High-throughput methods to analyze protein-DNA and protein-RNA interactions

Wei Zhou

A dissertation

Submitted in partial fulfillment of the

Requirement for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Stanley Fields, Chair

Jay A. Shendure

W. Conrad Liles

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

High-throughput methods to analyze protein-DNA and protein-RNA interactions

Wei Zhou

Chair of the Supervisory Committee:

Stanley Fields

Department of Genome Sciences

The genome of a cell contains the information to make thousands of RNA and protein molecules. However, a cell typically expresses only a fraction of its genes, with different phenotypes arising because of differentially expressed genes. A cell controls its gene expression at different levels. For example, it can control when and how often a gene is transcribed, or it can selectively destabilize mRNA molecules. These biological processes are controlled by DNA- and RNA-binding proteins. In this dissertation, I first discuss the current state of the field of transcription factor (TF) variation, and I describe *in vitro* and *in vivo* technologies for mapping TF-DNA interactions. I then describe studies that apply the "calling card" assay to analyze interactions between chromatin and TF variants. In Chapter 2, I use this assay to characterize the interactions with chromatin of six single amino acid variants of Ste12, a yeast transcription factor regulating mating and invasion. My study showed that while subtle changes in the coding region of this

transcription factor can result in large regulatory rewiring, the major determinants of organismal phenotype are the changes in the expression of a small, related set of genes. In Chapter 3, I use next generation sequencing technology to score the RNA-binding activity of variants in each of eight repeats present within a PUF domain. In this assay, in addition to the PUF variants, I generated a library of RNA variants in which each possible RNA base was present at the cognate position recognized by one of the PUF domain repeats. I identified many PUF domain variants with highly specific interactions by comparing their binding across the four RNA bases. This approach allows us to propose a complete code for RNA recognition by this PUF domain. In Chapter 4, I discuss a new method to concurrently profile the whole and newly synthesized transcriptome in each of many single cells, and use the data to quantify the dynamics of the cell cycle and glucocorticoid receptor activation. Finally, in Chapter 5, I discuss some of the outstanding questions for mapping protein-DNA and protein-RNA interactions, including research directions that I believe will be important in the future.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This work presented in this dissertation would not have been possible without the contributions and support from many people. I would like to thank the members of the Department of Genome Sciences as well as the Molecular and Cellular Biology program for their help and support throughout my graduate career.

I would like to thank my Ph.D. mentor, Stanley Fields, for being an amazing and outstanding advisor for the past five years. It was through his guidance that I became the scientist that I am today. His enthusiasm for science and technology development, depth of knowledge and incredible creativity have been my source of inspiration throughout my graduate school study. More importantly, his encouragement has given me the confidence to pursue science and develop novel techniques during the last five years. Also, I feel grateful to have joined the Fields lab and thank all the members of our lab for their support in my graduate school study, and for being wonderful colleagues and most importantly close friends. I thank Matt Rich, Ben Brandsen, Michael Dorrity, Stephanie Zimmermann and Josh Cuperus for their guidance and suggestions in all of my projects.

I would like to thank my committee members, Jay Shendure, Judit Villen, Conrad Liles and Christine Queitsch, for their generous support throughout my graduate research. I would like to specifically thank my rotation mentor Daniel Melamed, who provided invaluable instruction from experiment design to technique troubleshooting when I first entered the field.

And finally, I thank my parents and parents-in-law, who are the constant source of support in my life and career, as well as my dear sister and brother, who support me all the time. Their love is invaluable to me. I especially thank my husband Junyue Cao, who supports and shares every breath, and every heart-beating moment of my life and whom I will always admire. As a scientist himself, his enthusiasm for science and every novel idea lights up our everyday life. I also would like to thank my brilliant son, Jayden Cao, and my lovely daughter, Sonia Cao, for motivating me to be a great mother and for being prepared to make our future world a better place.

Chapter 1. Introduction

1.1 Massively assays for measuring TF-DNA interaction

Transcription factors (TFs) are one of the largest categories of DNA-binding proteins in the cell. Almost 10% of the protein-coding genes in most organisms encode TFs. Generally, a TF recognizes a short stretch of double-helical DNA of defined sequence and determines which of thousands of genes in a cell will be transcribed. TFs can make a series of contacts with DNA. While each individual contact is weak, many contacts are formed at the TF-DNA interface, such that altogether they form highly specific and tight interactions. To increase the binding affinity and specificity, some TFs form homodimers or heterodimers that can double the length of the *cis*-regulatory sequence that is recognized. In this way, these TFs can be reused to generate distinct DNA-binding modes or specificities.

1.1.1 *TF-DNA interactions are key drivers of phenotypic variation*

The discovery of *cis*-regulatory sequence by Jacob and Monod (1) initiated the debate over TF-DNA interactions as key drivers of phenotypic variation. This contention has been supported by many studies. Genome-wide association studies (GWAS) in multiple species have revealed that the vast majority of significantly associated genetic variants are located in non-coding regions of the genome (2–5). In addition, the majority of these non-coding GWAS single nucleotide polymorphisms (SNPs) are located within DNAseI hypersensitive sites (DHS) sites, which reflects the occupancy of DNA-binding proteins, including TFs (6). These types of evidence indicate that

GWAS loci may alter the binding of TFs and create variation in gene expression, driving phenotypic variation, evolution and disease.

Genetic variation in either a TF or its *cis*-regulatory sequences can induce variation in TF-DNA binding, gene exression and ultimately phenotypes. While recent studies have demonstrated the importance of variation in both TFs and regulatory sequence (7–10), emphasis has focused more on *cis*-regulatory variation; high-throughput *in vitro* assays for examining the binding of TFs to DNA sequences that use protein-binding microarrays (7) or sequencing-based technologies (11–13) are more amenable to testing large numbers of DNA sequences than large numbers of TF variants. However, recent findings have highlighted the need to evaluate the TF-DNA binding output of TF variants.

Human exome sequencing data have revealed abundant genetic variation in the coding sequences of TFs (14,15). Recent studies have found most individuals contain unique repertoires of variant factors (7). Compared to individual regulatory sequence variants, individual TF variants can exert a far larger phenotypic impact by changing the expression of multiple downstream target genes. Variation in the coding sequence of a TF may change DNA-binding affinity, specificity or both by altering the factor's structure of its oligomerization with itself or with cofactors. Although *in vitro* studies and modeling can predict the effects of variation on transcription factor-DNA binding (7), few *in vivo* studies connect altered TF binding to altered gene expression and whole-organism phenotypes.

1.1.2 *In vitro* and *in vivo technologies to measure TF-DNA interactions*

The development of high-throughput *in vitro* DNA binding technologies has significantly expanded TF motif catalogs. Protein-binding microarrays (PBM) (16) are one of the scalable methods to identify sequence specificities of DNA-binding proteins. In a PBM experiment, a TF of interest is purified and then applied to a DNA microarray. Non-specifically bound protein is removed through washing, and a primary antibody specific to the TF is applied to detect a binding signal. The dataset resulting from a PBM experiment is relatively quantitative since the signal within each spot on the microarray indicates the strength of TF-protein interaction. The sequence specificities of many yeast TFs, including Abf1, Rap1 and Mig1, have been identified through this technology (17).

Another *in vitro* method is the systematic evolution of ligands by exponential enrichment (SELEX) or high-throughput SELEX (13,18,19). In a typical SELEX experiment, a random pool of fixed-length oligonucleotides is synthesized. A TF of interest can bind to a subset of the oligonucleotides in the pool, which are retrieved and amplified. The new oligonucleotide pool is then used as the sequences for the next cycle. After several cycles, the binding sites for a TF of interest will be enriched. Recent studies have revealed that, in general, binding sites derived from both PBM and HT-SELEX methods mostly agree (20). As the PBM dataset is relatively more quantitative, it allows a better ranking of the binding of any k-mers, based on the signal detected in each spot of microarray. On the other hand, HT-SELEX datasets were found to be better concordant with *in vivo* binding profiles than PBM data. While PBM experiments are limited

to k−mers with k ≤ 8, owing to space constraints of the microarray (16), HT-SELEX has the capacity to measure binding of longer k-mers (k>20).

Although these *in vitro* methods have enabled an expansion of TF motif catalogs, they have technical limitations. For example, they mostly rely on purified DNA-binding domains of TFs, as the full-length versions of the proteins are difficult to work with in terms of cloning and expression (19). As many TFs do not bind DNA as single entities, but in cooperation with other cofactors, *in vitro* methods may not be accurate enough to reveal their genome binding sites. For example, ~3000 TF pairs have been predicted to be routinely formed in cells (21), and these cooperative pairs often show different binding sites preferences compared to their constituent individual TFs. There have also been discrepancies observed between *in vitro* derived DNA binding sites and *in vivo* DNA occupancy levels (22). Clearly, measuring binding events in the complex environment of a cell is critical and can be complementary to these *in vitro* assays.

Hybrid assays are one type of *in vivo* characterization methods, typically performed in yeast (23–25) or bacteria (26–28). The yeast two-hybrid system was developed to measure protein-protein interaction (29) by the activation of a reporter gene. The same principle was carried over to bacteria, with the activation of a reporter gene achieved through an interaction between RNA polymerase and a DNA-binding domain. The system was then optimized to a bacterial-one-hybrid (B1H) method to measure TF-DNA interaction (30). The binding motifs of DNA-binding domains, including homeodomains (31) and C2H2 Zinc fingers (32), have been identified by this method.

Other *in vivo* methods such as chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) have been used successfully to assess chromatin binding on a global scale (33–35). This approach requires the ability to express the TF of interest and chemically crosslink chromatin and proteins. After the DNA is sheared into fragments, an antibody targeting the TF of interest is used for immunoprecipitation, and the linked DNA is determined by high-throughput sequencing (36). Another genomic scale method for profiling protein-DNA interaction is DamID-seq (37), which relies on the expression of *E. coli* DNA adenine methyltransferase as a fusion protein with a TF of interest. When the TF binds to chromatin, the methyltransferase is recruited to the loci and deposits methyl groups on nearby GATC sequences. Genomic DNA is then digested by DpnI, a methylation sensitive restriction enzyme. In this way, methylated fragments are amplified and sequenced. DamID-seq has been used to map chromatin-binding profiles for many TFs with different physiological roles (38–40). While ChIP-seq and DamID-seq recover the actual genomic sequences bound by TFs on a global scale, they are still relatively low-throughput by requiring either a specific antibody that binds with high affinity to the protein, or a fusion of the methyltransferase with an individual TF of interest.

1.1.3 *The "calling card" approach to measure interactions between chromatin and TF variants*

As discussed in section 1.1.1, variation in the coding sequence of a transcription factor may change DNA-binding affinity, specificity or both. To evaluate TF-DNA binding variation, it is critical to identify individually the genomic targets for each TF variant . However, the *in vivo* techniques (*i.e.* ChIP-seq or DamID-seq) described in section 1.1.2 are relatively low-throughput for this purpose. Here, in Chapter 2, I describe our strategy based on the "calling card" approach (41). In this

method, a transposase is fused with a TF of interest.When the TF binds to chromatin, the fused transposase will recruit the corresponding transposon and direct its insertion into the genome near to where it binds. The original calling card approach used the yeast retrotransposon Ty5 system (41). In this system, a TF of interest is fused to Sir4 and expressed in yeast cells. The fusion directs the insertion of Ty5 into the genome near to where it binds. The same principle has been optimized in mammalian cells (42), taking advantage of the high transposition efficiency of the *PiggyBac* transposon system (43,44) in mammalian cells.

One potential advantage of this method is that it can be optimized to a more high throughput approach. In this way, a "calling card" can be inserted into the genome such that it not only labels a transcription factor "visit," but also serves as the identification of which variant of the TF was bound. This goal can be achieved by first matching each TF variant with a random barcode through subassembly, and then putting the random barcode within the transposon region. When a TF variant binds to genome, its fused transposase will drive the insertion of the paired barcode into genome. Therefore, the inserted barcode can reveal both TF-DNA binding locations and TF variant identity.

In Chapter 2, I describe a modified calling card approach with the *PiggyBac* transposon system to identify genomic sites for TF variants in the yeast *Saccharomyces cerevisiae*. With a combination of this *PiggyBac*-based calling card method and other technologies including RNA-seq and *HT-SELEX*, I examined how six single amino acid variants in the DNA-binding domain of Ste12 - a yeast transcription factor regulating mating and invasion - alter Ste12 genome binding, motif recognition and gene expression to yield markedly different phenotypes.

1.1.4 *Ste12: a model TF to study the genetic basis of phenotypic diversity*

Ste12, a key regulator of yeast pheromone response, was initially discovered in a collection of haploid yeast mutants with mating defects (45). Ste12 is a yeast transcription factor of 688 amino acids (46), with three functional domains, including an N-terminal DNA-binding domain, a central transactivation domain and a C-terminal regulatory domain (47). In the mating process, peptide pheromones bind to receptors on the cell surface and activate a signal transduction cascade that include a set of protein kinases. The actions of these protein kinases relieve the negative regulatory activity that the Dig1 and Dig2 proteins exert on Ste12, activating Ste12 to upregulate the expression of many genes involved in mating (48). Depending on the yeast cell type, the genes induced by pheromone-activated Ste12 can be classified into three distinct sets: **a**-specific genes (expressed only in **a** cells), α-specific genes (expressed only in α cells), and haploid-specific genes (expressed in both **a** and α-cells). Both **a**-specific and haploid-specific genes are regulated through binding of Ste12 to a sequence known as the pheromone response element (PRE). In **a**-specific genes such as *MFa1* and *MFa2*, Ste12 binds to the PRE in the upstream sequence cooperatively with a general transcriptional regulator Mcm1 (49), while in haploid-specific genes such as *FUS1*, expression is thought to be solely dependent on Ste12 binding (36). By comparison, the regulation of α-specific genes, such as *MFα1*, is carried out through a complex including Ste12, Mcm1 and α1. In this case, the PRE is not required for Ste12 activity (50).

Under conditions of nutrient limitation, some yeast strains undergo differentiation into elongated filamentous forms as a mechanism to enable foraging for nutrients. In collaboration with the

transcription factor Tec1, Ste12 is also an important regulator in the filamentous growth response. The Ste12/Tec1 complex binds cooperatively to the filamentation response element (FRE), which consists of closely spaced binding sites for Ste12 and Tec1, to activate filamentous response genes required for cell elongation and flocculation (51). Other filamentous genes, such as *FLO11* and *CWP1*, contain only a Tec1 binding site in their promoter regions, and can be activated by Tec1 in a Ste12-independent mechanism (52).

As described above, yeast mating and invasion pathways converge on the highly conserved fungal transcription factor Ste12, which interacts differentially with cofactors to activate either mating or invasion. Dorrity *et al.* (53) previously showed that single amino acid changes in a region of the Ste12 DNA-binding domain can shift the trait preference of yeast cells toward either mating or invasion, and that at least in some cases these changes result in altered *in vitro* DNA binding preferences of the Ste12 variants. Therefore, in Chapter 2, I use Ste12 as a model transcription factor to study the genetic basis of phenotypic diversity. I describe experiments to generate *in vivo* genome-binding profiles and expression profiles of wild-type Ste12 and six Ste12 variants that alter the mating or invasion phenotype. By integrating binding and expression data, we conclude that while subtle changes in the coding region of a transcription factor can result in a large reconfiguration of expression, the major determinants of organismal phenotype are the changes in the expression of a small, related set of genes.

1.1.5 Role of the trimerization domain of heat shock transcription factor 1 (Hsf1)

The heat shock transcription factor (Hsf1) is the central regulator of heat shock-inducible gene expression and the cytoplasmic unfolded protein response in eukaryotes. It has two highly conserved domains: a DNA binding domain and an oligomerization domain (54). The widely presented model is that Hsf1 exists as an inactive monomer under non-stress conditions. Upon heat shock stress, the monomers trimerize and activate heat shock target genes. In addition to its well-studied role in inducing rapid gene expression in response to heat stress, Hsf1 has basal functions. For example, it promotes development and fertility in mice (55,56) and serves as an essential gene in many model organisms, including *C. elegans, D. melanogaster, S. pombe and S. cerevisiae (57,58)*. However, the transcriptional programs mediated by Hsf1 appear to be different in stress and non-stress conditions (59). We are interested in understanding the structure of active Hsf1 and its role in differentiation between transcriptional programs in stress and basal conditions.

The high conservation of the Hsf1 oligomerization domain emphasizes its functional importance. The domain affects many aspects of Hsf1 activation, including its trimerization (60), intramolecular interactions (61), post-translational modification (62,63) and protein-protein interactions (64,65). Analysis of the oligomerization domain can furthermore inform our understanding of the mechanism by which Hsf1 senses and responds to stress.

There are many levels to Hsf1 regulation, both dependent on and independent from the cellular context. The importance of the cellular environment is illustrated by the fact that human Hsf1, for which 37°C is natively basal temperature, responds to 37°C as a heat stress when expressed in *Drosophila* or *Xenopus* (66,67). However, some level of temperature-sensing regulation occurs within the Hsf1 molecule itself. For example, Hsf1 is capable of induced DNA binding with heat shock *in vitro,* and purified monomer Hsf1 trimerizes with heat shock (68). To deepen our understanding of the role of the oligomerization domain in temperature discrimination, Michael

Dorrity and Elizabeth Morton carried out a deep mutational scan of a region of the oligomerization domain of *S. cerevisiae* Hsf1. They obtained the comparative fitness phenotypes of these Hsf1 variants at both basal and heat stress growth temperatures. Through additional RNA-sequencing analysis and the *PiggyBac*-based calling card method described in section 1.1.3, we elucidated the physiological link between Hsf1 mutations and their observed phenotype and demonstrated that point mutations in this domain can confer temperature-specific phenotypes.

1.2 High-throughput analysis of protein-RNA interactions

RNA-binding proteins can bind to specific sequences in mRNA and determine how long each mRNA will persist in the cell and be able to produce proteins, providing critical post-translational control of gene expression. They can thus be a useful platform for researchers to engineer protein-based RNA targeting systems, which offer unique opportunities to monitor gene expression and cellular function. In the following sections, I will first describe current efforts to engineering RNA targeting systems, and then specifically focus on engineering of an RNA-binding domain known as the PUF domain.

1.2.1 *RNA targeting technologies*

Significant progress has been made in the redesign of modular DNA-binding proteins in the past two decades (69,70). These systems include highly modular transcription activator-like (TAL) effectors (71,72), somewhat modular zinc finger proteins (73,74), as well as non-modular DNA-binding proteins or enzymes such as recombinases and restriction endonucleases (75–78). This

success has encouraged researchers to engineer protein-based RNA targeting systems (79,80), as RNA is involved in a more diverse range of biological function compared to DNA. In addition, manipulation of RNAs is usually transient and does not cause irreversible changes in the genome. The key to this engineering is the construction of an RNA-binding scaffold with the desired RNA-binding specificity, enabling specific recognition of an RNA with limited off-target effects. In the past decade, many studies have been conducted to uncover the specific binding code between RNA-binding domains (RBDs) and RNA at a molecular level. I will discuss three major classes of RBDs in the following sections.

The RNA recognition motif (RRM), the most abundant RBD in vertebrates, can recognize single-stranded RNA (ssRNA) in a sequence-specific fashion (81). A typical RRM consists of 80-90 amino acids, with its most conserved region commonly being two short stretches that are critical to interacting with RNA (82). The general rule for how different RRMs interact with RNA is largely unclear, restricting its potential application as an RNA-binding platform. However, as many RRM structures have been solved both in their free form or in their RNA-bound form (83), it has still been possible to engineer its specificity for certain RNA sequences. The KH domain is another abundant structure in vertebrates that binds to a 4-mer RNA sequence. These domains are often found in multiple copies, as an individual domain binds RNA with low affinity. Similar to the RRM, the RNA recognition specificity of the KH domain is still largely unknown (84).

Some RBDs contain tandem repeats that make a suitable surface for RNA interactions (85). These mostly include the PPR and PUF superfamilies. The PPR superfamily is found in most eukaryotes and involved in diverse roles, such as RNA editing, mRNA maturation or organelle biogenesis (86–88). PPR proteins bind to RNA in a parallel orientation. The amino acid in position 5 can

distinguish purines from pyrimidines while the amino acid in position 35 can distinguish between C/A and G/U (89–91). A code for base-specific recognition by PPR has been reported (92,93). However, as some structural evidence suggests two PPR domains commonly form a dimeric interface for RNA binding (92–94), it remains an open question whether the non-modular binding mode between PPR domains and RNA may affect the engineering of its specificity. The PUF superfamily is another RBD with tandem repeats. It functions to regulate mRNA stability by binding to the 3' untranslated region (3'UTR) of its target (95–97). The current efforts to engineer the PUF superfamily will be discussed in detail in section 1.2.2.

Other RNA targeting systems, such as the Cas13 family, have recently been engineered to bind and cleave RNA (98–100). Cas13 systems so far are the only prokaryotic CRISPR-Cas immune systems that target RNA molecules and are regarded as promising tools for RNA detection and manipulation (100). Similar to the Cas9 system, the RNase activities of the Cas 13-crRNA effector complex are triggered by target RNA binding (101). The activated complex cleaves both crRNA-bound target RNA, as well as nearby non-target RNA molecules. It was hypothesized that this collateral activity might be part of a programmed cell death pathway in bacteria; however, it has not been found in the mammalian system. The Cas13 system has been applied to transcript localization, knockdown or RNA editing (102,103), However, its base-pairing requirements for stable RNA association, as well as its potential off-target effects, are still under investigation.

1.2.2 *PUF superfamily: an attractive scaffold for RNA targeting*

PUF proteins, named for PUM (pumilio) and FBF (fem-3 binding factor), are RNA-binding proteins involved in regulating embryogenesis and development in most eukaryotes, including

flies, worms and yeast (104). In *Drosophila melanogaster*, a PUF protein acts to repress translation of maternal hunchback mRNA in the posterior half of *Drosophila* embryo and hence permit its abdominal development (105). In worms, the PUF proteins FBF-1 and FBF-2 bind the 3' UTR of the *fem-3* mRNA and repress its expression in order to mediate the sperm/oocyte switch in hermaphrodites (106). In yeasts, PUF protein can mediate mRNA degradation through binding to transcripts such as *COX17* or *HO* (107,108).

Classical PUF proteins contain a conserved RNA-binding domain, known as the Pumilio homology domain, that is generally composed of eight 36-amino-acid repeats. Each repeat displays three amino acid residues, known as the tripartite recognition motif (TRM), on the concave surface of the protein. Generally, an eight-base RNA sequence is bound as an extended strand to the concave surface. X-ray structural analysis of the complex indicates that the recognition is highly modular, as each repeat binds to a single RNA base (109). The length of bound RNA sequences is variable across species. For example, PUM1, the prototypical PUF protein found in humans, recognizes an 8-base RNA sequence 5'-UGUANAUA-3' (N indicates any nucleotide). In comparison, the yeast PUF protein Puf3 binds an 8-base RNA sequence 5'-UGUAHAUA-3' (H indicates A/C/U) and Puf4 binds a 9-base RNA sequence 5'-UGUAHAHUA-3' (110–112).

From a structure point of review, Wang *et al.* (113) revealed that the pumilio homology domain from the human Pumilio1 protein has eight copies of a single structural motif that pack together to form a right-handed superhelix, which can create a continuous hydrophobic core. The outer face of the domain binds with other proteins while the inner face of the domain binds RNA (113). A TRM combination in a typical repeat is comprised of three amino acid residues. Two residues at

position 12 and 16 contact the edge of the RNA base via hydrogen bonding, and a third residue at position 13 is sandwiched between two RNA bases to form stacking interactions.

The identities of the residues in a TRM combination play a key role in RNA-binding specificity (109,114,115). For example, cysteine and glutamine bind adenine, asparagine and glutamine bind uracil, and serine and glutamate bind guanine. Previous work has used this domain as a scaffold to engineer RNA-binding proteins with designed sequence specificity. Cheone *et al.* (115) created seven soluble mutant PUF proteins and measured their RNA recognition specificity through direct binding assays. They achieved guanine to uracil specificity through mutating the TRM combination *SNE* to *NNQ* (here, in *XXX* format, left to right indicates positions 12, 13, and 16). Similarly, they achieved adenine to guanine specificity through mutating *SRQ/CRQ* to *SRE*. Adenine to uracil specificity was achieved through mutating *CRQ* to *NRQ*. They also showed that by introducing two sets of mutations in PUF proteins, the mutant protein can bind to RNA sequences with two base changes (115). Ozawa *et al.* (79) introduced more than two sets of mutations into PUF proteins and created mutant proteins that recognize transcripts 4-nt different from the wild-type sequences (from 5'-UGUANAUA-3' to 5'-UGAUGGUU-3'). With the engineered RNA-binding protein as a probe, they localized mitochondrial RNA and visualized its dynamics in single living cells (79).

While the modular code for A/G/U-specific recognition has been identified in different PUF repeats, there is no code that specifically binds cytosine in any natural PUF protein. In 2011, two groups independently identified a code for cytosine binding. Dong *et al.* (116) generated a library with randomized residues in position 12 and 16 from repeat 6 of the Pumilio 1 PUF domain and

screened for their interaction with RNA using a yeast three-hybrid system. They found *SYR* as the TRM combination in a PUF mutant that specifically binds cytosine (116). While the code was detected using repeat 6, they found that the cytosine-recognition code could be transferred to other PUF repeats, although its specificity varied across repeats (116). In addition, they designed PUF mutants that can bind to multiple cytosines or sequences with multiple (CUG) repeats, demonstrating the potential to generate novel RNA-binding scaffolds that can be used for therapeutic applications. Independently, in 2011, Filipovska *et al.* (117) reported the identification of a set of cytosine-specific RNA recognition codes as well. They synthesized a library based on the PUM1 PUF domain and randomized the amino acid in positions 12 and 16 of repeat 6. Five PUF variants with an arginine at position 16, and an amino acid with a small or nucleophilic side chain (glycine, alanine, serine, threonine or cysteine) at position 12 were found to have cytosine-specific recognition. Moreover, they showed that the identified code has general applicability and can selectively bind cytosine at all eight positions (117).

Other groups have used an unbiased high-throughput-sequencing method that represents biochemical affinity *in vitro* (118). Campbell *et al.* (119) determined the specificities of 25 natural and engineered PUF variants in binding reactions with a library that included large numbers of RNA sequences. They introduced mutations into repeat 7 of the *C. elegans* PUF protein FBF-2 as a scaffold and examined the specificity of each variant based on a combination of *in vitro* selection and high-throughput sequencing of RNA (119). They found polar residues in position 12 and 16 can change specificity. For example, a change from *TYR* to *AYR* can result in altered specificity from guanine to uracil. In addition, while the stacking residue in position 13 does not make hydrogen-bonding interactions, the position can potentially have profound influence on specificity

as well. For example, a change from *TRQ* to *TWQ* changed specificity from uracil to adenine. Other examples include: uracil specificity for *CYR* versus adenine specificity for *CRR* and guanine specificity for *CNE* versus adenine preference for *CYE*.

Based on the data they described, for guanine and uracil recognition, the optimal TRMs follow the natural PUF code (G-code: *SHE* for *C. elegans* FBF-2 and *SNE* for human PUM1; U-code: *NxQ* with *x* denoting *T/H/F/Y*). For adenine recognition, the data indicate a more complex picture with *CFQ* being an adenine-specific code that behaved better than any natural A-binding TRMs. They also found that the previously identified cytosine-recognition code for PUM1 (117) could not be transferred to repeat 7 of FBF-2, used in their study (120). In addition, they demonstrated the *de novo* TRM codes can be broadly applied to different scaffolds and repeats, using the yeast three hybrid-system as validation (119).

While the studies discussed above provide a valuable resource for PUF engineering, we still lack a set of recognition codes that can specify all four RNA bases in any position. In Chapter 3, I describe experiments to score the binding activities of PUF variants of the TRM combination for each of the eight repeats of a PUF domain. I tested these variants for their ability to bind to each of the four possible RNA bases, using a combination of the yeast three-hybrid system with next generation sequencing technology. I carried out both randomized and targeted screens for a large number of PUF variants and identified many variants with highly specific interactions. I compared the base-specific recognition patterns across the eight repeats and found many highly specific PUF variants do not act as generic codes across all repeat locations. The dataset described allows me to propose a complete code for a random 8-mer RNA by swapping only the three key residues at the

contacting positions in a PUF domain. Additionally, the results highlight critical rules for PUF engineering.

1.3 Analyzing the temporal dynamics of gene expression in single cells

Gene expression programs are dynamics and a cell can control when and how often a given gene is transcribed, depending on its cell cycle, response to stimuli or differentiation process. Single cell genomics, together with pseudotime ordering techniques, has been widely applied to reconstructing cell state dynamics. The initial pseudotime analysis has recovered the development path for myoblast development (121) and similar strategies have been applied to other *in vitro* cell differentiation processes, including mesodermal lineage cell differentiation and human definitive endoderm cells differentiation (122–124). With the development of higher-throughput single cell techniques, as well as more advanced cell ordering techniques (125–127), the cell state dynamics of all major cellular lineages during *in vivo* development has been well characterized (128–130).

Despite the enormous power of single cell genomics in characterizing cell state transition, however, almost all current single cell genomic approaches capture only a "snapshot" of cell state. For example, most techniques profile gene expression at the moment of fixation or cell lysis. Although time-lapse microscopy is a distinct technology that overcomes some of the limitations, it is limited in throughput and scope. For instance, it allows only visualization of a few marker genes in a few cells and is insufficient to decipher the dynamics of cells in a transcriptome level. In Chapter 4, I describe a novel technique, sci-fate, to measure the dynamics of gene expression in single cells at the level of the whole transcriptome. Through combining 4sU labeling of newly

synthesized mRNA (131) with single cell combinatorial indexing RNA-seq (sci-RNA-seq)(132), we concurrently profile the whole and newly synthesized transcriptome in each of many single cells. We furthermore use the dataset described to develop a framework for inferring the distribution of cell state transitions. We anticipate sci-fate will be broadly applicable to quantitatively characterize transcriptional dynamics in diverse systems.

Chapter 2. Binding and regulation of transcription by yeast Ste12 variants to drive mating and invasion phenotypes

Genetic variation in either a transcription factor or the *cis*-regulatory sequences to which it binds creates variation in gene expression, driving phenotypic variation, evolution and disease. Emphasis has focused on *cis*-regulatory variation and its effect on gene expression, however, recent findings have highlighted the need to evaluate the functional consequences of transcription factor coding variants. Despite their high conservation overall, the coding sequences of human transcription factors show abundant genetic variation, suggesting that most individuals contain unique repertoires of variant factors. Compared to individual regulatory sequence variants, individual transcription factor variants can exert a far larger phenotypic impact by changing the expression of multiple downstream target genes. Variation in the coding sequence of a transcription factor may change DNA-binding affinity, specificity or both by altering the factor's structure or its oligomerization with itself or with cofactors. Here, we took advantage of *Saccharomyces cerevisiae* and its well-characterized mating and invasion pathways in an effort to explain how transcription factor variants alter motif recognition and gene expression, and ultimately organismal phenotypes.Using a combination of the calling card method, RNA sequencing and SELEX, we find that variants with dissimilar binding and expression profiles can converge onto similar cellular behaviors. Mating-defective variants down-regulated distinct subsets of genes necessary for mating. Hyper-invasive variants also down-regulated distinct subsets of genes involved in the mating process, but up-regulated others associated with the cellular response to osmotic stress, including ones induced by this stress. While single amino acid changes in the coding region of this transcription factor result in complex regulatory rewiring, the major phenotypic consequences for

the cell appear to depend on changes in the expression of a small number of genes with related functions.

Contributions Wei Zhou and Stanley Fields developed these ideas. Wei Zhou performed all experiments, with the exception of experiments from Michael Dorrity using HT-SELEX. Wei Zhou and Kerry Bubb performed computational analysis, with suggestions from Michael Dorrity and Christine Queitsch. Wei Zhou, Stanley Fields, and Christine Queitsch wrote the manuscript. We thank Robi Mitra for providing original plasmids and thank Josh Cuperus for comments on the manuscript. The work has been published (*Genetics* **214**: 397-407 (2020)).

2.1 Introduction

The binding of transcription factors to *cis*-regulatory sequence motifs controls the expression of target genes. Genetic variation in either a transcription factor or the *cis*-regulatory sequences to which it binds creates variation in gene expression, driving phenotypic variation, evolution and disease. Emphasis has focused on *cis*-regulatory variation and its effect on gene expression for multiple reasons. First, regulatory variation likely played a large role in evolution because it resulted in less severe pleiotropic effects on whole organism phenotype, allowing for subtle cell or tissue changes rather than interfering with overall body plan (133). Second, many trait-associated variants in human genome-wide association studies reside in or near regulatory DNA (4), with single nucleotide changes in a transcription factor binding site capable of altering factor occupancy and resulting gene expression (134–136). Third, high-throughput *in vitro* assays for examining the binding of transcription factors to DNA sequences, using protein-binding

microarrays (137) or sequencing-based technologies (11,12,138,139), are more amenable to testing large numbers of DNA sequences than large numbers of transcription factor variants.

Recent findings, however, have highlighted the need to evaluate the functional consequences of transcription factor coding variants. Despite their high conservation overall, the coding sequences of human transcription factors show abundant genetic variation, suggesting that most individuals contain unique repertoires of variant factors (7). Compared to individual regulatory sequence variants, individual transcription factor variants can exert a far larger phenotypic impact by changing the expression of multiple downstream target genes. Variation in the coding sequence of a transcription factor may change DNA-binding affinity, specificity or both by altering the factor's structure or its oligomerization with itself or with cofactors. Although *in vitro* studies and modeling can predict the effects of variation on transcription factor-DNA binding (7), few *in vivo* studies connect altered transcription factor binding to altered gene expression and whole-organism phenotypes.

Here, we took advantage of *Saccharomyces cerevisiae* and its well-characterized mating and invasion pathways in an effort to explain how transcription factor variants alter motif recognition and gene expression, and ultimately organismal phenotypes. Yeast mating and invasion pathways converge on the highly conserved fungal transcription factor Ste12, which interacts differentially with cofactors to activate either mating or invasion. In yeast mating, Ste12 binds at mating pheromone-responsive genes as a homodimer (140) or with the cofactors Mcm1 (141) or Matα1 (50). The Ste12 DNA-binding site is the pheromone response element (PRE), whose consensus sequence is TGAAACA. In invasion, which is triggered by increased temperature and lack of

nutrients, Ste12 and its cofactor Tec1 are both required to activate target genes, some of which contain a Ste12 binding site near the Tec1 consensus sequence (GAATGT), forming the filamentation response element (FRE) (142).

We have previously shown that single amino acid changes in a region of the Ste12 DNA-binding domain can shift the trait preference of yeast cells toward either mating or invasion, and that at least in some cases these changes result in altered *in vitro* DNA-binding preferences of the Ste12 variant (53). To interrogate how these *in vitro* preferences translate into markedly different organismal phenotypes, we generated *in vivo* genome-binding profiles and expression profiles of wild type Ste12 and six Ste12 variants previously shown to alter mating or invasion phenotypes. We used the "calling card" method (143) to identify genomic sites bound *in vivo* for each variant. Although each variant binds a set of genomic sites with high reproducibility, there is a comparatively low degree of overlap between the genome-wide profiles of Ste12 variants, even among variants with shared mating and invasion phenotypes. Nevertheless, we find examples of specific changes in binding sites and expression that likely contribute to the observed phenotypes. By integrating binding and expression data, we conclude that while subtle changes in the coding region of this transcription factor can result in large regulatory rewiring, the major determinants of organismal phenotype are the changes in the expression of a small, related set of genes.

2.2 Materials and Methods

Primers and other oligonucleotides. A list of oligonucleotides can be found in Table 7.1.

Construction of a deleted STE12 strain. We generated yeast strains in the BY4705 *MAT*α background whose copy of the *STE12* gene was deleted by site-specific genomic deletion (144). In the first step, yeast was transformed with a PCR fragment that contained the *URA3* gene flanked at each end with 40 bp of sequence corresponding to the flanking sequences of the *STE12* ORF. The *STE12* ORF was replaced with *URA3* by selecting for transformants in Ura- medium. In the second step, a PCR fragment was amplified that contains sequence flanking both sides of the deleted ORF, and this fragment was transformed into the strain generated in step 1 by selecting for loss of *URA3* on 5-fluoroorotic acid (5-FOA) medium. While *STE12* deletion resulted in completely sterile strains that were complemented by a plasmid-borne wild type *STE12* gene, we did not confirm that the *URA3* gene was fully deleted. Thus, in the following Ste12 Variant RNA-seq, we carried out differentially expressed gene analysis after filtering out *URA3*.

STE12 Variant RNA-Seq. The *STE12* locus from *S. cerevisiae* strain BY4705, including the intergenic regions, was introduced into the yeast vector pRS415 containing a *LEU2* marker (145). Individual point mutations were generated in wild type *STE12* plasmids by site-directed mutagenesis (Q5; New England Biolabs). Plasmids were transformed into yeast (BY4705 *MAT*α) with a deleted endogenous copy of *STE12* by high-efficiency lithium acetate transformation (146). Cells were grown into exponential phase (5 replicates). RNA was extracted using acid phenol extraction, as previously described (147). Total cDNA was generated using anchored oligo-dT primer and SuperScript IV (Life Technologies) (132). Second strand synthesis was carried out at 16°C for 180 min with NEBnext Second Strand Synthesis module (NEB). cDNA was tagmented with a Nextera tagmentation kit (Illumina) at 55°C for 5 min. The reaction was stopped by adding 1x DNA binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then

purified using 1.5x AMPure XP beads (Beckman Coulter), eluted in 16 μL of buffer EB (Qiagen). Each sample was then mixed with 2μL of 10 μM indexed P5 and P7 primers and 20 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 10 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. The library was purified with 0.8x AMPure XP beads (Beckman Coulter) and prepared for sequencing using Illumina Nextseq.

Construction of plasmids for a PiggyBac-based transposon-based calling card method. A donor plasmid carrying the PiggyBac transposon and an SP1-PBase helper plasmid were obtained from Robi Mitra (Washington University in St. Louis). To use G418 and 5-FOA selection in yeast, a *KanMX* gene was inserted into the transposon region. We added an 8 bp (NNNNNNNN) unique molecular identifier (UMI) sequence to the transposon region of each copy of the donor plasmid in order to quantify unique insertion events per cell. For the helper plasmid, we replaced SP1 with wild type and variant full length *STE12* fragment and fused it with the PiggyBac transposase to encode the Ste12-PBase, whose expression is under the control of the galactose-inducible *GAL1* promoter. Gibson assembly products were introduced into *Escherichia coli* and the plasmid was isolated. Constructs were confirmed by Sanger sequencing.

Transformation of cells and transposition of PiggyBac. Plasmids used for transformation were prepared using the Plasmid miniprep kit (Qiagen) following the manufacturer's protocol. Paired donor plasmid and helper plasmid were transformed into yeast (BY4705 *MAT*α) by high-efficiency lithium acetate transformation (146). After transformation, cells were collected and put on the induction plates with galactose. Cells were induced for 5 days to express the Ste12-PBase. Cells

were then collected, diluted back to OD_{600}=0.45, and cultured in rich medium for 6 hours recovery (to OD_{600}=~1.6). Cells were put onto selection plates with 5-FOA and G418 at varying dilutions and it takes around 2-3 days.

Inverse PCR. Cells were collected from selection plates and genomic DNA was extracted from each sample using the Smash and Grab method as described (148). Each DNA sample was divided into two 20-µg aliquots and digested by *Taq*I and *Rsa*I (NEB) individually. Digested products were purified by DNA purification column (Zymo Research) and ligated overnight at 15°C in dilute solution to encourage self-ligation. Self-ligated DNA was purified by DNA purification column (Zymo Research) and used as template in an inverse PCR. Primers that anneal to the PiggyBac donor sequences (primer 316: GATGTCCTAAATGCACAGCGAC and primer 317: GAGGCGTGCTTGTCAATGC) were used to amplify the genomic regions flanking the transposon, and then adaptor sequences that allow the PCR products to be sequenced on Illumina sequencing platforms were added by primer 367 (AATGATACGGCGACCACCGAGATCTACACCTCCATCGAGACACTCTTTCCCTACAC GACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC) and 377 (CAAGCAGAAGACGGCATACGAGATTGCTTGTCAATGCGGTAAG). The PCR products were purified using a DNA purification column (Zymo Research). For each sample, the same amount of PCR product from digestion with each restriction endonuclease was pooled and submitted for Illumina sequencing.

Processing of PiggyBac transposon-based calling card data. Sequencing was completed on Illumina's NextSeq platforms. The raw data included all fastq files from the PiggyBac transposase-

only control, wild type Ste12 and the six variants. Each sample had two replicates. In read1, the first segment is the universal primer sequence: CGTCAATTTTACGCAGACTATCTTTCTAGGG followed by the flanking genome sequence of 39 bp. The first 8 bp of read2 is the UMI sequence used to identify unique insertion events. We first filtered sequence reads with high quality and mapped them back to the yeast genome. Then we quantified independent PiggyBac insertions based on the UMIs detected in each sample. Finally, we called significant PiggyBac insertion peaks by MACS2, a peak-calling algorithm (149). Target genes were then assigned to insertion peaks that were within 1000 bp 5′ or 200 bp 3′ of the transcription start site for that gene (150).

Motif analysis for PiggyBac transposon-based calling card data. Insertions with counts above the 85th percentile were identified as "high count insertions." We identified 300 bp windows around each high count insertion, and then merged the windows (bedops -m) to generate high insertion count sites for each replicate. We extracted the sequence from these windows and attempted to identify *de novo* motifs using MEME (151). We also scanned these sequences for a set of known yeast motifs (152) using fimo (153). We tallied the coincidence of known motifs, normalizing by the number of merged high insertion count windows. We used DESeq2 (154) to identify motif pairs that appeared at different frequencies in the variants, taking advantage of the replicate data. We identified 529 pairs with log10 (mean motif-pair count across transcription factor variants) greater than one and dispersion levels greater than expected for that mean.

HT-SELEX. We purified fragments of Ste12(1-215) expressed from pGEX-4T-2 vectors. These protein fragments have been used previously (53) and are sufficient for binding *in vitro*. Fragments

of wild type and variant Ste12 proteins were purified using a GST tag and used for HT-SELEX. SELEX reactions with homogenous and mixed protein populations were performed identically to previous work (155). Briefly, a 50 μL reaction containing purified Ste12 (1:25 molar ratio with DNA), 200 ng nonspecific competitor double-stranded nucleic acid poly (dI/dC), and 100 ng selection ligand (36N) were incubated in binding buffer [140 mM KCl, 5 mM NaCl, 1 mM K2HPO4, 2 mM MgSO4, 20 mM Hepes (pH 7.05), 100 μM EGTA, 1 μM ZnSO4] for 2 h. GST Sepharose (GE) beads were added to each reaction, incubated for 30 min, and unbound ligand was removed using seven buffer washes. Output reactions were amplified by PCR after each round, and these products were subsequently used to prepare high-throughput sequencing libraries. SELEX motif enrichments were analyzed using Autoseed software (155). The pool of binding-selected output sequences was compared against a fully random input sequence pool to identify enriched motif sequences.

Motif analysis for HT-SELEX data. All possible 10mers were computed among the bound sequences observed in round five of SELEX for each Ste12 variant, a pool that should contain an enrichment of bound sequences. The 10mer counts for all output sequences were then normalized to the counts of each of those 10mers in the random oligo pool used as the input ligand for HT-SELEX. For each variant, the top enriched sequences were determined as those whose 10mer enrichment was three standard deviations above the mean enrichment of all 10mers. Weblogos were generated using the MEME Suite by searching for enriched motifs among these highly enriched 10mer sequences for each variant.

High-throughput sequencing reads, along with datasets showing calculated insertion scores for each variant (two replicates) and transcriptome data (five replicates) have been submitted to Gene Expression Omnibus(GEO) under accession number: GSE141713.

2.3 Results

2.3.1 *A PiggyBac based calling card identifies binding sites of Ste12 and Ste12 variants.*

We previously conducted a deep mutational scan of a segment of the Ste12 DNA-binding domain and subjected yeast cells carrying a library of these variant Ste12 proteins to selection for either mating or invasion (53). For the mating selection, *MAT**a*** cells with *STE12* variants were mixed with *MATα* cells and selected using auxotrophic markers that would be present only in mated diploids. For the invasion selection, the yeast were incubated on plates until invasion had occurred and then washed from the plate surface, such that only cells embedded in the agar should remain. We identified Ste12 variants with single amino acid changes in the DNA-binding domain that altered the preference of yeast cells in their mating or invasion trait. To determine how the genomic targets of Ste12 might differ depending on these amino acid changes, we chose six Ste12 variants with altered mating or invasion phenotype (Figure 2.1). We included three variants that cause reduced mating, one of which leads to wild type invasion (A160P) and two to hyper-invasiveness (K152L and K146D). We included three other variants that mate like wild type, two of which are hyper-invasive (K150A and K150I) and one defective for invasion (S158H).

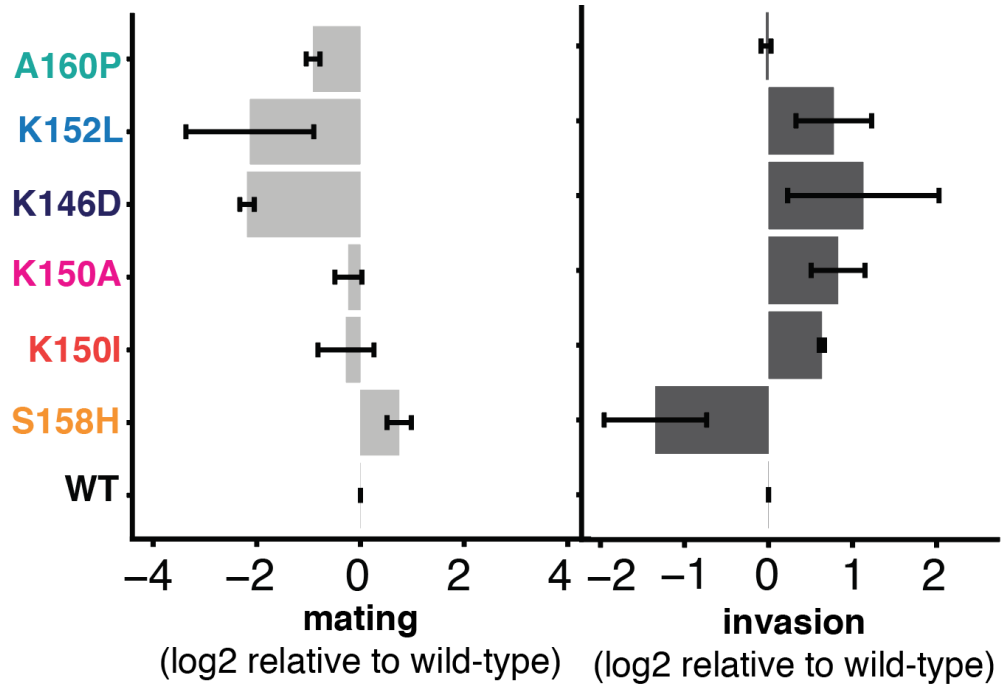


Figure 2.1. Mating and invasion phenotypes of selected Ste12 variants (53); the x-axis shows the mating or invasion score based on high-through trait selection assay; the y-axis shows six variants; error bar indicates SE of replicates; variants with different phenotypes are annotated by different colors.

We used the transposon-based "calling card" method (143) to identify the binding sites of these variants *in vivo* and genome-wide. This method fuses a transcription factor to a transposase such that the transcription factor directs transposon insertion into the genome at a TTAA site nearby to where it is bound. Following the transposition events, genomic sites that were used for transposition are amplified and characterized by high throughput DNA sequencing (150). We designed two plasmids for use of this method in yeast with the PiggyBac transposon (Figure 2.2). First, we constructed a donor plasmid that carries the PiggyBac transposon containing a G418 resistance marker (*KanMX*) and a *URA3* marker. We quantified unique insertion events per cell using a random 8 bp unique molecular identifier (UMI) sequence in the transposon region of each

copy of the donor plasmid. Second, we constructed a helper plasmid that encodes a fusion of the full-length Ste12 protein to the PiggyBac transposase (Ste12-PBase), whose expression is under the control of the galactose-inducible *GAL1* promoter. We co-transformed both of these plasmids into the *MATα STE*$^{+}$ BY4705 strain (156). This mating-competent strain was used because the Ste12-PBase acts dominantly in transposition and would be induced in cells that have an intact mating regulatory program.

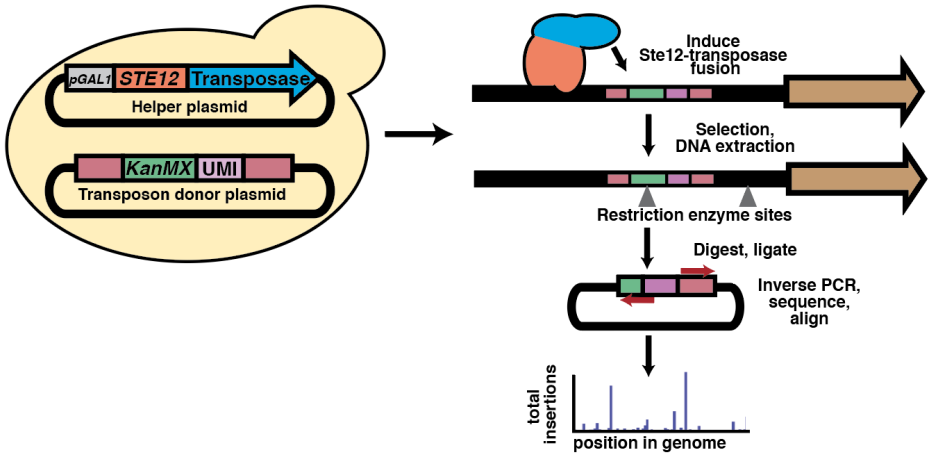


Figure 2.2. A PiggyBac transposon-based calling card method identifies genome-wide binding profiles for wild type Ste12 and its six variants.Workflow with key steps are illustrated in the figure.

We induced the Ste12-PBase, which results in the insertion of the transposon near sites bound by a variant Ste12 protein and the conversion of cells with these insertions to G418 resistance. We then measured the chromosomal acquisition of the transposon-borne *KanMX* marker in cells that had lost the donor plasmid. The cells were identified by selection in media with G418, which requires the *KanMX* marker, and with 5-fluoroorotoic acid (5-FOA), which is toxic for cells that have Ura3 activity (157). We found that the transposition efficiency of PiggyBac in yeast is ~10%

(Figure 2.3). To reveal the genomic DNA sequences flanking the end of the transposon, we isolated genomic DNA from colonies grown in 5-FOA and G418, cleaved it with *Taq*I or *Rsa*I, and re-circularized the resulting fragments through ligation in dilute solution. We carried out inverse PCR to amplify fragments containing the end of the PiggyBac transposon, and sequenced the PCR product (see Materials and Methods).

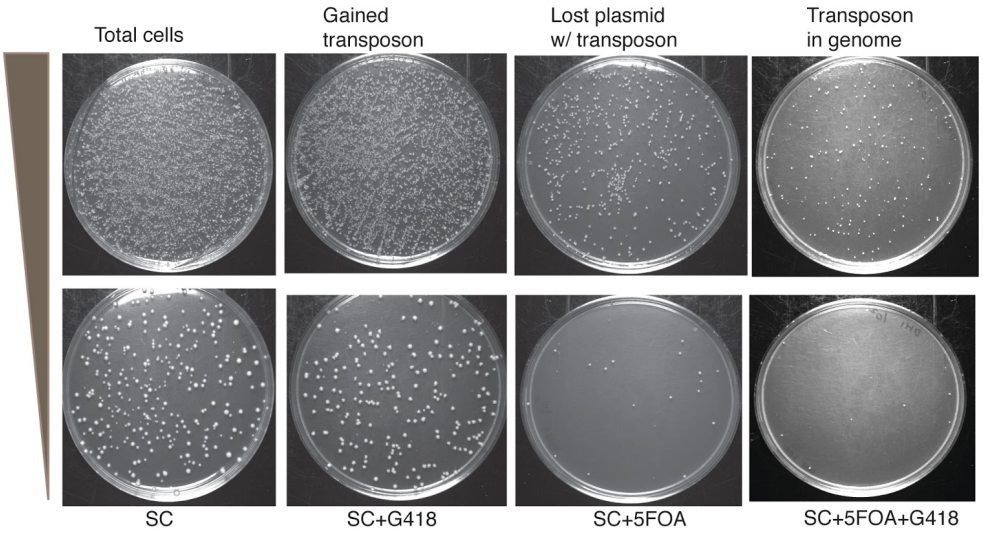
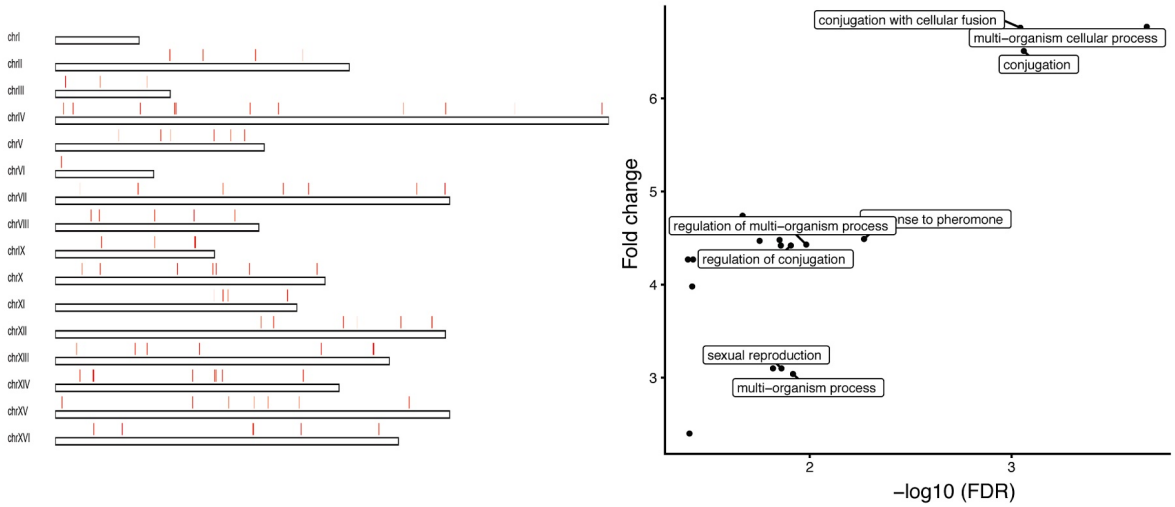


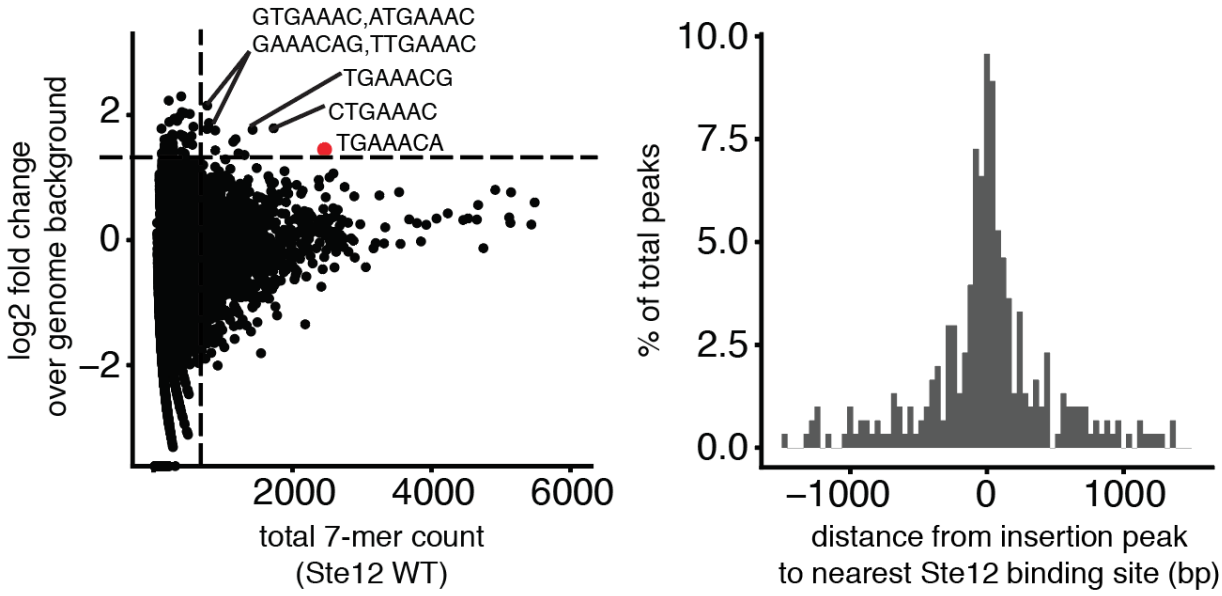
Figure 2.3. **Transposition efficiency for a PiggyBac transposon-based calling card method.** Transposition efficiency is calculated as the number of cells that survive the selection medium (SC+5FOA+G418) divided by the total number of cells plated on rich medium(SC). Two dilutions are shown.

For wild type Ste12, we obtained a total of 327 significant insertion peaks throughout the genome (Figure 2.4). A 7-mer analysis of the region 100 bp around each insertion found that highly enriched 7-mers include the common sequence TGAAAC (Figure 2.5), indicating that the Ste12-PBase fusion protein most frequently deposited the transposon near canonical Ste12 binding sites, as previously shown with the calling card method (143). Nearly half (49.2%) of the insertions had

such a canonical site within 200 bp (Figure 2.6), which is comparable to the detection resolution from the ChIP-seq method (158). By assigning the insertion peaks to nearby genes (see Methods), we obtained a total of 264 gene targets. GO analysis revealed that these genes are most enriched (FDR<0.001) for cell fusion or pheromone response pathway (Figure 2.7). Thus, while the Ste12-PBase was expressed under the control of the *GAL1* promoter, which is much stronger than the *STE12* promoter, the insertions fell into expected genes, suggesting that the high expression did not substantially affect Ste12 binding to the yeast genome.



(*Left*) Figure 2.4. The genome-wide PiggyBac insertion patterns of wild type Ste12 shown along the 16 *S. cerevisiae* chromosomes. (*Right*) Figure 2.7. Biological processes enriched through Gene Ontology (GO) annotations for genes assigned by insertion peaks.



(*Left*) Figure 2.5. 7-mer analysis of the region 100 bp around each insertion peak. Each point represents a unique 7-mer sequence; the x-axis shows the total count of each 7-mer and the y-axis shows the relative enrichment of each 7-mer over the genome background. Dashed lines represent the count threshold used to identify enriched 7-mers. (*Right*) Figure 2.6. The distribution of distances between each insertion peak and its closest Ste12 binding site. The *x*-axis specifies the distance from the center of the Ste12 binding site. The *y*-axis is the proportion of detected insertion peaks.

The genome-wide insertion patterns for the six Ste12 variants were also characterized by the calling card method (Figure 2.8). Individual Ste12 variants showed a high degree of overlap between their experimental replicates (Figure 2.9; Pearson correlation coefficient ranged from R=0.69-0.91, apart from A160P (R=0.37), which resulted in the detection of only a few calling card insertion peaks and thus had a relatively low correlation coefficient). That the variants yielded reproducible data also suggests that the high Ste12-PBase expression did not lead to adventitious

non-specific binding. However, pairwise comparisons revealed that the overlap among variants ranged only between 10% and 20%. Even for the two variants with the highest degree of overlap in target sites, K146D and K152L, 80% of the sites differed, despite their similar organismal phenotypes (Figure 2.10). For those sites that were common in all the variants, we found that their nearby genes are classical targets of Ste12 and critical to either the mating or invasion phenotype (*e.g. KAR4, FUS1* and *TEC1*). These results indicate that the small changes in the Ste12 DNA-binding domain led to many gains and losses of binding sites throughout the genome. However, even though the Ste12 variants bound overall to highly divergent sets of genomic sites, it is the binding sites that are in common that likely underlie the key phenotypic differences.

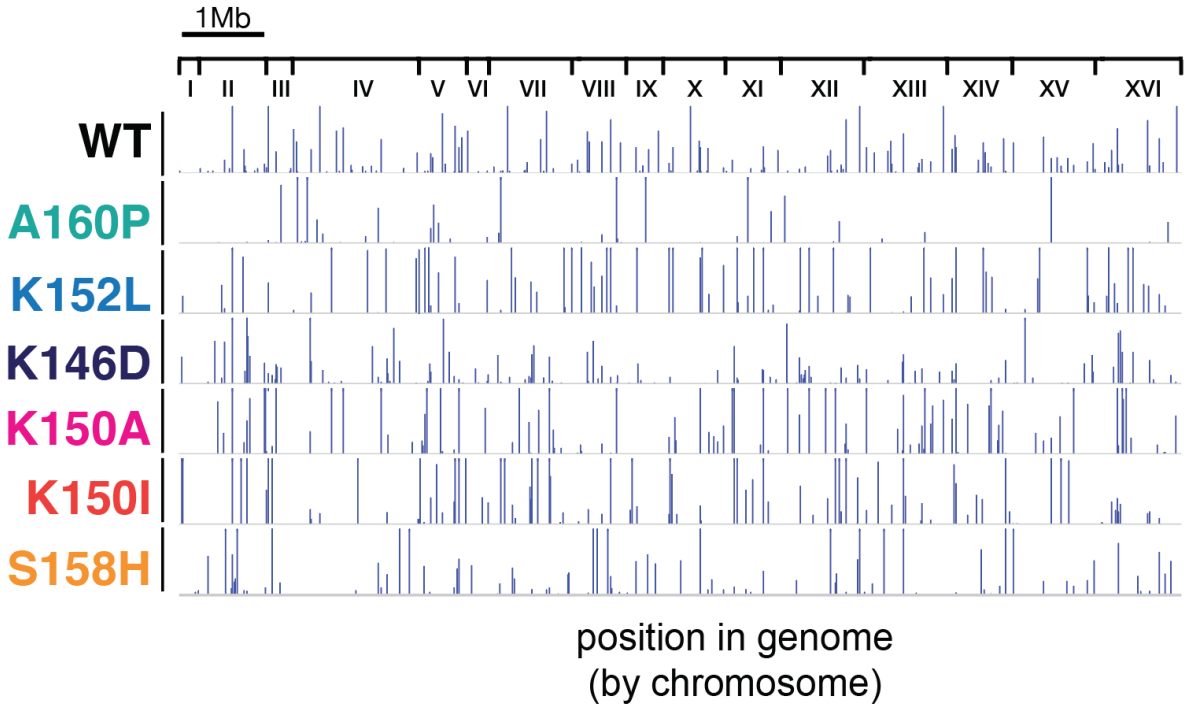


Figure 2.8. The genome-wide PiggyBac insertion patterns of each variant shown along the 16 *S. cerevisiae* chromosomes

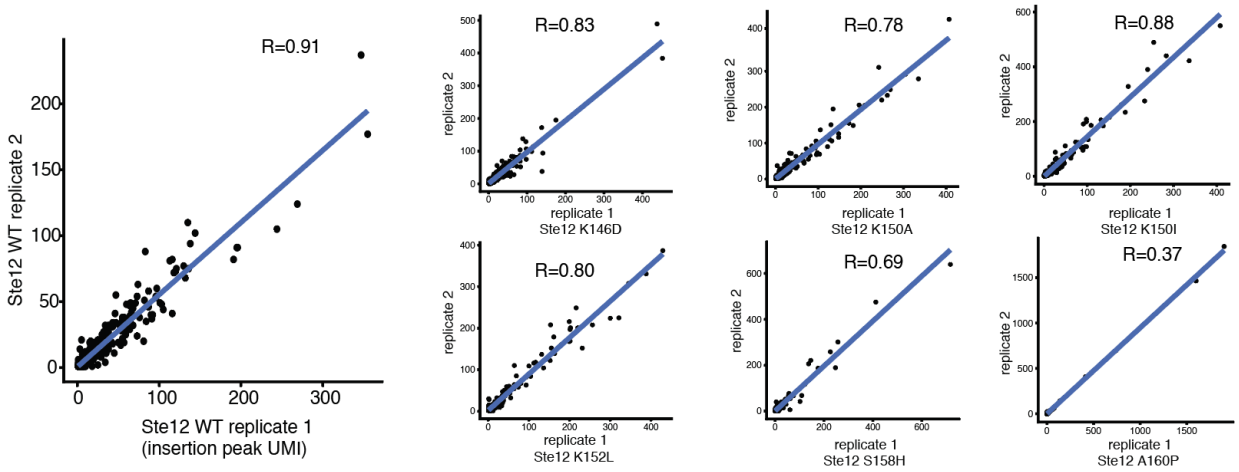


Figure 2.9. Correlation analysis between two replicates for wild type and each Ste12 variants replicates in the PiggyBac transposon-based calling card experiments.

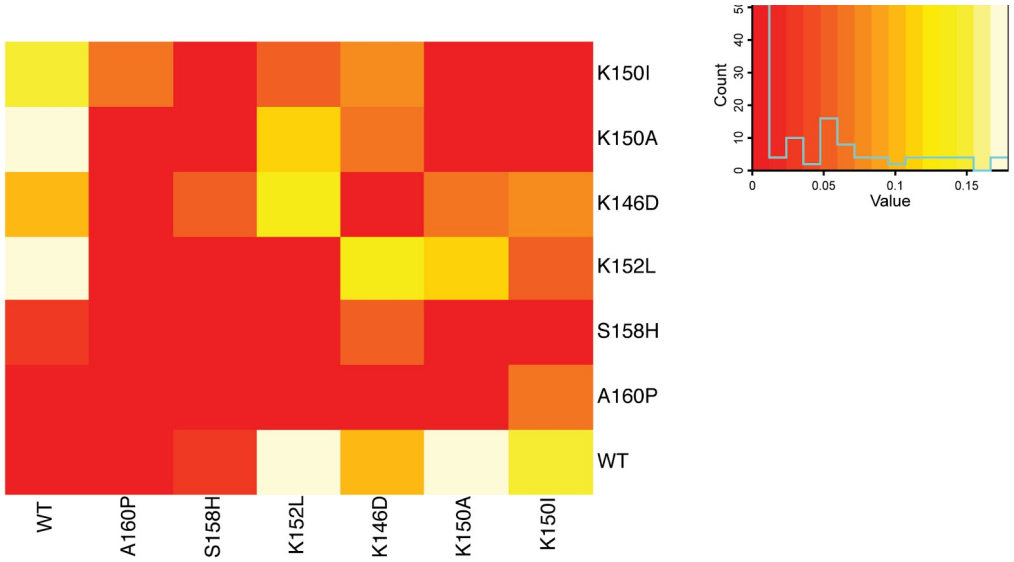


Figure 2.10. A heatmap showing proportions of overlapping binding sites among all variants. The overlap proportion represents the similarity between binding profiles between all pairs of variants; low overlap proportion values are shown in red (0-5% overlapping sites) while higher overlap proportion values are shown in yellow (10-20%).

2.3.2 *Transcriptome profiling of wild type and variant Ste12 proteins.*

We sought to compare the DNA-binding sites of the Ste12 variants with genome-wide expression patterns induced by these variant proteins. We generated variant Ste12 proteins by introducing single amino acid changes into the *STE12* gene under the control of its own promoter and carried on a centromere-based plasmid, and transformed the plasmids into a *ste12Δ* version of the BY4705 strain, which was generated through a site-specific genomic deletion method (see Materials and Methods). Cells were grown into exponential phase and total RNA was obtained. We carried out RNA-seq (5 replicates) to generate the transcriptome for each variant, which yielded a combined total of 145 differentially expressed genes from the six variant strains when compared with wild type Ste12, with a false discovery rate (FDR) of 5%. Replicability of the assay was high (Figure 2.11; Pearson correlation coefficient range R=0.97-0.99). The number of differentially expressed genes in the variant strains ranged from 15 to 84, but the number of these genes did not correlate with the severity of the altered mating or invasion phenotype.

Unsupervised hierarchical cluster analysis of the data revealed three main differentially expressed groups of genes (Figure 2.12). We used Gene Ontology (GO) annotations to identify significant biological process terms (FDR<0.1) for each cluster. For the largest cluster (bottom, Figure 2.12), the top enriched GO terms include mating, conjugation with cellular fusion, and agglutination. For the other two clusters, the top enriched GO terms are associated with primary metabolic process and cellular response to environmental stimulus. Thus, the examination of differential gene expression reveals major clusters of genes, with many of these concordant with changes in mating and invasion phenotypes.

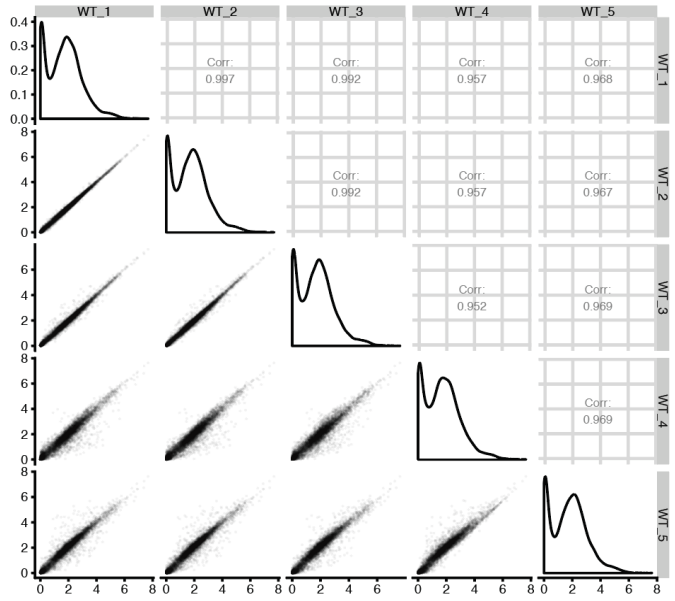


Figure 2.11. Correlation analysis between the replicates of the wild type Ste12 in the RNA-seq experiments. Left of the diagonal: correlation plots showing the Pearson's correlations (r) between pairs of the replicates of wild type Ste12 in the RNA-seq experiments. Diagonal: plots showing the distribution of the transcriptome of each replicate of the wild type Ste12. Right of the diagonal: correlation plots showing the Pearson's correlations (r) of the transcriptome of each pair of replicates of the wild type Ste12.

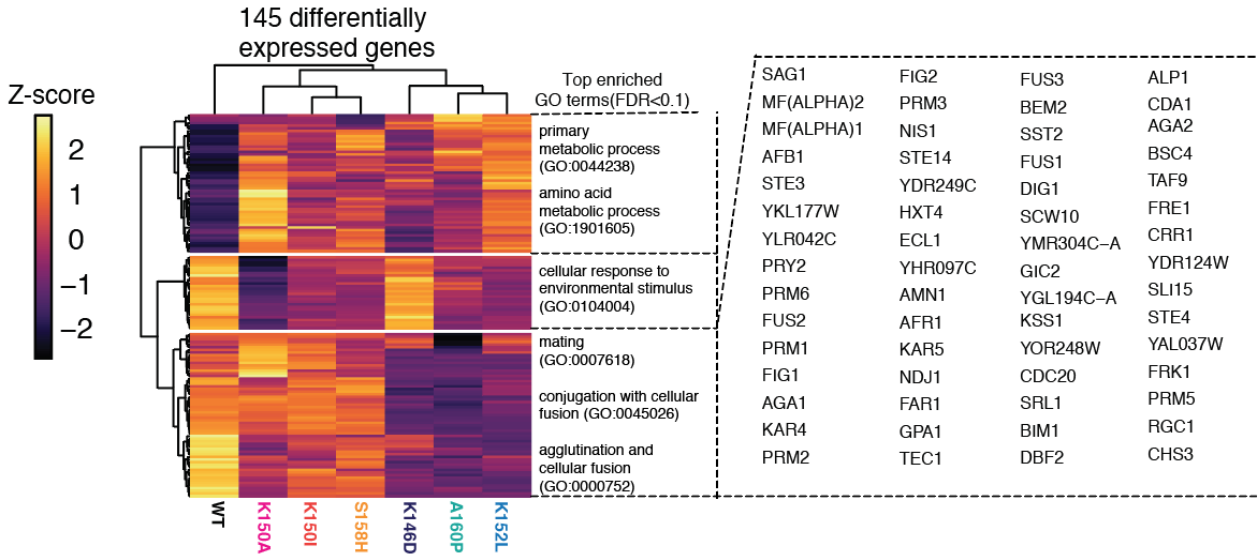


Figure 2.12 RNA-seq reveals gene targets of Ste12 and six variants; heatmap showing the clusters of differentially expressed genes across variants. The color in the heatmap indicates the expression level of each gene normalized by library size, log-transformed and then mapped to z-score. The top enriched Gene Ontology (GO) terms in each cluster are included. The differentially expressed genes of the bottom cluster are shown on the right.

2.3.3 *Loss of mating proficiency correlates with loss of DNA binding*

Amino acid changes in the DNA-binding domain of a transcription factor potentially disrupt its DNA-binding affinity, resulting in the down-regulation of common gene targets and phenotypic changes. To determine whether substitutions in Ste12 reduce its binding to genomic DNA, we compared DNA-binding activity across variants as the number of G418-resistant colonies (transposition efficiency); interaction between Ste12 and DNA provides the basis for the acquisition of the *KanMX* marker and G416 resistance. We found that the transposition efficiency of the three mating-defective variants was only ~10% of the wild type and mating-competent variants (Figure 2.13), suggesting that the mating-defective variants have decreased affinity for DNA.

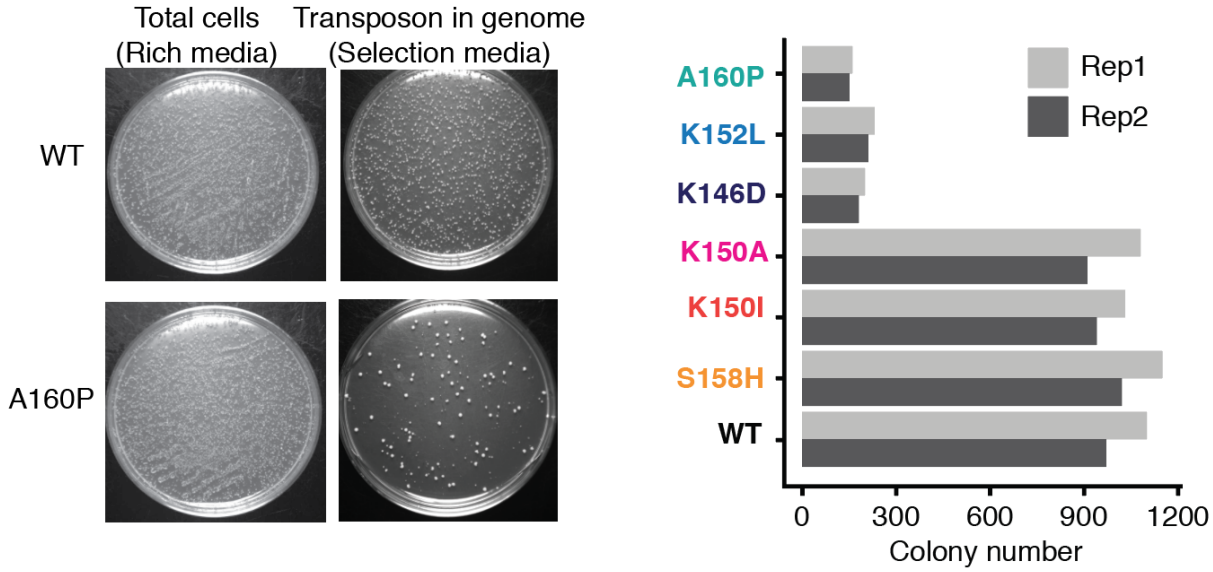


Figure 2.13 *Left:* cell growth in rich medium and selection medium for wild type Ste12 and the A160P variant. *Right*: bar plots showing colony number after selection across all variants. Variants with different phenotypes are annotated by different colors.

A reduction or loss of interaction between a transcription factor and the genome can result in the factor's inability to bind to canonical gene targets and activate their expression. For example, A160P, a mating-deficient variant, resulted in the fewest insertion sites in the calling card experiment and had significantly reduced expression of many genes compared to the wild type (50 of 63 differentially expressed genes; Figure 2.14). Many of the significantly down-regulated differentially expressed genes in the A160P variant were also expressed at a low level in the other mating-deficient variants (Figure 2.15) and these overlapped down-regulated genes act at every step of mating process. For example, *GPA1* encodes the α subunit of the G protein that mediates pheromone sensing at the initial step of the process. *FAR1* functions at an early step by mediating cell cycle arrest and stimulating the polarized growth of the cell towards its mating partner. *AGA1, SAG1* and *FIG2* act at a later step of cell agglutination, contributing to cell-cell contact and the

generation of an ultrastructure favorable for zygote formation. *FUS1, FUS2* and *KAR4* act at the terminal stage of the mating process of cell fusion and nuclear fusion/karyogamy formation. Approximately 78% (14 of 18) of the overlapping down-regulated genes were found to be direct targets of the wild type Ste12 based on the genome binding profile. Of the 18, 21% (including *KAR4, SST2* and *FUS1)* had a reduction in the DNA-binding signal (P-value <0.05; t-test) in all mating-deficient variants (Figure 2.15), and 57% (P-value <0.05; t-test) in at least one mating-deficient variant.

In summary, we conclude that amino acid changes in the DNA-binding domain of Ste12 that disrupt its interaction with the genome can be distinguished by the many fewer calling card insertions that were detected. This disruption results in loss of binding to critical gene targets and markedly reduced expression of mating-related genes, which would be expected to dramatically decrease the mating proficiency of the mating-deficient variants (K146D, K152L, A160P).

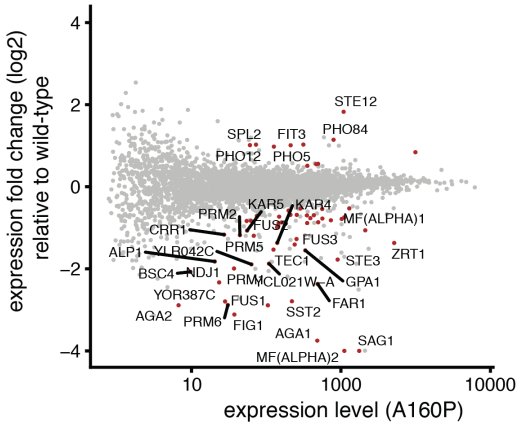


Figure 2.14. Plot showing differentially expressed genes between the A160P variant and wild type Ste12; the x-axis shows total expression level for each gene in A160P while the y-axis shows the expression fold change relative to wild type Ste12. Red color labels differentially expressed genes and these genes with log2(fold change) >1 are labelled with names.

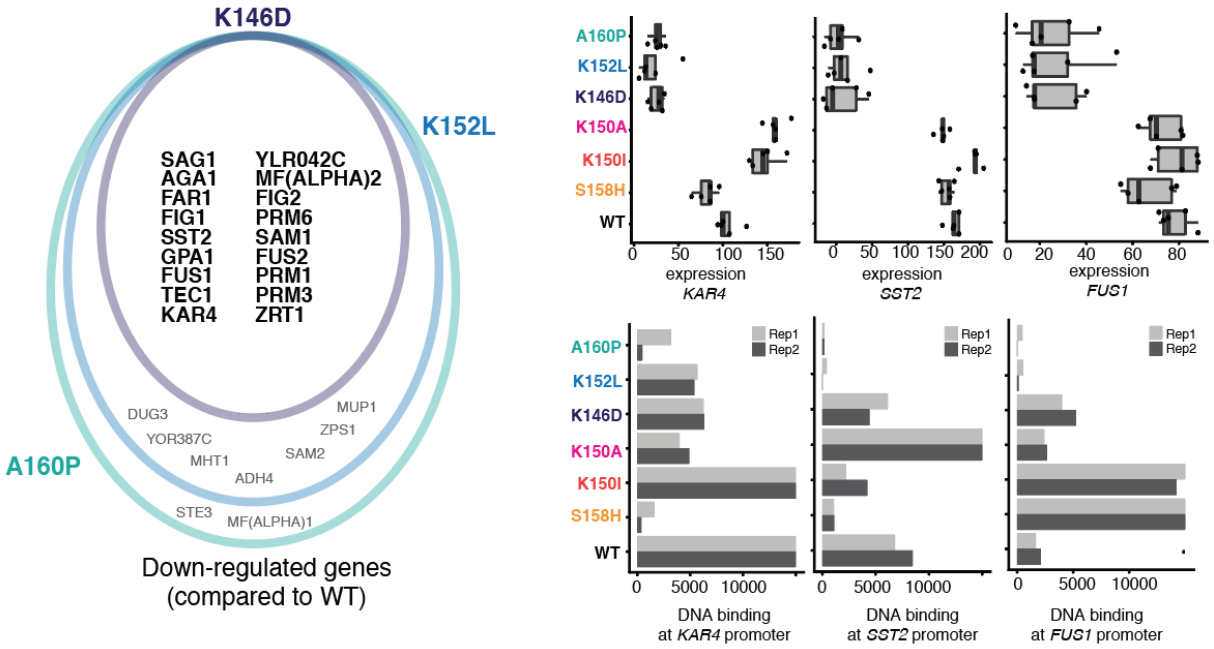


Figure 2.15. *Left*: Venn diagram showing overlap among down-regulated genes for the mating-deficient variants (K146D in purple; A160P in green; K152L in blue). *Right*: Box plots showing gene expression and DNA binding among all variants for the genes *KAR4*, *SST2* and *FUS1*. Upper: gene expression (FKM: fragments per million); Lower: DNA binding (FKM of total unique insertion events) .

2.3.4 *Altered DNA-binding and transcriptomes in hyperinvasive variants*

Two of the mating-defective variants (K146D, K152L) still invade as well or better than the wild-type, indicating that they are not completely defective proteins. One possibility is that they maintain interactions with some cofactors while losing interactions with others or directly with DNA, resulting in distinct variant-specific expression patterns. To test this possibility, we used the genome sequence nearby Ste12-PBase insertion peaks and examined the frequencies of Ste12

canonical sites across the variants. We found that the K152L and K146D variants showed a reduction in binding relative to the wild type Ste12 at canonical sites required for mating (Figure 2.16; *Left*). We calculated the frequencies of transcription factor motif pairs in wild type and the K146D and K152L variants. Pairs containing the well characterized Mcm1 binding site (159) were significantly enriched in K146D compared to the wild type, suggesting that interaction with Mcm1 may be implicated in altered Ste12 binding (Figure 2.16; *Right*). The K146D variant may preferentially bind to promoter regions containing the Mcm1 site to contribute to its unique genome-binding patterns.

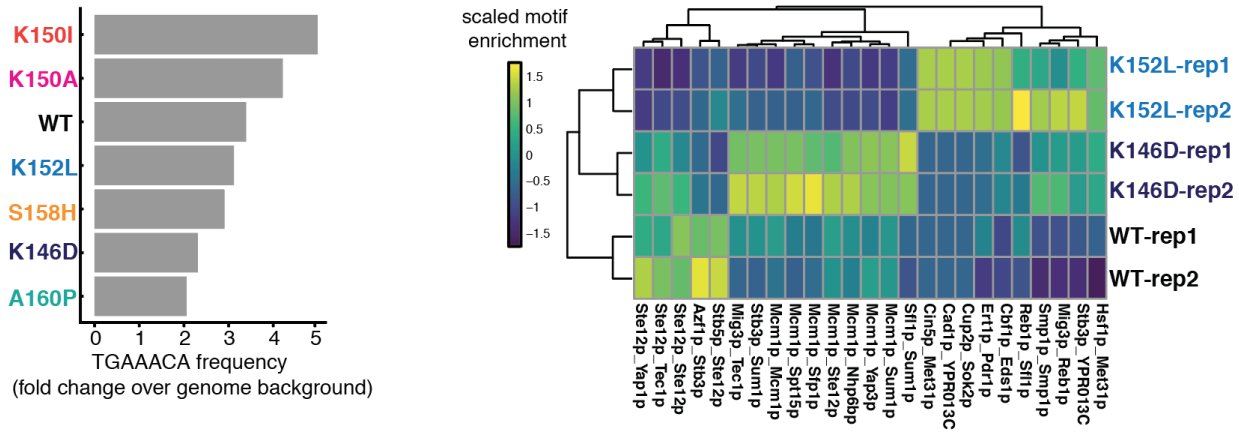


Figure 2.16 *Left*: bar plots showing enrichment of TGAAACA frequency relative to the genome background for the Ste12 variants. *Right*: heatmap showing enrichment of motif pairs found near binding sites for the wild type Ste12, K146D and K152L variants.

We integrated both genome-binding and transcriptome profiles in an effort to uncover the mechanism underlying the altered invasion phenotype seen in the hyper-invasive variants (K146D, K152L, K150I and K150A). Transcriptome analysis revealed that up-regulated genes overlapped in hyper-invasive variants are associated with the cellular response to osmotic stress (*e.g. SIP18, SPG4* and *GRE1*; Figure 2.17 *Left*). When yeast cells are exposed to a hyper-osmotic condition,

osmotic stress responses are triggered and cells form long projections and become hyper-invasive due to Kss1, a key MAP kinase in the invasion pathway (160)The signaling events that lead to osmotic stress response depend on the membrane environmental sensors Sho1 and Msb2 (161). Sho1 is one of the G protein-coupled receptors that act as transmembrane osmosensors (162)Msb2 is a membrane mucin protein (163). Based on the genome binding profile data, the *SHO1* and *MSB2* genes were direct targets of Ste12, suggesting a potential role for Ste12 upstream of osmotic stress response pathway.

For the hyper-invasive variants with reduced mating (K146D and K152L), while they failed to efficiently bind to the Ste12 consensus site and down-regulate critical genes involved in the mating process, both variants had an increased insertion signal (P-value <0.05; t-test) for *MSB2*, and K146D had a higher insertion signal for *SHO1* (although not at a significant level) (Figure 2.17 *Right*). Due to their low levels of transcription, the *SHO1* and *MSB2* genes were not different from wild type in the full transcriptome analysis. However, we detected higher *MSB2* and *SHO1* expression in K146D and K152L variants than wild type using RT-PCR, as well as higher expression of *FLO11*, a final target in the Sho1-sensing pathway (Figure 2.17 *Right*). In the promoter regions of *SHO1* and *MSB2*, an Mcm1 Site is found close to transposon insertion sites of K146D, suggesting that this co-factor interaction may partially underlie the hyper-invasive phenotype.

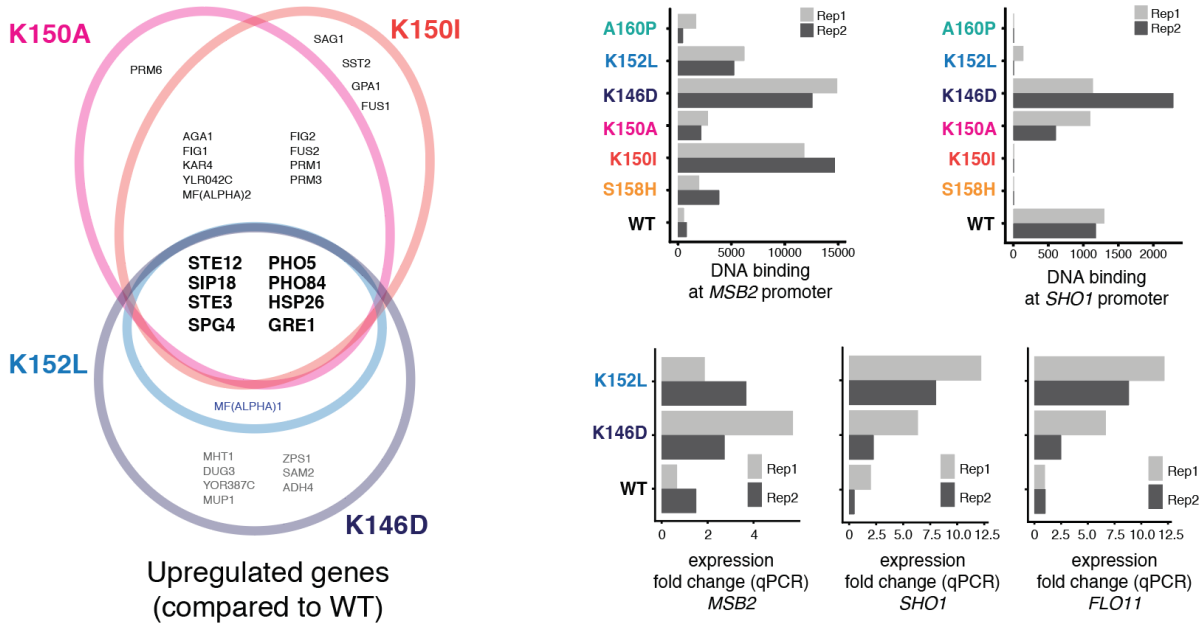


Figure 2.17 *Left*: Venn diagram showing overlap among down-regulated genes for the mating-deficient variants (K146D in purple; A160P in green; K152L in blue). *Right*: Upper: Box plots showing gene expression (FKM: fragments per million) among all variants for the genes *KAR4*, *SST2* and *FUS1*. Lower: Bar plot showing DNA binding (FKM of total unique insertion events) among all variants for the genes *KAR4*, *SST2* and *FUS1*.

2.3.5 *Binding patterns of Ste12 variants in vitro*

Because the binding patterns of the Ste12 variants *in vivo* can be influenced by cofactor interactions that affect the sequence contacted by Ste12, we asked whether direct binding specificity *in vitro* was altered. To characterize the *in vitro* DNA-binding preferences of the variants, we used the HT-SELEX method (155). In our use of this method, purified Ste12 protein was incubated with a large and random (N36) pool of DNA fragments. The DNA fragments bound to protein were isolated, amplified by PCR and incubated again with protein through five rounds,

with the PCR product from each round used for high throughput DNA sequencing. Motif analysis of the SELEX data for wild type Ste12 shows the canonical binding site was enriched over the rounds of selection, and web-logos of the most significant motifs for each variant show patterns similar to wild type Ste12. K146D and K152L were exceptions, showing either no enriched motif (K146D), or a weakly significant, degenerate version of the canonical Ste12 site (K152L).

The TGAAACA sequence was preferred for the K150I and K150A variants, but not for the mating-deficient variants, especially K152L and K146D, which showed no enrichment (Figure 2.18). The K150I variant showed greater enrichment *in vitro* for the TGAAACA sequence than wild type, and also showed a higher enrichment than wild type for this sequence in its *in vivo* genome binding pattern. The *in vitro* binding preferences of Ste12 variants for the canonical Ste12 binding site generally correlated with those preferences *in vivo*. Furthermore, no Ste12 variants showed novel binding specificity *in vitro,* suggesting that their divergent binding patterns *in vivo* are driven by altered affinity for the canonical Ste12 site and the presence of binding co-factors rather than changes in sequence specificity.

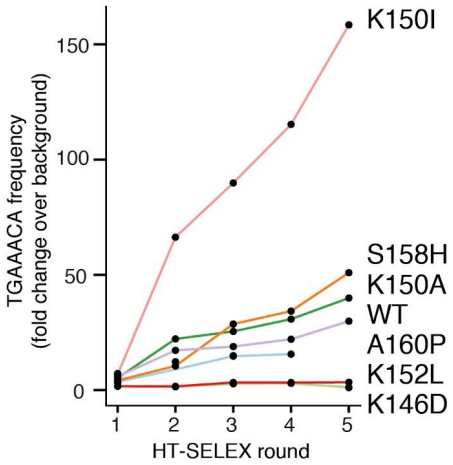


Figure 2.18 Line plot showing TGAAACA frequency in HT-SELEX for wild type Ste12 and six variants. The x-axis shows consecutive rounds of HT-SELEX binding assays. The y-axis shows the enrichment for the TGAAACA 7-mer over the random input sequence background for each variant.

2.4 Discussion

The common assumption that transcriptional regulation is orderly and hence predictable has been shaped by the conservation of gene expression patterns across species, conserved and modular transcription factor domains and a long history of experiments using simpler prokaryotic systems and model regulatory regions in eukaryotes. Our results add to the emerging argument that transcriptional regulation can be readily rewired, changing the underlying transcriptional circuits in various ways while preserving phenotypic outputs. In evolution, this extensive rewiring can occur even among closely related species, and is facilitated by the rapid gain and loss of short *cis*-regulatory sequences or by variation in transcription factors (164).

The *S. cerevisiae* Ste12 protein regulates the traits of mating and invasion by interacting with other transcription factors to bind and activate distinct sets of genes in response to mating pheromone or nutrients, respectively. Single amino acid substitutions in the Ste12 DNA-binding domain can result in dramatically altered phenotypes, such as a shift in preference toward either mating or invasion, or a hyper-invasive phenotype that is independent of the invasion cofactor Tec1. We demonstrate here that these variant Ste12 proteins lead to extensive changes in genome-wide binding patterns and transcriptional outputs.

Although individual Ste12 variants showed highly reproducible binding events in the calling card assay, there was little overlap in binding among variants, even for those with similar mating and invasion phenotypes. This lack of overlap indicates that the subtle changes in the Ste12 protein led to numerous gains and losses of individual binding sites throughout the genome. Nevertheless, it is difficult to pin down the mechanistic basis for gains and losses of sites without additional experimentation. In three cases (K146D, K152L and 160P), the mating deficiency caused by a variant could be explained by a marked reduction in DNA-binding affinity, based on many fewer calling card insertions and either failure to enrich or reduced enrichment of the consensus site in the SELEX experiment. This failure to efficiently bind to Ste12 consensus sites led to the down-regulation of a small set of common genes, several of which function in the mating process. However, that these variants maintain, or even increase, their invasiveness suggests that they are not completely defective proteins and that separation-of-function alleles can arise readily. Moreover, the maintenance of invasion function in variants defective for DNA binding implicates co-factors in the new binding and expression patterns. Variant proteins may maintain interactions with some cofactors while losing it with others, resulting in distinct variant-specific expression patterns. Further analyses of DNA-binding by these variants could use approaches such as gel mobility shift assays to obtain quantitative affinity data. In addition, carrying out these assays on a series of promoters from pheromone-responsive or invasion-specific genes in the presence of purified cofactors such as Tec1, Mcm1 and Matα1 might reveal protein interactions that are either enhanced or reduced by the Ste12 substitutions.

The three variants that resulted in near wild type mating (K150A, K150I and S158H) showed similar numbers of calling card insertions and similar enrichment of the consensus site in the SELEX experiment as the wild type protein. These results imply that DNA-binding affinity is likely to be intact in these variants. Variants that bind to the consensus site like wild type but result in a new phenotype also suggest a change in the interaction with cofactors. The K150A and K150I variants lead to a hyper-invasive phenotype, and these two uniquely up-regulated the expression of genes responsive to osmotic stress. The hyper-invasive variants that are mating-defective (K146D and K152L) up-regulated other genes, indicating that multiple paths to a hyper-invasive phenotype are possible.

Although the Ste12 transcription factor variants easily gained or lost binding sites, most variant-specific binding events seemed to have little to no effect on the organismal phenotypes of mating and invasion. We attribute this phenotypic robustness to the underlying regulatory network architecture which will amplify the effects of specific binding events and expression through feedback loops, motif degeneracy and binding site redundancy while canceling the effect of aberrant binding events. At the same time, the ease with which a single amino acid substitution in Ste12 shifts phenotype suggests that a few novel binding events and expression changes suffice as starting points for rewiring of regulatory programs in evolution(165).

A major goal in functional genomics is the prediction of variant effects from sequence alone. However, current variant effect prediction algorithms and computational structure-based approaches (7), provide minimal annotations such as loss of DNA-binding or likely pathogenesis in humans. In the case of Ste12 variation, it was not possible to attribute specific phenotypes of

mating or invasion based solely on the identity of the single amino acid changes in the Ste12 protein sequence. This failure to predict the *in vivo* effects of transcription factor variants calls for the systematic large-scale functional interrogation of human transcription factor variants *in vivo*. Thus far, there are few options to conduct such studies at sufficient scale in human cells. In this regard, it is notable that we detected genome-wide binding patterns of Ste12 variants in the presence of the wild type protein. Expressing libraries of transcription factor variants from a common genomic site in human cells containing their respective wild type proteins is far more feasible than the prospect of engineering many hundreds of endogenous loci. While most transcription factor variants are unlikely to contribute to phenotypic changes, identifying those with major downstream effects remains a critical challenge.

Chapter 3. Expanding the binding specificity for RNA recognition by a PUF domain

The ability to design a protein that can bind specifically to any RNA and regulate its fate would enable numerous research and therapeutic applications. However, decoding the RNA-binding specificity for most types of RNA-binding domains is challenging, as these domains associate with RNA via complex networks of interactions. The modular architecture of the PUF domain, with eight repeats that each contact one base, make this domain an ideal candidate to generate new RNA-binding specificities. For each repeat of the Pumilio-1 PUF domain, we generated a library of variants that contains the 8,000 possible combinations of amino acid substitutions at residues 12, 13 and 16 that are critical for RNA contact. We carried out yeast three-hybrid selections with

each library against the RNA recognition sequence for the Pumilio-1 domain, with any possible base present at the cognate position recognized by the repeat that was randomized. We used next generation sequencing technology to score the binding activity of each variant for its ability to bind to an RNA sequence containing each possible RNA base at the cognate position. We identified many variants with highly specific interactions by comparing their binding across the four RNA bases. This approach allows us to propose a complete code for RNA recognition by this domain.

Contributions: Daniel Melamed and Stanley Fields developed the idea. Daniel Melamed built the original libraries. Wei Zhou built all other libraries and performed all screen experiments. Wei Zhou performed all the analysis. Wei Zhou and Stanley Fields wrote the manuscript. This work is in preparation for a publication.

3.1 Introduction

Significant progress has been made in the redesign of modular DNA-binding proteins in the past two decades (69). Zinc finger proteins, transcription activator-like (TAL) effectors, recombinases and the CRISPR-Cas9 system have been engineered to display novel DNA recognition specificity (70,166). This success has encouraged researchers to engineer protein-based RNA targeting systems as well, as RNA is involved in a more diverse range of biological functions compared to DNA. RNA targeting in living cells can present a unique opportunity to monitor gene expression and cellular function (79,167). In addition, manipulation of RNAs is usually transient and does not cause irreversible changes in the genome. The key to this engineering is to use an RNA-binding

scaffold that incorporates the desired RNA-binding specificities, enabling specific recognition of a given RNA with limited off-target effects. However, decoding the RNA-binding specificity of most types of RNA-binding domains (RBDs) is challenging, as these domains associate with RNA via complex networks of interactions. For example, the RRM domain is the most abundant RBD in vertebrates to recognize ssRNA (81). However, different RRMs have different RNA recognition mechanisms. For proteins with multiple RRMs, the rule for how each RRM contributes to specificity is unknown, making it difficult to predict recognition specificity from their amino acid sequence alone. Other systems, such as the Cas13 family, have been recently engineered to bind and cleave RNA (98,99). However, their base-pairing requirements for stable RNA association, as well as their potential off-target effects, are still under investigation (100).

PUF proteins are involved in regulating embryogenesis and development in most eukaryotes (104). They contain a conserved RNA-binding domain, known as the Pumilio homology domain, that is generally composed of eight 36-amino-acid repeats (Figure 3.1). Each repeat displays three amino acid residues, called the tripartite recognition motif (TRM) combination, on the concave surface of the protein. An eight-nucleobase target RNA sequence is bound as an extended strand to the concave surface. X-ray structural analysis of the complex indicates that the recognition is highly modular, with each repeat binding to a single RNA base (168). Residues at position 12 and 16 in each repeat directly interact with a Watson-Crick edge of a base, whereas the residue at position 13 is involved in a stacking interaction with the base.

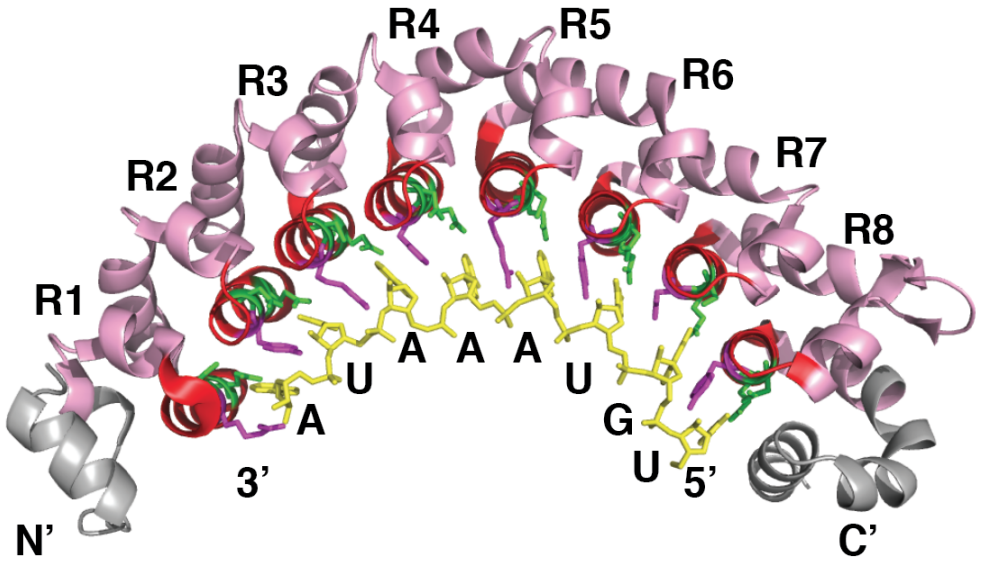


Figure 3.1 Crystal structure of the human PUM1 PUF domain bound to RNA (PDB 1MBY). Helices that carry functional RNA-binding residues (TRM residues) are colored in green and purple; green indicates polar residues (positions 12 and 16) and purple indicates a stacking residue (position 13). RNA bases of the NRE10 recognition sequence are colored in yellow.

The identities of the residues in a TRM combination play a key role in RNA-binding specificity (109,114,115). For example, cysteine and glutamine bind adenine, asparagine and glutamine bind uracil, and serine and glutamate bind guanine. Previous studies have engineered a PUF domain with designed sequence specificity. Cheone *et al.* (115) changed guanine to uracil specificity by mutating the TRM combination *SNE* to *NNQ* (here, in *XXX* format, left to right indicates positions 12, 13 and 16), adenine to guanine specificity by mutating *SRQ/CRQ* to *SRE*, and adenine to uracil specificity by mutating *CRQ* to *NRQ* (115). Ozawa *et al. (79)* introduced more than two sets of mutations described above into PUF proteins and created mutant proteins that recognize transcripts with sequences 4-nt different from the wild-type sequence (79).

While the modular code for A/G/U-specific recognition has been identified in different PUF repeats, there is no code that can specifically bind cytosine in any natural PUF proteins. To tackle this problem, *Dong et al.* (116) generated a library with randomized residues in position 12 and 16 from repeat 6 of the Pumilio 1 PUF domain, screened for their interaction with RNA through a yeast three-hybrid system, and found *SYR* as the TRM combination that specifically binds cytosine (116). Filipovska *et al.* (117) independently reported that five PUF variants with an arginine at position 16, and an amino acid with a small or nucleophilic side chain (glycine, alanine, serine, threonine or cysteine) at position 12 have cytosine-specific recognition (117). Both of these studies showed that the identified code has general applicability and can selectively bind cytosine at all eight positions of an RNA sequence.

Other groups have used an unbiased high-throughput-sequencing method that represents biochemical affinity *in vitro*. Campbell *et al.* (118) determined the specificities of 25 natural and engineered PUF variants in binding reactions with a library that included large numbers of RNA sequences and found, in contrast to previous reports, that the stacking residue in position 13 had a profound influence on specificity as well. For example, changing from *TRQ* to *TWQ* led to a change in specificity from uracil to adenine; changing from *CYR* to *CRR* resulted in a specificity change from uracil to adenine.

Based on this study of PUF engineering (119), for guanine and uracil recognition, the optimal TRMs follow the natural PUF code (G-code: *SHE* for *C. elegans* FBF-2, and *SNE* for human PUM1; U-code: *NXQ* with *X* denoting *T/H/F/Y*). For adenine recognition, the data indicate a more

complex picture. *CFQ* was the adenine-specific code that behaved better than any natural A-binding TRMs. However, the previously identified cytosine-recognition codes for PUM1 could not be broadly used and transferred across species (119).

In order to target any random 8-mer RNA sequence, we still lack a set of recognition codes that can specify the four RNA bases in any positions. Here, we focused on the human Pumilio-1 RNA-binding domain, and made a library of variants for the TRM combination in each repeat. By combining the yeast three-hybrid method with next generation sequencing technology, we scored the binding activities of these PUF variants for their ability to bind to each of the 4 possible RNA bases in the cognate position for each repeat. We identified many variants with highly specific interactions. The resulting dataset allows us to propose a complete code for random 8-mer RNA targeting by swapping the key residues at the contacting positions in each repeat of a PUF domain.

3.2 Materials and methods

Primers and other oligonucleotides. A list of oligonucleotides can be found in table 7.2.

Creation of "*all-in-one*" construct and library generation. A pAIO3H vector was created by cloning the *NotI*-RNA module-*NotI* fragment from p3HR2 into the *NotI* site of pACT2. The unique restriction sites for cloning the protein and the RNA elements were added into the construct. The plasmid uses a centromeric origin. For the randomized library, we used 32 primers (four primers for each PUF repeat) and each primer contained a random *NNK* in each TRM combination. We added a synonymous change in each primer to specify the identity of the RNA base in its cognate

binding sites. We completed the PUF module insert through Phusion overlapping PCR. For the targeted library, we ordered an oligo pool that contains 2000 fragments (Twist Bioscience) and incorporated them into the construct through Gibson assembly (169).

Yeast three-hybrid screen. Library plasmids were transformed into yeast strain *YBZ-1*. The yeast strain constitutively expresses the fusion protein LexA–MS2 coat and a *HIS3* reporter gene under the control of multiple LexA operators. Association of the PUF domain with an RNA sequence leads to the formation of a functional transcription factor that induces the expression of the *HIS3* reporter gene. As a result, yeast cells that carry functional PUF variants proliferate in media lacking histidine, while yeast cells that carry non-functional PUF variants are eliminated. We collected transformants from plates containing SC-Leu media. For selection, we put the transformants into two conditions: plates containing SC-Leu media and plates with SC-Leu-His + 0.5mM 3-AT media. The selection lasted for 4 days. We collected cells and extracted their plasmids (Zymoprep Yeast Plasmid Miniprep II kit; Cat 11-315).

Sequencing library preparation and analysis. The region including each PUF repeat was amplified through sequential reactions. Internal PCR was carried out through primers with sequences that anneal to each repeat (primers 623-654) and external PCR was carried out to add an Illumina sequencing adapter (primers 419-422). Phusion polymerase was used for all reactions, and each reaction was performed on a BioRad MiniOpticon and monitored to avoid over-amplification. The PCR products were sequenced using the Nextseq 550 platform. Downstream analyses were performed in R.

3.3 Results

3.3.1 *PUF variants with base-specific interactions identified with the yeast three-hybrid system.*

To quantify the interaction between a PUF domain and its cognate RNA, we applied the yeast three-hybrid system (170) (Figure 3.2). In this system, the interaction of the PUF domain with an RNA in a yeast cell leads to the activation of the reporter gene *HIS3*, such that the cells survive in selection media without histidine. We designed an "all-in-one" construct including both a protein module encoding the PUF-activation domain (PUF-AD) fusion protein and an RNA module encoding the cognate RNA sequence fused to the MS2 coat protein recognition sequence (Figure 3.3; *upper left*). We introduced the plasmid into yeast strain *YBZ-1*, which constitutively expresses the fusion protein LexA–MS2 coat protein. The *HIS3* reporter gene is under the control of multiple LexA operators. As expected, we found cells carrying each combination of RNA and protein grew equally well on media that selects only for the presence of the plasmid (Figure 3.3, *lower left*). In contrast, on media that selects for *HIS3* expression, only those cells that contained a cognate combination of PUF domain and RNA site grew (Figure 3.3, *lower left* -Leu-His + 0.5mM 3AT). The *in vivo* base-specific binding pattern for each PUF repeat could be recapitulated in this system (Figure 3.4), indicating that the yeast three-hybrid system can be used to identify the interaction and specificity between a PUF domain and its RNA target.

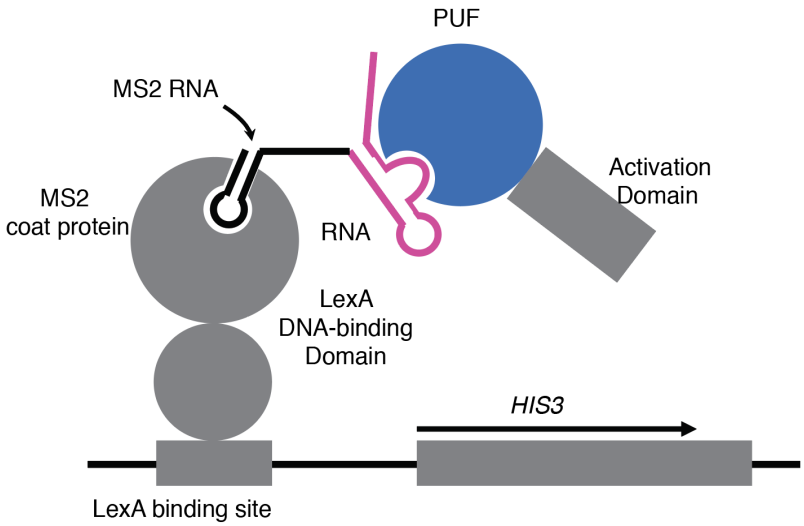


Figure 3.2 Illustration of the yeast three-hybrid screen. Interaction of a PUF domain and its cognate RNA sequence leads to the formation of a functional transcription factor that induces the expression of the *HIS3* reporter gene, such that yeast cells can survive in media without histidine. Elements illustrated here, including the LexA-MS2 coat protein and multiple LexA operators, were already incorporated into yeast strain *YBZ-1*.

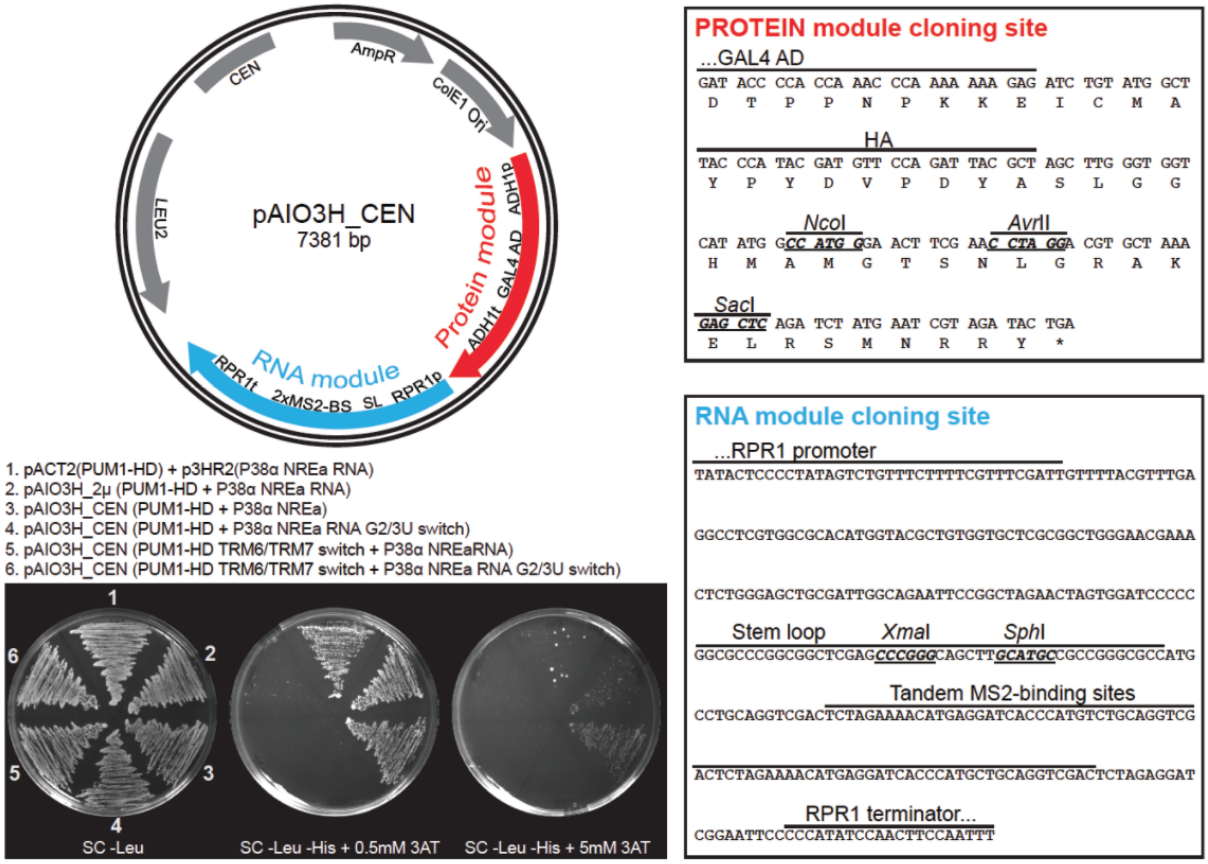


Figure 3.3 An "*all in one*" Y3H vector for analyzing a PUF domain-RNA interaction. *Upper left*, illustration of the pAIO3H construct, which incorporates both protein module and RNA module on a plasmid with a centromeric origin. *Lower left*, different conditions for testing the efficacy of the Y3H system: "1" indicates the classic two-plasmid system to test for PUF domain-RNA interaction; "2" indicates the "*all in one*" Y3H system with a 2-micron origin of replication; "3" indicates the "*all in one*" Y3H system with a centromeric origin of replication; "4/5/6" are negative controls with RNA bases switched so that the correct cognate base is not present. *Right*, description of restriction sites for manipulating the protein module and RNA module.

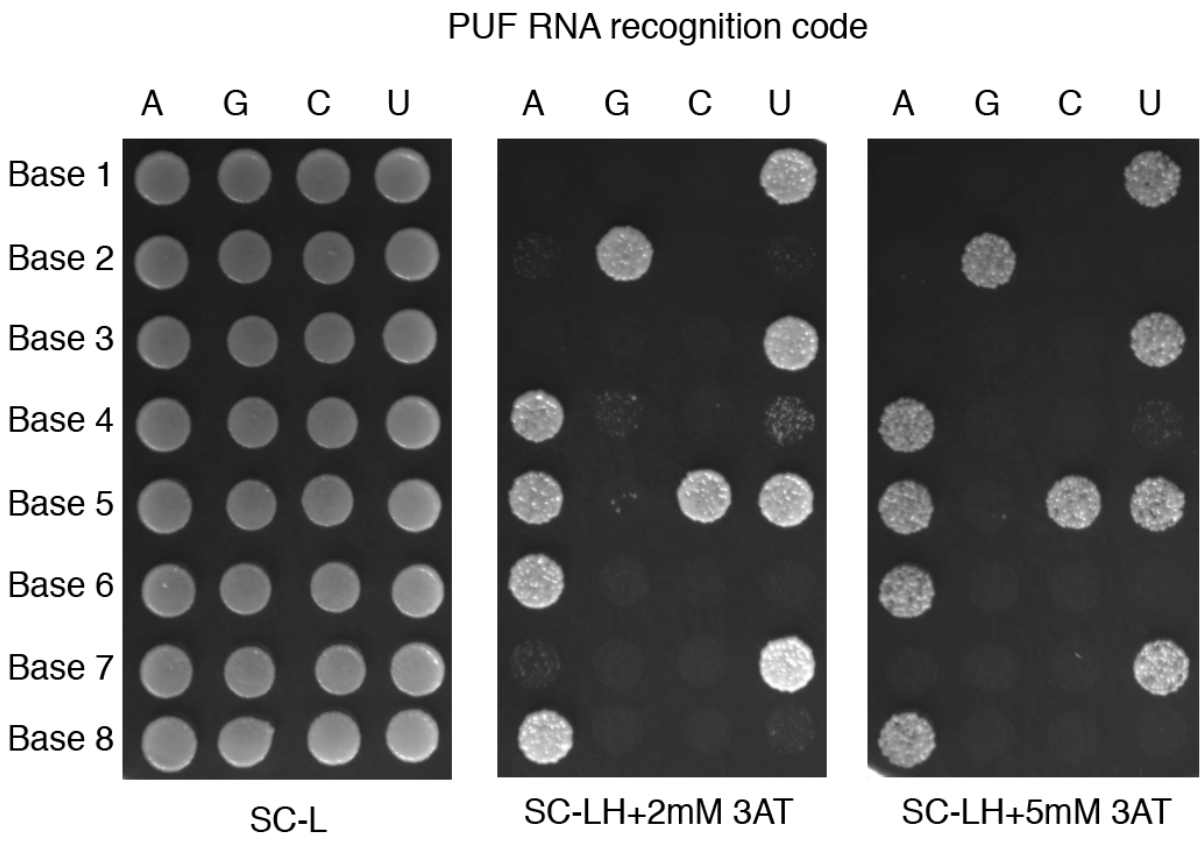


Figure 3.4 A colony spotting assay with or without histidine selection. The selection is carried out in different 3-AT concentrations (higher concentrations mean a more stringent selection). An RNA sequence with each of the possible single base substitutions in the 8-mer recognition sequence were tested for interaction with the wild-type PUF domain.

We then combined the yeast three-hybrid system with next generation sequencing technology to elucidate the RNA-binding preferences of a large number of PUF variants in a single culture. For each repeat of the PUF domain, we generated a library of variants that contains the 8,000 possible combinations of amino acid substitutions at residues 12, 13 and 16 that are critical for RNA contact (Figure 3.5). We screened the interaction of each repeat library against RNA sequences containing each of the four possible RNA bases present at the cognate position. Yeast cells that carry

functional interactions proliferate in media lacking histidine, while yeast cells that carry non-functional interactions are eliminated. We retrieved the library from both input and post-selection pools, and determined the frequency of each variant in both pools by high-throughput sequencing. The change in the frequency from input to selection pool serves as a measurement of binding activity for each PUF variant, which is calculated as a PUF domain-RNA interaction score in our assay.

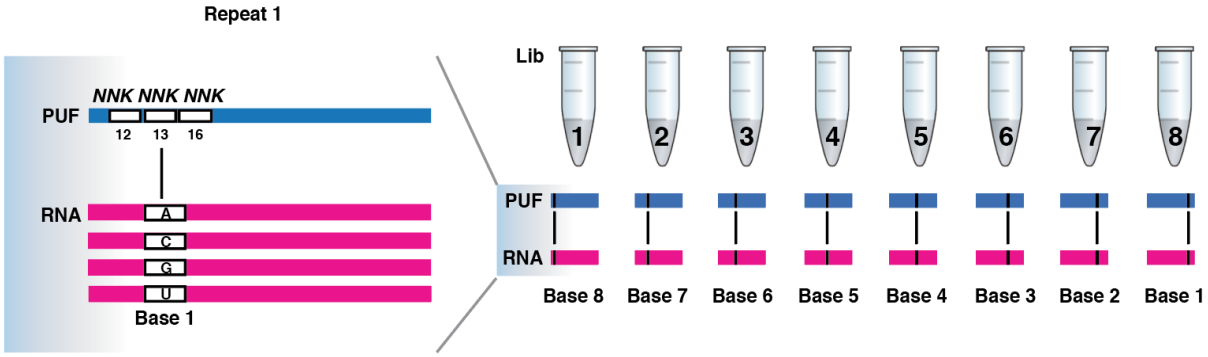


Figure 3.5 Creation of a randomized library of variants that contains the 8,000 possible combinations of amino acid substitutions for each PUF repeat at residues 12, 13 and 16, which are critical for RNA contact.

Based on the enrichment in the post-selection pool, the RNA-binding activity of 169,587 PUF domain variants were scored. The dataset contains 24,751 nonsense variants and 144,836 missense variants from the eight repeats. The interaction score distribution of all variants revealed that, in general, nonsense PUF variants were deleterious for interaction with any RNA sequence and missense PUF variants function in a bimodal distribution (Figure 3.6). The PUF domain-RNA interaction score for each PUF variant showed a high degree of overlap between two experimental replicates (Figure 3.7; Pearson correlation coefficient ranged from R=0.952-0.982).

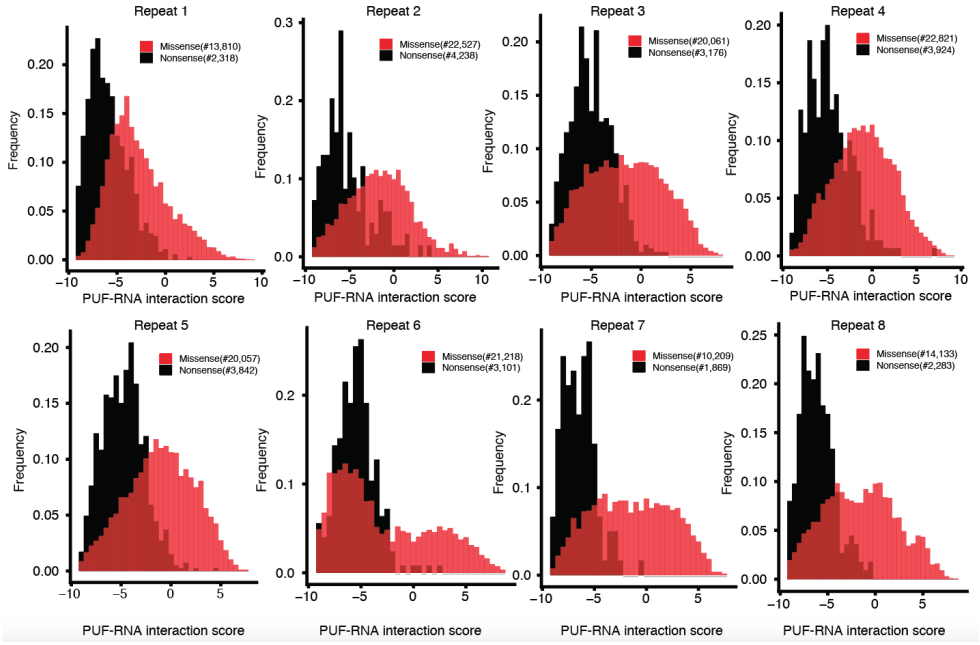


Figure 3.6. Histogram showing PUF domain-RNA interaction score distributions from repeat 1 to repeat 8. Nonsense variants are indicated in black; missense variants are indicated in red.

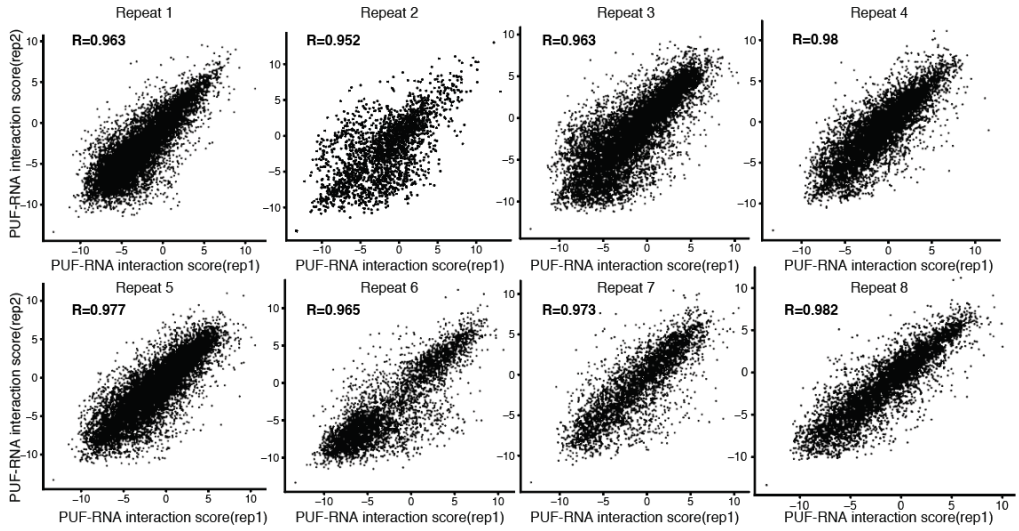


Figure 3.7. Correlation of PUF-RNA interaction score between two experimental replicates from repeat 1 to repeat 8.

We assigned a specificity score for each PUF variant by comparing its binding across the four RNA bases. For each PUF variant, the difference between the highest and second highest interaction score was regarded as a measure of its specificity. Using a threshold of interaction score greater than 5, and a specificity score greater than 4, we identified many PUF variants with highly specific interactions (Figure 3.8; number of enriched PUF variants ranged from 5 to 79 across the eight repeats). For example, in repeat 1 (Figure 3.9), we found nine PUF variants specific for uracil interaction (*e.g. NWS*, *NFS*), 11 PUF variants specific for cytosine recognition (*e.g. TFR*, *QFR*), one PUF variant specific for guanine recognition (*SGD*), and one PUF variant specific for adenine recognition (*IFV*). For uracil recognition, we found that asparagine is the most preferred amino acid in position 12 and a polar uncharged amino acid such as glutamine, serine or threonine is the most preferred in position 16. Position 13 is occupied by aromatic amino acids such as tryptophan and phenylalanine. While this pattern differs slightly from the optimal TRMs for uracil recognition (U-code: *NXQ* with *X* denoting *T/H/F/Y*), it recapitulates the general trend. As expected, in repeat 1, we found that arginine is the most preferred amino acid in position 16 and a polar uncharged amino acid (*e.g.* threonine, glutamine, asparagine) is preferred in position 12, which recapitulates previous findings (117,171).

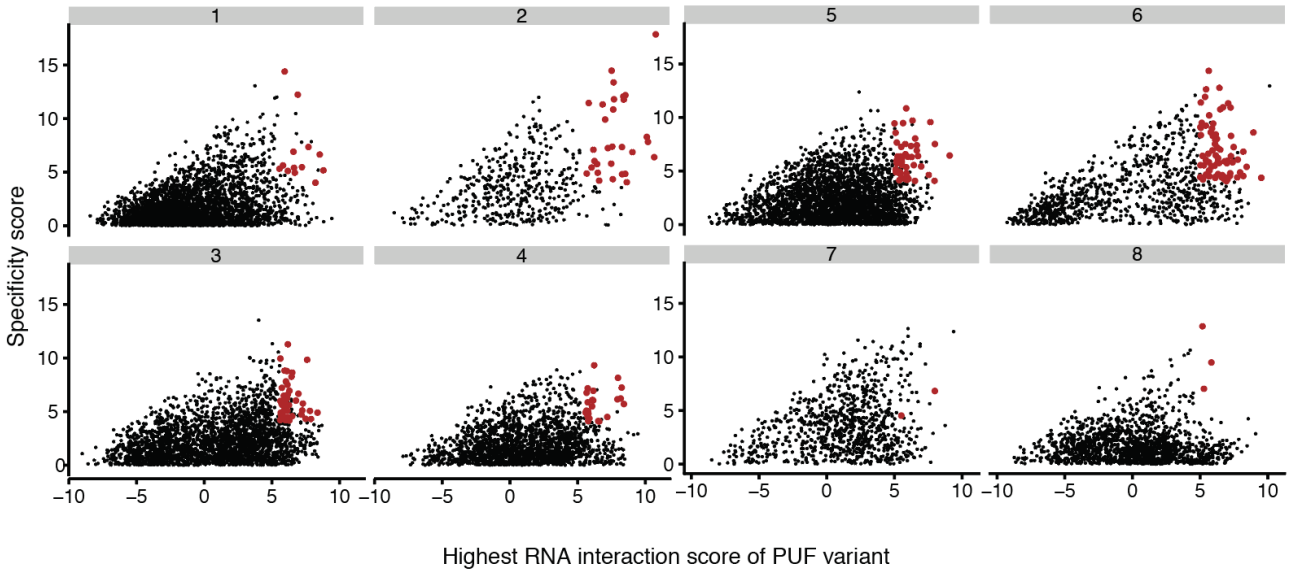


Figure 3.8. Dot plot indicates the highest PUF domain-RNA interaction score and the specificity score for each variant from repeat 1 to repeat 8. The X-axis indicates the highest RNA interaction score; the Y-axis indicates the specificity score. Those variants with an interaction score over than 5, as well as specificity score over than 4, are highlighted in red color.

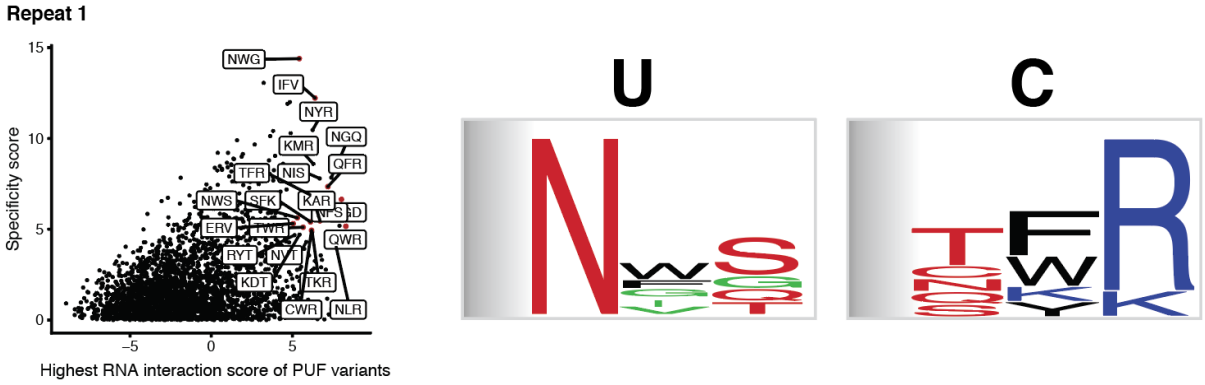


Figure 3.9. *Left*, dot plot indicates the highest PUF domain-RNA interaction score and the specificity score for each variant from repeat 1. *Right,* base-specific recognition patterns for uracil and cytosine in repeat 1.

3.3.2 *Targeted screening of candidate PUF variants*.

Due to the large size of the libraries of randomized PUF variants, for many variants, the initial yeast three-hybrid screen did not comprehensively recover a binding activity score against all four RNA bases and across all eight repeats. We thus conducted a targeted screen of promising candidate PUF variants. Using a threshold of interaction score greater than 5, as well as specificity score greater than 4, we chose about 250 candidate PUF variants for targeted oligonucleotide synthesis (Figure 3.10). We cloned the oligonucleotide pool into the eight PUF repeat locations for a comprehensive survey of the interaction of these candidates with the four possible RNA bases. For each repeat with a pool of synthesized variants, the selection for binding activity against the four RNA bases was carried out in a separate culture, with media that selected for *HIS3* expression (-Leu-His + 1mM 3AT). In order to compare its binding across the four RNA bases, we spiked the wild-type PUF domain into each culture for normalization. We collected plasmids from both input and post-selection pools and measured the change in frequency of each PUF variant by high-throughput sequencing.

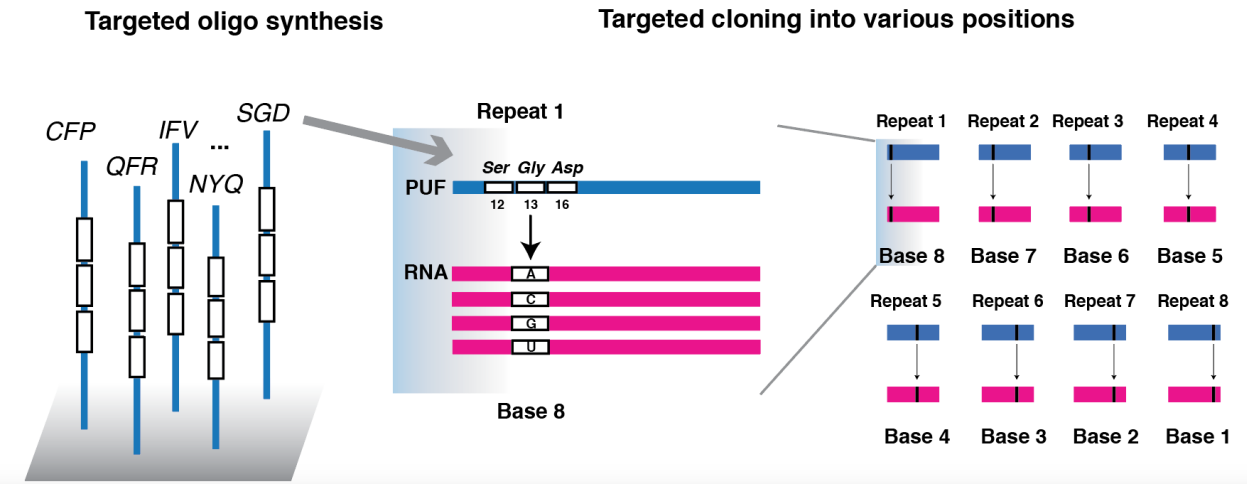


Figure 3.10. Workflow of targeted PUF variant screening. About 250 candidate PUF variants were chosen for oligonucleotide pool synthesis and these fragments were cloned into the eight different

PUF repeat locations for a comprehensive survey of their interaction with the four possible RNA bases.

For each repeat, we recovered the majority of the synthesized PUF variants (Figure 3.11, *left*). Each surveyed PUF variant was well-represented in the input pool, with a frequency centered on 0.1% (Figure 3.11, *right*). An interaction score was assigned to each PUF variant based on its enrichment in the post-selection pool. The distribution of interaction scores for all nonsense variants indicates that they were mostly deleterious, with an interaction score lower than -5. Compared with the initial screen, missense variants had a stronger enrichment in the post-selection pool (Figure 3.12).

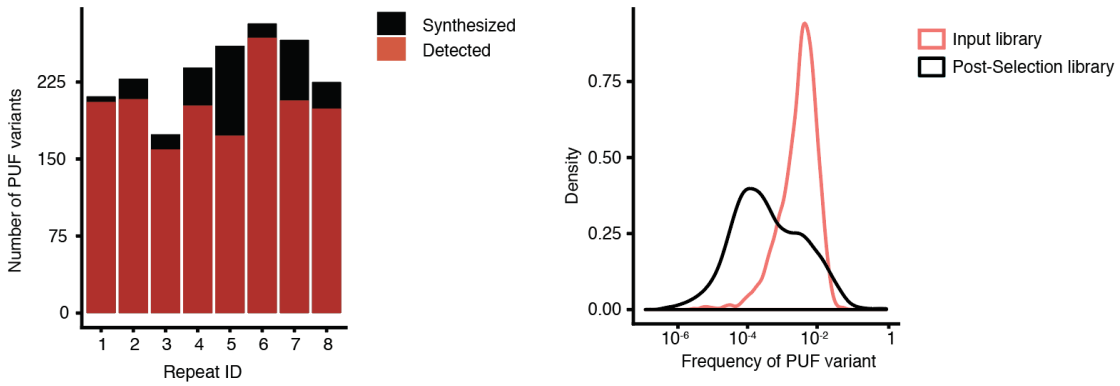


Figure 3.11 *Left*, Bar plot showing the recovery of synthesized PUF variants. X-axis indicates repeat identity from repeat 1 to repeat 8; Y-axis indicates the number of PUF variants in the synthesized pool and the detected pool. *Right*, density plot showing the frequency of each PUF variant in input library or selected library.

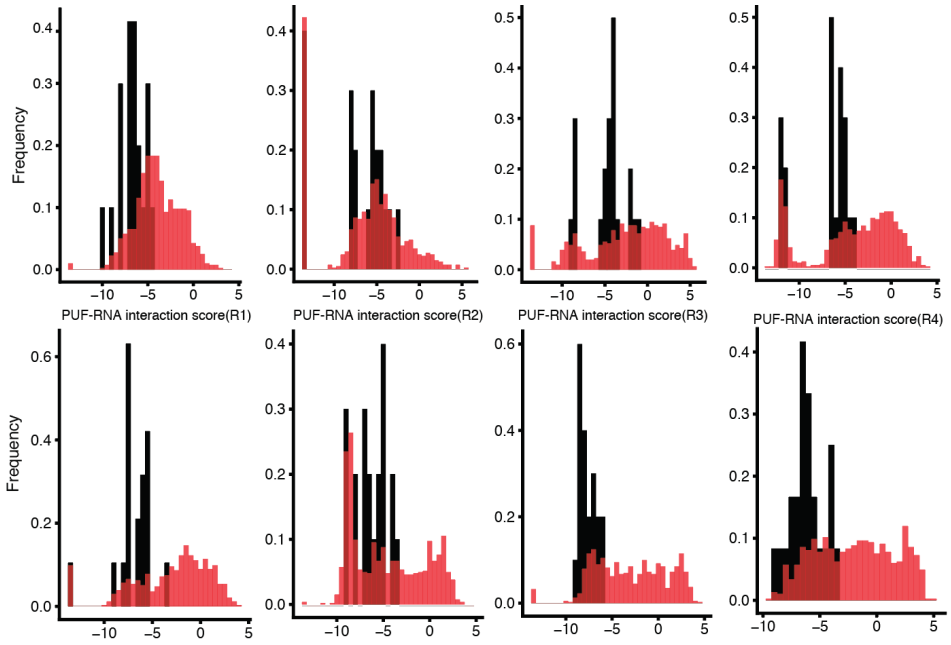


Figure 3.12 Histogram showing PUF domain-RNA interaction score distributions from repeat 1 to repeat 8. Nonsense variants are indicated in black; missense variants are indicated in red.

Based on the interaction scores against the four RNA bases, for each repeat, we clustered promising PUF variants with scores greater than 5. Remarkably, we identified many variants with highly base-specific interactions for each of the eight repeats. We generated sequence logos to summarize the base-specific recognition patterns for each repeat (Figure 3.13). For example, in repeat 1, consistent with the initial screen, *SGD* is one combination for binding guanine. While *SGD* differs from the natural guanine-specific binding code *SNE*, the combination of serine and a negatively charged amino acid (aspartate or glutamate) in position 12 and 16 was found as a general trend for G-specific binding across the majority of the repeats. Many U-specific variants were also identified in repeat 1 (*e.g. NFV*, *NFS*, *NWS*). As expected, asparagine is the preferred amino acid in position 12, and serine or glutamine is the preferred amino acid for position 16. For position 13, involved in a stacking interaction, aromatic amino acids such as phenylalanine or

tyrosine are preferred. Moreover, we found many combinations with non-specific binding in repeat 1; these combinations have positive charged amino acids in position 12 and 16, consistent with a previously characterized pattern for non-specific RNA recognition (113). In summary, the base-specific recognition codes summarized in Figure 3.13 can be used as a resource for future PUF engineering.

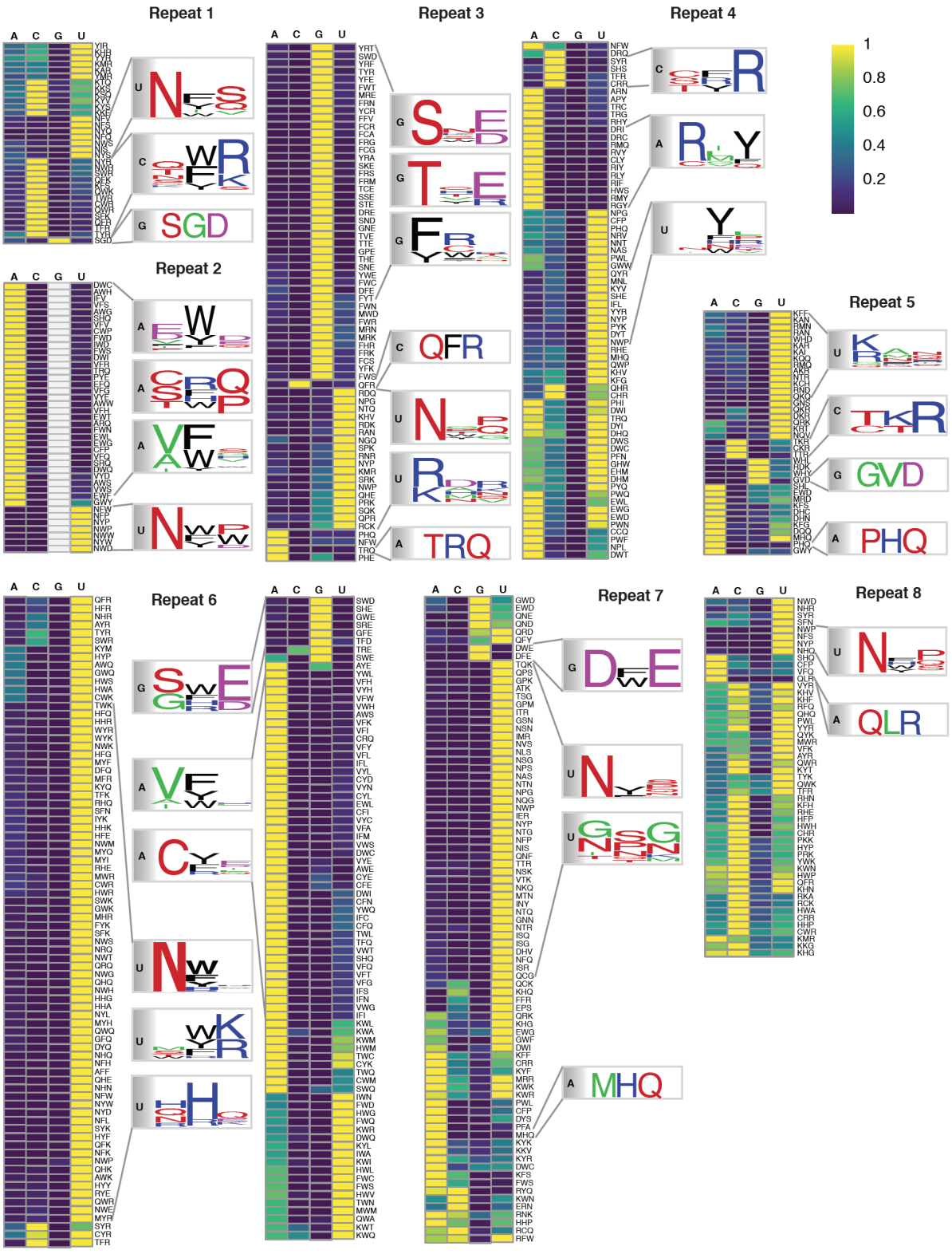


Figure 3.13 Base-specific recognition codes for each repeat are shown.

3.3.3 *Comparison of base-specific recognition patterns across the eight repeats.*

We compared the specificity of each PUF variant across the eight repeats. We found that many base-specific recognition codes are not generic for all repeat locations, as previously reported (117). For example, we found a polar, uncharged amino acid (*e.g.* glutamine or threonine) in position 12 and a positively charged amino acid (*e.g.* arginine) in position 16 is the preferred combination for C-specific binding. However, this preference was not uniform across the eight repeats. The high specificity for cytosine was found only in the more N-terminal PUF repeats, such as repeat 1 or repeat 3, and was markedly reduced in more C-terminal repeats (Figure 3.14, *A panel*). The same pattern was seen for the previously identified C-specific codes (*e.g. SYR)* as well (Figure 3.14, *A panel*), which potentially explains why C-specific codes were not transferable to other species (117,120). In addition, as expected, we found polar residues (position 12 and 16) are important in binding specificity. For example, *PHQ* was found to have an A-specific recognition pattern across the majority of repeats (Figure 3.14, *C panel*, *right*). When position 12 was instead asparagine, the specificity code of *NHQ* instead recognized uracil (Figure 3.14, *D panel*, *right*).

While most previous work suggests that the stacking residue (position 13) is critical to binding affinity only (172), Campbell *et al.* (118) showed that that stacking residue has a profound influence on specificity as well. In agreement with their report, we found that the stacking residue (position 13) plays a key role in binding specificity. For example, *SNE* is the natural code to recognize guanine in repeat 7. We found that this code can also be applied for G-specific binding in repeat 3 and repeat 5. If instead tryptophan is in position 13 (*SWD*), G-specific recognition can be achieved in repeat 6. Moreover, if position 13 is glycine (*SGD*), G-specific recognition can be

achieved in repeat 1, while *SNE* and *SWD* were not found (Figure 3.14, *B panel*). These results suggest that the presence of an aromatic amino acid (W/Y/F) or a non-aromatic amino acid in position 13 can greatly affect recognition specificity.

While most canonical base-specific recognition patterns were recapitulated in our screen, we found many examples of alternative codes. For example, for adenine recognition, (*C/S/T)RQ* are preferred for the majority of the repeats. However, in repeat 2 and repeat 6, the combination of valine and phenylalanine in position 12 and position 13 is an alternative way to achieve A-specific binding (Figure 3.14, *C panel*). In fact, *VFQ* is the best code to achieve A-specific recognition in repeat 1, repeat 2, repeat 4 and repeat 6 (Figure 3.14, *C panel*). For uracil recognition, asparagine is preferred in position 12 across the majority of repeats, except the middle repeats such as repeat 4 and repeat 5. For these middle repeats, we found a positively charged amino acid is more preferred than a polar residue in position 12 or 16 to achieve this specificity (Figure 3.13). Even for the positions that use canonical base-specific recognition pattern, each repeat has its own preferred code. For example, while *NWP* and *NHQ* bind to uracil across the majority of repeats, repeat 1 prefers *NFS* than these two canonical codes (Figure 3.14, *D panel*).

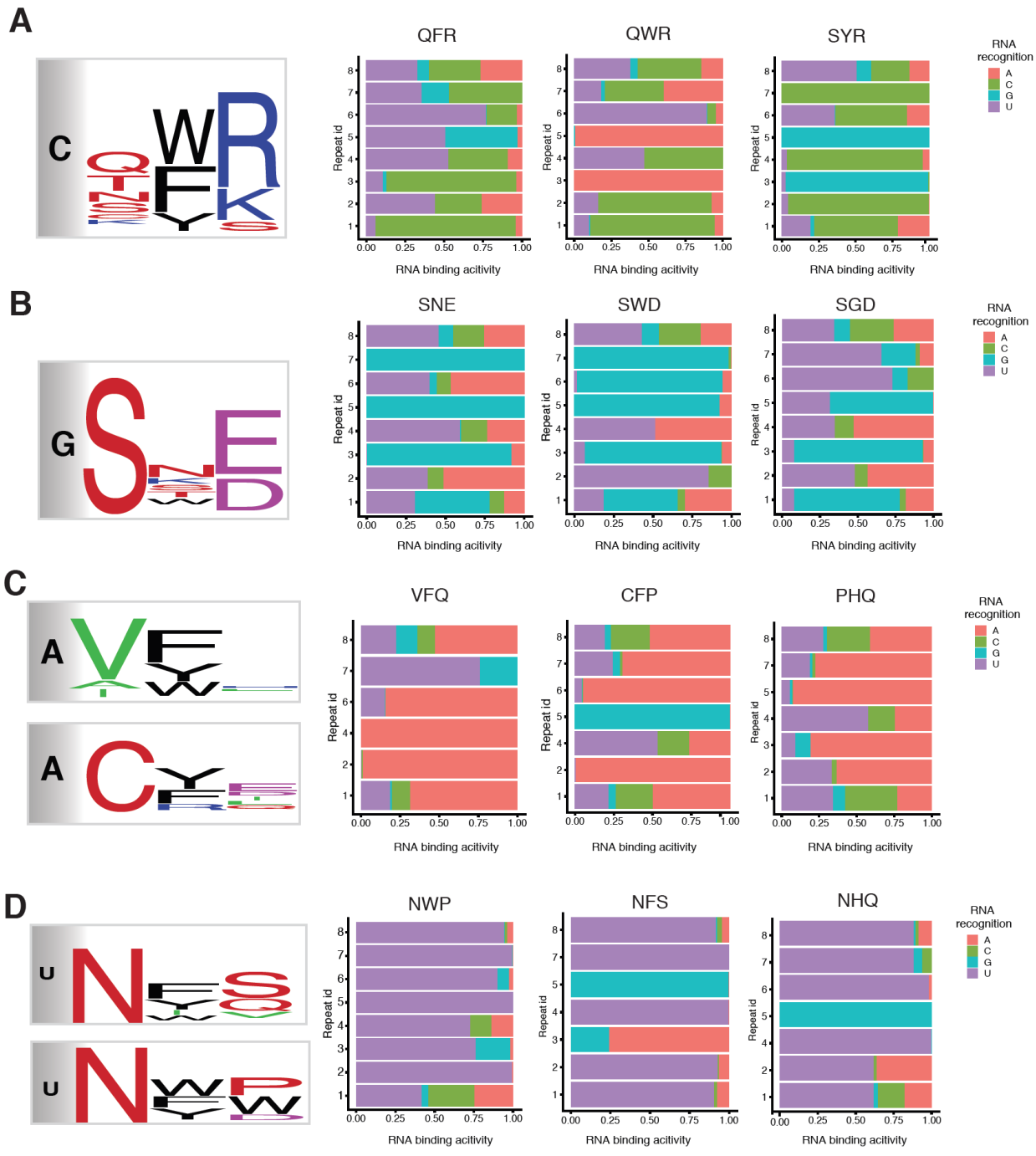


Figure 3.14 Comparison of base-specific recognition patterns across repeats. *Left*: Logo plots of base-specific patterns. *Right*: RNA recognition preferences for individual TRM combinations across repeats.

3.4 Discussion

The PUF domain is an ideal candidate for engineering a protein to bind to an arbitrary RNA sequence, as it uses a modular architecture for recognition and binds with high affinity and specificity to an 8-mer sequence. Using high-throughput sequencing of positives in the yeast three-hybrid assay, we carried out deep mutational scanning of all possible TRM combinations of each of the eight repeats of a PUF domain. By scoring the binding activity of variants of each repeat to an RNA sequence in which the cognate RNA position was occupied by any of the four possible bases, we elucidated the RNA-binding preferences of a large number of variants. We calculated scores both for binding interaction and for specificity, yielding the base-specific recognition patterns summarized in Figure 3.13. These patterns can be used as a resource for future PUF engineering. The enriched TRM combinations can be compared across the eight repeats to explore the molecular basis by which the PUF domain achieves its RNA binding specificity.

Stacking interactions pervade PUF domain-RNA complexes across species. These interactions result from long columns of stacking bases and amino acid side chains along the entire length of the PUF protein. However, how these interactions affect RNA binding specificity across repeats has not been well explored. Koh *et al.* (173) carried out an alanine substitution experiment in repeat 3 of *C. elegans* FBF-2 PUF domain and revealed that stacking amino acids contribute to the protein's specificity for RNA sequence. Further structural analysis supports the idea that different stacking arrangements can lead to different specificity for RNA recognition (173). Campbell *et al. (119)* provided additional examples through an *in vitro* protein-RNA binding assay. Here, through a comparison of base-specific recognition patterns across repeats, we found that stacking interactions in different positions have profound influence on the specificity of each repeat. For

example, *SWD* can specifically recognize guanine in repeat 5, 6, 7 while *SGD* can specify only guanine in repeat 1 and repeat 3. This results suggests that the stacking residue in each PUF repeat potentially functions in combination with the neighboring amino acids to specify RNA bases.

While cytosine-specific recognition was identified in variants of the human pumilio PUF domain (117,171), later studies found that the specificity of these codes could not be transferred across species (120). Here, our study suggests that these previously identified "universal" codes (*A/S/T/C/G-YR*) are not transferable across all repeats as well. For example, *SYR* can only specify cytosine in repeat 2, repeat 4 and repeat 7, but not the other repeats. We identified novel TRM combinations that work well in these other repeats, such as repeat 1 and repeat 3.

Natural existing PUF domains have been selected to increase or decrease specificity for each repeat while maintaining the binding affinity necessary for biological function. The current rational design of a PUF domain as an RNA-binding platform has mostly focused on polar residues in position 12 and 16; however, these hydrogen-bonding residues work together with the stacking residue. According to what we have found in our dataset, substitutions of stacking residues can either broaden or restrict the range of acceptable bases. For future rational designs of a PUF domain as an RNA-binding platform, our results suggest that repeat location, as well as stacking residues, should be important considerations.

Chapter 4. Characterizing the temporal dynamics of gene expression in single cells with sci-fate

Gene expression programs are dynamic, e.g. the cell cycle, response to stimuli, normal differentiation and development, etc. However, nearly all techniques for profiling gene expression in single cells fail to directly capture the dynamics of transcriptional programs, which limits the scope of biology that can be effectively investigated. Towards addressing this, we developed sci-fate, a new technique that combines 4sU labeling of newly synthesized mRNA with single cell combinatorial indexing (sci-), in order to concurrently profile the whole and newly synthesized transcriptome in each of many single cells. As a proof-of-concept, we applied sci-fate to a model system of cortisol response and characterized expression dynamics in over 6,000 single cells. From these data, we quantify the dynamics of the cell cycle and glucocorticoid receptor activation, and explore their intersection. We furthermore use these data to develop a framework for inferring the distribution of cell state transitions. We anticipate sci-fate will be broadly applicable to quantitatively characterize transcriptional dynamics in diverse systems.

Contributions: Junyue Cao and Jay Shendure developed the idea. Junyue Cao developed the technique and performed experiments. Junyue Cao performed computational analysis with suggestions from Wei Zhou. Junyue Cao and Jay Shendure wrote the manuscript. This work is in press for publication in *Nature Biotechnology*.

4.1 Introduction

During organismal development, as well as during myriad physiological and pathophysiological processes, individual cells traverse a manifold of molecularly and functionally distinct states. The accurate characterization of these trajectories is key to advancing our understanding of each such

process, for identifying the causal factors that drive them, and for rationally designing effective perturbations of them. However, although experimental methods for profiling various aspects of single cell biology have recently proliferated, nearly all such methods deliver only a static snapshot of each cell, e.g. of gene expression at the moment of fixation or cell lysis.

To recover temporal dynamics, several groups have developed computational methods that place individual cells along a continuous trajectory based on single cell RNA-seq data, *i.e.* the concept of pseudotime(121,124,126,174–176). However, such methods are inherently limited in several important ways, including that they are inferring rather than directly measuring dynamics, that they are dependent on sufficient representation across the trajectory, and that they may fail to capture the detailed dynamics of individual cells (*e.g.* directionality, multiple superimposed potentials, etc.). Although time-lapse microscopy is a distinct technology that overcomes some of these limitations, it is limited in throughput and scope (*e.g.* enabling visualization of a few marker genes in a few cells), and as such may be insufficient to decipher the complexity of many biological systems.

Here we describe a novel technique, sci-fate, to measure the dynamics of gene expression in single cells at the level of the whole transcriptome. In brief, we integrated protocols for labeling newly synthesized mRNA with 4-thiouridine (4sU)(131,177) with single cell combinatorial indexing RNA-seq(132). As a proof-of-concept, we applied sci-fate to a model system of cortisol response, and characterized expression dynamics in over 6,000 single cells. From these data, we quantify the dynamics of the transcription factor (TF) modules that underpin the cell cycle, glucocorticoid receptor activation, and other processes. We furthermore use these data to develop a framework

for inferring the distribution of cell state transitions. The experimental and computational methods described here may be broadly applicable to quantitatively characterize transcriptional dynamics in diverse systems.

4.2 Material and Methods

Mammalian cell culture. All mammalian cells were cultured at 37°C with 5% CO_2, and were maintained in high glucose DMEM (Gibco cat. no. 11965) for HEK293T and NIH/3T3 cells or DMEM/F12 medium for A549 cells, both supplemented with 10% FBS and 1X Pen/Strep (Gibco cat. no. 15140122; 100U/ml penicillin, 100 μg/ml streptomycin). Cells were trypsinized with 0.25% typsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times per week.

Sample processing for sci-fate. All cell lines (A549, HEK293T and NIH/3T3 cells) were trypsinized, spun down at 300x**g** for 5 min (4°C) and washed once in 1X ice-cold PBS. All cells were fixed with 4ml ice cold 4% paraformaldehyde (EMS) for 15 min on ice. After fixation, cells were pelleted at 500x**g** for 3 min (4°C) and washed once with 1ml PBSR (1 x PBS, pH 7.4, 1% BSA, 1% SuperRnaseIn, 1% 10mM DTT). After wash, cells were resuspended in PBSR at 10 million cells per ml, and flash frozen and stored in liquid nitrogen. Paraformaldehyde fixed cells were thawed in a 37°C water bath, spun down at 500x**g** for 5 min, and incubated with 500ul PBSR including 0.2% Triton X-100 for 3min on ice. Cells were pelleted and resuspended in 500ul nuclease free water including 1% SuperRnaseIn. 3ml 0.1N HCl were added into the cells for 5min incubation on ice. 3.5ml Tris-HCl (pH = 8.0) and 35ul 10% Triton X-100 were added into cells to neutralize HCl. Cells were pelleted and washed with 1ml PBSR. Cells were resuspended in 100ul

PBSR. 100ul PBSR with fixed cells were incubated with mixture including 40ul Iodoacetamide (IAA, 100mM), 40ul sodium phosphate buffer (500mM, pH = 8.0), 200ul DMSO and 20ul H2O, at 50°C for 15min. The reaction was quenched by 8ul DTT (1M) and 8.5ml PBS. Cells were pelleted and resuspended in 100ul PBSI (1 x PBS, pH 7.4, 1% BSA, 1% SuperRnaseIn). For all later washes, cells were pelleted by centrifugation at 500x**g** for 5 min (4°C).The following steps are similar with sci-RNA-seq protocol with paraformaldehyde fixed nuclei (178).

Read alignment and downstream processing. Read alignment and gene count matrix generation for the single cell RNA-seq was performed using the pipeline that we developed for sci-RNA-seq with minor modifications. Reads were first mapped to a reference genome with STAR/v2.5.2b (179), with gene annotations from GENCODE V19 for humans, and GENCODE VM11 for mouse. For experiments with HEK293T and NIH/3T3 cells, we used an index combining chromosomes from both human (hg19) and mouse (mm10). For the A549 experiment, we used human genome build hg19.The single cell sam files were first converted into alignment tsv file using sam2tsv function in jvarkit(180). Next, for each single cell alignment file, mutations matching the background SNPs were filtered out. For background SNP reference of A549 cells, we downloaded the paired-end bulk RNA-seq data for A549 cells from ENCODE(181) (sampled name: ENCFF542FVG, ENCFF538ZTA, ENCFF214JEZ, ENCFF629LOL, ENCFF149CJD, ENCFF006WNO, ENCFF828WTU, ENCFF380VGD). Each paired-end fastq files were first adaptor-clipped using trim_galore/0.4.1(182) with default settings, aligned to human hg19 genome build with STAR/v2.5.2b(179). Unmapped and multiple mapped reads were removed by samtools/v1.3(183). Duplicated reads were filtered out by MarkDuplicates function in picard/1.105(184). De-duplicated reads from all samples were combined and sorted with

samtools/v1.3(183). Background SNPs were called by mpileup function in samtools/v1.3(183) and mpileup2snp function in VarScan/2.3.9(185). For HEK293T and NIH/3T3 test experiment, a background SNP reference was generated in a similar pipeline above, with the aggregated single cell sam data from control condition (no 4sU labeling and no IAA treatment condition). The dimensionality of the data was first reduced with PCA (after selecting the top 2,000 genes with highest variance) on digital gene expression matrix on either full gene expression data or the newly synthesized gene expression data by Monocle 3/alpha (186,187).

Dimensionality reduction and clustering analysis.

For dimensionality reduction on single cell transcriptomes, the top 5 PCs for full transcriptomes and top 5 PCs for newly synthesized transcriptomes were selected for each state, and combined in temporal order along single cell state trajectory for UMAP analysis. Main cell trajectory types were identified by density peak clustering algorithm(188). With cell state proportion at the beginning time point (0 hour treatment) and cell state transition probabilities estimated from the data, we first predicted the cell state distribution after 2 hours, assuming the cell state transitions in DEX treatment are cell-autonomous, time-independent, Markovian processes. Similarly, the cell state distribution at later time points can be predicted from the cell state distribution 2 hours before.

4.3 Results

4.3.1 *Joint profiling of the total and newly synthesized transcriptome in cortisol response*

Briefly, sci-fate relies on the following steps (Figure 4.1, *a*): (i) Cells are incubated with 4-thiouridine (4sU), a thymidine analog, to label newly synthesized RNA(189–195). (ii) Cells are harvested, fixed with 4% paraformaldehyde, and then subjected to a thiol(SH)-linked alkylation reaction which covalently attaches a carboxyamidomethyl group to 4sU by nucleophilic substitution(131). (iii) Cells are distributed by dilution to 4 x 96 well plates. The first sci-RNA-seq molecular index is introduced via *in situ* reverse transcription (RT) with a poly(T) primer bearing both a well-specific barcode and a degenerate unique molecular identifier (UMI). During first strand cDNA synthesis, modified 4sU templates guanine rather than adenine incorporations. (iv) Cells from all wells are pooled and then redistributed by fluorescence-activated cell sorting (FACS) to multiple 96-well plates. Cells are gated on DAPI (4',6-diamidino-2-phenylindole) to distinguish singlets from doublets. (v) Double-stranded cDNA is generated by RNA degradation and second-strand synthesis. After Tn5 transposition, cDNA is PCR amplified via primers recognizing the Tn5 adaptor on the 5' end and the RT primer on the 3' end. These primers also bear a well-specific barcode that introduces the second sci-RNA-seq molecular index. (vi) PCR amplicons are subjected to massively parallel DNA sequencing. As with other sci-methods(132,196–202), most cells pass through a unique combination of wells, such that their contents are marked by a unique combination of barcodes that can be used to group reads derived from the same cell.(vii) The subset of each cell's transcriptome corresponding to newly synthesized transcripts is distinguished by T→ C conversions in reads mapping to mRNAs.

For quality control, we first tested sci-fate with a mixture of HEK293T (human) and NIH/3T3 (mouse) cells under four conditions: with vs. without 4sU labeling (200 nM, 6 hrs), and with vs. without the thiol(SH)-linked alkylation reaction. Under all four conditions, transcriptomes from

human/mouse cells were overwhelmingly species-coherent (>99% purity for both human and mouse cells, 2.7% collisions) with similar mRNA recovery rates (overall median 21,342 unique molecular identifiers (UMIs) per cell). However, only with 4sU labeling and the thiol(SH)-linked alkylation reaction did we observe a substantial proportion of reads bearing one or more T → C conversions, *i.e.* newly synthesized transcripts (46% for human and 31% for mouse cells, as compared with 0.8% for both species in untreated cells). The aggregated transcriptomes of cells derived from sci-fate and conventional sci-RNA-seq on the same cell types were highly correlated (Spearman's correlation r = 0.99), suggesting that the short term labeling and conversion process do not substantially bias transcript counts.

Cortisol influences the activity of almost every cell in the body, regulating genes involved in diverse processes including development, metabolism and immune response(203). To investigate the dynamics of cortisol response, we applied sci-fate to an *in vitro* model wherein dexamethasone (DEX), a synthetic mimic of cortisol, activates glucocorticoid receptor (GR), which binds to thousands of locations across the genome and significantly alters cell state within a rapid timeframe(204–207). Specifically, we treated lung adenocarcinoma-derived A549 cells for 0, 2, 4, 6, 8 or 10 hrs with 100 nM DEX. In each condition, cells were incubated with 4sU (200 nM) for the two hours immediately preceding harvest. We then performed a 384 x 192 sci-fate experiment (Figure 4.1, *b*). Each of the six conditions was represented by 64 wells during the first round of indexing, such that all samples could be processed in a single sci-RNA-seq experiment to minimize batch effects.

After filtering out low quality cells, potential doublets and a small subgroup of differentiated cells, we obtained single cell profiles for 6,680 cells (median of 26,176 UMIs corresponding to mRNAs detected per cell). A median of 20% of mRNA UMIs were labeled per cell (Figure 4.1, *c*). The proportion of newly synthesized mRNAs was markedly higher in reads mapping to intronic (65%) vs. exonic (13%) regions (p-value < 2.2e-16, two-sided Wilcoxon signed-rank test; Figure 4.1, *d*), consistent with the expectation that the intronic reads are more likely to have been recently synthesized. We also compared intronic reads and newly synthesised mRNA for RNA velocity analysis (208) and observed a subjectively consistent picture between the two methods, suggesting they convey similar information in our experiment.

In exploring these data, we first asked whether the newly synthesized vs. whole transcriptome data convey identical or distinct information with respect to cell state. For each condition, we generated pseudobulk transcriptomes for either the newly synthesized or whole transcriptomes (*i.e.* aggregating across cells), and compared these in a pairwise fashion between conditions (*e.g.* whole transcriptome at 0 hrs vs. 4 hrs; newly synthesized transcriptome at 2 hrs vs. 6 hrs, etc.). The lowest correlations corresponded to the newly synthesized transcriptome with no DEX treatment (0 hrs) vs. the newly synthesized transcriptomes of any DEX treated condition. Consistent with this, performing dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP)(209) on whole transcriptomes failed to separate DEX untreated (0 hrs) vs. treated (2+ hrs) cells (Figure 4.1, *e*). In contrast, applying UMAP to the newly synthesized subset of the single cell transcriptomes readily separated DEX untreated vs. treated cells (Figure 4.1, *e*). These patterns are likely a consequence of the fact that in DEX treated cells, the newly synthesized transcriptome more faithfully reflects the GR response itself. Illustrative of this, the classic markers for GR

response, *FGD4(204)* and *FKBP5(210)*, exhibited the highest fold induction in comparisons of the newly synthesized transcriptome at 0 hrs vs. 2 hrs, but the magnitude of their induction was dampened in comparisons of the whole transcriptome between the same time points.

To jointly make use of the information conveyed by the whole and newly synthesized transcriptomes, we combined their top principal components (PCs) for UMAP analysis. This approach separates cells that had experienced no DEX treatment (0 hrs), recent treatment (2 hrs) or extended treatment (4+ hrs) (Figure 4.1, *e*). Interestingly, with this joint approach, the cells corresponding to two clusters defined by analysis of whole transcriptomes (clusters 1 & 4; Figure 3.1, *f*) each split into two groups (Figure 4.1, *f*). By examining the levels of newly synthesized mRNAs corresponding to cell cycle markers(211), we found that one pair of these new groups corresponds to cells in G2/M phase (high levels of both overall and newly synthesized G2/M markers), and the other to early G0/G1 phase cells (high levels of overall but low levels of newly synthesized G2/M markers) (Figure 3.1, *g*). Of note, cells from the 2 hr time point exhibited a distribution of cell cycle states according to this joint information (Figure 4.1, *g*). Overall, these analyses illustrate how joint analysis of the newly synthesized and whole components of single cell transcriptomes can recover cell state information that is not easily obtained from the whole transcriptomes alone.

4.3.2 *TF module activity decomposes GR response, cell cycle, and other cellular processes*

Multiple dynamic gene regulatory processes are concurrently underway in this *in vitro* GR response system -- minimally, the GR response and the cell cycle. We speculated that these might

be disentangled, and their intersection probed, by first identifying the transcription factor (TF) modules driving new mRNA synthesis in relation to each such process.

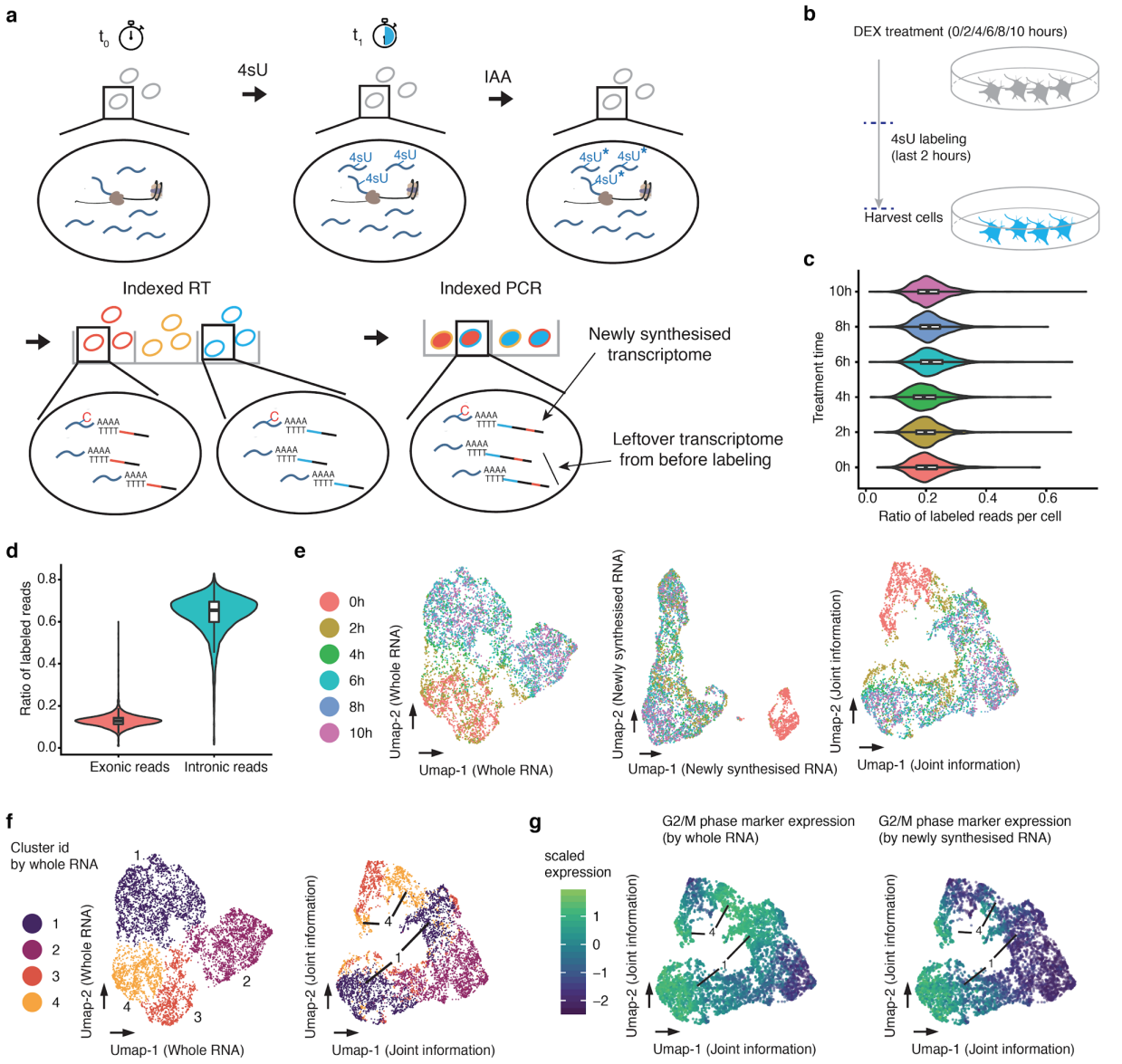


Figure 4.1. Sci-fate enables joint profiling of whole and newly synthesized transcriptomes. (**a**) The sci-fate workflow. Key steps are outlined in text. (**b**) Experimental scheme. A549 cells were treated with dexamethasone for varying amounts of time ranging from 0 to 10 hrs. Cells from all treatment conditions were labeled with 4sU two hours before harvest for sci-fate. (**c**) Violin

plot showing the fraction of 4sU labeled reads per cell for each of the six treatment conditions. For all violin plots in this figure: thick lines in the middle, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers. (**d**) Violin plot showing the fraction of 4sU labeled reads per cell, split out by the subsets that map to exons vs. introns. (**e**) UMAP visualization of A549 cells (n = 6,680) based on their whole transcriptomes (left), newly synthesized transcriptomes (middle) or with joint analysis, *i.e.* combining the top PCs from each (right). (**f**) Same as left and right of panel **e**, respectively, but colored by cluster id from UMAP based on whole transcriptomes. (**g**) Same as right of panel **e**, but colored by normalized expression of G2/M marker genes by their overall expression levels (left) or their levels of newly synthesized transcripts (right). UMI counts for these genes are scaled by library size, log-transformed, aggregated and then mapped to Z-scores.

Towards identifying such modules, candidate links between TFs and their regulated genes were identified as follows. For each gene, across the 6,680 cells, we computed correlations between the levels of newly synthesized mRNA for that gene and the overall expression level of each of 859 expressed transcription factors, using LASSO (least absolute shrinkage and selection operator) regression. Out of 1,086 links with TFs characterized by ENCODE(212), 807 were validated by TF binding sites near the genes' promoters(212), a 4.3-fold enrichment relative to background expectation (odds ratio for validation = 2.89 for links identified in LASSO regression vs. 0.67 for background, p-value < 2.2e-16, two-sided Fisher's Exact test). These covariance links were further filtered by ChIP-seq binding(181) and motif(213) enrichment analysis (Figure 4.2, *a*). In total, we identified 986 links between 29 TFs and 532 genes. As a control, we permuted the cell order of the cell x TF expression matrix used for this approach, and then performed the same analyses. No

links were identified after this permutation. Some of the identified TF and gene regulatory relationships are readily validated in a manually curated database of TF networks (TRRUST(214)), such as E2F1 (top enriched TF of E2F1 linked genes = E2F1, adjusted p-value = 8e-7)(215), NFE2L2 (top enriched TF of NFE2L2 linked genes = NFE2L2, adjusted p-value = 0.003)(215), and SREBF2 (top enriched TF of SREBF2 linked genes: SREBF2, adjusted p-value = 0.0006)(215).

The 29 TFs with one or more gene links included well-established GR response effectors such as *CEBPB(216)*, *FOXO1(217)*, and *JUNB(218)*. This group also included several TFs that have not previously been implicated in GR response, including *YOD1* and *GTF2IRD1*, both of which exhibited greater expression and activity in DEX treated cells. The main TFs driving cell cycle progression were also identified, including *E2F1*, *E2F2*, *E2F7*, *BRCA1*, and *MYBL2(219)*. Notably, the expression levels of TFs such as *E2F1* were more highly correlated with the levels of newly synthesized target gene mRNA than the overall levels of target gene mRNA. We also observed regulatory links corresponding to TFs involved in cell differentiation such as *GATA3(220)*, mostly expressed in a subset of quiescent cells, as well as TFs involved in oxidative stress response such as *NRF1(221)* and *NFE2L2* (*NRF2*)(222).

We calculated a measure of each of these 29 TFs' activities in each cell, based on the normalized aggregation of the levels of newly synthesized mRNA for all of its target genes. We then computed the absolute correlation coefficient between each possible pair of TFs with respect to their activity across the 6,680 cells. Hierarchical clustering of these pairwise correlations resulted in the identification of several major TF modules, *i.e.* sets of TFs that appear to be regulating the same

process (Figure 4.2, *b*). A first large TF module corresponds to all cell cycle-related TFs in the set, *e.g. E2F1* and *FOXM1(219)*. A second large TF module corresponds to GR response-related TFs including *FOXO1*, *CEBPB*, *JUNB* and *RARB(216–218)*. The other modules include one corresponding to GR-activated G1/G2/M phase cells (*KLF6*, *TEAD1*, and *YOD1*), and another corresponding to likely-differentiating GR-activated G1 phase cells *GATA3* and *AR(220,223)*. Additional TFs or TF modules appear to capture other processes that are heterogeneous in this population of cells, including *NRF1* and *NFE2L2* for stress response/apoptosis (top enriched pathway of *NFE2L2* linked genes: ferroptosis, adjusted p-value = 1e-5)(215,221,222,224); *KLF5* for DNA damage repair (top pathway: ATM signaling, adjusted p-value = 0.018)(215,225); and *SREBF2* for cholesterol homeostasis (top pathway: "SREBF and miR33 in cholesterol and lipid homeostasis", adjusted p-value = 9e-6)(215,226).

To assign cell cycle states to individual cells, we first ordered cells by their cell cycle-linked TF module activity. This resulted in a smooth, nearly circular trajectory, in which the levels of newly synthesized mRNA corresponding to known cell cycle markers was dynamic (Figure 4.2, *c*)(211). We observed a gap between late G2/M phase and early G1 phase, consistent with the dramatic cell state change during cell division. By unsupervised clustering of the activities of individual TFs within the cell cycle-linked TF module, we identified 9 cell cycle states spanning the early, middle and late cell cycle phases (Figure 4.2, *d*). Early G1 and late G2/M phase cells exhibited decreased synthesis of new RNA relative to other parts of the cell cycle, possibly due to chromosomal condensation during mitosis (Figure 4.2, *e*)(227–229). Other (*i.e.* non-cell-cycle) TF modules exhibited different dynamics in relation to cell cycle progression (Figure 4.2, *f*). For example, *GATA3* activity peaks in the early G1 phase, potentially reflecting a cell differentiation pathway

distinct from cell cycle reentry(220). In contrast, the modules of *KLF5* and *SREBF2*, associated with DNA repair and lipid homeostasis, respectively, exhibited greater activity from S to G2 phase, possibly related to roles in DNA replication and cell division, respectively(230).

With similar approaches, the cells can also be ordered into a smooth trajectory based on GR response-linked TF module activity. As expected, this trajectory correlates well with DEX treatment time, as well as the activity of GR response-related TFs (Figure 4.2, *g*). By unsupervised clustering of the activities of individual TFs within the GR response-linked TF module, we identified GR response states corresponding to no, low and high levels of activation (Figure 4.2, *g*).

We next sought to explore the intersection of the 9 cell cycle states (Figure 3.2, *d*) and the three GR response states (Figure 4.2, *g*). Each of their 27 possible state combinations were represented by some cells, with the smallest group corresponding to 1.1% of the overall dataset (n = 74 cells, intersection of "early G2/M" cell cycle state and "no GR activation" state). Although we observe several TF modules that appear specific to certain intersections of the cell cycle and GR response (*KLF6/TEAD1/YOD1* and *GATA3/AR*, discussed above), several observations support the conclusion that the dynamics of the cell cycle and GR response operate largely independently. First, we observe minimal correlation between the activities of the primary TF modules for cell cycle and GR response across the 6,680 cells (Pearson's correlation r = 0.004; Figure 4.2, *b*). Second, the relative proportions of each of the 27 possible state combinations are readily predicted by proportions of cell cycle and GR response states, *i.e.* with no interaction term.

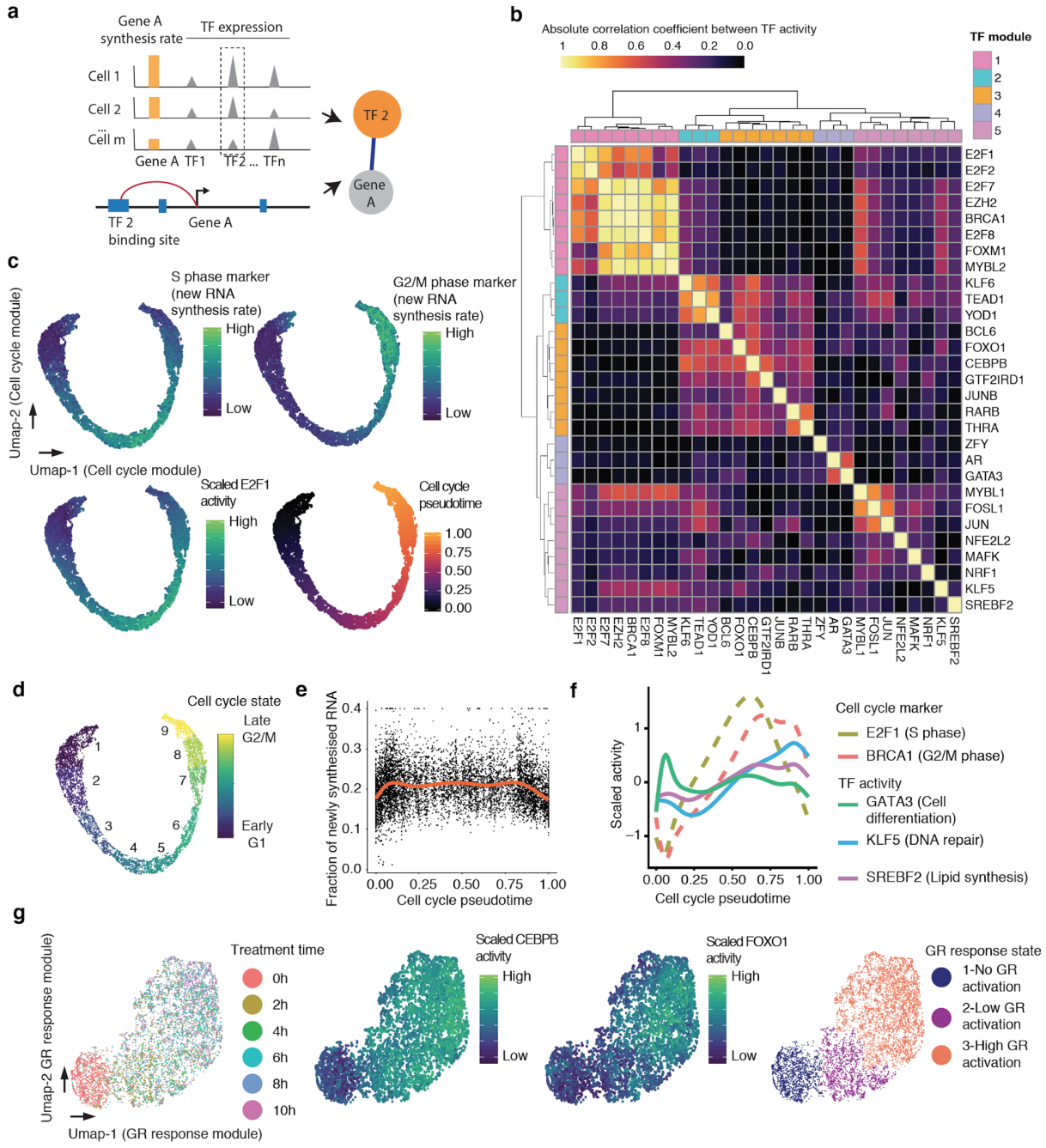


Figure 4.2. Characterizing TF modules driving concurrent, dynamic gene regulatory processes in populations of single cells. (**a**) Schematic of approach used to identify links between TFs and their regulated genes. (**b**) Heatmap showing the absolute Pearson's correlation coefficient between the activities of pairs of TFs. (**c**) UMAP visualization of A549 cells (n = 6,680) based on the activity of cell cycle-related TF module, colored by levels of newly synthesized mRNA

corresponding to S phase markers (top left), G2/M phase markers (top right), and *E2F1* activity (bottom left). Bottom right panel is colored by pseudotime based on point position on the principal curve estimated by princurve package(231). (**d**) Same as panel **c**, but colored according to nine cell cycle states defined by unsupervised clustering analysis. In broad terms, cell cycle states 1-3 correspond to G1 phase, 4-6 to S phase, and 7-9 to G2/M phase. (**e**) Scatter plot showing the changes in the fraction of newly synthesized mRNA in each cell (n = 6,680) along cell cycle progression. The red line is the smoothed curve estimated by the geom_smooth function(232). (**f**) Similar to panel **e**, but showing smoothed activity of selected TF modules as a function of cell cycle pseudotime. (**g**) UMAP visualization of A549 cells (n = 6,680) based on the activity of GR response-related TF module, colored by DEX treatment time (left), *CEBPB* or *FOXO1* activity (middle panels), or cluster id from unsupervised clustering (right). Throughout the figure, to calculate TF module activity, newly synthesized UMI counts for genes linked to module-assigned TFs are scaled by library size, log-transformed, aggregated and then mapped to Z-scores.

4.3.3 *Inferring single cell transcriptional dynamics with sci-fate*

We next sought to develop a strategy to use sci-fate data to infer the *past* transcriptional state of each cell, *i.e.* at the onset of 4sU labeling, which might in turn allow us to relate cells derived from different time points. The inference of this past transcriptional state requires knowledge of two parameters -- first, the detection rate of newly synthesized transcripts (*i.e.* the proportion of newly synthesised transcripts containing one or more detected T > C mutations), and second, the degradation rate of each mRNA species. Below, we discuss how each of these parameters can be estimated directly from the sci-fate data generated for this experiment.

Under the assumption that mRNA degradation rates are not affected by DEX treatment (this assumption is validated further below), it is relatively straightforward to estimate sci-fate's detection rate for newly synthesized transcripts. Each sci-fate transcriptome in this dataset consists of two components -- the newly synthesized transcriptome, whose detection rate we are hoping to estimate, and the 'leftover' transcriptome, *i.e.* transcripts that were present at the onset of 4sU labeling, minus any degradation over the course of the two hours. Comparing the 0 hr (untreated) and 2 hr DEX treatment groups, we expect that their leftover bulk transcriptomes (at the onset of 2 hour 4sU labeling) should be identical, as should sci-fate's detection rate for newly synthesized transcripts. As such, an equation can be constructed relating the transcriptomes of these treatment groups to one another. For each of 186 genes exhibiting the largest differences in new transcription between the two conditions, we solved this equation to estimate sci-fate's detection rate. As these estimates were largely consistent across genes and robust to sequencing depth, we used their median value (82%) as sci-fate's estimated detection rate for all subsequent analyses.

We next sought to estimate the degradation rate of each mRNA species. As noted above, the bulk transcriptome at each time point in our experiment can be decomposed into the newly synthesized transcriptome and the leftover transcriptome. Furthermore, the leftover transcriptome should equal the bulk transcriptome from the time point two hours earlier, provided that we correct for mRNA degradation over that interval. From these assumptions, an equation can be constructed and solved to estimate the mRNA half-life of each gene, which we did independently for each two hour interval of the experiment. As a first quality check, we simply compared these estimated mRNA degradation rates between time points, and found them to be both consistent and robust to

sequencing depth (median Pearson's $r = 0.92$). As a second quality check, we compared them to orthogonally generated estimates of mRNA half-lives from the literature(177). Despite the fact that different technologies were used on different cell lines (A549 vs. K562), the estimates of mRNA half-lives were reasonably consistent (Pearson's $r = 0.76$). Of note, the absolute differences in estimated mRNA half-lives between sci-fate and previous techniques could be due to the use of different cell lines or systematic differences between the techniques.

With these parameters in hand, we next estimated the *past* transcriptional state of each cell in our dataset, and sought to use these estimated states to link individual cells to one another across time points (Figure 4.3, *a*). Specifically, for each cell B (*e.g.* a cell from the 2 hr time point), we used a recently developed alignment method (211) to identify a cell A profiled at an earlier time point (*e.g.* a cell from the 0 hr time point), wherein A's current state was closest to B's estimated past state. In this framework, A can be regarded as the parent state of B. Applying this strategy to each of the five intervals comprising our experiment, we constructed a set of linkages spanning the entire dataset and time course (Figure 4.3, *b*).

A key contrast with conventional pseudotime is that with sci-fate, each cell is now characterized not only by its present state, but also by specific linkages to a series of distinct cells matching its predicted past and/or future states (Figure 4.3, *c*). To evaluate whether these mini-trajectories contain structure, we applied UMAP and unsupervised clustering, which resulted in three distinct trajectory clusters (Figure 4.3, *d*). To annotate these, we checked the proportions of each of the aforementioned three GR response states and nine cell cycle states in each of them, as a function of time. As expected, all three trajectories exhibited a rapid transition from no GR activation to

low/high GR activation (Figure 4.3, *e*). However, each trajectory appears to correspond to a different starting point with respect to the cell cycle (Figure 4.3, *f*). Trajectory 1 corresponds to cells that transition from G2/M to G1 phase over the course of the 10 hr experiment. Trajectory 2 corresponds to cells that transition from late S phase to G2/M phase over the course of the experiment. Finally, trajectory 3 corresponds to cells that transition from G1 to either S phase or G1 arrest over the course of the experiment. The inference of G1 arrest subsequent to DEX treatment is consistent with the dynamics of cell state proportions in this experiment as well as with previous research (233,234). As a control, we clustered the cell state transition trajectories by simply aligning neighboring time points without knowledge of newly synthesized mRNA; this failed to recover expected cell cycle dynamics.

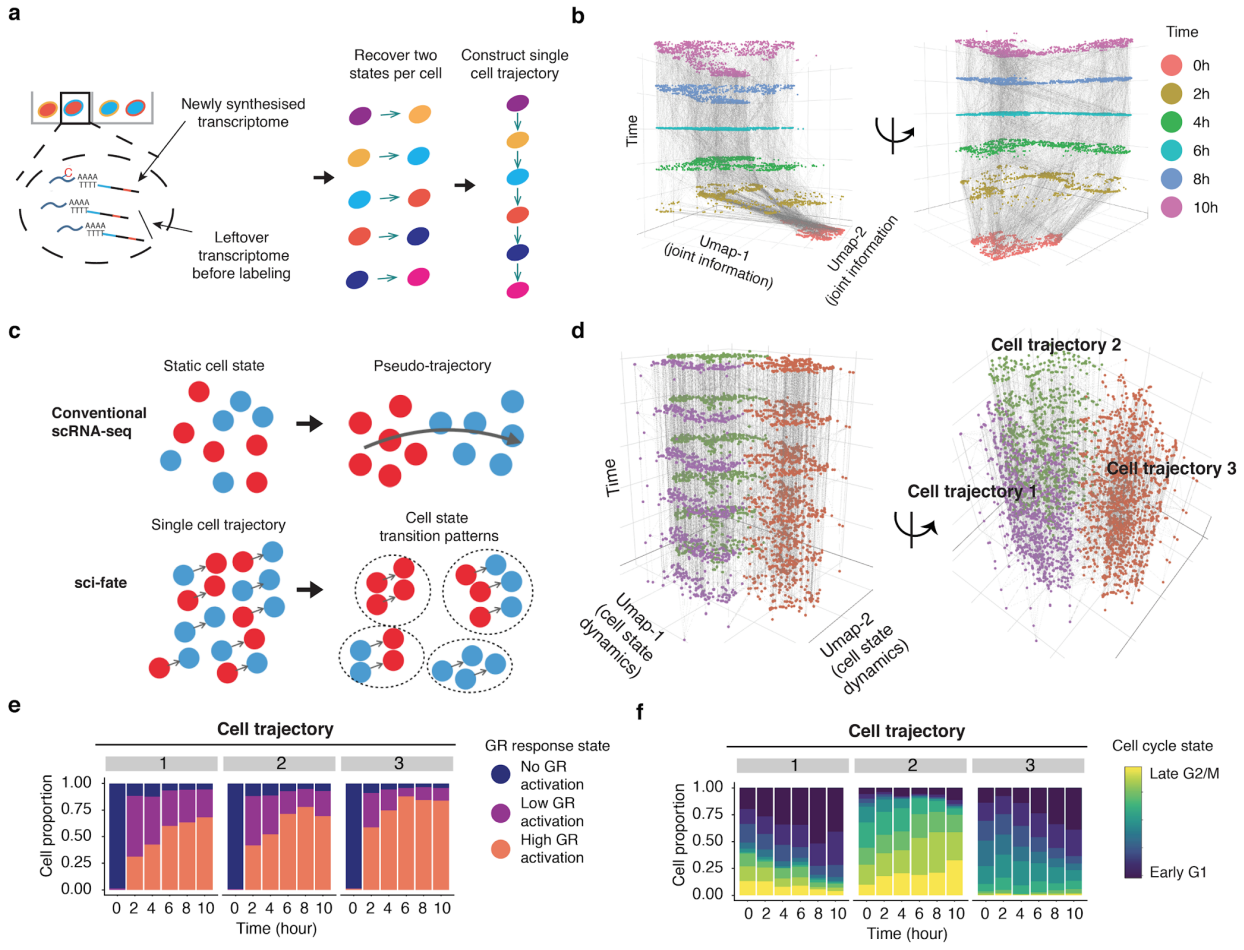


Figure 4.3. Inferring single cell transcriptional dynamics with sci-fate. (**a**) Schematic of approach for linking cells based on estimated past transcriptional states to reconstruct single cell transition trajectories. (**b**) 3D plot of all cells (cell number n = 6,680). The x and y coordinates correspond to the joint information UMAP space shown in the rightmost panel of **Fig. 4.1e**. The z coordinate as well as colors correspond to DEX treatment time. Linked parent and child cells are connected with grey lines. (**c**) Schematic comparing conventional scRNA-seq and sci-fate for cell trajectory analysis. (**d**) Similar to panel **b**, except the x and y coordinates correspond to the UMAP space based on the single cell transition trajectories across the six time points (cell number n = 6,680). (**e-f**) Barplots showing the contributions of the 3 GR response states (**e**) and the 9 different cell cycle states (**f**) to each of three cell trajectory clusters.

4.3.4 *Inferred single cell state transition links recapitulate expected dynamics*

We next sought to evaluate whether the distribution of cell state transitions inferred by sci-fate are consistent with the expected dynamics. We assigned each cell into one of the 27 states (3 GR response x 9 cell cycle states) and computed a cell state transition network (Figure 4.4, *a*), with the assumption that the cell state transitions in this experiment follow a Markov process with transition probabilities that do not change over time. This assumption is validated in part by the observation that the distribution of predicted cell state transitions estimated from the last three time intervals (4 hrs to 6 hrs, 6 hrs to 8 hrs, 8 hrs to 10 hrs) are highly correlated with each other despite of varied cell state proportions at 4 hrs vs. the later time-points. Consistent with DEX treatment, transitions are highly biased from G1 to S, S to G2/M, and G2/M to G1 phase of the cell cycle (Figure 4.4, *a*). As a control analysis, cell state transition networks were similarly derived, but based either on randomly permuted cell state transition links, or on links derived from mature mRNAs only; these both failed to recapitulate the expected pattern of cell cycle transitions.

The 27 states shown in Figure 4.4*a* each correspond to subsets of cells whose transcriptomes are similar, making use of the joint information provided by distinguishing between old (> 2 hrs) vs. new (< 2 hrs) transcripts. Importantly, the distribution of transitions are inferred, rather than explicitly known, but supported by the fact that they correspond to expected phenomena, *e.g.* irreversible progression through GR response, as well as irreversible progression through the cell cycle. As an example, S phase cells without GR activation (0 hrs treatment) mostly transit into cell state in S phase with GR activation (2 hrs treatment), while G2/M phase cells with no GR activation (0 hrs treatment) mostly transit to G2/M or G1 phase cells with GR activation (2 hrs treatment) (Figure 4.4, *b*). For comparison, overlaying the same UMAP coordinates with the RNA velocity vectors(208), which infers transcriptional dynamics in single cell data from the proportion of intronic vs. exonic reads, recovered similar patterns, but only when treatment time information was incorporated into the RNA velocity analysis.

Can we use this framework to better understand the characteristics of transcriptional states that govern their dynamics? As a first approach, we calculated the pairwise Pearson's distance between the aggregated transcriptomes of each of the 27 states. As expected, the greater the distance between any pair of states, the lower the proportional representation of that transition in the network (Spearman's correlation coefficient = -0.38; Figure 4.4, *c*). As a second approach, we computed "instability" as the proportion of cells inferred to be moving out of a given state between time points (Figure 4.4, *d*). As expected, states corresponding to no GR activation were the least stable by this metric. Furthermore, amongst high GR activation states, states corresponding to early G1 were the most stable. These representations of the data are consistent with the transition

network, wherein the states corresponding to high GR activation and early G1 are a frequent "destination" of all nearby states (purple triangles in Figure 4.4, *a*).

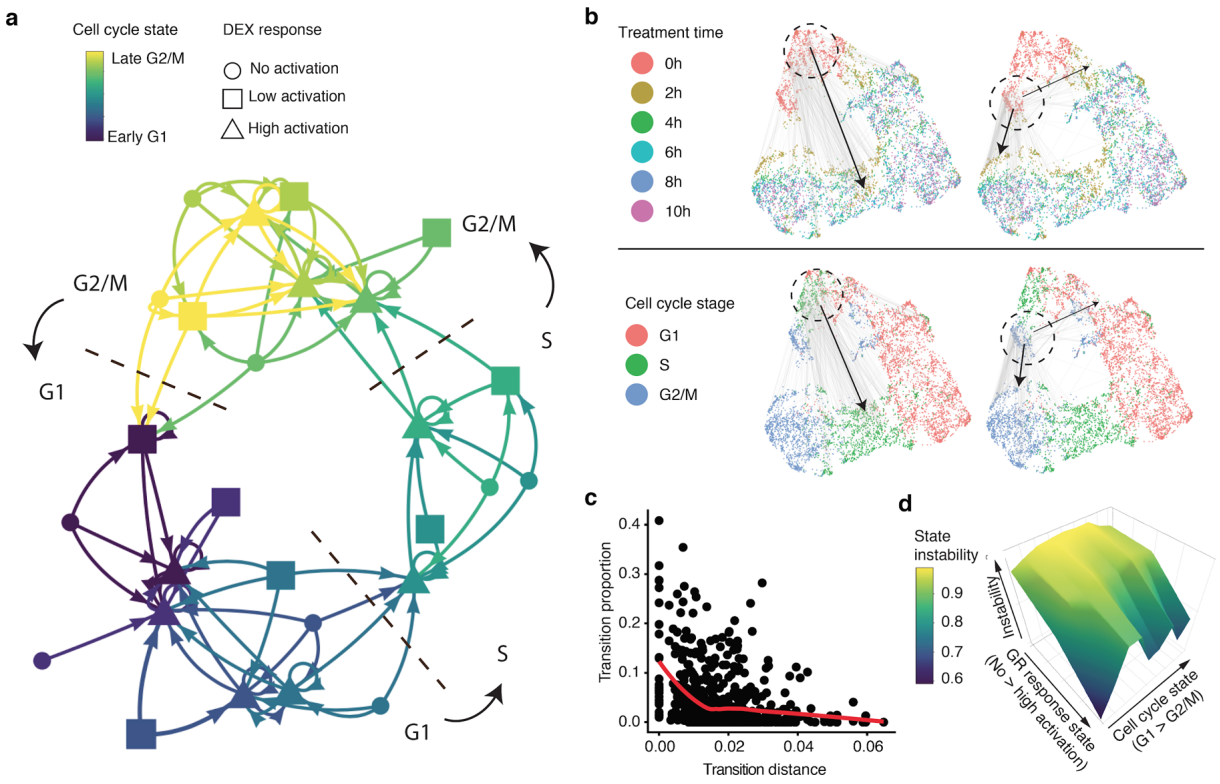


Figure 4.4. Constructing a state transition network for GR response and cell cycle. (**a**) Cell state transition network. The nodes are 27 cell states characterized by combinations of cell cycle and GR activation states. The links represent frequent cell state transition trajectories (transition proportion > 10%) between cell states. This threshold for defining a link corresponds to approximately two standard deviations from the mean transition proportion calculated after permuting cell transition links. (**b**) The x and y coordinates correspond to the joint information UMAP space shown in the rightmost panel of **Fig. 1e**, colored by DEX treatment time (top) or inferred cell cycle state (bottom). Grey lines represent inferred cell state transition links between parent and child cells (left: cell state transition links starting from cells at the S phase and no GR activation stage; right: cell state transition links starting from cells at G2/M phase and no GR activation stage). Black arrows show main cell state transition directions. (**c**) Scatter plot showing

the relationship between transition distance (Pearson's distance) and transition proportion (n = 729), together with the red LOESS smoothed line by ggplot2(232). (**d**) 3D plot showing the cell state stability landscape. X-axis represents GR response states (from no to low to high activation state). Y-axis represents the cell cycle states ordered from G1 to G2/M. Z-axis represents cell state instability, defined as the proportion of cells inferred to be moving out of a given state between time points.

4.4 Discussion

Experimental methods that recover not only the current state of any given cell, but also its velocity vector, are distinct and potentially more powerful than computational methods for inferring such vectors(235), *e.g.* pseudotime. To that end, we developed sci-fate, a novel method for concurrently profiling the whole and newly synthesized transcriptome in each of many single cells. In applying sci-fate to a model system of cortisol response, we found that the joint analysis of whole and newly synthesized single cell transcriptomes enabled greater discrimination of cell states than was possible with whole transcriptomes alone. Most notably, it became straightforward to distinguish between the dynamic transcriptional modules underlying GR activation vs. progression through the cell cycle. By analyzing covariance between TF expression and *new* RNA synthesis across many single cells, we identified nearly one thousand regulatory links between 27 TFs and 532 target genes. These separated into several modules, including the GR response, cell cycle and others, reflecting cellular processes that were heterogeneous across this population of cells and that appeared to operate largely independently of one another. We were also able to infer the past state of each single cell in the experiment, as well as to use links between cells based on these inferences to construct a cell state transition network. Thus, sci-fate could in principle help

overcome limitations of conventional single-cell RNA-seq when inferring causal regulatory networks and may facilitate the development of computational methods serving this end (236).

Sci-fate captures information that is analogous to RNA velocity(208), which distinguishes 'older' vs. 'newer' transcripts based on their splicing status. On one hand, RNA velocity is more straightforward than sci-fate, as it makes use of information that is indirectly captured by many single cell profiling technologies, whereas sci-fate requires 4sU labeling steps that cannot necessarily be used in all contexts. On the other hand, sci-fate lends itself to experimental control in a way that RNA velocity does not, as the timing and length of 4sU labeling can be specified whereas with RNA velocity it is a product of endogenous splicing dynamics. Furthermore, as we show, an experimental design that couples the labeling of newly synthesized mRNA to a time series (*i.e.* wherein the labeling and sampling intervals are matched), enables the quantitative analysis of cells with complex transcriptional histories and futures.

While our manuscript was under review, two methods directed at the same goal, scSLAM-seq and NASC-seq, were reported(237,238). Although there are broad similarities, including the labeling strategy, we note major differences with respect to performance, accuracy and scalability: (1) Because sci-fate uses combinatorial indexing, we were able to measure newly synthesized mRNA in over 6,000 cells in a single experiment, compared with less than 200 cells for scSLAM-seq or NASC-seq. Given that sci-fate is easily adaptable to three levels of combinatorial indexing(178), it should already be possible to profile newly synthesized mRNA is >1 million cells per sci-fate experiment. (2) sci-fate costs less than $0.20 per cell for library preparation with two-level indexing, and less than $0.01 per cell with three-level indexing(178). By comparison, both

scSLAM-seq and NASC-seq utilize smart-seq which costs about $11 per cell for library preparation(239). On a related point, sci-fate required an order of magnitude fewer raw reads per cell (~200,000 per cell with sci-fate vs. ~2 million reads per cell with scSLAM-seq), but achieved a greater number of genes detected per cell (discussed further below). (3) A key feature of sci-fate is that we performed *in situ* 4sU chemical conversion in bulk, fixed cells, resulting in a high reaction efficiency and low mRNA loss. In contrast, scSLAM-seq and NASC-seq require extracting mRNA from each cell followed by bead-based purification and chemical conversion, which results in low reaction efficiency and a high rate of loss. As a result, sci-fate exhibits higher efficiency to detect low abundance transcripts (median 6,500 genes detected per cell with sci-fate vs. ~4,000 with scSLAM-seq, despite of 1/10 of the raw sequencing depth). Furthermore, sci-fate exhibits a higher detection rate of newly synthesised mRNA (82% in sci-fate vs. <50% in scSLAM-seq). (4) The signal-to-noise ratio (labeled vs. unlabeled cells) of sci-fate is 38- to 58-fold , compared with only ~10-fold for scSLAM-seq or NASC-seq. This is partly due to the fact that the sci-fate library preparation is strand-specific, whereas smart-seq is not. (5) sci-fate enables direct counting of newly synthesised vs. pre-existing mRNA via 3' tagged unique molecular identifiers (UMIs). This is essential for removing PCR duplicates and for accurate modeling of gene expression(240). In contrast, scSLAM-seq and NASC-seq do not have UMIs incorporated in their design, which undoubtedly introduces additional noise from duplicates given the large number of PCR amplification cycles used when amplifying material derived from a single cell. (6) Additional major advantages of sci-fate over scSLAM-seq and NASC-seq include compatibility with fixed cells and the ability to concurrently process multiple independent biological samples within a single experiment. On a related point, it is notable that *in situ* 4sU chemical conversion requires cell permeabilization and, at least in our experience, PFA fixation, neither of which is

straightforward to introduce on droplet-based scRNA-seq platforms such as 10x Genomics. It is also potentially compatible with concurrent profiling of the epigenomes from the same cells(197), although ideally, we would be able to profile not only nascent transcription but also nascent epigenetic events at the single cell level.

We note that while sci-fate enables quantification of mRNA synthesis in single cells, we are still in need of methods for measuring mRNA degradation rates in single cells. Related to that, our simplifying assumption that gene-specific degradation rates are constant across our DEX time course might not be a good choice in other systems. Specifically, in systems where the gene-specific degradation rates are expected or observed to substantially vary over time, these should be estimated for each time interval separately.

Sci-fate can be broadly applied to most *in vitro* systems to quantitatively characterize cell state dynamics within short time windows (*e.g.* one to several hours). For even shorter time frames, a concern is that signal-to-noise will drop as the rate of labeling falls towards the background rate of 0.8%. For longer time frames, a time series approach can be adopted as in the main experiment described here.

A major limitation of sci-fate is that 4sU labeling experiments are generally performed within *in vitro* cell culture models. However, recent studies have shown that 4sU can be used in conjunction with transgenic *UPRT*-expressing mice to stably label cell type-specific nascent RNA transcription *in vivo(241–243)*, suggesting that sci-fate, with further optimizations to enhance 4sU incorporation

and detection rate, can potentially be used to profile single cell transcriptional dynamics *in vivo* and at scale.

Chapter 5. The future of high-throughput methods for profiling protein-DNA and protein-RNA interactions.

Many high-throughput methods have been developed to profile the interactions between proteins and nucleic acids to explore critical biological questions, including the binding profiles of

transcription factors or RNA-binding proteins. In this dissertation, I describe my work to identify the binding sites of variants of the yeast transcription factor Ste12 in a genome-wide manner to understand how variation in a transcription factor affects its *in vivo* genome binding, the gene expression program it drives, as well as the ultimate phenotype of the cell. I found subtle changes in the coding region of the transcription factor can result in large regulatory rewiring; however, the major determinants of organismal phenotype are the changes in the expression of a small, related set of genes. I also describe another project in which I combined a yeast-three hybrid system with next-generation sequencing to score the binding activity of each of a large number of PUF variants to an RNA sequence containing each possible RNA base at the cognate position recognized by a PUF repeat. From this work, I identified a complete code for RNA recognition by this domain. In the following section, I will describe some areas that I believe will be important for the field to consider over the coming years, as more high-throughput methods are continuously being developed to tackle these questions.

5.1 Is a complete DNA motif catalog feasible for transcription factors as well as their variants?

The transcription factor motif catalog has been greatly expanded with the recent development of high-throughput *in vitro* technologies to characterize DNA binding, including protein-binding microarrays (137), high-throughput *SELEX (12,137)* and bacterial one-hybrid systems (30). However, the DNA-binding specificity of several hundreds of transcription factors still remains uncharacterized, not to mention variants of these factors. The lack of a motif catalog for these transcription factor variants limits our ability to explore the effects of genetic variation on TF-

DNA binding behavior. The scenario is complicated even more by the fact that many TFs do not bind to the genome individually, but in cooperation with other DNA-binding proteins. In fact, studies have already shown a great discrepancy between the *in vivo* DNA binding profile and the *in vitro* derived binding motif (22). Therefore, these studies emphasize the need for new high-throughput *in vivo* technologies to close this gap.

Existing *in vivo* technologies, including *ChIP-seq (33)*, *DamID-seq (244)*, or *Calling Card-seq (42)*, are still low-throughput in terms of exploring the effects of genetic variation on TF-DNA binding behavior and achieving a complete catalog for each TF variant. However, it is highly possible that an optimized calling card approach could achieve this goal. In this optimized approach, a transposon would be inserted into the genome near to where the TF binds and a barcode within the transposon would not only label the "visit" of the TF to that location, but also serve as the identity for a given transcription factor variant. Therefore, such an approach may enable multiplexed identification of genomic targets for a large number of transcription factor variants at one time.

Another concern for achieving a complete DNA motif catalog is that the current catalog mostly relies on position-weight matrix (PWM) representation (136), as PWMs allow the scoring of a given DNA sequence based on its similarity to a motif. However, many recent studies have shown that the sequence environment may involve features that promote DNA binding, such as high GC content or a higher similarity to the core motif (245,246), which may be critical to predict the DNA binding events for a TF. In fact, TF motifs tend to occur in clusters, and these clustered sites may buffer genetic perturbations that may affect one site (247). To increase the robustness of motif

definition, further efforts should aim at incorporating more datasets at genomic scale and applying machine learning methods to predict a comprehensive catalog of TF binding sequences. One recent example are studies that apply machine learning methods to k-mer vocabularies trained on *ChIP-seq* datasets in order to predict binding preferences of TFs. In the future, more input features, including k-mer vocabulary, structure, chemical or epigenomic features, based on both *in vivo* and *in vitro* derived datasets, should be applied to train machine learning models in order to generate a complete catalog of TF-binding sequences.

5.2 How can RNA recognition and the manipulation of RNA-related processes be optimized?

Nature has provided simple rules for different proteins to recognize RNA sequences, which affect various processes of RNA biology. The work I present in Chapter 3 demonstrated how we can apply next-generation sequencing techniques to comprehensively explore the rules of RNA recognition. Advances in structural biology and biochemistry will further unravel these rules and allow the engineering of RNA-binding scaffolds.

For any method that specifically targets endogenous gene expression, one important concern is to control its off-target effects. To best recognize RNA, the first opportunity is in the design stage. For many RNA-binding platforms, such as the PUF or PPR scaffold, it is possible that a look-up table that includes comprehensive base-specific recognition codes can be achieved through high-throughput specificity screen experiments, similar to the work in chapter 4. As the genome sequences of many organisms have been characterized, it will be useful to search the transcriptome

and estimate the frequency of possible off-target sites. In addition, it is important to consider the effect of RNA structure, as many studies have shown that structures such as extensive hairpin can greatly decrease binding affinity (248). For example, we can select target sites in less structured regions, with off-target sites that are embedded in more structured regions. We can increase recognition specificity by using an RNA-binding scaffold that contains more repeats. In fact, recent studies have engineered a PUF domain with a capacity to recognize RNA sequences of more than eight bases (172,249). Lastly, we can control the expression of the engineered RNA-binding domain to occur at the same time and in the same space as its target RNA. For example, studies have successfully silenced mitochondrial RNA by engineering an RNA-binding protein to be expressed only in mitochondria, thereby eliminating off-target effects against nuclear/cytoplasmic RNA (250).

In the future, the manipulation of gene expression at the RNA level may become easily accessible by fusing functional domains with RNA-binding platforms of high specificity. Compared with manipulation at the DNA level, this manipulation will be non-permanent and reversible, and therefore may have good potential as a therapeutic reagent. For example, by fusing a PUF domain with Arg/Ser-rich domains of SRSF1, Wang *et al.* (251) developed novel splicing activators or inhibitors, which allow splicing to change from an antiapoptotic long isoform into a proapoptotic short isoform. This change sensitizes several cancer cell lines to anticancer drugs. In addition, Choudhury *et al. (250)* combine a general RNA cleavage domain with a PUF domain to achieve a function analogous to DNA restriction enzymes *in vivo*. Moreover, through combining a PUF domain with a translational activator or a repressor, the engineered domain can specifically recognize its RNA target and then activate or repress translation (80). Other manipulations,

including RNA editing (252), modification (253), translocation, as well as degradation (254), can be achieved through fusing an RNA-binding domain with a functional domain. Together, the same design principles can be applied by fusing various functional domains to RNA-binding platforms to achieve novel biological activities.

Overall, significant advances in understanding biological processes can be achieved through targeting and manipulation of gene expression at the RNA level. I believe RNA-binding platforms will be continuously optimized, in terms of their affinity, specificity, as well as their broad applications.

REFERENCES

1. Jacob F, Monod J. Genetic Regulatory Mechanisms in the Synthesis of Proteins [Internet]. Molecular Biology. 1989. p. 82–120. Available from: http://dx.doi.org/10.1016/b978-0-12-131200-8.50010-1

2. Consortium T 1000 GP, The 1000 Genomes Project Consortium. A global reference for human genetic variation [Internet]. Vol. 526, Nature. 2015. p. 68–74. Available from: http://dx.doi.org/10.1038/nature15393

3. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011 Sep 14;477(7364):289–94.

4. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012 Sep 7;337(6099):1190–5.

5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009 Jun 9;106(23):9362–7.

6. Maurano MT, Haugen E, Sandstrom R, Vierstra J, Shafer A, Kaul R, et al. Large-scale

identification of sequence variants influencing human transcription factor occupancy in vivo. Nat Genet. 2015 Dec;47(12):1393–401.

7. Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, et al. Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science. 2016 Mar 25;351(6280):1450–4.

8. Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 2007 May;61(5):995–1016.

9. Lynch VJ, Wagner GP. Resurrecting the role of transcription factor change in developmental evolution. Evolution. 2008 Sep;62(9):2131–54.

10. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007 Mar;8(3):206–16.

11. Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev Genet. 2010 Nov;11(11):751–60.

12. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 2010 Jun;20(6):861–73.

13. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, et al. Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins [Internet]. Vol. 147, Cell. 2011. p. 1270–82. Available from: http://dx.doi.org/10.1016/j.cell.2011.10.053

14. Coffey AJ, Kokocinski F, Calafato MS, Scott CE, Palta P, Drury E, et al. The GENCODE exome: sequencing the complete human exome [Internet]. Vol. 19, European Journal of Human Genetics. 2011. p. 827–31. Available from: http://dx.doi.org/10.1038/ejhg.2011.28

15. Veraksa A, Del Campo M, McGinnis W. Developmental patterning genes and their conserved functions: from model organisms to humans. Mol Genet Metab. 2000 Feb;69(2):85–100.

16. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities [Internet]. Vol. 24, Nature Biotechnology. 2006. p. 1429–35. Available from: http://dx.doi.org/10.1038/nbt1246

17. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet. 2004 Dec;36(12):1331–9.

18. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990 Aug 3;249(4968):505–10.

19. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding

specificities of human transcription factors. Cell. 2013 Jan 17;152(1-2):327–39.

20. Orenstein Y, Shamir R. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. Nucleic Acids Res. 2014 Apr;42(8):e63.

21. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010 Mar 5;140(5):744–52.

22. Biggin MD. Animal transcription networks as highly connected, quantitative continua. Dev Cell. 2011 Oct 18;21(4):611–26.

23. Fields S, Song O-K. A novel genetic system to detect protein–protein interactions [Internet]. Vol. 340, Nature. 1989. p. 245–6. Available from: http://dx.doi.org/10.1038/340245a0

24. Deplancke B. A Gateway-Compatible Yeast One-Hybrid System [Internet]. Vol. 14, Genome Research. 2004. p. 2093–101. Available from: http://dx.doi.org/10.1101/gr.2445504

25. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, Martinez NJ, et al. A gene-centered C. elegans protein-DNA interaction network. Cell. 2006 Jun 16;125(6):1193–205.

26. Dove SL, Joung JK, Hochschild A. Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature. 1997 Apr 10;386(6625):627–30.

27. Joung J, Koepp D, Hochschild A. Synergistic activation of transcription by bacteriophage lambda cI protein and E. coli cAMP receptor protein [Internet]. Vol. 265, Science. 1994. p. 1863–6. Available from: http://dx.doi.org/10.1126/science.8091212

28. Joung JK, Ramm EI, Pabo CO. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc Natl Acad Sci U S A. 2000 Jun 20;97(13):7382–7.

29. Bartel PL, Fields S. The Yeast Two-hybrid System. Oxford University Press, USA; 1997. 344 p.

30. Meng X, Brodsky MH, Wolfe SA. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol. 2005 Aug;23(8):988–94.

31. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell. 2008 Jun 27;133(7):1277–89.

32. Gupta A, Christensen RG, Bell HA, Goodwin M, Patel RY, Pandey M, et al. An improved predictive recognition model for Cys2-His2 zinc finger proteins [Internet]. Vol. 42, Nucleic Acids Research. 2014. p. 4800–12. Available from: http://dx.doi.org/10.1093/nar/gku132

33. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, et al. Genome-

wide analysis of transcription factor binding sites based on ChIP-Seq data [Internet]. Vol. 5, Nature Methods. 2008. p. 829–34. Available from: http://dx.doi.org/10.1038/nmeth.1246

34. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004 Sep 2;431(7004):99–104.

35. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007 Jun 8;316(5830):1497–502.

36. Ren B. Genome-Wide Location and Function of DNA Binding Proteins [Internet]. Vol. 290, Science. 2000. p. 2306–9. Available from: http://dx.doi.org/10.1126/science.290.5500.2306

37. van Steensel B, Henikoff S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase [Internet]. Vol. 18, Nature Biotechnology. 2000. p. 424–8. Available from: http://dx.doi.org/10.1038/74487

38. Pym ECG, Southall TD, Mee CJ, Brand AH, Baines RA. The homeobox transcription factor Even-skipped regulates acquisition of electrical properties in Drosophila neurons. Neural Dev. 2006 Nov 16;1:3.

39. Wolfram V, Southall TD, Gunay C, Prinz AA, Brand AH, Baines RA. The Transcription Factors Islet and Lim3 Combinatorially Regulate Ion Channel Gene Expression [Internet]. Vol. 34, Journal of Neuroscience. 2014. p. 2538–43. Available from: http://dx.doi.org/10.1523/jneurosci.4511-13.2014

40. Southall TD, Davidson CM, Miller C, Carr A, Brand AH. Dedifferentiation of Neurons Precedes Tumor Formation in lola Mutants [Internet]. Vol. 28, Developmental Cell. 2014. p. 685–96. Available from: http://dx.doi.org/10.1016/j.devcel.2014.01.030

41. Wang H, Heinz ME, Crosby SD, Johnston M, Mitra RD. “Calling Cards” method for high-throughput identification of targets of yeast DNA-binding proteins. Nat Protoc. 2008;3(10):1569–77.

42. Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. “Calling Cards” for DNA-Binding Proteins in Mammalian Cells [Internet]. Vol. 190, Genetics. 2012. p. 941–9. Available from: http://dx.doi.org/10.1534/genetics.111.137315

43. Cary LC, Goebel M, Corsaro BG, Wang HG, Rosen E, Fraser MJ. Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology. 1989 Sep;172(1):156–69.

44. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell. 2005 Aug 12;122(3):473–83.

45. Hartwell LH. Mutants of Saccharomyces cerevisiae unresponsive to cell division control by polypeptide mating hormone. J Cell Biol. 1980 Jun;85(3):811–22.

46. Dolan JW, Kirkman C, Fields S. The yeast STE12 protein binds to the DNA sequence mediating pheromone induction. Proc Natl Acad Sci U S A. 1989 Aug;86(15):5703–7.

47. Kirkman-Correia C, Stroke IL, Fields S. Functional domains of the yeast STE12 protein, a pheromone-responsive transcriptional activator. Mol Cell Biol. 1993 Jun;13(6):3765–72.

48. Tedford K, Kim S, Sa D, Stevens K, Tyers M. Regulation of the mating pheromone and invasive growth responses in yeast by two MAP kinase substrates [Internet]. Vol. 7, Current Biology. 1997. p. 228–38. Available from: http://dx.doi.org/10.1016/s0960-9822(06)00118-7

49. Sengupta P, Cochran BH. The PRE and PQ box are functionally distinct yeast pheromone response elements. Mol Cell Biol. 1990 Dec;10(12):6809–12.

50. Yuan YO, Stroke IL, Fields S. Coupling of cell identity to signal response in yeast: interaction between the alpha 1 and STE12 proteins. Genes Dev. 1993 Aug;7(8):1584–97.

51. Madhani HD, Fink GR. Combinatorial control required for the specificity of yeast MAPK signaling. Science. 1997 Feb 28;275(5304):1314–7.

52. Heise B, van der Felden J, Kern S, Malcher M, Brückner S, Mösch H-U. The TEA transcription factor Tec1 confers promoter-specific gene regulation by Ste12-dependent and -independent mechanisms. Eukaryot Cell. 2010 Apr;9(4):514–31.

53. Dorrity MW, Cuperus JT, Carlisle JA, Fields S, Queitsch C. Preferences in a trait decision determined by transcription factor variants. Proc Natl Acad Sci U S A. 2018 Aug 21;115(34):E7997–8006.

54. Goff PL, Le Goff P, Michel D. HSF1 Activation Occurs at Different Temperatures in Somatic and Male Germ Cells in the Poikilotherm Rainbow Trout [Internet]. Vol. 259, Biochemical and Biophysical Research Communications. 1999. p. 15–20. Available from: http://dx.doi.org/10.1006/bbrc.1999.0729

55. Xiao X, Zuo X, Davis AA, McMillan DR, Curry BB, Richardson JA, et al. HSF1 is required for extra-embryonic development, postnatal growth and protection during inflammatory responses in mice. EMBO J. 1999 Nov 1;18(21):5943–52.

56. Metchat A, Akerfelt M, Bierkamp C, Delsinne V, Sistonen L, Alexandre H, et al. Mammalian heat shock factor 1 is essential for oocyte meiosis and directly regulates Hsp90alpha expression. J Biol Chem. 2009 Apr 3;284(14):9521–8.

57. Hajdu-Cronin YM, Chen WJ, Sternberg PW. The L-type cyclin CYL-1 and the heat-shock-factor HSF-1 are required for heat-shock-induced protein expression in Caenorhabditis elegans. Genetics. 2004 Dec;168(4):1937–49.

58. Morton EA, Lamitina T. Caenorhabditis elegansHSF-1 is an essential nuclear protein that forms stress granule-like structures following heat shock [Internet]. Vol. 12, Aging Cell. 2013. p. 112–20. Available from: http://dx.doi.org/10.1111/acel.12024

59. Li J, Labbadia J, Morimoto RI. Rethinking HSF1 in Stress, Development, and Organismal Health [Internet]. Vol. 27, Trends in Cell Biology. 2017. p. 895–905. Available from: http://dx.doi.org/10.1016/j.tcb.2017.08.002

60. Sorger PK, Nelson HCM. Trimerization of a yeast transcriptional activator via a coiled-coil motif [Internet]. Vol. 59, Cell. 1989. p. 807–13. Available from: http://dx.doi.org/10.1016/0092-8674(89)90604-1

61. Rabindran SK, Giorgi G, Clos J, Wu C. Molecular cloning and expression of a human heat shock factor, HSF1 [Internet]. Vol. 88, Proceedings of the National Academy of Sciences. 1991. p. 6906–10. Available from: http://dx.doi.org/10.1073/pnas.88.16.6906

62. Soncin F, Zhang X, Chu B, Wang X, Asea A, Ann Stevenson M, et al. Transcriptional activity and DNA binding of heat shock factor-1 involve phosphorylation on threonine 142 by CK2. Biochem Biophys Res Commun. 2003 Apr 4;303(2):700–6.

63. Hashikawa N, Yamamoto N, Sakurai H. Different mechanisms are involved in the transcriptional activation by yeast heat shock transcription factor through two different types of heat shock elements. J Biol Chem. 2007 Apr 6;282(14):10333–40.

64. Satyal SH, Chen D, Fox SG, Kramer JM, Morimoto RI. Negative regulation of the heat shock transcriptional response by HSBP1. Genes Dev. 1998 Jul 1;12(13):1962–74.

65. Tan K, Fujimoto M, Takii R, Takaki E, Hayashida N, Nakai A. Mitochondrial SSBP1 protects cells from proteotoxic stresses by potentiating stress-induced HSF1 transcriptional activity [Internet]. Vol. 6, Nature Communications. 2015. Available from: http://dx.doi.org/10.1038/ncomms7580

66. Baler R, Dahl G, Voellmy R. Activation of human heat shock genes is accompanied by oligomerization, modification, and rapid translocation of heat shock transcription factor HSF1. Mol Cell Biol. 1993 Apr;13(4):2486–96.

67. Rabindran SK, Haroun RI, Clos J, Wisniewski J, Wu C. Regulation of heat shock factor trimer formation: role of a conserved leucine zipper. Science. 1993 Jan 8;259(5092):230–4.

68. Hentze N, Le Breton L, Wiesner J, Kempf G, Mayer MP. Molecular mechanism of thermosensory function of human heat shock transcription factor Hsf1. Elife [Internet]. 2016 Jan 19;5. Available from: http://dx.doi.org/10.7554/eLife.11576

69. Bogdanove AJ, Bohm A, Miller JC, Morgan RD, Stoddard BL. Engineering altered protein-DNA recognition specificity. Nucleic Acids Res. 2018 Jun 1;46(10):4845–71.

70. Wood AJ, Lo T-W, Zeitler B, Pickle CS, Ralston EJ, Lee AH, et al. Targeted genome editing across species using ZFNs and TALENs. Science. 2011 Jul 15;333(6040):307.

71. Streubel J, Blücher C, Landgraf A, Boch J. TAL effector RVD specificities and efficiencies [Internet]. Vol. 30, Nature Biotechnology. 2012. p. 593–5. Available from: http://dx.doi.org/10.1038/nbt.2304

72. Christian ML, Demorest ZL, Starker CG, Osborn MJ, Nyquist MD, Zhang Y, et al. Targeting G with TAL Effectors: A Comparison of Activities of TALENs Constructed with NN and NK Repeat Variable Di-Residues [Internet]. Vol. 7, PLoS ONE. 2012. p. e45383. Available from: http://dx.doi.org/10.1371/journal.pone.0045383

73. Desjarlais, Berg JM. Redesigning the DNA-binding specificity of a zinc finger protein: a data base-guided approach. Proteins. 1992 Jul;13(3):272.

74. Desjarlais JR, Berg JM. Toward rules relating zinc finger protein sequences and DNA binding site preferences [Internet]. Vol. 89, Proceedings of the National Academy of Sciences. 1992. p. 7345–9. Available from: http://dx.doi.org/10.1073/pnas.89.16.7345

75. Lambert AR, Hallinan JP, Shen BW, Chik JK, Bolduc JM, Kulshina N, et al. Indirect DNA Sequence Recognition and Its Impact on Nuclease Cleavage Activity. Structure. 2016 Jun 7;24(6):862–73.

76. Flick KE, Jurica MS, Monnat RJ Jr, Stoddard BL. DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI. Nature. 1998 Jul 2;394(6688):96–101.

77. Seligman LM, Chisholm KM, Chevalier BS, Chadsey MS, Edwards ST, Savage JH, et al. Mutations altering the cleavage specificity of a homing endonuclease. Nucleic Acids Res. 2002 Sep 1;30(17):3870–9.

78. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL, et al. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature. 2006 Jun 1;441(7093):656–9.

79. Ozawa T, Natori Y, Sato M, Umezawa Y. Imaging dynamics of endogenous mitochondrial RNA in single living cells. Nat Methods. 2007 May;4(5):413–9.

80. Cooke A, Prigge A, Opperman L, Wickens M. Targeted translational regulation using the PUF protein family scaffold [Internet]. Vol. 108, Proceedings of the National Academy of Sciences. 2011. p. 15870–5. Available from: http://dx.doi.org/10.1073/pnas.1105151108

81. Tsai YS, Gomez SM, Wang Z. Prevalent RNA recognition motif duplication in the human genome. RNA. 2014 May;20(5):702–12.

82. Auweter SD, Oberstrass FC, -T. Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? [Internet]. Vol. 34, Nucleic Acids Research. 2006. p. 4943–59. Available from: http://dx.doi.org/10.1093/nar/gkl620

83. Afroz T, Cienikova Z, Cléry A, Allain FHT. One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence. Methods Enzymol. 2015 Mar 12;558:235–78.

84. Nicastro G, Taylor IA, Ramos A. KH–RNA interactions: back in the groove [Internet]. Vol. 30, Current Opinion in Structural Biology. 2015. p. 63–70. Available from: http://dx.doi.org/10.1016/j.sbi.2015.01.002

85. Main ERG, Jackson SE, Regan L. The folding and design of repeat proteins: reaching a consensus. Curr Opin Struct Biol. 2003 Aug;13(4):482–9.

86. Saha D, Prasad AM, Srinivasan R. Pentatricopeptide repeat proteins and their emerging roles in plants. Plant Physiol Biochem. 2007 Aug;45(8):521–34.

87. Delannoy E, Stanley WA, Bond CS, Small ID. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles [Internet]. Vol. 35, Biochemical Society Transactions. 2007. p. 1643–7. Available from: http://dx.doi.org/10.1042/bst0351643

88. Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, et al. Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell. 2004 Aug;16(8):2089–103.

89. Barkan A, Rojas M, Fujii S, Yap A, Chong YS, Bond CS, et al. A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. PLoS Genet. 2012 Aug 16;8(8):e1002910.

90. Takenaka M, Zehrmann A, Brennicke A, Graichen K. Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. PLoS One. 2013 Jun 6;8(6):e65343.

91. Yagi Y, Hayashi S, Kobayashi K, Hirayama T, Nakamura T. Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. PLoS One. 2013 Mar 5;8(3):e57286.

92. Yin P, Li Q, Yan C, Liu Y, Liu J, Yu F, et al. Structural basis for the modular recognition of single-stranded RNA by PPR proteins. Nature. 2013 Dec 5;504(7478):168–71.

93. Gully BS, Cowieson N, Stanley WA, Shearston K, Small ID, Barkan A, et al. The solution structure of the pentatricopeptide repeat protein PPR10 upon binding atpH RNA. Nucleic Acids Res. 2015 Feb 18;43(3):1918–26.

94. Ke J, Chen R-Z, Ban T, Zhou XE, Gu X, Tan MHE, et al. Structural basis for RNA recognition by a dimeric PPR-protein complex. Nat Struct Mol Biol. 2013 Dec;20(12):1377–82.

95. Goldstrohm AC, Hook BA, Seay DJ, Wickens M. PUF proteins bind Pop2p to regulate messenger RNAs. Nat Struct Mol Biol. 2006 Jun;13(6):533–9.

96. Goldstrohm AC, Seay DJ, Hook BA, Wickens M. PUF protein-mediated deadenylation is catalyzed by Ccr4p. J Biol Chem. 2007 Jan 5;282(1):109–14.

97. Weidmann CA, Goldstrohm AC. Drosophila Pumilio Protein Contains Multiple Autonomous Repression Domains That Regulate mRNAs Independently of Nanos and Brain Tumor [Internet]. Vol. 32, Molecular and Cellular Biology. 2012. p. 527–40. Available from: http://dx.doi.org/10.1128/mcb.06052-11

98. Gootenberg JS, Abudayyeh OO, Kellner MJ, Joung J, Collins JJ, Zhang F. Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science. 2018 Apr 27;360(6387):439–44.

99. Freije CA, Myhrvold C, Boehm CK, Lin AE, Welch NL, Carter A, et al. Programmable Inhibition and Detection of RNA Viruses Using Cas13. Mol Cell. 2019 Dec 5;76(5):826–37.e11.

100. O'Connell MR. Molecular Mechanisms of RNA Targeting by Cas13-containing Type VI CRISPR-Cas Systems. J Mol Biol. 2019 Jan 4;431(1):66–87.

101. Liu T, Pan S, Li Y, Peng N, She Q. Type III CRISPR-Cas System: Introduction And Its Application for Genetic Manipulations. Curr Issues Mol Biol. 2018;26:1–14.

102. Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, et al. RNA targeting with CRISPR-Cas13. Nature. 2017 Oct 12;550(7675):280–4.

103. Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, et al. RNA editing with CRISPR-Cas13. Science. 2017 Nov 24;358(6366):1019–27.

104. Quenault T, Lithgow T, Traven A. PUF proteins: repression, activation and mRNA localization. Trends Cell Biol. 2011 Feb;21(2):104–12.

105. Lehmann R, Nüsslein-Volhard C. hunchback, a gene required for segmentation of an anterior and posterior region of the Drosophila embryo. Dev Biol. 1987 Feb;119(2):402–17.

106. Zhang B, Gallegos M, Puoti A, Durkin E, Fields S, Kimble J, et al. A conserved RNA-binding protein that regulates sexual fates in the C. elegans hermaphrodite germ line [Internet]. Vol. 390, Nature. 1997. p. 477–84. Available from: http://dx.doi.org/10.1038/37297

107. Olivas W, Parker R. The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. EMBO J. 2000 Dec 1;19(23):6602–11.

108. Tadauchi T. Post-transcriptional regulation through the HO 3'-UTR by Mpt5, a yeast homolog of Pumilio and FBF [Internet]. Vol. 20, The EMBO Journal. 2001. p. 552–61. Available from: http://dx.doi.org/10.1093/emboj/20.3.552

109. Wang X, McLachlan J, Zamore PD, Tanaka Hall TM. Modular Recognition of RNA by a Human Pumilio-Homology Domain [Internet]. Vol. 110, Cell. 2002. p. 501–12. Available from: http://dx.doi.org/10.1016/s0092-8674(02)00873-5

110. Gerber AP, Herschlag D, Brown PO. Extensive Association of Functionally and Cytotopically Related mRNAs with Puf Family RNA-Binding Proteins in Yeast [Internet]. Vol. 2, PLoS Biology. 2004. p. e79. Available from: http://dx.doi.org/10.1371/journal.pbio.0020079

111. Lapointe CP, Wilinski D, Saunders HAJ, Wickens M. Protein-RNA networks revealed

through covalent RNA marks. Nat Methods. 2015 Dec;12(12):1163–70.

112. Lapointe CP, Preston MA, Wilinski D, Saunders HAJ, Campbell ZT, Wickens M. Architecture and dynamics of overlapped RNA regulatory networks. RNA. 2017 Nov;23(11):1636–47.

113. Wang X, Zamore PD, Hall TM. Crystal structure of a Pumilio homology domain. Mol Cell. 2001 Apr;7(4):855–65.

114. Opperman L, Hook B, DeFino M, Bernstein DS, Wickens M. A single spacer nucleotide determines the specificities of two mRNA regulatory proteins [Internet]. Vol. 12, Nature Structural & Molecular Biology. 2005. p. 945–51. Available from: http://dx.doi.org/10.1038/nsmb1010

115. Cheong C-G, Hall TMT. Engineering RNA sequence specificity of Pumilio repeats. Proc Natl Acad Sci U S A. 2006 Sep 12;103(37):13635–9.

116. Hook B, Bernstein D, Zhang B, Wickens M. RNA-protein interactions in the yeast three-hybrid system: affinity, sensitivity, and enhanced library screening. RNA. 2005 Feb;11(2):227–33.

117. Filipovska A, Razif MFM, Nygård KKA, Rackham O. A universal code for RNA recognition by PUF proteins. Nat Chem Biol. 2011 May 15;7(7):425–7.

118. Campbell ZT, Bhimsaria D, Valley CT, Rodriguez-Martinez JA, Menichelli E, Williamson JR, et al. Cooperativity in RNA-Protein Interactions: Global Analysis of RNA Binding Specificity [Internet]. Vol. 1, Cell Reports. 2012. p. 570–81. Available from: http://dx.doi.org/10.1016/j.celrep.2012.04.003

119. Campbell ZT, Valley CT, Wickens M. A protein-RNA specificity code enables targeted activation of an endogenous human transcript. Nat Struct Mol Biol. 2014 Aug;21(8):732–8.

120. Hall TMT. Expanding the RNA-recognition code of PUF proteins. Nat Struct Mol Biol. 2014 Aug;21(8):653–5.

121. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381–6.

122. Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol. 2016 Aug 17;17(1):173.

123. Loh KM, Chen A, Koh PW, Deng TZ, Sinha R, Tsai JM, et al. Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types. Cell. 2016 Jul 14;166(2):451–67.

124. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H, et al. Reversed graph embedding

resolves complex single-cell developmental trajectories [Internet]. Available from: http://dx.doi.org/10.1101/110668

125. Welch JD, Hartemink A, Prins JF. SLICER: Inferring Branched, Nonlinear Cellular Trajectories from Single Cell RNA-seq Data [Internet]. Available from: http://dx.doi.org/10.1101/047845

126. Haghverdi L, Büttner M, Alexander Wolf F, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching [Internet]. Vol. 13, Nature Methods. 2016. p. 845–8. Available from: http://dx.doi.org/10.1038/nmeth.3971

127. Campbell KR, Yau C. Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers. Wellcome Open Res. 2017 Mar 15;2:19.

128. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science. 2018 Jun 1;360(6392):981–7.

129. Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science [Internet]. 2018 Jun 1;360(6392). Available from: http://dx.doi.org/10.1126/science.aar3131

130. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science [Internet]. 2018 May 25;360(6391). Available from: http://dx.doi.org/10.1126/science.aaq1723

131. Herzog VA, Reichholf B, Neumann T, Rescheneder P, Bhat P, Burkard TR, et al. Thiol-linked alkylation of RNA to assess expression dynamics. Nat Methods. 2017 Dec;14(12):1198–204.

132. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017 Aug 18;357(6352):661–7.

133. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence [Internet]. Vol. 13, Nature Reviews Genetics. 2012. p. 59–69. Available from: http://dx.doi.org/10.1038/nrg3095

134. Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression [Internet]. Vol. 22, Genome Research. 2012. p. 860–9. Available from: http://dx.doi.org/10.1101/gr.131201.111

135. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, et al. Divergence of Transcription Factor Binding Sites Across Related Yeast Species [Internet]. Vol. 317, Science. 2007. p. 815–9. Available from: http://dx.doi.org/10.1126/science.1140748

136. Deplancke B, Alpern D, Gardeux V. The Genetics of Transcription Factor DNA Binding Variation [Internet]. Vol. 166, Cell. 2016. p. 538–54. Available from: http://dx.doi.org/10.1016/j.cell.2016.07.012

137. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006 Nov;24(11):1429–35.

138. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, et al. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell. 2011 Dec 9;147(6):1270–82.

139. Hughes AEO, Myers CA, Corbo JC. A massively parallel reporter assay reveals context-dependent activity of homeodomain binding sites in vivo. Genome Res. 2018 Oct;28(10):1520–31.

140. Wong Sak Hoi J, Dumas B. Ste12 and Ste12-like proteins, fungal transcription factors regulating development and pathogenicity. Eukaryot Cell. 2010 Apr;9(4):480–5.

141. Mead J, Bruning AR, Gill MK, Steiner AM, Acton TB, Vershon AK. Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. Mol Cell Biol. 2002 Jul;22(13):4607–21.

142. Bardwell L, Cook JG, Voora D, Baggott DM, Martinez AR, Thorner J. Repression of yeast Ste12 transcription factor by direct binding of unphosphorylated Kss1 MAPK and its regulation by the Ste7 MEK. Genes Dev. 1998 Sep 15;12(18):2887–98.

143. Wang H, Johnston M, Mitra RD. Calling cards for DNA-binding proteins. Genome Res. 2007 Aug;17(8):1202–9.

144. Gray M, Piccirillo S, Honigberg SM. Two-step method for constructing unmarked insertions, deletions and allele substitutions in the yeast genome. FEMS Microbiol Lett. 2005 Jul 1;248(1):31–6.

145. Mumberg D, Müller R, Funk M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds [Internet]. Vol. 156, Gene. 1995. p. 119–22. Available from: http://dx.doi.org/10.1016/0378-1119(95)00037-7

146. Gietz RD, Daniel Gietz R, Woods RA. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method [Internet]. Guide to Yeast Genetics and Molecular and Cell Biology - Part B. 2002. p. 87–96. Available from: http://dx.doi.org/10.1016/s0076-6879(02)50957-5

147. Cuperus JT, Lo RS, Shumaker L, Proctor J, Fields S. A tetO Toolkit To Alter Expression of Genes in Saccharomyces cerevisiae. ACS Synth Biol. 2015 Jul 17;4(7):842–52.

148. Hoffman CS, Winston F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformaion of Escherichia coli [Internet]. Vol. 57, Gene. 1987.

p. 267–72. Available from: http://dx.doi.org/10.1016/0378-1119(87)90131-4

149. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008 Sep 17;9(9):R137.

150. Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. Genome Res. 2011 May;21(5):748–55.

151. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W202–8.

152. Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, dos Santos SC, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res. 2014 Jan;42(Database issue):D161–6.

153. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.

154. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

155. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015 Nov 19;527(7578):384–8.

156. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, et al. Designer deletion strains derived fromSaccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications [Internet]. Vol. 14, Yeast. 1998. p. 115–32. Available from: 3.0.co;2-2">http://dx.doi.org/10.1002/(sici)1097-0061(19980130)14:2<115::aid-yea204>3.0.co;2-2

157. Boeke JD, Trueheart J, Natsoulis G, Fink GR. 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. Methods Enzymol. 1987;154:164–75.

158. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. Genetic analysis of variation in transcription factor binding in yeast [Internet]. Vol. 464, Nature. 2010. p. 1187–91. Available from: http://dx.doi.org/10.1038/nature08934

159. Hwang-Shum JJ, Hagen DC, Jarvis EE, Westby CA, Sprague GF Jr. Relative contributions of MCM1 and STE12 to transcriptional activation of a- and alpha-specific genes from Saccharomyces cerevisiae. Mol Gen Genet. 1991 Jun;227(2):197–204.

160. Davenport KD, Williams KE, Ullmann BD, Gustin MC. Activation of the Saccharomyces cerevisiae filamentation/invasion pathway by osmotic stress in high-osmolarity glycogen pathway mutants. Genetics. 1999 Nov;153(3):1091–103.

161. Tatebayashi K, Yamamoto K, Tanaka K, Tomida T, Maruoka T, Kasukawa E, et al. Adaptor functions of Cdc42, Ste50, and Sho1 in the yeast osmoregulatory HOG MAPK pathway. EMBO J. 2006 Jul 12;25(13):3033–44.

162. Posas F, Saito H. Osmotic activation of the HOG MAPK pathway via Ste11p MAPKKK: scaffold role of Pbs2p MAPKK. Science. 1997 Jun 13;276(5319):1702–5.

163. Tatebayashi K, Tanaka K, Yang H-Y, Yamamoto K, Matsushita Y, Tomida T, et al. Transmembrane mucins Hkr1 and Msb2 are putative osmosensors in the SHO1 branch of yeast HOG pathway. EMBO J. 2007 Aug 8;26(15):3521–33.

164. Sorrells TR, Booth LN, Tuch BB, Johnson AD. Intersecting transcription networks constrain gene regulatory evolution. Nature. 2015 Jul 16;523(7560):361–5.

165. Dalal CK, Johnson AD. How transcription circuits explore alternative architectures while maintaining overall circuit output. Genes Dev. 2017 Jul 15;31(14):1397–405.

166. Gupta RM, Musunuru K. Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9. J Clin Invest. 2014 Oct;124(10):4154–61.

167. Cooke A, Prigge A, Opperman L, Wickens M. Targeted translational regulation using the PUF protein family scaffold. Proc Natl Acad Sci U S A. 2011 Sep 20;108(38):15870–5.

168. Wang X, McLachlan J, Zamore PD, Hall TMT. Modular recognition of RNA by a human pumilio-homology domain. Cell. 2002 Aug 23;110(4):501–12.

169. Biolabs NE, New England Biolabs. Gibson Assembly® Master Mix – Assembly (E2611) v1 [Internet]. protocols.io. Available from: http://dx.doi.org/10.17504/protocols.io.cjxupm

170. Bernstein D. Analyzing mRNA–protein complexes using a yeast three-hybrid system [Internet]. Vol. 26, Methods. 2002. p. 123–41. Available from: http://dx.doi.org/10.1016/s1046-2023(02)00015-4

171. Dong S, Wang Y, Cassidy-Amstutz C, Lu G, Bigler R, Jezyk MR, et al. Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. J Biol Chem. 2011 Jul 29;286(30):26732–42.

172. Adamala KP, Martin-Alarcon DA, Boyden ES. Programmable RNA-binding protein composed of repeats of a single modular unit. Proc Natl Acad Sci U S A. 2016 May 10;113(19):E2579–88.

173. Koh YY, Wang Y, Qiu C, Opperman L, Gross L, Tanaka Hall TM, et al. Stacking interactions in PUF-RNA complexes. RNA. 2011 Apr;17(4):718–27.

174. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol. 2019 Mar 19;20(1):59.

175. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016 Jun;34(6):637–45.

176. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018 Jun 19;19(1):477.

177. Schofield JA, Duffy EE, Kiefer L, Sullivan MC, Simon MD. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. Nat Methods. 2018 Mar;15(3):221–5.

178. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019 Feb;566(7745):496–502.

179. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21.

180. Lindenbaum P. JVarkit: java-based utilities for Bioinformatics. figshare. 2015;

181. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636–40.

182. FelixKrueger. FelixKrueger/TrimGalore [Internet]. GitHub. [cited 2019 Jan 28]. Available from: https://github.com/FelixKrueger/TrimGalore

183. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

184. Picard Tools - By Broad Institute [Internet]. [cited 2019 Jan 27]. Available from: http://broadinstitute.github.io/picard/

185. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568–76.

186. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017 Oct;14(10):979–82.

187. cole-trapnell-lab. cole-trapnell-lab/monocle-release [Internet]. GitHub. [cited 2019 Jan 28]. Available from: https://github.com/cole-trapnell-lab/monocle-release

188. Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science. 2014 Jun 27;344(6191):1492–6.

189. Cleary MD, Meiering CD, Jan E, Guymon R, Boothroyd JC. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of

mRNA synthesis and decay. Nat Biotechnol. 2005;23(2):232–7.

190. Dolken L, Ruzsics Z, Radle B, Friedel CC, Zimmer R, Mages J, et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA. 2008;14(9):1959–72.

191. Miller C, Schwalb B, Maier K, Schulz D, Dumcke S, Zacher B, et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol Syst Biol. 2014;7(1):458–458.

192. Duffy EE, Rutenberg-Schoenberg M, Stark CD, Kitchen RR, Gerstein MB, Simon MD. Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. Mol Cell. 2015 Sep 3;59(5):858–66.

193. Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, et al. TT-seq maps the human transient transcriptome. Science. 2016;352(6290):1225–8.

194. Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, et al. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. Nat Biotechnol. 2011;29(5):436–42.

195. Miller MR, Robinson KJ, Cleary MD, Doe CQ. TU-tagging: cell type–specific RNA isolation from intact complex tissues. Nat Methods. 2009;6(6):439–41.

196. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015 May 22;348(6237):910–4.

197. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science. 2018 Sep 28;361(6409):1380–5.

198. Ramani V, Deng X, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, et al. Massively multiplex single-cell Hi-C [Internet]. 2016. Available from: http://dx.doi.org/10.1101/065052

199. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol. 2018;36(5):428–31.

200. Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, et al. Sequencing thousands of single-cell genomes with combinatorial indexing. Nat Methods. 2017;14(3):302–8.

201. Yin Y, Jiang Y, Berletch JB, Disteche CM, Noble WS, Steemers FJ, et al. High-throughput mapping of meiotic crossover and chromosome mis-segregation events in interspecific hybrid mice [Internet]. 2018. Available from: http://dx.doi.org/10.1101/338053

202. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018 Apr 13;360(6385):176–82.

203. Buckingham JC. Glucocorticoids: exemplars of multi-tasking. Br J Pharmacol. 2006 Jan;147(Suppl 1):S258.

204. Reddy TE, Pauli F, Sprouse RO, Neff NF, Newberry KM, Garabedian MJ, et al. Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res. 2009 Dec;19(12):2163–71.

205. John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet. 2011 Mar;43(3):264–8.

206. Reddy TE, Gertz J, Crawford GE, Garabedian MJ, Myers RM. The Hypersensitive Glucocorticoid Response Specifically Regulates Period 1 and Expression of Circadian Genes. Mol Cell Biol. 2012;32(18):3756–67.

207. Vockley CM, D'Ippolito AM, McDowell IC, Majoros WH, Safi A, Song L, et al. Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. Cell. 2016 Aug 25;166(5):1269–81.e19.

208. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018 Aug 8;560(7719):494.

209. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29):861.

210. Binder EB. The role of FKBP5, a co-chaperone of the glucocorticoid receptor in the pathogenesis and therapy of affective and anxiety disorders [Internet]. Vol. 34, Psychoneuroendocrinology. 2009. p. S186–95. Available from: http://dx.doi.org/10.1016/j.psyneuen.2009.05.021

211. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018 Jun;36(5):411–20.

212. Dataset - ENCODE Transcription Factor Binding Site Profiles [Internet]. [cited 2019 Jan 27]. Available from: http://amp.pharm.mssm.edu/Harmonizome/dataset/ENCODE+Transcription+Factor+Binding+Site+Profiles

213. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017 Nov;14(11):1083–6.

214. Han H, Cho J-W, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded

reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res. 2018 Jan 4;46(D1):D380–6.

215. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update [Internet]. Vol. 44, Nucleic Acids Research. 2016. p. W90–7. Available from: http://dx.doi.org/10.1093/nar/gkw377

216. Boruk M, Savory JGA, Haché RJG. AF-2-Dependent Potentiation of CCAAT Enhancer Binding Proteinβ -Mediated Transcriptional Activation by Glucocorticoid Receptor. Mol Endocrinol. 1998;12(11):1749–63.

217. Qin W, Pan J, Qin Y, Lee DN, Bauman WA, Cardozo C. Identification of functional glucocorticoid response elements in the mouse FoxO1 promoter. Biochem Biophys Res Commun. 2014;450(2):979–83.

218. Sheela Rani CS, Elango N, Wang S-S, Kobayashi K, Strong R. Identification of an Activator Protein-1-Like Sequence as the Glucocorticoid Response Element in the Rat Tyrosine Hydroxylase Gene. Mol Pharmacol. 2009 Mar;75(3):589.

219. Fischer M, Müller GA. Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. Crit Rev Biochem Mol Biol. 2017 Dec;52(6):638–62.

220. Chou J, Provot S, Werb Z. GATA3 in development and cancer differentiation: cells GATA have it! J Cell Physiol. 2010 Jan;222(1):42–9.

221. Madhurima Biswas JYC. Role of Nrf1 in antioxidant response element-mediated gene expression and beyond. Toxicol Appl Pharmacol. 2010 Apr 1;244(1):16.

222. Ryoo I-G, Kwak M-K. Regulatory crosstalk between the oxidative stress-related transcription factor Nfe2l2/Nrf2 and mitochondria. Toxicol Appl Pharmacol. 2018 Nov 15;359:24–33.

223. Heer R, Robson CN, Shenton BK, Leung HY. The role of androgen in determining differentiation and regulation of androgen receptor expression in the human prostatic epithelium transient amplifying population. J Cell Physiol. 2007 Sep;212(3):572–8.

224. Meixner A, Karreth F, Kenner L, Penninger JM, Wagner EF. Jun and JunD-dependent functions in cell proliferation and stress response. Cell Death Differ. 2010 Sep;17(9):1409–19.

225. Li M, Gu Y, Ma Y-C, Shang Z-F, Wang C, Liu F-J, et al. Krüppel-Like Factor 5 Promotes Epithelial Proliferation and DNA Damage Repair in the Intestine of Irradiated Mice. Int J Biol Sci. 2015 Dec 2;11(12):1458–68.

226. Eberlé D, Hegarty B, Bossard P, Ferré P, Foufelle F. SREBP transcription factors: master regulators of lipid homeostasis. Biochimie. 2004 Nov;86(11):839–48.

227. Shermoen AW, O'Farrell PH. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. Cell. 1991 Oct 18;67(2):303–10.

228. Palozola KC, Donahue G, Liu H, Grant GR, Becker JS, Cote A, et al. Mitotic transcription and waves of gene reactivation during mitotic exit. Science. 2017 Oct 6;358(6359):119–22.

229. Parsons GG, Spencer CA. Mitotic repression of RNA polymerase II transcription is accompanied by release of transcription elongation complexes. Mol Cell Biol. 1997 Oct;17(10):5791–802.

230. Sanchez-Alvarez M, Zhang Q, Finger F, Wakelam MJO, Bakal C. Cell cycle progression is an essential regulatory component of phospholipid metabolism and membrane homeostasis. Open Biol. 2015 Sep;5(9):150093.

231. Hastie T, Stuetzle W. Principal Curves [Internet]. Vol. 84, Journal of the American Statistical Association. 1989. p. 502. Available from: http://dx.doi.org/10.2307/2289936

232. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2016. 260 p.

233. Harmon JM, Norman MR, Fowlkes BJ, Thompson EB. Dexamethasone induces irreversible G1 arrest and death of a human lymphoid cell line. J Cell Physiol. 1979 Feb;98(2):267–78.

234. Greenberg AK, Hu J, Basu S, Hay J, Reibman J, Yie T-A, et al. Glucocorticoids inhibit lung cancer cell growth through both the extracellular signal-related kinase pathway and cell cycle regulators. Am J Respir Cell Mol Biol. 2002 Sep;27(3):320–8.

235. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. Fundamental limits on dynamic inference from single-cell snapshots. Proc Natl Acad Sci U S A. 2018 Mar 6;115(10):E2467–76.

236. Qiu X, Rahimzamani A, Wang L, Mao Q, Durham T, McFaline-Figueroa JL, et al. Towards inferring causal gene regulatory networks from single cell expression measurements [Internet]. Available from: http://dx.doi.org/10.1101/426981

237. Erhard F, Baptista MAP, Krammer T, Hennig T, Lange M, Arampatzi P, et al. scSLAM-seq reveals core features of transcription dynamics in single cells. Nature. 2019 Jul;571(7765):419–23.

238. Hendriks G-J, Jung LA, Larsson AJM, Lidschreiber M, Andersson Forsman O, Lidschreiber K, et al. NASC-seq monitors RNA synthesis in single cells. Nat Commun. 2019 Jul 17;10(1):3138.

239. Baran-Gale J, Chandra T, Kirschner K. Experimental design for single-cell RNA sequencing [Internet]. Vol. 17, Briefings in Functional Genomics. 2018. p. 233–9. Available from: http://dx.doi.org/10.1093/bfgp/elx035

240. Chen W, Li Y, Easton J, Finkelstein D, Wu G, Chen X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. Genome Biol. 2018 May 31;19(1):70.

241. Matsushima W, Herzog VA, Neumann T, Gapp K, Zuber J, Ameres SL, et al. SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. Development [Internet]. 2018 Jul 11;145(13). Available from: http://dx.doi.org/10.1242/dev.164640

242. Sharma U, Sun F, Reichholf B, Herzog V, Ameres S, Rando O. Small RNAs are trafficked from the epididymis to developing mammalian sperm [Internet]. 2017. Available from: http://dx.doi.org/10.1101/194522

243. Gay L, Miller MR, Ventura PB, Devasthali V, Vue Z, Thompson HL, et al. Mouse TU tagging: a chemical/genetic intersectional method for purifying cell type-specific nascent RNA. Genes Dev. 2013 Jan 1;27(1):98–115.

244. Aughey GN, Cheetham SW, Southall TD. DamID as a versatile tool for understanding gene regulation. Development [Internet]. 2019 Mar 15;146(6). Available from: http://dx.doi.org/10.1242/dev.173666

245. Lowe WL, Reddy TE. Genomic approaches for understanding the genetics of complex disease [Internet]. Vol. 25, Genome Research. 2015. p. 1432–41. Available from: http://dx.doi.org/10.1101/gr.190603.115

246. White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci U S A. 2013 Jul 16;110(29):11952–7.

247. Gotea V, Visel A, Westlund JM, Nobrega MA, Pennacchio LA, Ovcharenko I. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 2010 May;20(5):565–77.

248. Zhang W, Wang Y, Dong S, Choudhury R, Jin Y, Wang Z. Treatment of type 1 myotonic dystrophy by engineering site-specific RNA endonucleases that target (CUG)(n) repeats. Mol Ther. 2014 Feb;22(2):312–20.

249. Zhao Y-Y, Mao M-W, Zhang W-J, Wang J, Li H-T, Yang Y, et al. Expanding RNA binding specificity and affinity of engineered PUF domains. Nucleic Acids Res. 2018 May 18;46(9):4771–82.

250. Choudhury R, Tsai YS, Dominguez D, Wang Y, Wang Z. Engineering RNA endonucleases with customized sequence specificities. Nat Commun. 2012;3:1147.

251. Wang Y, Cheong C-G, Hall TMT, Wang Z. Engineering splicing factors with designed specificities. Nat Methods. 2009 Nov;6(11):825–30.

252. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.

253. Montiel-Gonzalez MF, Vallecillo-Viejo I, Yudowski GA, Rosenthal JJC. Correction of mutations within the cystic fibrosis transmembrane conductance regulator by site-directed RNA editing [Internet]. Vol. 110, Proceedings of the National Academy of Sciences. 2013. p. 18285–90. Available from: http://dx.doi.org/10.1073/pnas.1306243110

254. Maekawa S, Imamachi N, Irie T, Tani H, Matsumoto K, Mizutani R, et al. Analysis of RNA decay factor mediated RNA stability contributions on RNA abundance [Internet]. Vol. 16, BMC Genomics. 2015. p. 154. Available from: http://dx.doi.org/10.1186/s12864-015-1358-y

Chapter 6. Appendices

APPENDIX A

Table 7.1 Oligonucleotides use in Chapter 2

Oligo #	Oligo Sequence

316(internal_fwd)	GATGTCCTAAATGCACAGCGAC
317(internal_rev)	GAGGCGTGCTTGTCAATGC
376(external_fwd_1)	AATGATACGGCGACCACCGAGATCTACACCTCCATCGAGACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
377(external_rev)	CAAGCAGAAGACGGCATACGAGATTGCTTGTCAATGCGGTAAG
378(external_fwd_2)	AATGATACGGCGACCACCGAGATCTACACTTGGTAGTCGACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
429(external_fwd_3)	AATGATACGGCGACCACCGAGATCTACACCCTAGACGAGACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
430(external_fwd_4)	AATGATACGGCGACCACCGAGATCTACACTCGTTAGAGCACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
431(external_fwd_5)	AATGATACGGCGACCACCGAGATCTACACCGTTCTATCAACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
432(external_fwd_6)	AATGATACGGCGACCACCGAGATCTACACCGGAATCTAAACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
532(external_fwd_7)	AATGATACGGCGACCACCGAGATCTACACATGACTGATCACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC
533(external_fwd_8)	AATGATACGGCGACCACCGAGATCTACACTCAATATCGAACACTCTTTCCCTA CACGACGCTCTTCCGATCTCGTCAATTTTACGCAGACTATC

APPENDIX B

Table 7.2 Oligonucleotides use in Chapter 3

Oligo #	Oligo sequence
623(internal_fwd_R1_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAAGGAATTTTCCCAAGACCAGCATGG
624(internal_fwd_R1_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTGGAATTTTCCCAAGACCAGCATGG
625(internal_fwd_R1_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGGGAATTTTCCCAAGACCAGCATGG
626(internal_fwd_R1_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCGGAATTTTCCCAAGACCAGCATGG
627(internal_fwd_R2_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAACCAACTCATGGTGGATGTGTTTGG
628(internal_fwd_R2_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTCCAACTCATGGTGGATGTGTTTGG
629(internal_fwd_R2_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGCCAACTCATGGTGGATGTGTTTGG

630(internal_fwd_R2_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCCCAACTCATGGTGGATGTGTTTGG
631(internal_fwd_R3_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAAGTCATTGGCACTACAGATGTATGG
632(internal_fwd_R3_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTGTCATTGGCACTACAGATGTATGG
633(internal_fwd_R3_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGGTCATTGGCACTACAGATGTATGG
634(internal_fwd_R3_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCGTCATTGGCACTACAGATGTATGG
635(internal_fwd_R4_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAAGAAGTGTGTGAAAGATCAGAATGG
636(internal_fwd_R4_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTGAAGTGTGTGAAAGATCAGAATGG
637(internal_fwd_R4_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGGAAGTGTGTGAAAGATCAGAATGG
638(internal_fwd_R4_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCGAAGTGTGTGAAAGATCAGAATGG
639(internal_fwd_R5_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAATGCCTTATCCACACATCCTTATGG
640(internal_fwd_R5_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTTGCCTTATCCACACATCCTTATGG
641(internal_fwd_R5_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGTGCCTTATCCACACATCCTTATGG
642(internal_fwd_R5_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCTGCCTTATCCACACATCCTTATGG
643(internal_fwd_R6_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAAGCAGCTTGTACAGGATCAATATGG
644(internal_fwd_R6_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTGCAGCTTGTACAGGATCAATATGG
645(internal_fwd_R6_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGGCAGCTTGTACAGGATCAATATGG
646(internal_fwd_R6_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCGCAGCTTGTACAGGATCAATATGG
647(internal_fwd_R7_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAATGTATTGAGTCAGCACAAATTTGC
648(internal_fwd_R7_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTTGTATTGAGTCAGCACAAATTTGC
649(internal_fwd_R7_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGTGTATTGAGTCAGCACAAATTTGC
650(internal_fwd_R7_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCTGTATTGAGTCAGCACAAATTTGC
651(internal_fwd_R8_A)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNAACACCATGATGAAGGACCAGTATGC
652(internal_fwd_R8_U)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNTTCACCATGATGAAGGACCAGTATGC
653(internal_fwd_R8_G)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNGGCACCATGATGAAGGACCAGTATGC
654(internal_fwd_R8_C)	GCAAGCGGTGCGGCAGGAGGTCGTGNNNNCCCACCATGATGAAGGACCAGTATGC
418(internal_rev)	GCTGGACGACGCTGCACGGAGCTGCTTGGTGTGGCACGCTCCAGTTTCAG
419(external_fwd)	AATGATACGGCGACCACCGAGATCTACACGCAAGCGGTGCGGCAGGAGGTCGTG
420(external_rev_1)	CAAGCAGAAGACGGCATACGAGATAGCTTACGGCTGGACGACGCTGCACGGAGCTGC T
421(external_rev_2)	CAAGCAGAAGACGGCATACGAGATATCGCGATGCTGGACGACGCTGCACGGAGCTGC T
422(external_rev_3)	CAAGCAGAAGACGGCATACGAGATTACGGTCAGCTGGACGACGCTGCACGGAGCTGC

	T
423(external_rev_4)	CAAGCAGAAGACGGCATACGAGATCTCAAGGTGCTGGACGACGCTGCACGGAGCTGC T

VITA

Education

UNIVERSITY OF WASHINGTON, Seattle, WA, USA Sept. 2014 – March 2020

Ph.D. candidate, *Molecular & Cellular program*, Genome Sciences Department

SHANDONG UNIVERSITY, Jinan, China Sept. 2007 – July 2012

***Bachelor of medicine* (*MD equivalent*)**

Publications

1. **Wei Zhou**, Daniel Melamed, and Stanley Fields. Expanding the binding specificity for RNA recognition by a PUF domain (2020). *Manuscript in preparation*.

2. Junyue Cao, **Wei Zhou**, Frank Steemers, Cole Trapnell, Jay Shendure (2020). Characterizing the temporal dynamics of gene expression in single cells with sci-fate. (*Nature Biotechnology*, *in press*)

3. **Wei Zhou**, Michael Dorrity, Kerry Bubb, Christine Queitsch, Stanley Fields (2019). Binding and Regulation of Transcription by Yeast Ste12 Variants To Drive Mating and Invasion Phenotypes. ***Genetics***. 214. genetics.302929.2019. 10.1534/genetics.119.302929.

4. Zijian Tang*, Siyuan Dai*, Junyue Cao*, **Wei Zhou*,** Xiongtao Ruan, Huawen Li, Xin Wang, Stephen Sampson, and Chengkai Dai (2014). Suppression of the HSF1-mediated proteotoxic stress response by the metabolic stress sensor AMPK. ***EMBO J.*** (2015) **34**, 275-293 (**Co-first author**)