pool (Figure A.4), based on qRT-PCR of sucrose gradient fractions. Thus, we observed a striking lack of translational effects of isolated 100bp 3'-UTR segments in the HEK293 cell line. We caution that because we defined polysomal RNA as any polyribosomal-RNA complex containing 3 or more ribosomes, we would not be able to detect translational effects that impact the number of ribosomes per transcript. Rather, our observations indicate that the 3'-UTR segments that we tested did not impact the proportion of RNA that is translated.

The high correlation that we observed between total RNA and polysomal RNA levels is evidence that the 3'-UTR may not play a significant role in translational buffering, a controversial yet widely reported phenomenon in which

noisy mRNA expression changes are “buffered” by translation[30,55–57]. However, we note that during assay development we were able to induce a buffering-like effect by placing the gene-specific primer used during reverse transcription directly adjacent to the oligo (Figure 2.6). Oligos that formed secondary structures close in vicinity to the gene-specific primer’s binding site were depleted from total RNA but not polysomal RNA.

3.3.3 Sensitive Detection of Functional 3'-UTR Variants

Having characterized the regulatory dynamics of 100bp sequences surrounding common 3’-UTR variants, we sought to identify variants with regulatory differences *between alleles*. To quantify differences between alleles, we incorporated an allelic interaction term into our DESeq model (see 3.5.9 Expression and Allelic Skew Quantification). DESeq outputs this interaction term as the log2 fold change between alleles for total RNA expression (which we term the “RNAskew”) as well as polysomal RNA expression (“POLYskew”). Using Bonferroni correction and a p-value cutoff of 0.1, only 115 variants (1.2%) had a significant RNAskew, with 38 variants demonstrating greater than 2-fold allelic differences and 77 variants demonstrating greater than 1.5-fold allelic differences (Figure 3.4A). Similar results were obtained for POLYskew, with 41 and 80 variants respectively (Figure 3.4B). Only one variant had a significant effect on translation, and this was marginally significant (adjusted p=0.023).

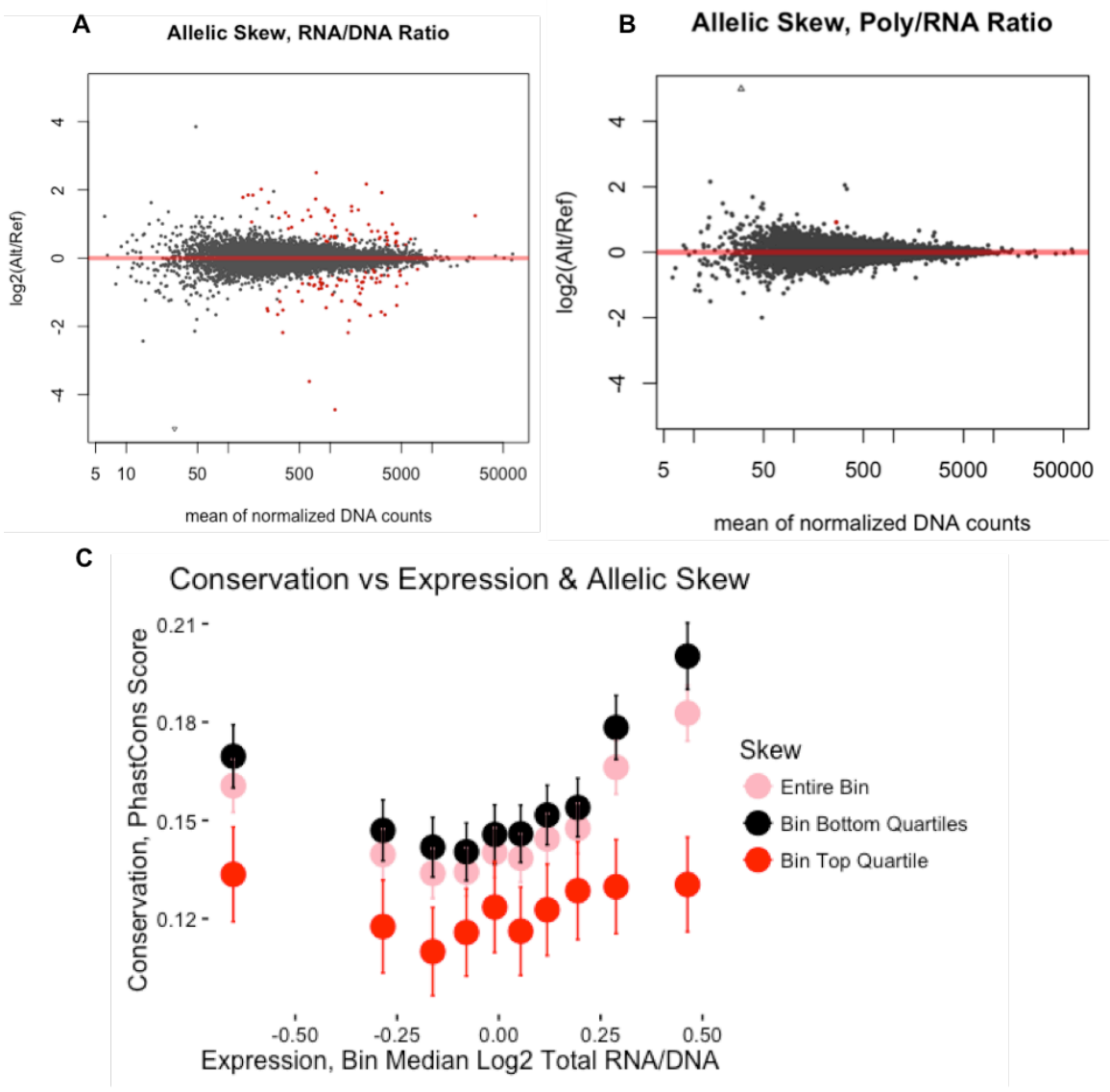


Figure 3.4: Unbiased detection of functional 3'-UTR variants. (A) M-A plot for steady-state expression (RNA/DNA) allelic skew, e.g. the log-fold expression difference between the alternate allele and the reference allele. Statistically significant variants are highlighted in red. **(B)** M-A plot for translation (Poly/RNA) allelic skew. Only one variant was marginally significant (padj=0.02). **(C)** PhastCons conservation of UTR oligos, binned by plasmid-normalized total RNA expression. Points correspond to median expression and mean conservation. Error bars correspond to 95% confidence intervals of the mean. Each expression bin is further subdivided by absolute allelic skew (top quartile – red, bottom quartiles – black, entire bin – pink).

The low proportion of variants with allelic effects may be the result of purifying selection acting to remove functional variants with potentially deleterious effects from the population before they reach high frequency[58]. To evaluate this hypothesis, we calculated conservation scores for the 3'-UTR sequences that we tested (Figure 3.4C). Conservation was highest for 100bp sequences at the extremes of expression, suggesting that although the largest regulatory effects that we detected attenuated expression, MPRAu also has the ability to detect oligos that amplify expression. We then compared conservation of sequences for in the top quartile of RNAskew, to sequences for which no allelic effects were observed (bottom three quartiles). This loose definition included variants with very subtle allelic effects. Variants with allelic effects were significantly depleted in conservation compared to variants without allelic effects, indicating that purifying selection is indeed acting to deplete allelic effects detected by MPRAu from common variation.

3.3.4 Allelic Effects are Congruent with Complementary Approaches

Current efforts to characterize 3'-UTR variation are often focused on well-annotated regulatory elements[59,60]. This is based on the premise that variants predicted to disrupt such elements are enriched for regulatory function. We tested enrichments for allelic effects of variants predicted to disrupt several well-studied regulatory elements (Figure 3.5A). For example, microRNAs are ~22nt RNAs which have been shown to cause mRNA degradation and/or translational repression when bound to the 3'-UTR[21]. microRNA targeting nearly always

requires complementarity to nucleotides 2-7 of the microRNA (termed the “seed site”), and improves with additional requirements for target bases across from nucleotides 1 and 8. As expected, microRNA target sites were enriched for attenuating effects on expression, and this enrichment increased with extent of seed match. Importantly, enrichment was drastically greater for targets of the 10 most abundant microRNAs in the HEK293 cell line than for targets of the top 100 most abundant microRNAs. This highlights the importance of predicting targets of microRNAs abundant in a cell type of interest rather than agnostically predicting targets of all annotated microRNAs, a list which continues to grow exponentially[61].

Similarly, RNA binding proteins (RBPs) bind to the 3'-UTR and regulate both steady-state expression and translation[22]. The binding preferences of most RBPs are poorly defined, and vary with respect to the importance of secondary structure. However, we examined enrichments for some well-defined RBP binding sites, such as AU-rich elements (AREs) and the canonical Pumilio motif, both of which have been associated with decreased mRNA stability[62,63]. As expected, both AREs and Pumilio motifs were enriched for attenuating effects on expression.

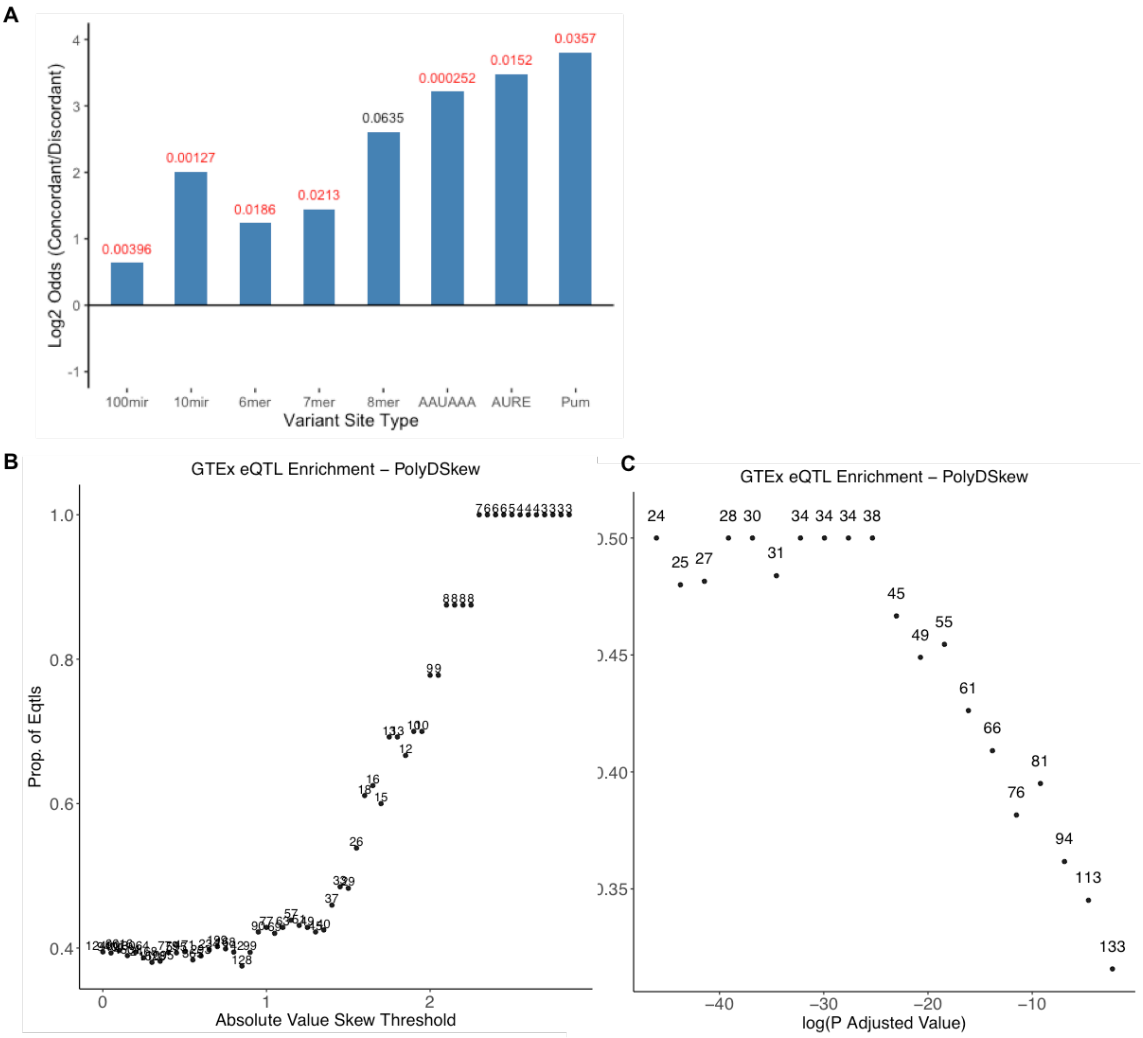


Figure 3.5: Enrichment of known regulatory elements. (A) Variants disrupting several known 3'-UTR regulatory motifs were identified in the top and bottom quartiles of RNAskew. For each motif, an odds ratio was calculated that the undisrupted regulatory element resided on the expected allele. p-values were calculated using Fisher's exact test and are shown above each bar, with significant values in red. **(B)** Proportion of variants called as eQTLs in the Genotype-Tissue Expression Consortium, increasing a minimum threshold of POLYskew or **(C)** p-value. The number of variants contributing to each point is labeled.

Functional regulatory variants may also be identified in an unguided fashion via expression quantitative trait locus (eQTL) analysis, in which

expression in a given cell type is correlated with genotype across many individuals. eQTL analysis is limited to transcripts expressed in the cell type studied, requiring many tissues to be sampled in order to capture regulatory effects throughout the genome. The Genotype-Tissue Expression consortium (GTEx) is the largest multi-tissue eQTL database currently available, with over 500 nonliving tissue donors in over 50 tissues[46]. We determined which of our tested variants were eQTLs in at least one tissue catalogued by GTEx. Approximately 40% of our tested set met this criterion, likely due to large amounts of linkage disequilibrium in European populations. Nonetheless, as we applied an increasing threshold of MPRAu allelic skew magnitude, the proportion of variants called as eQTLs increased and eventually reached 100% (Figure 3.5B, p=0.0303 for correlation). In addition, for most tissues eQTL directionality was concordant with allelic skew (Figure A.5), suggesting that allelic skews detected by MPRAu are also present in physiological contexts. By comparison, when we used MPRAu p-value as the threshold the proportion of eQTLs plateaued at 50% (Figure 3.5C, p=0.0052). This suggests that MPRAu may be better-powered to detect subtle allelic effects than eQTL analysis, for which large sample sizes would be needed.

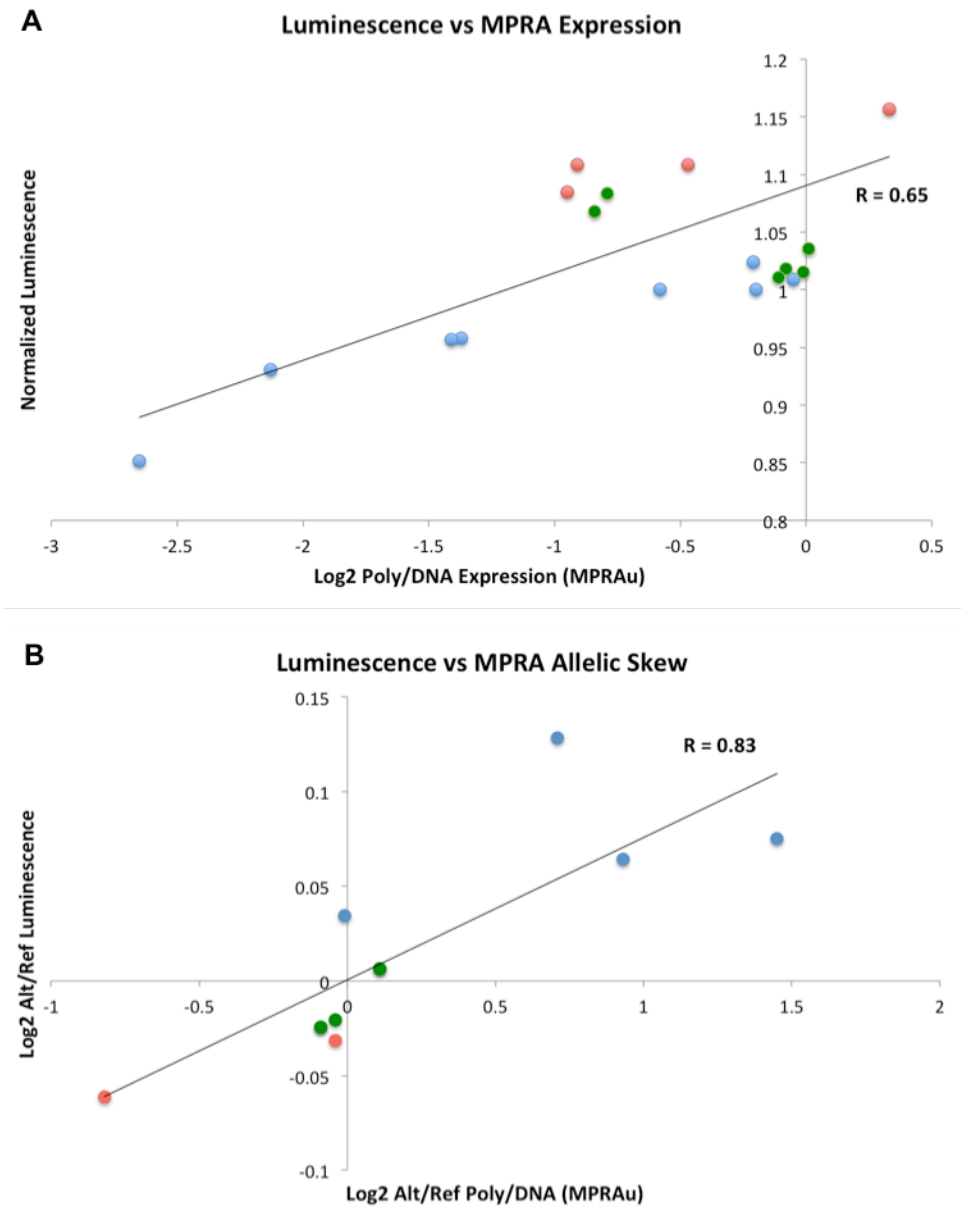


Figure 3.6: Low-throughput validation of functional variants. (A) The 101bp sequences for the reference and alternate alleles of 4 variants with allelic effects detected by MPRAu as well as 5 no-skew control variants were cloned downstream of firefly luciferase. Firefly luminescence was normalized to Renilla luciferase luminescence. Correlation between normalized luminescence and MPRAu steady-state expression is shown (represented as Pearson's R). Red points correspond to oligos with 40-45% uridine. Green points correspond to oligos with the canonical polyadenylation signal, AAUAAA**. (B)** Correlation between allelic differences in luminescence with MPRAu allelic skews.

In addition to testing concordance of MPRAu allelic skews with eQTL data, we also validated the regulatory effects that we observed using a complementary low-throughput method, luciferase assays. 18 oligos (Figure 3.6A) corresponding to the reference and alternate alleles of 9 variants (Figure 3.6B) were cloned downstream of Firefly Luciferase and individually transfected into HEK293 cells to measure relative expression. To keep growth conditions as consistent with MPRAu as possible, cells were split 24 hours after transfection and expression was quantified 48 hours after transfection. Furthermore, to avoid overexpression of a single oligo, each transfection was diluted with a “dummy plasmid” encoding GFP under the control of a defective promoter. Overall, oligo luminescence was well-correlated with MPRAu expression ratios (R=0.65). Five oligos had higher than average luminescence despite lower than average expression in MPRAu. This would be expected for sites that lead to early termination and generate short UTRs which prevent amplification in MPRAu but are detectable via the luciferase assay. When we examined these oligos’ sequences further, we found that four contained between 40-45% uridine, while the other two contained a PAS site near the 5’ end, in line with a role in termination. Overall, a high correlation was observed between luminescence and MPRAu skew (R=0.83), although effect sizes for the luciferase assay were notably smaller than those observed by MPRAu. Thus, subtle allelic effects detected by MPRAu would likely be missed by standard luciferase assays.

3.3.5 A Prioritized Set of Functional 3'-UTR Variants for Selection and Disease

As a highly sensitive and parallelized method for detecting functional UTR variation, MPRAu generates a shortlist of potential causal variants in selection and disease. We focused on the 115 variants whose allelic effects reached statistical significance by a rigorous Bonferroni multiple testing threshold. In order to explore the role of these 115 variants in selection and disease, we aligned these regions with CMS positively selected loci as well as GWAS trait-associated loci (Table 3.1).

To determine potential causal roles in positive selection, we identified overlap with CMS loci identified from Phase 3 Thousand Genomes data. 17 variants had potential causal roles in selection, such as rs34448361, which creates an AU-rich element in the Leptin Receptor (LEPR), as well as rs1049508 and rs199645276, variants in Arginine:Glycine Amidinotransferase (GATM), a vital enzyme in the creatine synthesis pathway which is thought to have been selected for during the transition from hunter-gatherer to farming cultures[64]. The allelic effect of rs199645276 was validated by our luciferase assay. We note that several loci, including LEPR and GATM, contained missense variants. However, GWAS studies have failed to convincingly link these missense variants with a trait underlying the signal of selection, warranting consideration of variants that act in a regulatory manner.

Table 3.1: MPRAu identifies promising candidates for follow-up. Variants with significant allelic expression differences implicated in disease and selection. LFC: MPRAu expression difference between alternate and reference alleles expressed as log2(alt/ref); padj: Bonferroni-adjusted MPRAu p-value; GWAS: LD with SNP in NHGRI-EBI GWAS Catalog and p-value of SNP in GWAS study; CMS Pop: population in which positive selective pressure was detected (see 3.5.13 Selection and GWAS Associations).

Gene	Variant	LFC	padj	Annotations	GWAS (p-value;LD)
HLA-DQA1	rs3667	-4.502	1.90E-183		
GLMN	rs2046620	-1.377	2.13E-115		
SLC35E4	rs2240002	-1.880	5.21E-95		
TXN2	rs139999	2.312	2.02E-58		
NAA40	rs202234263	-1.297	8.59E-38		
NAA40	rs200234485	-1.416	3.58E-34		
APPBP2	rs75062707	1.586	6.56E-29		
GATM	rs199645276	-0.821	1.81E-24		Serum Creatinine (1e-42; 1)
GATM	rs1049508	0.873	1.51E-22		Serum Homoarginine (1e-45; 1)
HMGA2	rs10878350	1.315	8.22E-19		
PRR23A	rs1553877	-0.949	8.76E-17		Keloids (2e-13; 0.99)
APPBP2	rs11289661	-1.000	5.64E-12		
SFPQ	rs200831313	-0.890	2.92E-11		
HIF1AN	rs2495760	0.873	2.93E-08		Plasma Palmitoleic Acid (6e-9; 0.99)
SEMA3C	rs201521853	-0.371	5.31E-07	AU-rich element	
SEMA3C	rs1527474	-0.650	2.91E-06		
NAA40	rs200402503	1.070	1.88E-04		
SLC12A1	rs78562329	0.705	6.24E-04		
ADAM17	rs189060518	1.352	1.73E-03		
CCDC66	rs73081823	-0.699	4.74E-03		Height (4e-22; 0.98)
LEPR	rs34448361	-0.601	7.22E-03	AU-rich element	
RGMA	rs3752103	0.931	9.86E-03		
KIF3A	rs4705963	0.526	1.75E-02		Atopic dermatitis (1e-10; 0.91)

To determine potential causal roles in disease traits, we identified variants in close linkage disequilibrium ($r^2 = 0.8$) with variants in the NHGRI-EBI GWAS Catalog[4]. 6 variants were associated with disease traits. This included rs1049508 and rs199645276 in GATM, which were associated with increased levels of the cardioprotective metabolite homoarginine[65].

Finally, we observed that MPRAu identified variants of potential

physiologic interest that might not be detected in GWAS or eQTL studies due to population constraints. We observed that the minor allele of rs78562329, which is virtually absent from European populations, was significantly upregulated and may therefore be the first gain-of-function variant identified in the Furosemide receptor SLC12A1[66]. The allelic effect of this variant was independently validated by our luciferase assay.

As previously noted, MPRAu allelic skews are enriched for known regulatory annotations such as microRNA and RBP binding sites. Despite this enrichment, regulatory annotations explained only a small fraction of our shortlist. Of the 115 variants, only 13 were concordant with known regulatory annotations. Thus, many regulatory effects identified by MPRAu are mediated by novel, unannotated mechanisms.

3.4 Discussion

In order to identify functional 3'-UTR variants from the thousands of variants implicated in selection and disease, we developed MPRAu, a high-throughput tool that directly tests 3'-UTR variants for regulatory effects on transcription, termination, stability, and/or translation. Using this novel tool we identified 115 variants with regulatory molecular phenotypes, which we provide as a resource of candidate causal variants in selection and disease.

MPRAu quantifies steady-state RNA expression levels of a reporter protein containing a variable 101bp oligonucleotide corresponding to one allele of a given variant in its 3'-UTR (Figure 3.1). As a result, it is capable of quantifying

several distinct but important regulatory mechanisms that take place in the 3'-UTR. We showed that MPRAu can detect elements associated with decreased mRNA stability, including microRNA binding sites, AU rich elements, and Pumilio binding sites. In addition, MPRAu detects elements involved in termination/polyadenylation, such as the polyadenylation signal sequence (PAS). Interestingly, we showed that MPRAu is especially sensitive to the position of the PAS within the oligo (Figure 3.3C). Finally, as MPRAu measures steady-state RNA expression, it is also likely sensitive to transcriptional effects. However, similar approaches suggest that transcriptional effects may be rare in the 3'-UTR compared to post-transcriptional effects such as mRNA destabilization[67] and termination[68].

We caution against speculation on the overall contribution of each of these mechanisms in MPRAu steady-state RNA expression, as sensitivity may vary between mechanisms. For example, microRNA-mediated destabilization involves shortening of the poly(A) tail and eventual exosomal degradation[21], while early termination may prevent amplification and thus immediately remove target mRNA. As another example, a weak promoter may sensitize MPRAu to stabilizing elements, while a strong promoter may promote detection of destabilizing elements. Rather, we propose MPRAu as a frontline tool to prioritize functional 3'-UTR variants for independent validation studies that characterize their regulatory mechanism and physiological role(s).

On the other hand, by simultaneously quantifying total RNA and polysomal RNA expression levels, MPRAu does allow the distinction of translational effects

from changes in steady-state RNA expression. This differs from previous approaches, which have relied on cell sorting by reporter protein expression[67–70] ("Flow-Seq"), an output that is cumulatively affected by RNA quantity and translation efficiency. We observed that under standard growth conditions in HEK293FT cells, polysomal RNA expression was well-correlated with steady-state RNA expression (Figure 3.3E), suggesting that translational effects of 101bp 3'-UTR oligonucleotides were marginal at best. This does not take into consideration regulatory effects that impact translation at the level of the number of ribosomes per transcript rather than the proportion of transcripts being actively translated. Nonetheless, this result is in line with the concordance of RNA-level changes and marker bin enrichment in Flow-Seq studies, as well as the concordance of RNA expression and ribosome occupancy in eQTL/rQTL studies[45]. The paucity of translational effects that we observed for the 101bp 3'-UTR segments that we tested is surprising given the abundance of evidence implicating microRNAs and other 3'-UTR elements in translation[54]. 3'-UTR sequences may have an impact on translation within specific contexts. For example, microRNAs have been shown to primarily act on mRNA levels in HeLa cells[71], but to predominantly impact translation during Zebrafish development until shortly after the onset of zygotic transcription[72]. Furthermore, translational effects may involve complex secondary structures that require greater than 101bp of sequence context.

By comparing steady-state RNA expression between alleles, we established MPRAu as an effective tool for identifying the small fraction of

common polymorphisms that are functional. Out of 9,754 tested variants, only 115 variants met our stringent Bonferroni multiple testing threshold. This echoed our previous work[20], in which out of 32,373 eQTL-linked variants, only 842 were functional. The low proportion of common (MAF>5%) variants that are functional may be due to purifying selection, which acts to prune variants with deleterious functions from the population before they reach a high allele frequency. Lending support to this idea, we demonstrated that functional variants are depleted from regions under a high degree of purifying selection (Figure 3.4C). Similarly, although we tested regions under positive selection, the majority of our 412 loci are expected to contain one causal variant which may not reside in the 3'-UTR. The small proportion of variants with allelic differences in expression highlights the utility of high-throughput validation approaches such as MPRA and MPRAu.

Although we set a stringent threshold for defining functional variants, we showed enrichment for known regulatory elements within the top quartile of skew (Figure 3.5A). With higher coverage, it is possible that many of these variants would have met our significance threshold. Despite enrichment for regulatory elements among variants with subtle skews, only variants with a high magnitude of skew were enriched in eQTLs above background (Figure 3.5B). Thus, current eQTL studies lack the sensitivity to detect subtle regulatory changes. In the future, this may be improved through greater sample sizes and/or better normalization of sample covariates.

The 115 variants that met our significance threshold are prime candidates for functional follow-up and include putative causal variants for selection (such as

rs34448361 in LEPR) and disease (such as rs1049508 and rs199645276 in GATM), as well as variants whose population demographics might have allowed them to escape detection by GWAS and eQTL studies (such as rs78562309 in SLC12A1). We validated rs78562329 as well as rs199645276 in luciferase assays, and showed that while luciferase assays may be more accurate for measuring the regulatory effects of termination, they also may be less sensitive to detect subtle allelic effects (Figure 3.6). This may be due to difficulty obtaining sufficient signal without overexpression of a single sequence and consequent saturation of endogenous regulatory factors, an issue that may be overcome by simultaneous overexpression of the regulatory factor(s) mediating the allelic effect. However, only a very small proportion of our candidates could be explained by known regulatory elements such as microRNA binding sites, AU-rich elements, or RBP binding sites, an observation that highlights the importance of unbiased validation approaches.

MPRAu allows for direct and high-throughput identification of functional 3'-UTR variants. Although we piloted MPRAu by testing common variants under standard growth conditions, its high-throughput nature is well-suited to testing rare variants, a class whose importance in regulatory variation and disease is being increasingly recognized[73]. Furthermore, minimal adaptations would be needed to test 5'-UTR variants, which are thought to play a large role in translation initiation[74]. Finally, by assaying short variable sequences with an attached unique molecular barcode, we identified a bias in reverse transcription that may play a role in the discrepancy in translational buffering observed in the

literature[45,75–78] (Figure 2.6). This sheds light on another utility of engineered oligonucleotide libraries: quality control. MPRAu provides a framework for prioritizing regulatory variation in untranslated regions, bringing us closer to linking regulatory processes that occur in this highly active portion of the genome with specific phenotypic traits.

3.5 Experimental Procedures

3.5.1 Variant Selection

Variants were taken within 1Mb tiles which attained genome-wide significance for positive selection when applying CMS[8] to pilot phase data of the Thousand Genomes Project (1kG). Single Nucleotide Polymorphisms (SNPs) and indels up to 10bp were taken from Phase 1 of 1kG using individuals from pilot populations, e.g. CEU (Utah residents with ancestry from Northern and Western Europe), YRI (Yoruba in Ibadan, Nigeria), CHB (Han Chinese in Beijing, China), and JPT (Japanese in Tokyo, Japan). To select 3'-UTRs, we searched Gencode v19 for UTR-annotated entries downstream of CDS-annotated entries, filtering entries annotated for nonsense-mediated decay and non-stop decay. Variants with a global minor allele frequency, MAF ≥ 5% were aligned with Gencode v19 3'-UTR entries.

3.5.2 Oligo Synthesis and Amplification

Oligos were synthesized by CustomArray as 133 bp sequences containing

100 bp of UTR context and adapters on either end (5'CGAGCTCGCTAGCCT [101 bp oligo] AGATCGGAAGAGCGTCG3') (Figure 3.1A). Post synthesis, 6bp random hexamer barcodes and additional adapter sequences were added by performing 20 PCR reactions each 50µl in volume containing 5.7ng of oligo, 25 µL of Q5 NEBNext MasterMix (NEB, M0541S), 1 unit Q5 HotStart polymerase (NEB, M0493S), 0.5 µM oligo_BAR_Bmt_F (CTAGTTGTTTAAAGCCCAACGCTAGTCNNNNNNCGAGCTCGCTAGCCT) and oligo_pmir_R_min (CGACGCTCTTCCGATCT) primers and 20 µg BSA (NEB, B9000). PCR reaction was purified by performing a 2.5X SPRI purification using Agencourt AMPure XP SPRI (Beckman Coulter, A63881) beads according to manufacturer instructions.

3.5.3 MPRAu Vector Assembly

To create our MPRAu plasmid pool (Figure 3.1B), barcoded oligos were inserted into BmtI/XbaI digested pmirGLO:Δluc::gfp ΔAmp^R::Kan^R by gibson assembly (NEB, E2611) using 1µg of digested vector and a 10:1 molar ratio (415ng) of amplified oligos in a 40µl reaction incubated for 60 min at 50°C followed by SPRI purification and eluted in 25µl EB, then concentrated to 10µl by vacuum centrifugation. The ligated vector was then split into 5µl aliquots, each of which was transformed into 100 µL of 10-beta e.coli (NEB, C3020K) by electroporation (2kV, 200 ohm, 25 µF). Electroporated bacteria were immediately split into five 1mL aliquots of SOC (NEB, B9020S) and recovered for 1 hour at 37°C then independently expanded in 1L of LB supplemented with 50µg/mL of

kanamycin (Teknova, K2125) on a floor shaker at 37°C for 12 hours. After outgrowth aliquots were pooled prior to plasmid purification (Qiagen, 12991). For each of the aliquots we plated serial dilutions after SOC recovery and estimated a library size of $>10^8$ CFUs.

3.5.4 MPRAu Experiments

HEK293FT cells were grown in DMEM (ThermoFisher, 10564) supplemented with 10% FBS (ThermoFisher, 26140). For all 5 transfections (Figure 3.1C), cells were plated in a 15cm plate and grown to a density of ~80-90%. Cells were transfected with 80µl Lipofectamine 2000 (ThermoFisher, 11668027) and 20µg DNA. Cells were incubated with transfection reagents for 24 hours, monitored by fluorescent microscopy, and then split 1:3 into 2 new 15cm plates. After 24 hours (48 hours post-transfection), media from each plate was replaced with 30mL DMEM + 100µg/mL cycloheximide (CHX) (Sigma-Aldrich, C4859) and incubated for 5 minutes at 37°C. Both plates were then washed with 10mL cold PBS (ThermoFisher, 14040) + 100µg/mL CHX, scraped in 1mL cold PBS (ThermoFisher, 14090) + 100µg/mL CHX, pooled and centrifuged for 5 minutes at 500xg, and finally resuspended in 350µl lysis buffer composed of 5mM Tris pH 7.5 (ThermoFisher, 15567), 2.5mM Magnesium Chloride (ThermoFisher, AM9530G), 1.5mM Potassium Chloride (ThermoFisher, AM9640G), 2µM DTT (VWR, 97061), 100µg/mL CHX, 5 mg/mL Triton-X (Sigma-Aldrich, T8787), 5 mg/mL Sodium Deoxycholate (Sigma-Aldrich, 30970). Cell lysate was then centrifuged for 5 minutes at 12,000xg and supernatant was flash-

frozen in vapor phase nitrogen. For polysome profiling, approximately 200µl lysate was loaded onto 10-50% sucrose gradients in 15mM Tris pH 7.5, 15mM Magnesium Chloride, 150mM Sodium Chloride (ThermoFisher, AM9760G), and 100µg/mL CHX. Gradients were centrifuged in SW41Ti rotor at 35,000 rpm for 2.5 hours at 4°C and 0.5mL fractions were collected.

3.5.5 RNA Extraction and cDNA Synthesis

RNA was extracted with TRIzol LS (ThermoFisher, 10296) according to manufacturer instructions. Total RNA was purified from lysate. Polysomal RNA was purified from sucrose fractions, pooling fractions corresponding to three or more ribosomes. 7.5µg GlycoBlue (ThermoFisher, AM9515) was added each sample in order to visualize the pellet. mRNA was purified from total RNA using oligo $d(T)_{25}$ magnetic beads (NEB, S1419S) according to manufacturer's instructions, and eluted at 80°C. Purified total and polysomal mRNA were then subjected to Turbo DNase treatment (ThermoFisher, AM2239). The reaction was terminated in 1 mg/mL SDS (ThermoFisher, AM9822) and purified by performing a 2X SPRI purification using Agencourt RNAClean XP SPRI (Beckman Coulter, A63987) beads according to manufacturer instructions. DNase-treated mRNA for each of the five replicates was diluted to the concentration of the lowest concentration sample, and first-strand cDNA was synthesized from concentration-normalized mRNA with SuperScript III (ThermoFisher, 18080) and a gene-specific primer 162bp downstream of the oligo (oligo_RT_R, GCATTCTAGTTGTGGTTTGTCCA) using the manufacturer’s recommended

protocol, modifying the total reaction volume to 40 µL and performing the elongation step at 55°C for 80 minutes. Single-stranded cDNA was purified by performing a 2X SPRI purification using Agencourt RNAClean XP SPRI beads.

The translational “buffering” artifact that we described was observed using a gene-specific primer directly downstream to the oligo (oligo_pmir_R_min, CGACGCTCTTCCGATCT). To study how this artifact was affected by reverse transcription conditions, we also performed reverse transcription using an oligo(dT) primer.

3.5.6 qPCR and Library Construction

To minimize amplification bias during the creation of cDNA sequencing libraries, samples were amplified by qPCR to estimate relative concentrations of GFP cDNA using 1 µL of sample diluted 1:100 in a 20 µL PCR reaction containing 10 µL PowerSYBR master mix (ThermoFisher, 4367659), and 0.5 µM of PE_PCR_P1 (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT) and PE_PCR_P2_Bmt (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTTTAAAGCCCAACGCTAGTC) primers, which add adapters for Illumina sequencing. Samples were amplified with the following conditions: 95°C for 10 minutes, 40 cycles (95°C for 15 sec, 65°C for 1 min). For GFP quantification across the sucrose gradient (Figure A.4), qPCR was performed on cDNA prepared from each fraction as described above.

The highest Ct across all MPRAu cDNA samples was 20. Therefore, to add Illumina sequencing adapters, 10µl of cDNA from each replicate was amplified in a 50µl PCR reaction containing 25µL Q5 NEBNext master mix (NEB, M0541S) and 0.5µM of PE_PCR_P1 and PE_PCR_P2_Bmt with the following conditions: 98°C for 30 seconds, 15 cycles (98°C for 10 sec, 65°C for 30 sec, 72°C for 30 sec), 72°C for 5 minutes, following by 2.5X AMPure SPRI purification and another round of PCR (as above, with 5 cycles) using PE_PCR_P1 and a TruSeq index primer. For approximation of plasmid pool counts, the plasmid pool was diluted 1:30,000 and amplified as above in five independent PCR reactions (technical replicates). Indexed libraries were SPRI purified and pooled according to molar estimates from Agilent TapeStation quantifications. Samples were sequenced using 2x150 bp chemistry on an Illumina HiSeq through the Broad Institute's walk-up sequencing service.

3.5.7 Luciferase Assays

To validate the expression values obtained by MPRAu, we selected 18 oligos consisting of 9 ref/alt pairs. Five of the oligos were selected as no-skew controls for having uncorrected p-values of greater than 0.01. We designed the same 101 bp sequence that was tested by MPRAu as a gBlock (IDT) and cloned each into the pmirGLO dual-luciferase reporter vector (Promega, E1330). We co-transfected HEK293FTs with 500pg of the cloned dual-luciferase vector and 49.5ng of pGL4.23, a promoterless control vector (Promega, E8411) in a 96 well plate, split the cells 1:3 after 24 hours, and read luminescence at 48 hours post-

transfection. We performed five transfection replicates per per oligo. Firefly luciferase luminescence for each well was normalized to the renilla luciferase luminescence for the same well and each experiment was normalized as a log-ratio value relative to the mean of a control oligo with an MPRAu RNA/DNA ratio of -0.2.

3.5.8 Oligo Read Alignment

Paired-end 150 bp reads were merged into single amplicons using Flash[43] v1.2.11 (flags: -M 150, -O). Amplicon sequences were kept if the sequence surrounding the barcode met the following conditions: (1) a perfect match was found to the 10bp sequence on either the left or right side of the barcode, (2) the 10bp sequences on both the left and right sides of the barcode matched with a Levenshtein distance of 3 or less, and (3) the 2bp immediately surrounding each side of the barcode matched perfectly. Oligo sequences from the passing reads were then mapped back to the expected oligo sequences using BWA-MEM[44] version 0.7.12 (flags: -M). Alignment scores were calculated as matching bases divided by the expected oligo size and reads with alignment scores of less than 0.95 were discarded. Oligo libraries were extremely complex, with each oligo/barcode pair being observed on average 2-3 times in the plasmid replicates and 4-5 times in the total RNA and polysomal RNA replicates. As a result, oligo reads were pooled across barcodes for oligo analyses and barcode reads were pooled across oligos for barcode analyses.

3.5.9 Expression and Allelic Skew Quantification

Oligo counts from all 15 samples (5 plasmid PCR replicates, 5 total RNA, 5 polysomal RNA) were passed into DESeq2 and a median- of-ratios method was used to normalize samples for varying sequencing depths[79]. Normalized read counts of each oligo were then modeled by DESeq2 as a negative binomial distribution (NB). DESeq2 estimates variance for each NB by pooling all oligo counts across samples and fitting a trend line to model the relationship between oligo counts and observed dispersion. It then applies an empirical Bayes shrinkage by taking the observed relationship as a prior and performing a maximum a posteriori estimate of the dispersion for each oligo.The overall result is that DESeq2 can obtain an estimate for dispersion of each oligo with greatly reduced bias by pooling information from all oligos. We excluded plasmid samples from the dispersion calculation, as these samples were technical replicates and were observed to follow a Poisson distribution (Figure A.3).

We then used DESeq2 to estimate oligo fold changes between conditions, e.g. the baseline counts in the plasmid pool, total RNA counts in the lysate, and polysomal RNA counts (design = ~ Replicate+Condition). We estimated fold changes between the reference and alternate alleles (design = ~ Replicate+Condition+Variant+Condition:Variant) using a Wald test with Bonferroni multiple test correction and a significance threshold of 0.1. As expected, p-values were uniformly distributed (Figure A.2). In all models, a replicate term was added in order to pair samples from the same transfection.

3.5.10 Annotations Used for Enrichment Analysis

To predict microRNA binding sites, we used TargetScan6[80] to identify 7mer and 8mer seed matches to the top 10 and top 100 most abundant microRNAs in HEK293 cells lysate according to previously published Argonaute PAR-CLIP data in HEK293 cells[81]. 6mer seed matches were determined manually for the top 10 most abundant microRNAs. To predict AU-rich elements, we used the ARE-mRNA Database criteria[62] for class I and class II AU-rich elements. To predict Pumilio binding sites, we used the canonical motif[63] UGUANAUA.

3.5.11 Conservation Analysis

Phastcons scores for the 100 vertebrate species alignment were downloaded from the UCSC genome browser. For each oligo, the average Phastcons score was calculated in the 100bp context centered around the variant position. Some oligos did not have a full 100bp context due to a splice site lying close to the variant - for those sites, we simply took the maximum amount of sequence context until reaching the boundary. To assess whether conservation scores were related to expression levels, we binned RNA expression levels for all oligos in increments of 10% and plotted the mean conservation for each corresponding bin (Figure 3.4C).

3.5.12 Concordance of Allelic Skews with GTEx Data

The file "GTEx_Analysis_v6p_eQTL.tar," which contained significant snp-gene associations across 44 human tissues, was used from the Single Tissue cis-eQTL dataset in GTEx. An oligo was considered an eQTL if it was classified as having a significant snp-gene association in any of the 44 tissues. To determine whether variants were more likely to be eQTLs given higher allelic skew or significance cutoffs, we calculated the proportion of eQTLs across all variants passing a range of allelic skew or p-adjusted value cutoffs (Figure 3.5B & Figure 3.5C).

After labeling variants as eQTLs from the GTEx dataset, we performed a permutation test to assess the significance of the upward trend observed in Figure 3.5B. For all variants, we permuted the labels of which variant is called an eQTL, then took the slope coefficient from a linear regression performed on the proportion of variants classified as eQTLs against the absolute value of the skew threshold cutoff values. We performed this permutation 10,000 times, then took the fraction of permuted slope coefficients greater than the observed slope coefficient for the trend in Figure 3.5B as a p-value. A p-value of 0.0303 was derived for such a test, which provides further evidence that variants with higher skew are more likely to be enriched to be GTEx eQTLs. We observed a similar significance on the downward trend observed in Figure 3.5C (p-value=0.0052).

To assess whether or not the effect size of GTEx eQTLs were concordant to POLYskew, we took variants that were considered GTEx eQTLs and had a significant allelic skew (p-adj<0.1), then we investigated the effect size of the

eQTL against PolyDSkew for these variants in each tissue (Figure A.5). Because most tissues had very few variants that were both GTEx eQTLs and had high allelic skew, we only considered tissues which had greater than 6 points. Although none of the individual R^2 values were highly significant, across almost all of the tissues we observed a positive correlation between between GTEx eQTL slope vs. PolyDSkew (15/17 tissues).

3.5.13 Selection and GWAS Associations

To determine putative causal variants for positive selection, we overlapped differentially expressed variants with positively selected regions identified by the Composite of Multiple Signals test[8] in 26 populations genotyped by the Thousand Genomes Project Consortium[3]. CMS localizes positively selected regions to loci of median size 27kb. Variants which overlapped a selected locus in any of the genotyped populations were noted in Table A.1 (CDX: Chinese Dai in Xishuangbanna, CEU: Utah Residents with European Ancestry, CHB: Han Chinese in Bejing, CHS: Southern Han Chinese, GWD: Gambian in Western Divisions, ITU: Indian Telugu from the UK, KHV: Kinh in Ho Chi Minh Vietnam, MSL: Mende in Sierra Leone, STU: Sri Lankan Tamil from the UK, YRI: Yoruba in Ibadan).

To determine overlap with disease-associated regions, the 02/13/2017 release of the NHGRI-EBI GWAS Catalog was downloaded from EBI (https://www.ebi.ac.uk/gwas). PLINK v1.9 was used to identify differentially-expressed variants in linkage disequilibrium (r^2>0.8) with genome-wide significant

($p<5x10^{-8}$) associations.

3.6 Contributions

James Xue performed eQTL and conservation analyses. Joseph Vitti performed CMS analysis. Dustin S. Griesemer performed all experiments, read alignment, and remaining analyses. Drs. Ryan S. Tewhey and Cailin Joyce provided scientific and technical guidance. Dr. Stephen Schaffner provided critical review during manuscript preparation. Drs. Carl D. Novina and Pardis C. Sabeti provided scientific guidance.

3.7 Acknowledgements

We would like to acknowledge Dr. David P. Bartel for providing invaluable scientific guidance. This work was supported by the National Institutes of Health under Ruth L. Kirschstein National Research Service Award 5F30GM114940 from the National Institute of General Medical Sciences, as well as by the National Science Foundation under Doctoral Dissertation Improvement Grant DEB-1401237 from the Division of Environmental BIology.

Chapter 4: Perspectives

4.1: Utility of MPRAu

4.1.1 Massively Parallel Validation

We utilized a novel high-throughput screening tool, MPRAu, to test 9,754 variants for allelic differences in steady-state expression and translation. Out of this large set, we identified 115 variants (Table A.1) with significant differences in steady-state expression. The high-throughput nature of MPRAu is perhaps its most important utility. Current approaches to identify functional 3'-UTR variants are limited, relying on either in-silico prediction of functional elements[59,60], QTL analysis[45], or low-throughput validation. None of these approaches are as powerful as MPRAu for prioritization of the thousands of 3'-UTR variants implicated in selection[8] and disease[4]. While in-silico prediction and QTL analysis are capable of assessing millions of variants, these approaches lack specificity[82]. On the other hand, low-throughput testing requires molecular cloning of each sequence of interest, and is not feasible on a large scale. The creation of a repository of human 3'-UTRs would facilitate parallelization of low-throughput approaches, and is underway with Gateway® cloning for 1,461 full-length 3'-UTRs having already been reported[83]. However, even after the creation of a "UTRome," MPRAu and array-based synthesis methods will likely retain utility in variant testing.

4.1.2 Sensitivity to Subtle Allelic Differences

We also demonstrated that MPRAu is remarkably sensitive for detecting subtle allelic differences in steady-state RNA expression. While few of the 115

variants meeting our stringent Bonferroni multiple correction testing threshold overlapped annotated regulatory elements, annotated regulatory elements were highly enriched amongst the approximately 2500 variants in the top quartile of allelic skew (Figure 3.5A), suggesting that allelic effects that did not meet our stringent threshold were nonetheless meaningful. In contrast, only the strongest allelic effects were detected by eQTL analysis (Figure 3.5B,C), despite general agreements in directionality. Furthermore, we observed much smaller effect sizes for luciferase assays, despite a strong overall correlation in allelic skew (Figure 3.6B). We posited that the reduction in effect size that we observed in our luciferase assays resulted from increased baseline expression in singleplexed transfections.

4.1.2 Unbiased Testing

The dearth of annotated elements amongst variants with large effect sizes highlights another important utility of MPRAu: its ability to screen 3'-UTR variants in an unbiased fashion. Approaches that rely on perturbation of known regulatory factors have often found that while the distributions of predicted targets appear to validate known targeting rules, even the best models suffer from poor predictive power[82]. Furthermore there remains much to be learned about the target specificity of each of the hundreds of trans-acting factors that may bind to 3'-UTR sequences[84]. In the future, MPRAu may help contribute to this knowledge gap by identifying regulatory motifs responsible for functional variation. This may be carried out through a separate MPRAu experiment testing every single

nucleotide variant (SNV) within 10-20bp around each functional variant (Figure 4.1), and/or performing deletion tiling along the oligonucleotide (Figure 4.2) to identify critical sites within the variant's sequence context (such as the PAS for a variant that affects termination).

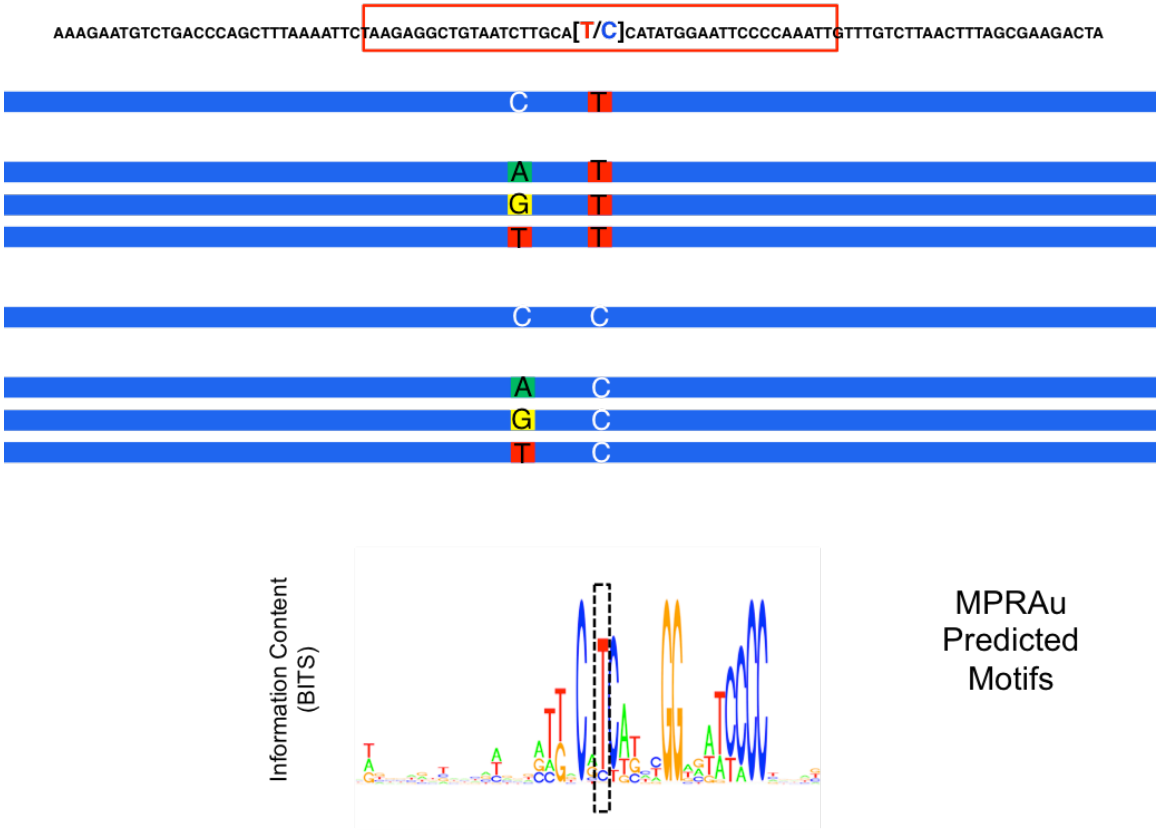


Figure 4.1: Mutagenesis of MPRAu-identified functional variants. A sample oligonucleotide sequence pair, centered about a variant, is shown above. Reference alleles appear in white. Oligonucleotides may be generated to test every possible single nucleotide variant (SNV) within 20bp of the variant. Each oligonucleotide's expression may be compared to the reference sequence in order to obtain an optimal sequence motif.

Beyond allowing validation without prior knowledge of the trans-acting factor involved, MPRAu is capable of detecting multiple post-transcriptional

processes, including mRNA decay via microRNA targeting, termination and polyadenylation due to the presence of a polyadenylation signal (PAS), or potential transcriptional effects due to transcriptional enhancers or repressors. We made several observations suggesting that many of the 101bp 3'-UTR segments we tested created early termination sites, including a position-dependent effect of the polyadenylation signal (PAS) (Figure 3.3C), a reduction in mRNA read counts of uridine-rich oligonucleotides (Figure 3.3B), and discordance in luciferase assay readout for uridine-rich oligonucleotides as well as those containing a PAS (Figure 3.6A). A large degree of genetic variation in alternative polyadenylation has recently been demonstrated[27,45], and individual differences in polyadenylation have been shown to impact health and adaptation[24] An important next step that we are currently pursuing is the preparation of PolyA-seq libraries from MPRAu samples to definitively identify the presence of early termination sites[53]. We are also pursuing evidence for early termination sites, as well as allelic differences in transcript abundance, in RNA expression data from the Genotype-Tissue Expression Consortium (GTEx).

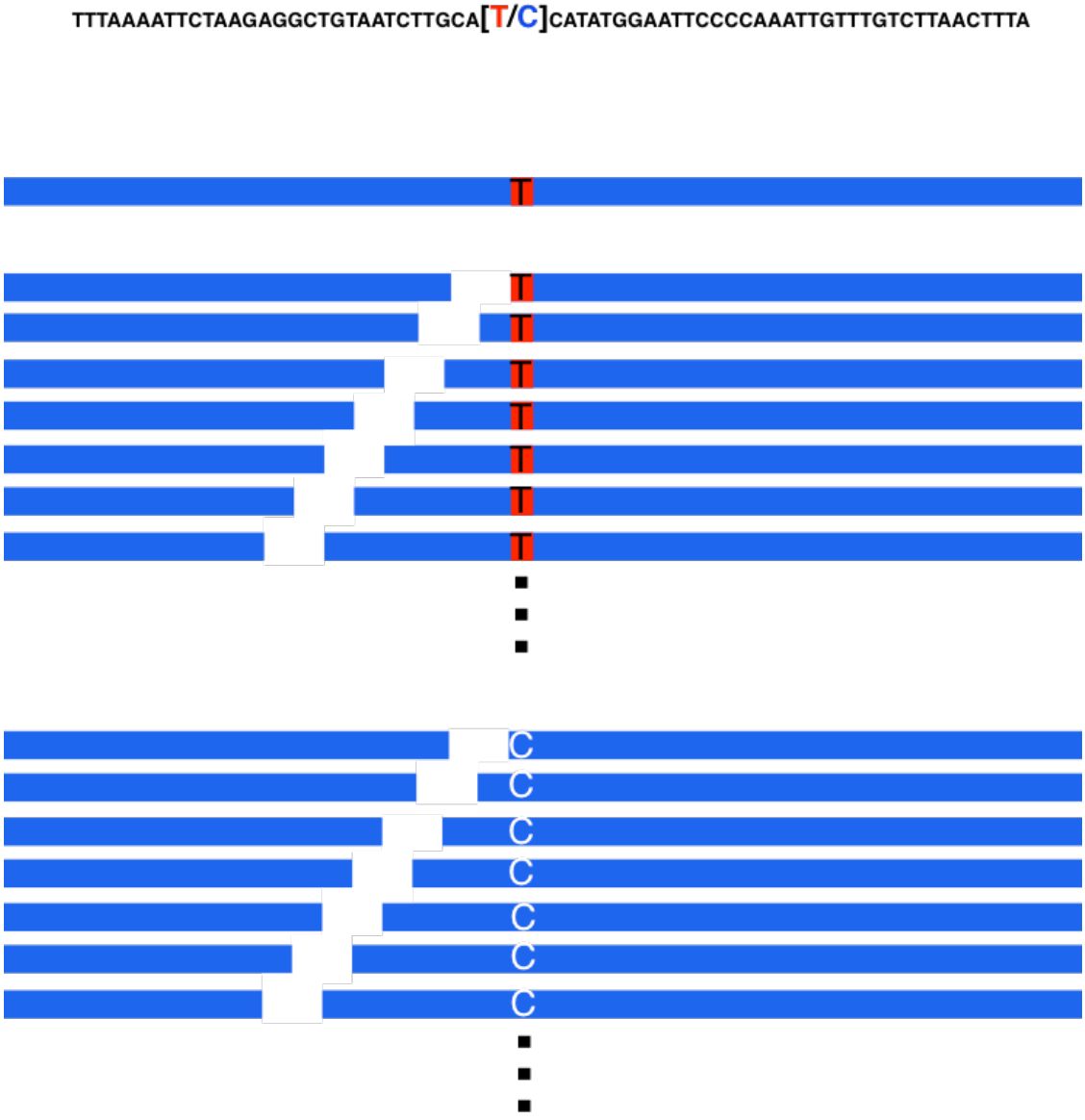


Figure 4.2: Tiling deletions of MPRAu-identified functional variants. A sample oligonucleotide sequence pair, centered about a variant, is shown above. Reference alleles appear in white. Oligonucleotides may be generated to test deletions within the sequence context of the oligo. Deletions may be tiled in order to reduce the number of oligonucleotides tested. This allows detection of regions in the variant's sequence context that help mediate its observed allelic effect.

4.2 Limitations of MPRAu

4.2.1 Technical Limitations

MPRAu also has important limitations, both technically and conceptually. Because each oligo is not equally represented in the plasmid library, RNA read counts must be normalized to plasmid read counts. This creates the potential for bias during the reverse transcription step. An oligonucleotide that inhibits reverse transcription would artificially appear to have an attenuating effect on expression. We observed this artifact when we placed the gene-specific primer used during reverse transcription directly adjacent to the variable oligonucleotide (Figure 2.6). Oligos that formed secondary structures with the GSP annealing site appeared to have attenuating effects on expression, creating the illusion of translational "buffering" due to observed allelic differences in steady-state expression that were not present in polysomal RNA. We tied the "buffering" phenomenon to the GSP by demonstrating that the phenomenon disappeared using a downstream GSP. It is interesting that reverse transcription of polysomal RNA was more robust to secondary structure than reverse transcription of steady-state RNA. It is unlikely that this is carried forward from an in-vivo difference in secondary structure formation, because prior to reverse transcription, isolated RNA undergoes oligo(dT) mRNA purification involving elution at 80°C.

In most cases, however, we would expect biases in reverse transcription to affect reverse transcription of steady-state RNA and polysomal RNA equally. This helped us make the observation that 101bp 3'-UTR segments have strikingly little impact on the proportion of mRNA that is translated (Figure 3.4B).

We confirmed that the subset of translated GFP mRNA represented only a fraction of the total (Figure A.4), and thus GFP translation initiation is not saturated at baseline. Nonetheless, we are limited in drawing conclusions about the impact of 3'-UTR sequences on overall protein production due to another limitation of MPRAu, its inability to detect differences in the number of ribosomes per translated mRNA molecule. Better resolution may be obtained by subsetting the polysomal isolate into light and heavy polysomal pools. However we caution against overly subsetting the polysomal product, as this may result in reduction of library diversity. Flow-seq approaches, which quantify reporter protein levels as their output, may be complementary to polysome profiling, which quantifies translation initiation in isolation. Recent Flow-seq studies studying the effects of short 3'-UTR segments have found strong concordance between steady-state expression and protein output, further suggesting that short 3'-UTR segments exert little impact at the level of translation[67,69].

4.2.2 Conceptual Limitations

MPRAu, as well as other methods that test short oligonucleotide sequences outside of their genomic context, also suffers from an important conceptual limitation: the UTR variant is placed within a limited (101bp) sequence context, outside of its physiologically relevant context. Some sites may rely heavily on orientation within the transcript. For example, microRNA target site efficacy decreases with distance from the 3'-UTR ends[85]. 3'-UTR variation may also act through transcriptional enhancers[86] and repressors,[87] which

oftentimes act through highly specific looping interactions with the endogenous promoter or depend on the surrounding chromatin state. These features would not be recapitulated by an episomal assay, in which the oligonucleotide maintains a constant position within a plasmid containing the non-native promoter, protein-coding sequence, and polyadenylation site.

This limitation is not limited to elements with strict dependencies on endogenous context; it is also relevant to large elements that may not be fully captured by the oligonucleotide sequence. This may be especially important in the 3'-UTR, where large secondary structures may form[88]. For example, adenosine deaminases acting on RNA (ADAR) family enzymes, which edit adenosine to inosine in mRNA, predominantly act on long (>100bp) dsRNA structures formed by inverted *Alu* repeats[89].

We observed an interesting example of context dependence during the detection of alternative polyadenylation sites. Alternative polyadenylation sites cause early termination, the effects of which are thought to vary depending the cumulative impact of regulatory elements downstream of the early termination site[90]. Critically, the MPRAu expression readout depends on the presence of annealing sites for sequencing and reverse transcription (GSP) adapters. Termination upstream to the required adapters presumably results in an mRNA product that cannot be amplified by PCR, and therefore a reduction in mRNA read counts. In contrast, termination downstream of the required adapters would result in a PCR-competent product. Furthermore, this has the potential to remove destabilizing elements between the oligonucleotide and the plasmid

polyadenylation signal, resulting in an increase in mRNA read counts. Indeed, we observed that the PAS had a position-dependent effect on MPRAu readout (Figure 3.3C). Early termination signals appeared to increase expression in the luciferase assay, which differs from MPRAu in that its readout should not be impeded by early termination. The true effect of early termination would strongly depend on regulatory elements downstream of the termination site within the endogenous 3'-UTR, and thus even the luciferase assay is not capable of accurately quantifying the end-result of early termination using 101bp of sequence context.

4.3 Future Directions of MPRAu

4.3.1 Application to Rare Variants

Despite its limitations, MPRAu is a sensitive tool for high-throughput detection of functional variation in UTR regions. Out of 9,574 common variants tested, e.g. those with a minor allele frequency of 5% or greater, we identified 115 variants with significant allelic differences in expression. The proportion of variants with a significant allelic expression difference was low (approximately 1%), although this likely partially resulted from the stringent Bonferroni cutoff that we used to call allelic differences. We observed that variants in the top quartile of skew (a much less stringent threshold) were depleted of conserved regions. This would be expected under a model of purifying selection acting to deplete variants with functional effects from conserved regions, where they likely have deleterious effects, before they reach high allele frequencies in the population[58].

Under this model, large functional effects, especially those that are deleterious, might be expected to be more prevalent amongst rare variants. Indeed, Mendelian diseases are typically associated with rare variants[11], and recent work has shown that individuals who are expression outliers for a particular gene in GTEx often have rare variants nearby the affected gene[73]. Thus, an MPRAu assay testing rare variants may identify a greater number of variants with allelic expression differences than we observed for common variants. Notably, there is an exponentially greater number of rare variants than common variants, and very large sample sizes are needed to capture them[91] in population studies. MPRAu, which allows each variant to be generated on an oligonucleotide array for high-throughput testing, is therefore especially well suited to the testing of rare variants.

4.3.2 Application to 5'-UTR Variants

One adaptation that we incorporated into MPRAu to adapt it for the testing of UTRs was the incorporation of the polysome profiling technique, a classical method involving isolation of polyribosomal complexes on a sucrose gradient (Figure 3.1E). This allows allelic differences in translation initiation that affect the proportion of translated RNA to be detected alongside allelic differences in steady-state RNA expression.

Little effort would be needed to repurpose MPRAu for the testing of 5'-UTRs, which are known to play a large role in translation[92]. The insights that we drew from optimization of MPRA for testing 3'-UTRs (Chapter 1), such as the

importance of including a barcode as a unique molecular identifier and the sensitivity of reverse transcription to secondary structure, would aid in potential technical issues that may arise in the development of a similar assay. Furthermore 5'-UTR sequences may impact both transcription and translation, sometimes through interacting mechanisms[93]. Previous work testing randomized sequences within 5'-UTRs have relied upon a protein expression readout based either on Flow-Seq[94] or selection[95], and thus have been unable to separate transcriptional effects from translation. However, the separation of polysomal RNA from total RNA during polysome profiling allows the isolation of translational effects, which may be uniquely suited to 5'-UTR validation.

As for any 5'-UTR validation approach, the context in which the tested sequence is placed may be especially relevant. 5'-UTRs impact translation through multiple mechanisms (Figure 4.3), including complex secondary structures, upstream start codons with an in-frame stop codon upstream of the main coding region (uORFs), and upstream start codons without an in-frame upstream stop codon (uAUGs)[92]. These elements are particularly sensitive to sequence context. For example, uORFs may become uAUGs if the in-frame stop codon is not captured. This challenge may be mitigated by the short length of 5'-UTRs compared to 3'-UTRs.

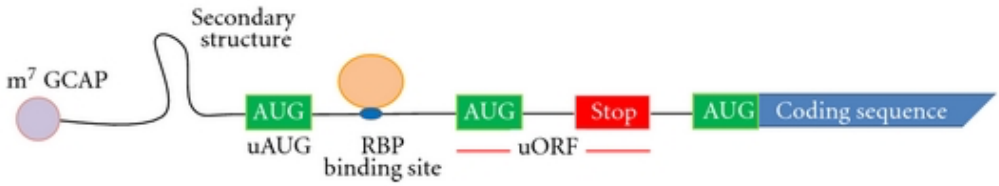


Figure 4.3: 5'-UTR Regulatory Mechanisms. Many regulatory mechanisms may act on 5'-UTR sequences, including ribosome recruitment to complex secondary structures, uAUGs and uORFs, and the binding of RNA-binding proteins (RBPs)[92].

4.3.3 Application to Specific Post-transcriptional Mechanisms

The MPRA technique, which was originally developed to quantify the effects of promoter/enhancer single nucleotide variants on steady-state RNA expression[19], is not capable of discerning the many distinct stages at which post-transcriptional regulation may occur. We adapted MPRA by incorporating polysome profiling, which allowed us to detect allelic differences in translation initiation distinctly from differences in steady-state RNA expression. Similarly, MPRA may be adapted to dissect specific post-transcriptional regulatory mechanisms in an isolated manner.

For example, TAIL-seq is a technique that allows simultaneous sequencing of the 3'-UTR end to determine the identity of a given transcript alongside sequencing of the poly-A tail to determine its length (number of adenosines)[96]. By incorporating engineered oligonucleotides into the 3'-UTR and performing TAIL-seq, one could dissect the sequence determinants of poly-A tail length, which in turn may be linked with translation efficiency and mRNA stability[97,98]. Notably, this assay setup would also be capable of detecting

alternative polyadenylation sites that may arise as a result of the engineered oligonucleotides.

In addition, a repressible promoter may be utilized in the plasmid backbone to directly test mRNA stability, as has been previously reported[67]. According to this approach, transfection replicates are harvested after a range of time points following transcriptional repression using an antibiotic. Afterwards, each oligonucleotide's mRNA stability is modeled based on its RNA expression levels during each point in the time course. Notably, as this involves comparison of RNA samples, it would be less likely to suffer from a reverse transcription bias. It may be particularly interesting to compare variant effects on mRNA stability to overall effects on steady-state expression, as QTL studies have found evidence for discordance[26].

4.3.4 Application to Relevant Physiological Contexts

We initially tested MPRAu in the HEK293 cell line as a result of its high transfection efficiency and high degree of overlap with other studies. Testing MPRA in additional physiological contexts may uncover cell-type specific regulatory effects. When using MPRA to test eQTL variants for promoter/enhancer function, we found a high degree of overlap between the lymphoblastoid (LCL) and hepatocarcinoma (HepG2) cell lines[20]. Nonetheless, eQTL studies suggest that while a large proportion of regulatory variation is concordant between two given tissues, tissue-specific regulatory effects may cumulatively outnumber multi-tissue regulatory effects[46]. Thus, while testing

MPRAu in a handful of cell lines may not yield many additional variants with significant allelic effects, testing MPRAu in many cell lines has the potential to uncover many additional effects.

Some physiological contexts may be particularly relevant to post-transcriptional regulation. One special context is early development, prior to the onset of zygotic transcription (the maternal-to-zygotic transition or MZT). Prior to the MZT, all regulation of protein levels occurs through post-transcriptional mechanisms, including widespread translational activation and repression[99]. As might be expected, post-transcriptional regulatory mechanisms behave differently during this period than in fully differentiated adult cells. For example, microRNAs were found to predominantly act by translational repression prior to the MZT, but to predominantly act on mRNA stability afterwards[72]. This observation was elaborated on by a later observation that poly(A) tail length undergoes a similar transition during the MZT[97].

Erythropoiesis, the process of red blood cell generation, is another special context in which the cell is especially dependent on translation due to extrusion of the nucleus during reticulocyte formation[100] as well as the concurrent high demand for hemoglobin production. In fact, defects in the ribosomal machinery, such as is observed due to haploinsufficiency of ribosomal proteins in Diamond-Blackfan Anemia[31], are known to affect erythropoiesis. An MPRA-based approach was previously used to screen 2,756 variants linked with red blood cell traits[101], however this study focused on transcriptional effects. MPRAu is poised to test common variants linked with rare blood cell traits or variants linked with

rare forms of anemia, in which it has the potential to uncover allelic differences in translation unique to this important physiological function.

Appendix A: Supplementary Tables and Figures

Table A.1: Differentially expressed variants identified by MPRAu. Variants with significant allelic expression differences identified by MPRAu. LFC: log2-fold change between alternate and reference alleles; padj: Bonferroni-adjusted MPRAu p-value; GWAS: LD with SNP in NHGRI-EBI GWAS Catalog and p-value of SNP in GWAS study; CMS Pop: population in which positive selective pressure was detected (see **3.5.13 Selection and GWAS Associations**).

Gene	Variant	Ref	Alt	LFC	padj	Annotations	GWAS (p-value;LD)	CMS Pop
HLA-DQA1	rs3667	A	G	-4.502	1.90E-183			
TTC30A	rs13020788	G	T	2.286	3.39E-163			
GLMN	rs2046620	G	C	-1.377	2.13E-115			CHB
FAM92B	rs16975240	T	C	2.070	2.40E-113	hsa-miR-92a		
PRKAG2	rs200222228	GT	T	-1.466	2.85E-99			
RP11-169F17.1	rs1532747	T	A	-1.720	3.56E-96			
SLC35E4	rs2240002	A	C	-1.880	5.21E-95			YRI
VPS8	rs58290679	C	T	-2.451	1.09E-76			
RBBP4	rs398500	T	C	1.109	1.47E-70			
TXN2	rs139999	G	T	2.312	2.02E-58			GWD
NAA40	rs202234263	TATATATA	T	-1.297	8.59E-38			CHS
NAA40	rs200234485	TATATA	T	-1.416	3.58E-34			CHS
ZNF641	rs2732479	T	G	1.466	6.63E-34			
GNB1L	rs5748423	A	G	1.540	1.65E-32			
APPBP2	rs75062707	T	A	1.586	6.56E-29			STU
NEDD4	rs1042477	G	T	-1.594	3.02E-28			
HLA-DQA1	rs9272962	G	T	-1.079	1.98E-25			
GATM	rs199645276	G	CAG	-0.821	1.81E-24		Serum Creatinine (1e-42; 1)	CEU
RAPGEF2	rs200077320	TG	T	-0.668	1.99E-24			
CCDC62	rs897393	C	T	-1.160	5.47E-24			
RBBP4	rs634787	C	T	0.760	8.03E-24			
GATM	rs1049508	T	C	0.873	1.51E-22		Serum Homoarginine (1e-45; 1)	CEU
TAP2	rs73410775	A	G	1.145	5.03E-20			
HMGA2	rs10878350	A	C	1.315	8.22E-19			CDX, KHV
STX3	rs482356	C	T	1.453	2.34E-18			
PRR23A	rs1553877	G	C	-0.949	8.76E-17		Keloids (2e-13; 0.99)	
TAP2	rs114247673	G	T	2.531	1.08E-15			
SLC30A9	rs4861160	A	T	-0.498	1.41E-15			
SRP54	rs11540758	C	T	-0.619	3.39E-12			
SLC8A3	rs8020841	A	G	0.636	5.19E-12			
APPBP2	rs11289661	CA	A	-1.000	5.64E-12			STU
SLC30A9	4_42091049_TTTGA	T	TTTGA	0.519	8.69E-12			
PGM3	rs150637901	ATTTT	T	0.618	1.15E-11	AU-rich element		
GPATCH2L	rs11398506	T	TC	-0.722	1.86E-11			
GRB10	rs2240498	T	G	-1.410	2.41E-11			
SFPQ	rs200831313	GT	T	-0.890	2.92E-11			ITU
MRM1	rs4796229	A	G	-1.371	3.45E-11			
TTLL11	rs41277110	T	G	-1.016	5.56E-11	hsa-miR-10a		
KIAA0513	rs3751756	T	G	-1.515	1.64E-10			
RAB3GAP2	rs200375338	ACTT	T	0.550	1.91E-10			
TAPBP	rs71536179	G	TTTTTTCTG	-1.581	7.46E-10			
CD164L2	rs2273269	G	A	1.131	8.84E-10			
NAA30	rs3825621	T	C	-0.627	1.25E-09			
SLC15A5	rs200804946	TAAA	A	0.516	4.73E-09	PAS		
FRA10AC1	rs4919398	G	A	-0.771	5.59E-09			
CNTLN	rs2780207	G	C	-0.861	8.00E-09			
DAPL1	rs10016	G	A	-1.294	1.50E-08			
RAB3GAP2	rs3217443	CTTA	A	0.414	1.89E-08			
HIF1AN	rs2495760	T	C	0.873	2.93E-08		Plasma Palmitoleic Acid (6e-9; 0.99)	
GATA4	rs3203358	C	G	1.871	6.17E-08			
TLDC1	rs422598	T	G	1.098	1.35E-07			
DKFZP779L1853	rs11419833	A	AT	0.792	1.72E-07			
DCLRE1B	rs67257357	AC	A	-0.681	3.31E-07			
SEMA3C	rs201521853	T	ATTTT	-0.371	5.31E-07	AU-rich element		YRI
KLHL17	rs142545439	CTTAT	C	0.792	6.82E-07			
POU2F3	rs78221919	T	C	1.713	1.14E-06			
PRR23C	rs12637621	G	C	-0.318	1.97E-06			
LCLAT1	rs199959171	T	TTTTTC	1.284	2.66E-06			

Table A.1 (Continued)

Gene	Variant	Ref_allele	Alt_allele	LFC	padj	Annotations	GWAS (p-value;LD)	CMS Pop
PKP1	rs7541805	G	C	1.194	2.80E-06			
SEMA3C	rs1527474	G	T	-0.650	2.91E-06			YRI
ST8SIA1	rs10841971	T	C	1.201	2.98E-06			
WDR90	rs3752490	G	T	-0.395	3.66E-06			
SPRY1	rs36125802	CCTT	C	0.555	3.91E-06			
MARK2	11_63678128_TA	T	TA	0.876	4.44E-06			
PIK3C2G	rs142896323	GTTGT	G	0.358	7.70E-06			
GCLM	rs34571118	C	AC	0.629	3.21E-05			
LINS	rs28438569	T	C	0.541	5.84E-05			
RPE65	rs3118416	C	G	0.551	7.34E-05			
COTL1	rs1046623	T	C	0.931	1.03E-04			
TAP2	rs73410774	C	G	-0.707	1.15E-04			
VGLL3	rs9828396	A	T	-0.530	1.18E-04			
FBN1	rs201948323	ATG	G	-0.636	1.84E-04			
NAA40	rs200402503	TTTTTC	T	1.070	1.88E-04			CHS
LEMD3	rs117013092	G	T	-0.517	1.97E-04	PAS		
LCLAT1	rs67960987	CT	C	0.829	2.11E-04			
CNTLN	rs632255	T	C	0.776	3.38E-04	AU-rich element		
EMC4	rs7167503	A	T	-1.161	3.82E-04	AU-rich element		
SLC12A1	rs78562329	T	C	0.705	6.24E-04			
TMEM244	rs9375666	C	T	-0.355	8.05E-04			
AC105020.1	rs57163932	T	C	-0.723	8.24E-04			
TAP2	rs112927456	G	A	1.622	1.16E-03			
ALK	rs80152976	C	A	-0.657	1.28E-03	hsa-miR-19b		
CMPK2	rs13308	A	G	-0.415	1.42E-03			
SLC2A1	rs55728431	G	T	-1.007	1.50E-03			
CNDP1	rs2887	A	G	-0.840	1.58E-03			
GABRA2	rs497068	C	T	-0.395	1.66E-03			
ADAM17	rs189060518	T	G	1.352	1.73E-03			CHB
RNF141	rs1993818	T	G	0.445	1.76E-03			
MRAP2	rs45583848	T	G	0.307	2.66E-03			
PKP1	rs201366172	G	GC	0.878	2.73E-03			
MAF	rs1055708	T	C	-0.372	2.87E-03			
PYHIN1	rs35746133	C	CT	-0.381	3.13E-03			
ARAP2	rs12645801	T	A	0.266	4.43E-03			
CCDC66	rs73081823	C	T	-0.699	4.74E-03		Height (4e-22; 0.98)	
LEPR	rs34448361	T	TTTTA	-0.601	7.22E-03	AU-rich element		CDX
NF1	rs200002326	TA	T	-0.708	7.80E-03			
ZNF720	rs192992586	T	A	-0.224	9.67E-03	PAS		
RGMA	rs3752103	T	C	0.931	9.86E-03			GWD, MSL
ZNF714	rs200383268	T	TA	0.338	1.12E-02			
RP1L1	rs12547192	G	C	-0.900	1.15E-02			
COL6A3	rs200357374	GT	T	-0.323	1.64E-02			
KIF3A	rs4705963	T	C	0.526	1.75E-02		Atopic dermatitis (1e-10; 0.91)	
RP11-766F14.2	rs35320285	TA	A	-0.275	2.28E-02			
MTHFD2	rs828853	T	G	0.638	2.48E-02			
CCNT2	rs13389642	T	C	0.551	2.59E-02	hsa-miR-93		
ZBTB37	rs12063818	G	C	-0.538	2.86E-02			
ITGB8	rs1045957	A	T	0.447	3.42E-02	hsa-miR-19b		
RAB27A	rs1050931	C	T	0.585	3.83E-02			
NPM1	rs117701982	A	G	-0.314	4.40E-02			
NAPEPLD	rs13232194	A	T	-0.259	5.00E-02			
MAPK8IP3	rs2076432	A	T	1.050	5.52E-02			
HLA-DQA1	rs1065043	C	T	-0.508	6.85E-02			
PKP1	rs947375	T	G	-0.730	8.15E-02			
C2CD4D	rs12122596	C	G	0.385	8.52E-02			
TNRC6B	rs41281265	A	G	0.249	9.49E-02			

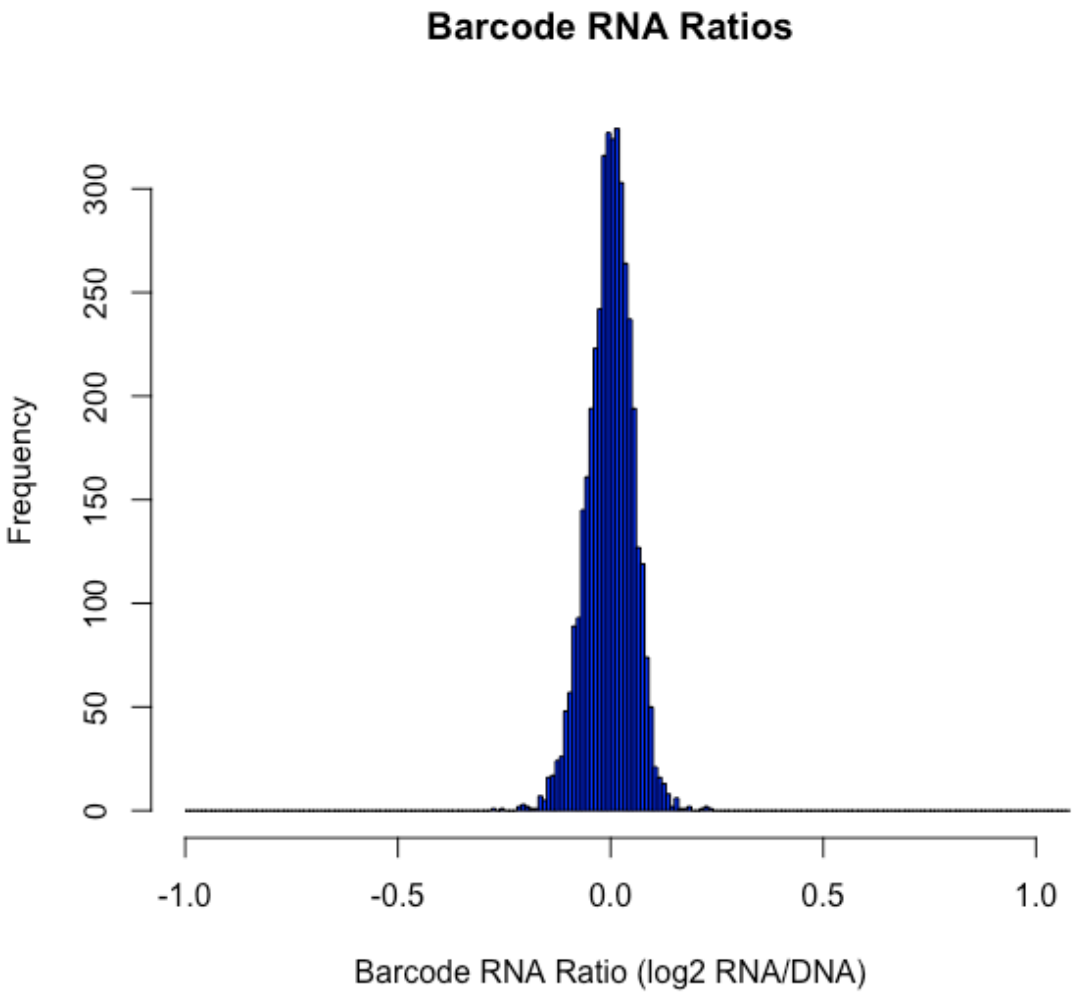


Figure A.1: Random hexamer effects on total RNA expression. Histogram of random hexamer barcode effects on plasmid-normalized total RNA expression.

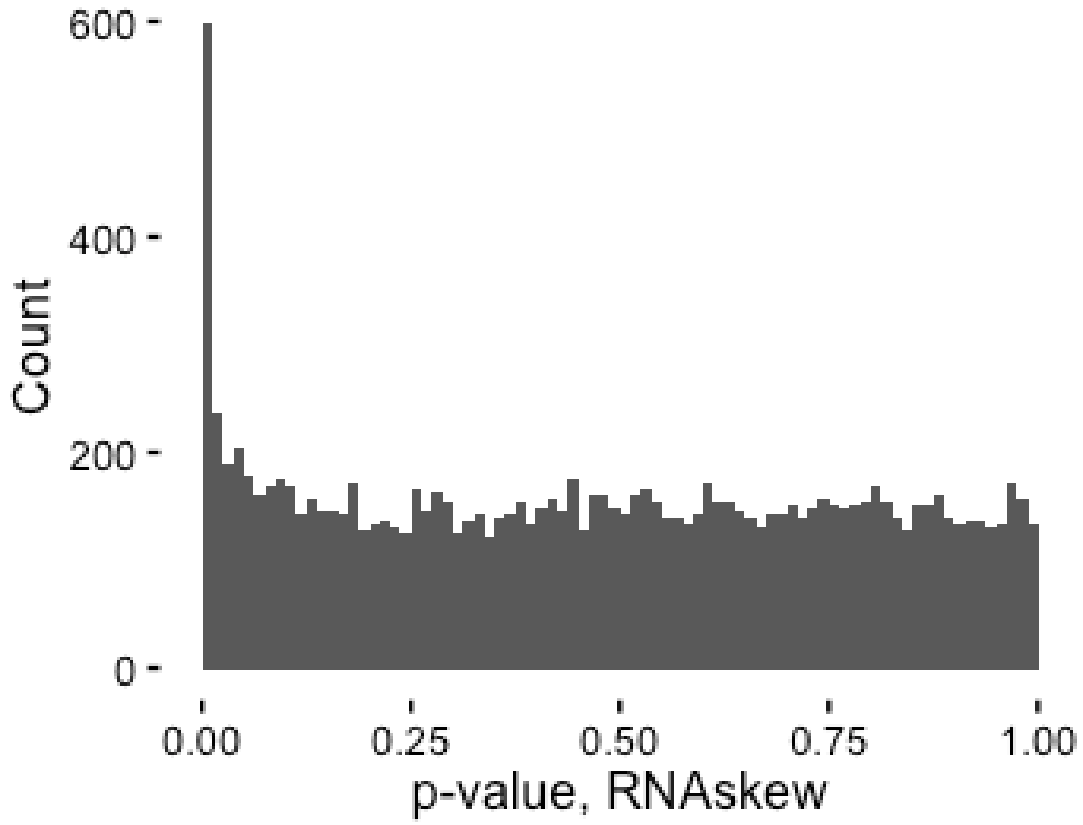


Figure A.2: Allelic skew p-value distribution. Distribution of RNAskew p-values from MPRAu.

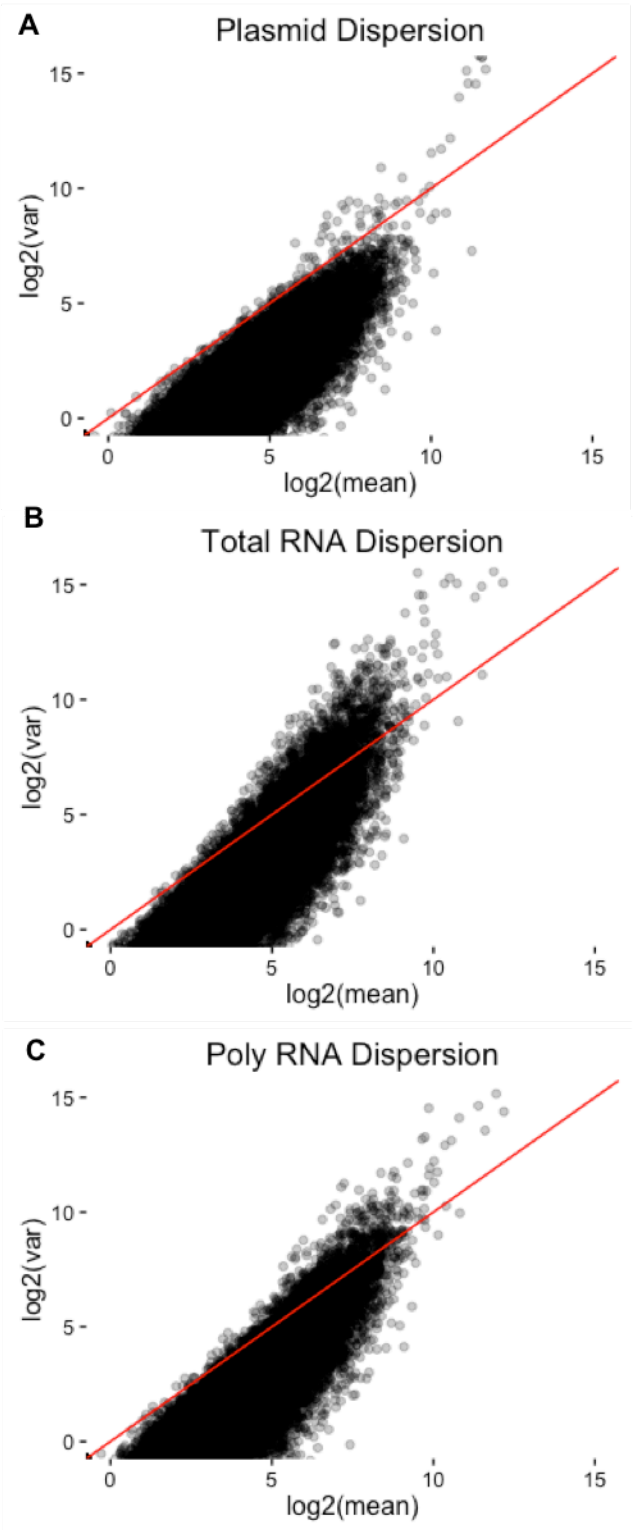


Figure A.3: Plasmid vs RNA dispersions. (A) Variance of plasmid replicates is linearly correlated with mean coverage, as expected by a Poisson distribution. This contrasts with variance of **(B)** total RNA and **(C)** polysomal RNA replicates, for which over-dispersion is observed.

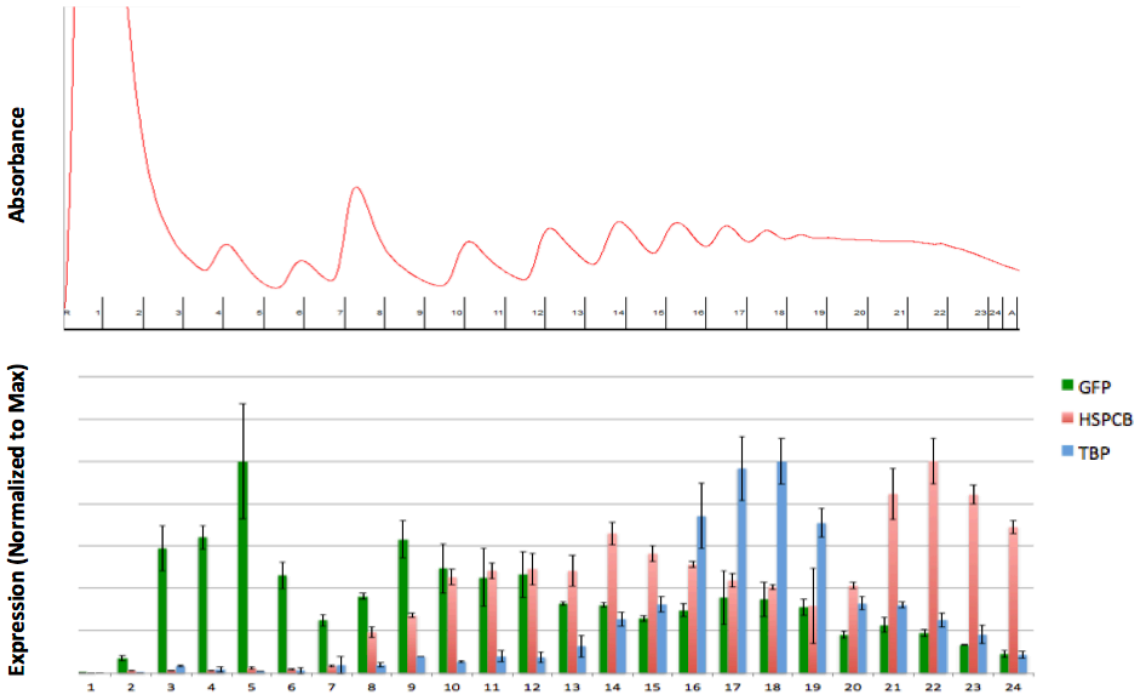


Figure A.4: Sucrose gradient qPCR. Quantitative PCR (qPCR) was performed on each individual sucrose fraction to quantify overall RNA levels across the gradient. A UV spectrograph is shown above, with the end-point of each fraction demarcated on the x-axis. RNA expression, as determined by qPCR, is shown below. In addition to GFP, we quantified RNA levels of two housekeeping genes, HSPCB and TBP. Each gene's expression is normalized to the fraction with the highest expression.

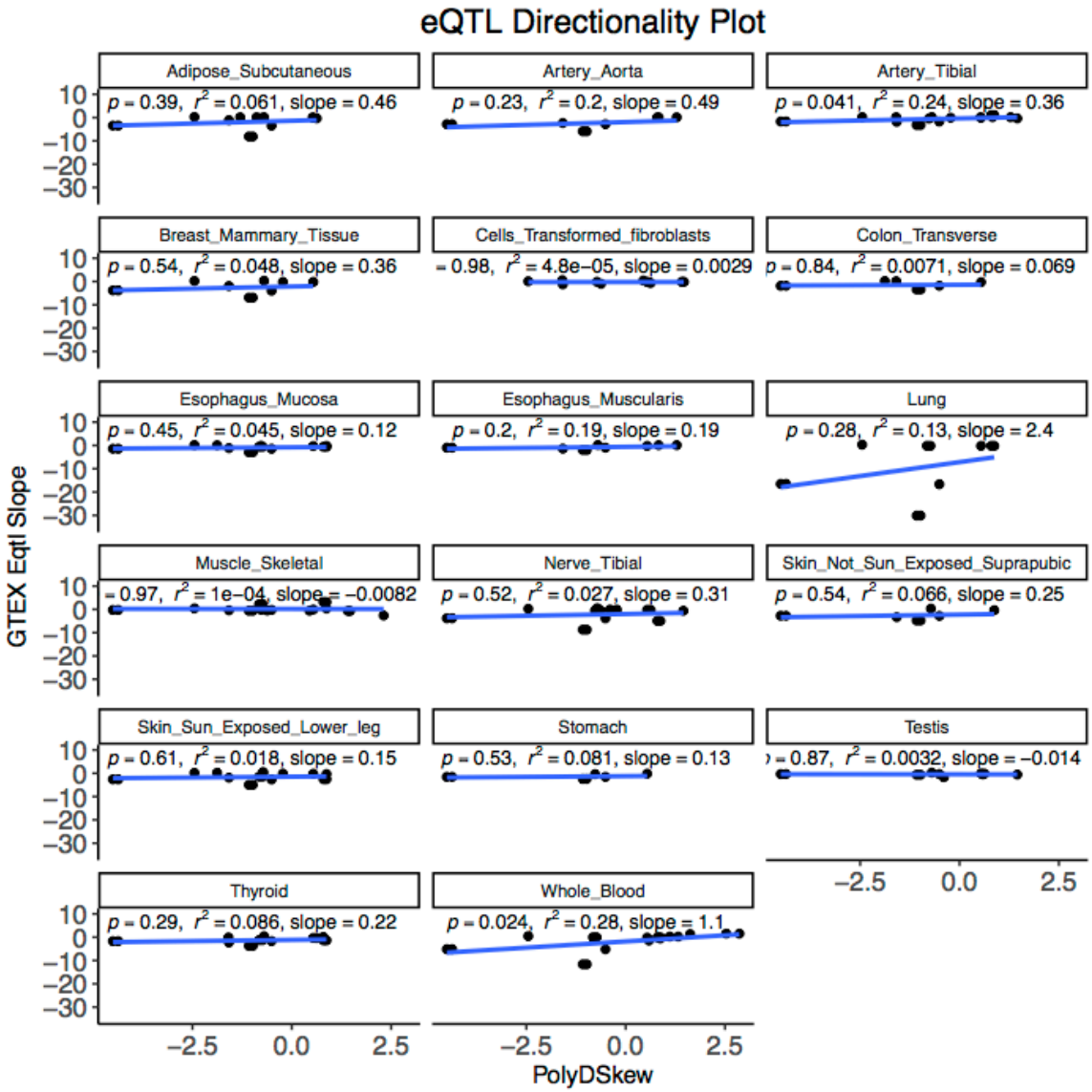


Figure A.5: eQTL directionality. Correlation of POLYskew and eQTL effect size for variants with significant differential expression in MPRAu that were also called as eQTLs in GTEx. Tissues with greater than six overlapping points are shown.

Bibliography

1. Mendel, G. Experiments on Plant Hybridisation,(Tr) Harvard University Press (1946). (1870).

2. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

3. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

4. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

5. Genome-wide association studies. *yourgenome.org* (2016). Available at: about:reader?url=http%3A%2F%2Fwww.yourgenome.org%2Fstories%2Fgenome-wide-association-studies. (Accessed: 12th May 2017)

6. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).

7. Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327,** 883–886 (2010).

8. Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152,** 703–713 (2013).

9. Schaffner, S. & Sabeti. Evolutionary Adaptation in the Human Lineage. *Nat. Educ.* **1,** 1–14 (2008).

10. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

11. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl,** 228–237 (2003).

12. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).

13. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

14. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

15. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).

16. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

17. Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316,** 889–894 (2007).

18. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373,** 895–907 (2015).

19. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30,** 271–277 (2012).

20. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165,** 1519–1529 (2016).

21. Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215–233 (2009).

22. Rinn, J. L. & Ule, J. 'Oming in on RNA–protein interactions. *Genome Biol.* **15,** 401 (2014).

23. Schork, A. J. *et al.* All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLoS Genet* **9,** e1003449 (2013).

24. Steri, M. *et al.* Overexpression of the Cytokine BAFF and Autoimmunity Risk. *N. Engl. J. Med.* **376,** 1615–1626 (2017).

25. Kulkarni, S. *et al.* Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20705–20710 (2013).

26. Pai, A. A. *et al.* The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet.* **8,** e1003000 (2012).

27. Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.* **8,** e1002882 (2012).

28. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).

29. Artieri, C. G. & Fraser, H. B. Evolution at two levels of gene expression in yeast. *Genome Res.* **24,** 411–421 (2014).

30. McManus, C. J., May, G. E., Spealman, P. & Shteyman, A. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24,** 422–430 (2014).

31. Thiadens, K. A. M. H. & Lindern, M. von. Selective mRNA translation in erythropoiesis. *Biochem. Soc. Trans.* **43,** 343–347 (2015).

32. Koritzinsky, M. & Wouters, B. G. Hypoxia and regulation of messenger RNA translation. *Methods Enzymol.* **435,** 247–273 (2007).

33. Provenzani, A. *et al.* Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis* **27,** 1323–1333 (2006).

34. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107,** 1–8 (2016).

35. Baker, M. Microarrays, megasynthesis. *Nat. Methods* **8,** 457–460 (2011).

36. Tian, J., Ma, K. & Saaem, I. Advancing high-throughput gene synthesis technology. *Mol. Biosyst.* **5,** 714–722 (2009).

37. Ma, S., Saaem, I. & Tian, J. Error correction in gene synthesis technology. *Trends Biotechnol.* **30,** 147–154 (2012).

38. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38,** 2522–2540 (2010).

39. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* **23,** 800–811 (2013).

40. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).

41. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431 (2003).

42. Zhang, L., Kasif, S., Cantor, C. R. & Broude, N. E. GC/AT-content spikes as genomic punctuation marks. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 16855–16860 (2004).

43. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinforma. Oxf. Engl.* **27,** 2957–2963 (2011).

44. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

45. Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352,** 600–604 (2016).

46. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).

47. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).

48. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6,** e1000895 (2010).

49. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).

50. Pichon, X. *et al.* RNA binding protein/RNA element interactions and the control of translation. *Curr. Protein Pept. Sci.* **13,** 294–304 (2012).

51. Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138,** 673–684 (2009).

52. Spies, N., Burge, C. B. & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23,** 2078–2090 (2013).

53. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22,** 1173–1183 (2012).

54. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11,** 113–127 (2010).

55. Cenik, C. *et al.* Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* **25,** 1610–1621 (2015).

56. Hou, J. *et al.* Extensive allele-specific translational regulation in hybrid mice. *Mol. Syst. Biol.* **11,** 825 (2015).

57. Wang, Z. *et al.* Evolution of gene regulation during transcription and translation. *Genome Biol. Evol.* **7,** 1155–1167 (2015).

58. Hartl, D. L. *Principles of population genetics*. (Sinauer Associates, 2007).

59. Ziebarth, J. D., Bhattacharya, A., Chen, A. & Cui, Y. PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Res.* **40,** D216-221 (2012).

60. Bakheet, T., Williams, B. R. G. & Khabar, K. S. A. ARED 3.0: the large and diverse AU-rich transcriptome. *Nucleic Acids Res.* **34,** D111-114 (2006).

61. Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* **39,** D152-157 (2011).

62. Bakheet, T., Frevel, M., Williams, B. R., Greer, W. & Khabar, K. S. ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res.* **29,** 246–254 (2001).

63. White, E. K., Moore-Jarrett, T. & Ruley, H. E. PUM2, a novel murine puf protein, and its consensus RNA-binding site. *RNA N. Y. N* **7,** 1855–1866 (2001).

64. Khan, A., Tian, L., Zhang, C., Yuan, K. & Xu, S. Genetic diversity and natural selection footprints of the glycine amidinotransferase gene in various human populations. *Sci. Rep.* **6,** 18755 (2016).

65. Kleber, M. E. *et al.* Genome-wide association study identifies 3 genomic loci significantly associated with serum levels of homoarginine: the AtheroRemo Consortium. *Circ. Cardiovasc. Genet.* **6,** 505–513 (2013).

66. Ares, G. R., Caceres, P. S. & Ortiz, P. A. Molecular regulation of NKCC2 in the thick ascending limb. *Am. J. Physiol. Renal Physiol.* **301,** F1143-1159 (2011).

67. Zhao, W. *et al.* Massively parallel functional annotation of 3' untranslated regions. *Nat. Biotechnol.* **32,** 387–391 (2014).

68. Shalem, O. *et al.* Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet.* **11,** e1005147 (2015).

69. Oikonomou, P., Goodarzi, H. & Tavazoie, S. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Rep.* **7,** 281–292 (2014).

70. Wissink, E. M., Fogarty, E. A. & Grimson, A. High-throughput discovery of post-transcriptional cis-regulatory elements. *BMC Genomics* **17,** 177 (2016).

71. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466,** 835–840 (2010).

72. Bazzini, A. A., Lee, M. T. & Giraldez, A. J. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336,** 233–237 (2012).

73. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *bioRxiv* 074443 (2016). doi:10.1101/074443

74. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.* **14,** 1787–1799 (2016).

75. Albert, F. W., Muzzey, D., Weissman, J. S. & Kruglyak, L. Genetic influences on translation in yeast. *PLoS Genet.* **10,** e1004692 (2014).

76. Csárdi, G., Franks, A., Choi, D. S., Airoldi, E. M. & Drummond, D. A. Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.* **11,** e1005206 (2015).

77. Battle, A. *et al.* Impact of Regulatory Variation from RNA to Protein. *Science* **347,** 664–667 (2015).

78. Schafer, S. *et al.* Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat. Commun.* **6,** 7200 (2015).

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

80. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).

81. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141,** 129–141 (2010).

82. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4,**

83. Kotagama, K., Babb, C. S., Wolter, J. M., Murphy, R. P. & Mangone, M. A human 3′UTR clone collection to study post-transcriptional gene regulation. *BMC Genomics* **16,** (2015).

84. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–177 (2013).

85. Grimson, A. *et al.* MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing. *Mol. Cell* **27,** 91–105 (2007).

86. Jash, A., Yun, K., Sahoo, A., So, J.-S. & Im, S.-H. Looping Mediated Interaction between the Promoter and 3′ UTR Regulates Type II Collagen Expression in Chondrocytes. *PLOS ONE* **7,** e40828 (2012).

87. Awad, T. A. *et al.* Negative transcriptional regulation mediated by thyroid hormone response element 144 requires binding of the multivalent factor CTCF to a novel target DNA sequence. *J. Biol. Chem.* **274,** 27092–27098 (1999).

88. Ding, Y. *et al.* In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505,** 696–700 (2014).

89. Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.* **17,** 83–96 (2016).

90. An, J., Zhu, X., Wang, H. & Jin, X. A dynamic interplay between alternative polyadenylation and microRNA regulation: Implications for cancer (Review). *Int. J. Oncol.* **43,** 995–1001 (2013).

91. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

92. Araujo, P. R. *et al.* Before It Gets Started: Regulating Translation at the 5′ UTR. *Int. J. Genomics* **2012,** e475731 (2012).

93. Dikstein, R. Transcription and translation in a package deal: the TISU paradigm. *Gene* **491,** 1–4 (2012).

94. Evfratov, S. A. *et al.* Application of sorting and next generation sequencing to study 5´-UTR influence on translation efficiency in Escherichia coli. *Nucleic Acids Res.* **45,** 3487–3502 (2017).

95. Dvir, S. *et al.* Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci.* **110,** E2792–E2801 (2013).

96. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: genome-wide determination of poly(A) tail length and 3’ end modifications. *Mol. Cell* **53,** 1044–1052 (2014).

97. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508,** 66–71 (2014).

98. Eichhorn, S. W. *et al.* mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. *eLife* **5,** (2016).

99. Schier, A. F. The maternal-zygotic transition: death and birth of RNAs. *Science* **316,** 406–407 (2007).

100. Migliaccio, A. R. Erythroblast enucleation. *Haematologica* **95,** 1985–1988 (2010).

101. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165,** 1530–1545 (2016).