High throughput determination of sequence-function relationships in protein and RNA

David L. Young

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

Doctor of Philosophy

University of Washington
2016

Reading Committee:
Stan Fields, Chair
Charles Chavkin
Jay Shendure

Program authorized to offer degree:
Department of Genome Sciences

University of Washington

ABSTRACT

High throughput determination of sequence-function relationships in protein and RNA

David L. Young

Chair of the Supervisory Committee:
Professor Stan Fields, PhD
Department of Genome Sciences

As more individuals have their genomes sequenced, more genetic variation is discovered. The problem of interpretation of this variation has become intractable using established methods of linking phenotype to genotype, due to the low throughput of these methods and the exponential increase in newly discovered genetic variants. My graduate studies have revolved around the development and application of high throughput methods for functionally characterizing genetic variation within synthesized libraries of mutants. I have applied these methods in both yeast and mammalian cells to study the functions of a diverse set of proteins and an RNA: PAB1, a poly(A) binding protein necessary for translation of mRNA into protein; BRCA1, a DNA repair protein associated with familial breast cancer; SUP4oc, a yeast tyrosyl tRNA that has been altered to suppress the ochre stop codon; and MOR1, the human Mu Opioid receptor, a G-protein Coupled Receptor important for pain relief. Each assay yielded a genotype-phenotype map that showed the functional effect of all mutations present in the library. Through analysis of these maps, we found that as many as 20-30% of single mutations were tolerated without loss of function and that peripheral, core, and RNA-binding positions could be distinguished from each other by the pattern of fitness effects across amino acid substitutions. In addition, we found that evolutionary conservation will often fail to predict whether a position will tolerate mutation, demonstrating the importance of measuring function directly rather than relying on conservation alone. Further demonstrating this point, we found that models using functional data from these assays outperform current computational methods for predicting the pathogenicity of mutations in BRCA1. However, we also showed that the combination of evolutionary data and high throughput mutational data can be useful

for identifying sites where a protein interacts with other molecules, since almost all deleterious substitutions that are nevertheless present in one of a protein's homologs are likely involved in intermolecular interactions. Finally, by analyzing variants containing two or more mutations, we found a significant and varied role for epistasis in gene function, with intramolecular interactions being much more prevalent and detrimental in tRNA than in PAB1. A closer look at individual instances of epistasis in PAB1 and SUP4oc showed that certain mutations are epistasis hot-spots, rescuing or pulling down the fitness of many other mutations, and that positive epistasis can occur via changes in conformation that accommodate otherwise detrimental changes. High throughput mutagenesis screens are potentially useful for both basic and clinical research, and will likely be an integral part of deciphering the ever-growing collection of genetic data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Many people contributed to the results described in this dissertation. The Phizicky lab and especially Eric Phizicky and Michael Guy were outstanding collaborators, and it was really rewarding to explore a tRNA fitness landscape with experts who were in such a great position to make sense of some of it. The Fields lab has been an enjoyable and productive place to study during all of graduate school and I think everyone has at some point contributed to either data analysis or experimental design through frequent spontaneous discussions. I'm indebted to Russell Lo for facilitating it all by running such an uncommonly smooth lab operation. I especially want to thank Daniel Melamed and Lea Starita for enlisting me in two very interesting projects, and more importantly, for their friendship and mentorship.

I'd also like to acknowledge Tamas Ordog, who kindled my early interest in research, and my wife Laura, who I've leaned on more than anyone else. Finally, I want to thank Stan Fields, who is, in most ways, the ideal mentor. He can both foster creativity and troubleshoot experimental specifics with surprising efficiency, often both within the same five minute discussion. He's been a great role model whose attitude toward science I'll be attempting to replicate for the rest of my life.

Chapter 1. INTRODUCTION

Geneticists have been mapping the relationship between genotype and phenotype since before Avery, Macleod, and McCarty first demonstrated in 1944 that DNA is the heritable material (Avery *et al.*, 1944). There are many reasons why such maps are desirable. By finding the effect of a genetic change on an organism's development or behavior, we can gain mechanistic insights into biological function. For example, many biological processes require a number of interacting genes, use redundant genes for added robustness to mutation, or make use of pleiotropic genes, which affect other processes that may have previously seemed to be unrelated. A gene's involvement in organismal function can be found only by connecting a change in genotype to a change in function. Additionally, these maps can inform our understanding of the process of evolution, as the evolvability or robustness of a phenotype depends on the shape of these maps, with robust phenotypes having flatter maps with smaller mutational effects (Wagner, 2012; Stiffler *et al.*, 2015). Finally, knowledge of the effects of genetic variation is central to the idea of personalized medicine, in which a patient's unique set of genetic variants is used to help determine his or her susceptibility to disease and potential response to treatment (Simon and Roychowdhury, 2013).

For most of the last century, genotype-phenotype maps were created through forward genetic screens, in which random mutations were introduced into an organism's genome using radiation or mutagenic chemicals. Mating and backcrossing yielded the F2 generation, which was examined for interesting phenotypes. By crossing these new mutants to a collection of already characterized mutants, researchers could map the location of the new mutation based on its recombination frequency with the already-mapped mutations. This process, though effective at identifying genes involved in many different processes, may take months to years, depending on the organism. During the past several decades, the speed of characterizing mutational effects has increased somewhat through the use of reverse genetics, in which a known gene is targeted for mutation or deletion and the effects are examined. By dispensing with the need to localize a mutation, reverse genetics has the potential to be much faster than forward genetics, but each mutation still must be made and assayed separately.

Though some high value targets have been studied by mutating all positions to alanine in a method called alanine-scanning (Cunningham and Wells, 1989), it has remained impractical to mutate every position in a large protein to every alternative amino acid. As projects such as the 1000 Genomes project (1000 Genomes Project Consortium *et al.*, 2015) and the Exome Aggregation Consortium (Consortium *et al.*, 2015) collect sequence data from more and more individuals, it has become clear that

the speed of functional characterization of genetic variants via traditional mutagenesis techniques is insufficient to keep pace with the discovery of new variants. It seems likely now that every gene in the human genome has hundreds or even thousands of single base changes present somewhere in the population. To date, there are more than 97 million Single Nucleotide Polymorphisms (SNPs) catalogued in the public SNP database dbSNP (dbSNP build 144 www.ncbi.nlm.nih.gov/news/06-09-2015-dbsnp-build-144), and the growth of this database is still accelerating. Though the majority of this variation is rare, any given individual will have hundreds of such rare mutations, including 40-50 de novo mutations not present in either parent (Besenbacher *et al.*, 2015). Though it would be useful to know what all these rare variants do, obtaining this information has proven to be very difficult. Genome-wide association studies (GWAS) and linkage studies can be used to associate a SNP with a disease or phenotype, but GWAS are significantly underpowered to find associations for rare SNPs(Moutsianas *et al.*, 2015), and linkage studies, though they are the gold standard in disease association, are more laborious even than reverse genetics. These features mean that the vast majority of genomic variants—unique insertions, deletions, and substitutions in our DNA—have been inaccessible to functional annotation until recently.

The same improvements in next generation sequencing that led to the discovery of huge numbers of genetic variants across individuals have also allowed the development of deep mutational scanning (DMS), a new method for analyzing mutational effects at much higher throughput, allowing tens of thousands of mutations to be functionally characterized in a single experiment (Fowler and Fields, 2014).

The central feature that all DMS experiments share is that the change in abundance, or frequency, of each variant of a gene in a library of pooled variants depends on its corresponding activity. When the library is placed under an appropriate selective pressure, those variants functionally better equipped to withstand the selective pressure increase in frequency, while the less functional variants decrease in frequency. For essential genes, such as PAB1, the WT version of the gene can be knocked out so that the yeast depend on the variant library for growth. As a consequence, only those yeast that express functional variants of the protein are able to grow (Melamed, 2013). Differences in growth rate between yeast cells harboring diverse variants lead to different changes in frequency over time, and these changes can be measured by using high-throughput sequencing to count the numbers of each variant before and after selection. A number of variations of this experimental approach have been developed, using different organisms, selection conditions, and phenotypic outputs (Patwardhan *et al.*, 2009, 2012; Fowler *et al.*, 2010a; McLaughlin Jr *et al.*, 2012; Sharon *et al.*, 2012; Traxlmayr *et al.*, 2012; Forsyth *et al.*, 2013; Kim *et al.*, 2013; Roscoe *et al.*, 2013; Starita *et al.*, 2013, 2015; Firnberg *et al.*, 2014; Melnikov *et al.*, 2014; Thyagarajan and Bloom, 2014; Shalem *et al.*, 2015), but have so far been limited to yeast and phage.

Deep mutational scanning enables the functional characterization of not only every possible single mutant in most target genes, but also large numbers of double mutants, allowing for the first time the empirical study of the interactions between mutations in a real biological setting. This interaction, called epistasis, was previously studied mainly in silico in the context of computationally predicted RNA folds. Epistasis determines to what extent a particular phenotype can be modeled using the functional effects of single mutations alone, and it is an important potential hurdle in the prediction of phenotype from genotype.

My graduate work has been focused on expanding the types of systems in which DMS can be used, developing statistical and heuristic methods for analyzing DMS data, studying epistasis in several different genes, and using this method to study protein and RNA function in both yeast and human cell-lines.

1.1 POLY(A) BINDING PROTEIN

The yeast *Saccharomyces cerevisiae* Poly(A) Binding Protein (PAB1) has several features that make it a good target for deep mutational scanning. First, PAB1 was the first member identified (Dreyfuss *et al.*, 1988) of a family of RNA binding proteins containing the RNA Recognition Motif (RRM), which is one of the most common protein domains in eukaryotes and is found in all kingdoms of life. It is present, often in multiple copies, in as many as 2% of human genes (Maris *et al.*, 2005). As such, the RRM is involved in a wide array of essential cellular functions, including translation, intracellular signaling, and nearly all post-transcriptional processes (Lunde *et al.*, 2007). Though RRM domains display a range of binding mechanisms (Clery *et al.*, 2008), functional insights into an RRM-containing protein may be broadly applicable. Second, PAB1 is an essential protein that functions in both the nucleus and cytoplasm, where it helps in polyadenylation and translation initiation via its interactions with the poly(A) tail and other proteins including the eukaryotic Initiation Factor 4G (eIF4G). Since PAB1 is an essential protein, PAB1 variants can be assayed via growth, simplifying the selection. Third, the structure of PAB1 is known, allowing us to compare genetic interactions to structural interactions on a scale not previously possible. Finally, at least one of its 4 RRM domains is involved in interactions with both protein (eIF4G) and poly(A) RNA and limited mutagenesis had previously identified several residues involved in both these interactions. Coupled with the fact that PAB1 is present in all eukaryotes, this allowed us to compare the functional effects of genetic variation to evolutionary conservation and to show that the relationship between function and conservation differs at sites of molecular interactions. This finding highlights a major limitation in using evolutionary conservation in the computational prediction of the effects of genetic perturbation (Grantham, 1974; Ng and Henikoff, 2003). The deep

mutational scan of PAB1 was the first done on an essential gene and was the first to be done on a protein domain while still in the context of a larger protein. It provided estimates of important parameters of the genotype-phenotype map for an essential gene, including the distribution of functional effects of single mutations and the prevalence and magnitude of epistasis.

1.2 $tRNA^{Tyr}$

tRNAs are of fundamental importance to all life, as all life forms must translate mRNA into functioning proteins. In order to facilitate the translation from mRNA to protein, each tRNA must fold into a highly conserved 3-dimensional structure, allowing it to interact with its amino acid tRNA synthetase, with the ribosome, and with other proteins involved in translation, including elongation factors. Before becoming functional, each tRNA must also undergo a considerable amount of processing, including numerous types of base modifications. With all these structural requirements, and all the energy that goes into processing, it seems unlikely that a tRNA would tolerate many mutations. In support of this hypothesis, there are over 200 mutations in tRNAs that are known to cause disease in humans (Ruiz-Pesini *et al.*, 2007). On the other hand, there is also an appreciable amount of sequence diversity among tRNAs, especially in the D and TΨC loops, even among different tRNAs from the same species (Goodenbour and Pan, 2006).

In order to test the robustness of tRNA to mutation, we generated a library of mutant tRNA suppressors and adapted the DMS method for determining tRNA function. The anticodon of the tRNA recognizing the tyrosine-encoding codon UAC can be mutated so that it instead binds to the stop codon UAA, creating a suppressor tRNA (SUP4oc) that allows read-through of UAA stop codons. By introducing a GFP with a premature stop codon to the same cells that carry SUP4oc, we can determine the function of the suppressor by whether or not full length GFP is expressed. By introducing one SUP4oc mutant into each yeast cell and sorting the cells into bins based on GFP fluorescence, we can ascertain the function of each mutant. This same assay can be modified to study other aspects of tRNA function and to further narrow down the reasons for functional changes among tRNA variants by performing the flow-based selection in a different genetic or physical background. We used this approach to study temperature sensitivity by performing the assay at both 28C and 37C, and to study the Rapid tRNA Decay (RTD) pathway, a tRNA quality control system in yeast, by performing the assay in a yeast strain with an essential member of this pathway (met22) removed from the genome. These additional selections showed that the RTD pathway will target mutations throughout the tRNA, whereas only T-stem targets had previously been identified, and that temperature sensitivity cannot be perfectly predicted computationally, but that we can define a cutoff in the change in folding free energy above which nearly all mutations are temperature-sensitive.

1.3 BRCA1

Evidence for an inherited condition conferring an increased risk of breast and ovarian cancers was first identified in 1971 (Lynch and Krush, 1971), but it wasn't until 1990 that the laboratory of Mary-Claire King localized the gene responsible for this condition to chromosome 17q21 (Hall *et al.*, 1990). The breast cancer associated gene BRCA1 was cloned 4 years later and several different loss of function mutations were identified in families with hereditary breast and ovarian cancers (Miki *et al.*, 1994). Since then, it's been estimated that 5-15% of ovarian cancers and 20-25% of breast cancers are inherited (Lynch *et al.*, 2013). This translates to ~12,000-36,000 new cases of ovarian cancer and 334,000-417,000 new cases of breast cancer per year with a genetic cause (Ferlay *et al.*, 2015). 65-85% of familial ovarian cancer cases are due to mutations in two breast cancer genes (BRCA1/2), while the genetic causes of breast cancer are much more heterogeneous, with the most common high-risk variants in BRCA1/2, p53, PTEN, and MMR accounting for only around 25% of cases of familial breast cancer (Lynch *et al.*, 2013). For carriers of known loss of function mutations in BRCA1 in particular, the risks of ovarian cancer (54%) and breast cancer (67%) are much higher than in the general population (King *et al.*, 2003), and carriers tend to develop cancer much earlier in life. However, the risk of ovarian cancer can be lowered by around 80% by prophylactic salpingo-oophorectomy (Finch A *et al.*, 2006) and the risk of breast cancer can be lowered by nearly 100% by bilateral mastectomy (Heemskerk-Gerritsen *et al.*, 2013). Even without surgery, increased screening frequency can more than double the chance that cancer will be identified at an earlier, more manageable stage (Saadatmand *et al.*, 2014). For these reasons, genetic screening for a predisposition to breast and ovarian cancers is becoming more and more popular.

A recent call for the genetic screening of BRCA1 and BRCA2 mutations in all women over 30, regardless of family history, based on the efficacy of population wide screening of Askenazi Jews (King M *et al.*, 2014), has elicited a discussion of the costs and benefits of implementing population-wide screening for inherited genetic risk factors for cancers and other diseases (Yurgelun *et al.*, 2015; Foulkes *et al.*, 2016). Because of the genetic heterogeneity of breast cancer risk factors across populations, one of the chief concerns when considering implementing population wide screening is the potential to discover a large number of variants of uncertain significance (VUS), or new mutations for which the pathogenicity is unknown. Through a large effort, the rate of VUS detection in BRCA1 and BRCA2 tests has decreased from 10-40% in 2002 to 2-6% in 2013, depending on the population, requiring the individual assessment of over tens of thousands of VUS (Eggington *et al.*, 2014). However, only a minority of these assessments were made using co-segregation within pedigrees or the results of biochemical assays, with the majority being classified based on a logistic regression model using aspects of family history as covariates (Easton *et al.*, 2007). This model cannot classify rare variants and makes some assumptions about variant effect

that limit its accuracy. Though it is unclear whether the rate of novel VUS accumulation per test will stay the same as the test is applied more broadly, there will likely be tens of thousands of new VUS discovered when screening all 10 million women, aged 30-34, in the US. This number becomes even more intractable when expanding the tests to panels of genes or whole exome and whole genome sequencing. One strategy when implementing a population wide screen would be to ignore VUS, at least initially, which would increase the number of false negatives in those cases in which the VUS was truly pathogenic, but would mitigate the potential emotional distress that can be caused by receiving an equivocal test result. Such a strategy may differentially affect different populations, with less well studied and historically underserved populations receiving the majority of these false negatives (Hall *et al.*, 2009), and we would likely want a system in place for informing patients when or if a VUS is later determined to be pathogenic. Another strategy is to make a computational prediction of pathogenicity (Kumar *et al.*, 2009; Adzhubei *et al.*, 2010; Kircher *et al.*, 2014) at the time of discovery, but the accuracy of these methods can be inconsistent across genes and is likely not sufficient to guide clinical decisions.

Deep mutational scanning enables a third strategy: the prospective functional assessment of all possible mutations in genes that have been targeted for screening. We have shown that a model incorporating the measured functional effects of mutations made to the N terminal RING domain of BRCA1 predict damage to the DNA repair function BRCA1 better than any computational method. Furthermore, the results of these assays suggest that a lot higher proportion of VUS may be damaging to BRCA1 function (~16%) than some conservative estimates (1-2%)(Yurgelun *et al.*, 2015), though this result is in line with other estimates (20%) (Easton *et al.*, 2007). Though additional assays would be necessary to assess effects on transcript processing, these results show the potential utility of DMS for cancer and disease susceptibility genes that have selectable effects on cellular phenotypes. This DMS also provides the first data set in which a large number of single mutations are simultaneously assessed for their effects on two different biological functions, allowing us to determine the extent to which these functions are under simultaneous selection.

1.4 Mu Opioid Receptor

G Protein Coupled Receptors (GPCRs) are a diverse family of plasma membrane-bound proteins that all share 7 transmembrane helices, 3 intracellular loops, and 3 extracellular loops. The human genome encodes close to 800 human GPCR genes which are responsible for a large proportion of the cellular communication in our species (Venkatakrishnan *et al.*, 2013). Approximately 369 of these are non-sensory, making them current or potential drug targets. An estimated 40-60% of current therapeutic drugs target at least one GPCR, so advances in our understanding of signal transduction through GPCRs

have potentially widespread clinical ramifications (Lagerström and Schiöth, 2008; Unal and Karnik, 2012; Stevens *et al.*, 2013).

For any given GPCR, multiple signaling pathways might be activated and multiple mechanisms might lead to receptor internalization at different rates (Pupo *et al.*). Both the activation and desensitization of GPCRs have been shown to depend on both the cell type and the specific ligand. This phenomenon is often called "functional selectivity" or "biased agonism," and its molecular and structural basis is only just starting to be elucidated. The structural basis of such selectivity is difficult to examine, as crystallizing GPCRs has proved very challenging. In fact, the 2012 Nobel Prize in Chemistry was awarded to Lefkowitz and Kobilka for their research on GPCRs, which included considerable efforts into stabilizing GPCRs for crystallization (Lefkowitz *et al.*, 1970; Cherezov *et al.*, 2007). These advances in stabilizing GPCRs have allowed for a near exponential increase in the number of crystal structures, but there are still relatively few structures currently available, and it is unclear how exhaustively conformational states might be sampled through crystallographic techniques. To complement the structural data, I generated a library of all single mutants of the Mu Opioid Receptor and developed a set of related high throughput screening assays to probe function. By examining the functional impact of mutations in various locations of a GPCR on various aspects of GPCR function, I sought to not only identify residues important for all GPCR function, but also to identify residues which might be important in mediating functional selectivity in signaling or internalization events. Specifically, I hypothesize that a subset of mutations will have a differential impact on signaling and internalization in different genetic backgrounds and in response to different ligands.

So far, DMS experiments have been limited to non-mammalian systems, as there are considerable barriers related to plasmid copy number, integration efficiency, and growth rate in mammalian cells. The use of overlap extension PCR during library construction to control library uniformity and the use of Bxb recombinase, followed by a flow cytometry based selection, have enabled the integration and collection of large number of variants of the MOR into a modified human embryonic kidney cell line, but there is still significant room for improvement in integration efficiency.

Chapter 2. A DEEP MUTATIONAL SCAN OF THE POLY(A)-BINDING PROTEIN

Chapter 2 appeared in this form in the journal *RNA* (Melamed *et al.*, 2013). My contributions were in the processing, filtering, and scoring of the variants from the sequence data, as well as in the analyses of epistasis and evolutionary conservation. I contributed to the creation of Figure 2.2, Figure 2.3, Figure 2.5, Figure 2.6, Figure S2.7, Figure S2.8, and Figure S2.9.

2.1 ABSTRACT

The RNA recognition motif (RRM) is the most common RNA-binding domain in eukaryotes. Differences in RRM sequences dictate, in part, both RNA and protein-binding specificities and affinities. We used a deep mutational scanning approach to study the sequence-function relationship of the RRM2 domain of the *Saccharomyces cerevisiae* poly(A)-binding protein (Pab1). By scoring the activity of more than 100,000 unique Pab1 variants, including 1246 with single amino acid substitutions, we delineated the mutational constraints on each residue. Clustering of residues with similar mutational patterns reveals three major classes, composed principally of RNA-binding residues, of hydrophobic core residues, and of the remaining residues. The first class also includes a highly conserved residue not involved in RNA binding, G150, which can be mutated to destabilize Pab1. A comparison of the mutational sensitivity of yeast Pab1 residues to their evolutionary conservation reveals that most residues tolerate more substitutions than are present in the natural sequences, although other residues that tolerate fewer substitutions may point to specialized functions in yeast. An analysis of ~40,000 double mutants indicates a preference for a short distance between two mutations that display an epistatic interaction. As examples of interactions, the mutations N139T, N139S, and I157L suppress other mutations that interfere with RNA binding and protein stability. Overall, this study demonstrates that living cells can be subjected to a single assay to analyze hundreds of thousands of protein variants in parallel.

2.2 INTRODUCTION

The RNA recognition motif (RRM) is one of the most common protein domains in eukaryotes, encoded in ~2% of all human genes (Maris *et al.*, 2005). This ~90-amino acid domain is present in proteins with roles in post-transcriptional processes such as pre-mRNA processing, mRNA nuclear export, translational regulation, and mRNA decay(Mangus *et al.*, 2003; Erkmann and Kutay, 2004; Deschenes-Furry *et al.*, 2006; Kühn *et al.*, 2009). About half of the proteins containing an RRM have

multiple copies of this domain (Maris *et al.*, 2005; Clery *et al.*, 2008), with the spatial arrangement of the domains, their sequence variation, and the presence of auxiliary domains dictating the affinity, specificity, and function of these proteins (Lunde *et al.*, 2007).

A typical RRM folds into a four-stranded antiparallel β sheet, packed against two α helices, with RNA binding usually achieved by contacts made between the β sheet surface and a single-stranded RNA (Maris *et al.*, 2005; Clery *et al.*, 2008; Muto and Yokoyama, 2012). Two highly conserved motifs, RNP1 (consensus K/R-G-F/Y-G/A-F/Y-V/I/L-X-F/Y, where X is any amino acid) and RNP2 (consensus I/V/L-F/Y-I/V/L-X-N-L), in the central two β strands, are the primary mediators of RNA binding (Adam *et al.*, 1986; Swanson *et al.*, 1987; Dreyfuss *et al.*, 1988).

The poly(A)-binding protein (PABP) is a well-characterized RRM-containing protein (Dreyfuss *et al.*, 2002; Maris *et al.*, 2005; Lunde *et al.*, 2007; Muto and Yokoyama, 2012) and was the first member of the RRM family to be identified (Adam *et al.*, 1986; Sachs *et al.*, 1986). There are two major forms of PABP, which differ both in structure and in function. A nuclear poly(A)-binding protein (PABPN) is required for efficient polyadenylation of mRNA tails in the nucleus (Kühn *et al.*, 2009). A cytoplasmic poly(A)-binding protein (PABPC) plays roles in mRNA translation and decay, with each protomer associating with ∼27 nucleotides of poly(A) (Baer and Kornberg, 1983).

The *PAB1* gene of the yeast *Saccharomyces cerevisiae* encodes an essential cytoplasmic poly(A)-binding protein of 577 amino acids (Adam *et al.*, 1986; Sachs *et al.*, 1986). Pab1 consists of four tandem RRM domains that are highly conserved among cytoplasmic PABP members, as well as a proline-rich linker and a C-terminal domain (Adam *et al.*, 1986; Sachs *et al.*, 1986). The RRM domains associate directly with the RNA molecule, while the C-terminal region is not required for RNA binding or yeast viability (Sachs *et al.*, 1987; Burd *et al.*, 1991). In addition to poly(A) binding, all Pab1 RRM domains mediate protein–protein interactions (Kessler and Sachs, 1998; Yao *et al.*, 2007; Richardson *et al.*, 2012). In particular, binding of Pab1 RRM2 to the eukaryotic initiation factor 4G (eIF4G) (Kessler and Sachs, 1998) is presumed to promote the formation of a closed-loop structure between the mRNA cap and the poly(A) tail (Jacobson and Favreau, 1983; Wells *et al.*, 1998; Amrani *et al.*, 2008) and to stimulate mRNA translation (Tarun *et al.*, 1997; Imataka, 1998; Park *et al.*, 2011).

The modular arrangement of Pab1 RRM domains shows functional redundancy. Fragments composed of RRM1–RRM2, RRM2–RRM3, and RRM3–RRM4 can bind independently to RNA in vitro (Sachs *et al.*, 1987; Burd *et al.*, 1991). In vivo, yeast survive most Pab1 deletions that remove large parts from either single or two adjacent RRM domains (Sachs *et al.*, 1987), and a mutation in each RNP1 motif

of the four RRM domains is required to reduce poly(A) binding sufficiently to abolish growth of yeast (Deardorff and Sachs, 1997).

We sought to define the determinants of an RRM domain of the yeast Pab1 protein by the use of a method known as deep mutational scanning (Fowler *et al.*, 2010b; Araya and Fowler, 2011). This method allows a large number of mutant versions of a protein to be scored for function in a single experiment. It combines high-throughput DNA sequencing with a selection in which a physical association is maintained between each protein variant and the DNA that encodes it. The sequence analysis provides the frequency of each variant in an input population and in a population after selection, with this ratio serving as a proxy for the function of each variant (Fowler *et al.*, 2010b). We demonstrated that a Pab1 construct carrying the first three RRM domains was sufficient for near wild-type growth of yeast, yet was highly sensitive to a point mutation in RRM2. This result allowed us to generate plasmid libraries containing mutations in RRM2, and to score more than 100,000 unique variants, including 1246 with single amino acid substitutions, for their ability to support the growth of yeast. Using these data, we measured the contribution of each structural element in RRM2 to Pab1 performance, dissected the in vivo effects of mutations at known RNA-binding residues and other interaction sites, and identified non-RNA-binding residues essential to RRM2 function.

2.3 RESULTS

2.3.1 *Mutagenesis of the Pab1 RRM2 domain*

We sought to establish an in vivo assay for scoring the function of variants of the Pab1 RRM2 domain based on complementation of the *pab1*Δ mutation. We deleted the endogenous wild-type *PAB1* gene from the BY4741 strain, and because *PAB1* is essential, the cells were maintained via expression of this gene from a plasmid under the control of a tetracycline-off promoter (Figure 2.1A). We transformed these cells with a second plasmid that constitutively expressed a variant of *PAB1* (Figure 2.1A). Addition of a tetracycline analog (doxycyline) to the culture shut off the expression of the wild-type gene, making the cells completely dependent on the mutated *PAB1* for their growth.

We required a Pab1 construct in which single amino acid changes in RRM2 would affect the activity of the protein. Point mutations in one of the RRM domains, designed to disrupt the domain's ability to bind RNA, are suppressed by the redundant function of the other three RRM domains (Deardorff and Sachs, 1997). Therefore, we tested a series of Pab1 C-terminal truncations both for their ability to support cell growth and for their sensitivity to a single amino acid substitution, F170V, which disrupts RNA binding (Deardorff and Sachs, 1997). Like the full-length protein Pab1(1–577), a construct

lacking most of the C-terminal domain, Pab1(1–469), was sufficient for growth and was insensitive to the F170V mutation (Figure 2.1B). However, a further truncation, Pab1(1–343), which includes RRM1–RRM2–RRM3 and the N-terminal 25 amino acids of RRM4, resulted in good growth upon doxycycline treatment only when F170 was present; the substitution F170V in this construct resulted in almost no growth (Figure 2.1B). In liquid culture, cells carrying the Pab1(1–343) fragment grew at a slightly decreased rate, which might be due to the loss of RRM4 rather than to the absence of the C-terminal region, as deletion of the C-terminal region alone did not affect growth (Figure 2.1B, bottom). Based on these observations, we chose the Pab1(1–343) fragment as the scaffold into which mutations would be introduced. To avoid unwanted PCR amplification of the wild-type RRM2 domain encoded on the tetracycline-off promoter plasmid, we introduced a total of 18 synonymous changes including BamHI and XbaI sites in the eight codons N-terminal and C-terminal to RRM2, and designated this construct as Pab1(1–343BX). These mutations had no effect on growth (Figure 2.1B).

We created three separate libraries from DNA oligonucleotides that were made double stranded and cloned into the Pab1(1–343BX) construct (Figure 2.1C). Each library spanned 75 bases (i.e., 25 amino acids) in RRM2, with an average of three mutations per variant. Yeast carrying each one of the library pools were grown to logarithmic phase and then diluted into doxycycline-containing media. We collected samples before (input) and after 22 h of growth in the presence of doxycyline (selected), extracted plasmids, PCR amplified the segment that had been mutated, and carried out sequence analysis. The enrichment score for each variant, based on the change in frequency from input to selection, serves as a proxy for the function of the variant (Figure 2.1A).

Enrichment scores were generated for hundreds of thousands of DNA and protein variants (Supplemental Table S2.1). Input read counts ranged from a single read for variants with multiple base substitutions to tens of thousands of reads for variants with a single-base substitution. Assuming that most synonymous mutations have a negligible effect on Pab1 activity in this assay, we used the enrichment score distribution of ~5000 synonymous variants (carrying either single or multiple synonymous substitutions) to assess the effect of input read depth on the reliability of the enrichment scores. Based on a variance cutoff of 0.25 (Figure S2.7), we required a read depth of at least 40 input reads for inclusion of a variant in further analysis. This cutoff provided data for 110,745 protein variants, including 1246 single amino acid substitutions (~83% of all possible ones in each library); 39,912 double amino acid substitutions (~11% of all possible ones in each library); and many other variants with three or more mutations. The enrichment score distribution of all variants (normalized to the wild-type enrichment score) revealed that, in general, most mutations were deleterious for RRM2 function. Unlike variants with missense mutations, variants with synonymous mutations had enrichment scores that were concentrated

around the wild-type score (Figure S2.8). Assuming that neutral variants with missense mutations have an enrichment score distribution similar to that of synonymous mutants, we used the enrichment score distribution of synonymous variants to calculate an empirical false discovery rate (FDR) of neutral variants among variants carrying missense mutations. For all three segments, the distribution of synonymous variants suggested that all missense variants with an enrichment score >1 are likely to be neutral (Figure S2.8). For variants with an enrichment score <1, the FDR of neutral variants dropped sharply as enrichment scores decreased (Figure S2.8), with an average estimate of ~25%, 8.5%, and 3.25% of variants with $\log_2$ enrichment scores of −0.5, −1.0, and −2.0 being neutral, respectively, for the three library segments. These distributions indicate that low enrichment scores arise mostly from the failure of Pab1 to function rather than from stochastic variation in measurements. After correcting for the enrichment score distribution of neutral variants with missense mutations, we estimated the fraction of variants carrying deleterious mutations (i.e., enrichment scores <1) to be ~83%, 81%, and 63% of total variants for libraries 1, 2, and 3, respectively.

2.3.2 *Effect of single amino acid substitutions*

We generated a mutational sensitivity map for single mutations that shows the enrichment scores of 1190 missense and 56 nonsense mutations (Figure 2.2A; Supplemental Table S2.2). Several observations suggest that these enrichment scores correlate with the function of the Pab1 RRM2 domain. First, missense mutations led to a wide range of growth, whereas nonsense mutations uniformly resulted in extremely poor growth (median enrichment score of 0.06). Second, proline was the most harmful missense substitution (median enrichment score of 0.22) compared with all other missense substitutions (median enrichment score of 0.8), consistent with the disruptive nature of a proline mutation on α helices and β sheet structures (Chou and Fasman, 1978). Third, we found a good correspondence between the effect of previously characterized RRM2 mutations and enrichment scores; RRM2 F170V (enrichment score of 0.04) reduces binding of the protein to poly(A) by >97% (Deardorff and Sachs, 1997) and RRM2 K166Q (enrichment score of 0.65), if combined with mutations to the equivalent residues in the other three RRMs, reduces binding to poly(A) by >70% (Deardorff and Sachs, 1997).

While the enrichment scores of the single amino acid substitutions indicate that most mutations were deleterious for RRM2 function, a few mutations had enrichment scores that were greater than that of the wild type. In particular, the enrichment score for Q194C was 2.9. However, we measured the growth rate of yeast cells carrying Q194C and found that it was the same as those carrying the wild-type version

(data not shown). This observation agrees with our finding that enriched variants follow the distribution of synonymous mutants and therefore are likely to be neutral (Figure S2.8).

The distribution of the single amino acid substitution enrichment scores along the sequence and structure of RRM2 points to the β sheet as the element most sensitive to mutation (Figure 2.2A,B). In particular, strands β1 and β3, which carry the RNP motifs, show the highest sensitivity to mutation (both with a median enrichment score of 0.09), suggesting that RNA binding mediated by these two motifs is the most important in vivo function of the RRM2 domain. Strands β2 and β4, which assist poly(A) binding in vitro (Deo *et al.*, 1999), appear to contribute less to this function (median enrichment scores 0.38 and 0.90, respectively). In helices α1 and α2, residues with side chains oriented toward the core showed greater sensitivity to mutation than surface residues (Figure 2.2B). Additionally, helix α2 was less sensitive to mutation (median enrichment score 0.90) than helix α1 (median enrichment score 0.76). In particular, mutations at residues 180–181 (KE) and 184–186 (DAL), which are part of the eIF4G binding site in helix α2 (Otero, 1999), had only minor effects on cell growth (median enrichment score 0.89). In vitro, mutations at these sites result in complete loss of eIF4G binding and diminished mRNA translation (Otero, 1999), but in vivo, a weak affinity of RRM1 for eIF4G may compensate for loss of eIF4G binding by RRM2 (Kessler and Sachs, 1998; Richardson *et al.*, 2012). Lastly, of the loop regions, L2, which connects helix α1 to strand β2 by a four amino acid turn, was the least resistant to mutation (median enrichment score 0.19), making it the most sensitive element after strands β1 and β3.

2.3.3 *Clustering mutation sensitivity profiles identifies structurally related residues*

We clustered both RRM2 positions (along the *x*-axis) and the substituting amino acids (along the *y*-axis) by the similarity in their sensitivity profiles (Figure 2.3A). This clustering grouped together amino acids that have similar chemical properties such as hydrophobic, aromatic, positively charged, and negatively charged. That replacements of amino acids with similar ones resulted in correspondingly similar enrichment scores argues that the deep mutational scanning assay provides a sensitive and accurate readout of mutational sensitivity.

The clustering revealed three major groups of RRM2 positions that have distinct profiles. The first group showed sensitivity to nearly all amino acid substitutions (Figure 2.3A, cluster I). Seven of the 11 residues in this group (N127, F129, K131, S154, F168, F170, and H172) interact directly with RNA, based on the structure of the human protein in complex with poly(A) (Deo *et al.*, 1999). Three other residues whose human equivalents were shown to associate with poly(A), N132, Y197, and P200, did not cluster with this group, displaying lower sensitivities to mutations. N132 and P200 tolerated multiple

amino acid substitutions without affecting Pab1 function, while Y197 showed moderate sensitivity to mutations and could not be substituted by any amino acid without reducing function by >10%. Unlike the mutation-sensitive RNA-binding residues that make substantial contacts with the adenine bases and the backbone phosphates, these three residues are situated more peripherally to the RNA path (Figure 2.3B, left and right). N132 is part of the RNP2 motif (Maris *et al.*, 2005) and is highly conserved. The equivalent residue in the human protein forms a hydrogen bond with an adenine base (Deo *et al.*, 1999), but this base is also specified and stabilized by the residues equivalent to K131 and F168, which also form hydrogen bonds with the RNA phosphate groups that surround the adenine base. Similarly, the residues in the human protein equivalent to Y197 and P200 make contacts with another adenine base, which also interacts with three other residues (Deo et al. 1999). Thus, the minimal effect of substitutions to N132, Y197, and P200 may be due to their limited contributions to RNA binding relative to other residues that associate with the same adenine bases.

The second group of clustered residues showed sensitivity to most amino acid substitutions except for hydrophobic ones (Figure 2.3A, cluster II). Ten of the 12 residues in this group are aliphatic, with most of them inaccessible to solvent (average Accessible Surface Area = 2.6 $Å^2$) and constituting part of the RRM2 core structure (Figure 2.3B, middle). I152 in loop L2 was an exception as it showed the highest sensitivity to hydrophobic substitutions and the highest solvent accessibility area (ASA = 24.2 $Å^2$) relative to the other aliphatic residues within this group. These observations suggest a specialized role for I152 that requires features other than hydrophobicity. All other aliphatic residues that were not clustered with this group were more exposed to solvent (average ASA = 51.1 $Å^2$) (Figure 2.3B, middle). Taken together, these results show that mutational profiles can accurately distinguish different classes of aliphatic residues.

We found K156 to cluster together with core residues, with leucine, isoleucine, and arginine being its least detrimental replacements, although none of them was able to fully compensate for loss of K156. These results suggest that both polarity and the nonpolar neck of K156 are required for its function (Dyson *et al.*, 2006). In the human PABP-1 RRM2, the residue equivalent to K156 is involved in packing interactions between RRM1 and RRM2, which stabilize the RNA-binding trough (Deo *et al.*, 1999). Compaction of RRM2 against RRM1 buries the large surface at their interface from solvent access, likely providing an explanation for why hydrophobic residues at this position are tolerated.

The third cluster is composed of the remaining positions, which show moderate to low mutation sensitivity. These residues are found mostly at the outer shell of RRM2 (Figure 2.3A, cluster III) and contribute to Pab1 activity by functions that do not produce a clear mutational profile.

2.3.4 *Mutation at G150 destabilizes RRM2 structure*

Four positions other than the RNA-binding residues were clustered as highly sensitive to most amino acid substitutions (Figure 2.3A, cluster I). Of these, G169 and F173 are adjacent to RNA-binding residues in RNP1 and, therefore, the deleterious effect of mutations at these positions could result at least in part from interference with RNA binding. Of the other two, F149 is situated at the end of helix α1 and G150 is in a β-turn structure that follows this helix, remote from the RNA-binding trough and in close proximity to the side chain of F173 (Figure 2.4A). The high sensitivity may suggest an additional function for F173, together with F149 and G150, that does not involve RNA binding.

Based on their sensitivity to mutation, we examined the roles of F149, G150, G169, and F173 in more detail. From the RRM2 structure (Figure 2.4A), we hypothesized that F149 and G150 act with F173 to stabilize the RRM structure by bridging helix α1 and helix α2 to bury the hydrophobic core from solvent. Mutations at these positions should therefore destabilize RRM2, a phenotype that might be suppressed at low temperature. Yeast expressing a variant with any of the F149T, G150T, G169M, F173A mutations or with F170V, which interferes with RNA binding (all with enrichment scores <0.1) did not grow at 30°C (Figure 2.4B). However, the growth defects due to the F149, G150, and F173 mutations were suppressed at 20°C (Figure 2.4B). In support of the role of G150 in RRM2 stability, the G150T protein showed higher sensitivity to protease cleavage and a different cleavage pattern from that of the wild-type protein (Figure 2.4C). In contrast, the protease sensitivity and cleavage pattern of the RNA-binding defective mutant F170V were similar to that of the wild-type protein. A comparison of RRM sequences from various RRM-containing proteins present in the protein database revealed that F149, G150, and F173 are highly conserved (Figure 2.4D), with G150 (glycine in 102 of the 119 RRM sequences) and F173 (phenylalanine or tyrosine at 99 of the 119) having the highest conservation score among all RRM residues. Taken together, the temperature-sensitive phenotype, protease sensitivity, and conservation suggest a role for G150 and the two phenylalanines in stabilizing RRM structure.

2.3.5 *A comparison of functional data to evolutionary conservation*

The mutational sensitivity and the evolutionary conservation of a residue are strongly correlated (Bottema *et al.*, 1991; Stone and Sidow, 2005), but discordances between these two properties may help to characterize function. For example, a residue with low evolutionary conservation, but high mutational sensitivity, may participate in a function specific to the organism being tested, or these discordant properties may suggest that the assay conditions were harsher than the forces applied by natural selection.

As a first approach, we sought to compare neutral amino acid substitutions from this study with naturally occurring substitutions in other PABP sequences. The degree of tolerance to homologous substitutions may indicate the functional constraint on each residue in yeast. To this end, we created function-based logo plots for the four β strands of the RNA-binding surface and flanking residues (Figure 2.5A), which display wild-type residues and amino acid substitutions that resulted in a neutral effect on Pab1 RRM2 function (defined as >0.95 of the wild-type residue performance, see Materials and Methods). The logo plots show that of the RNA-binding residues, N127, K131, S154, F168, H172, and Y197 could not be replaced by any other amino acids without loss of Pab1 activity. Tyrosine substitution of F129 and F170 and multiple substitutions of N132 and P200 were tolerated.

A comparison of the function-based logo plot to the logo plot derived from 306 poly(A)-binding protein homologs (listed in Supplemental Table S1.4) reveals a general agreement between the two, with 24 out of the 32 positions sharing at least one amino acid of the two most frequent amino acids that occupy these positions in each plot. However, most positions of the yeast Pab1 functionally tolerated more substitutions than are present in the natural sequences, a feature that might be due to the limited selective force applied on yeast Pab1 function in this assay.

For 11 positions, the amino acid in the yeast RRM2 sequence was different from the most frequent amino acid in the naturally occurring sequences. Of these, in nine cases (D151N, S155C, I157V, G193D, I196V, A199G, P200H, Q194K, and E195K) a change of the yeast amino acid to the most frequent amino acid in natural PABP sequences had a neutral effect. These observations suggest that adaptation of poly(A)-binding proteins to various eukaryotes involves minor functional consequences of single amino acid substitutions. However, some substitutions to amino acids that are present in natural PABP proteins were less tolerated by yeast Pab1. For example, the RNA-binding residue H172 is a glutamine in some poly(A)-binding proteins (Figure 2.5A), but H172Q cannot fully complement for the loss of H172 (enrichment score 0.73). This result indicates that poly(A)-binding proteins with histidine at position 172 differ in how they bind RNA compared with PABP proteins with glutamine at this position.

To further study the correlation between our functional data set and the evolutionary record, we sought to compare the degree of conservation for each RRM2 position. To this end, we used the "property entropy" method (Capra and Singh, 2007), which scores the variation (Shannon entropy) at each position in a multiple sequence alignment with respect to the stereochemical similarity between the amino acids that populate it (Williamson, 1995; Mirny and Shakhnovich, 1999; Capra and Singh, 2007). This measure allowed us to assess the conservation of natural as well as engineered variants by applying identical criteria for the two data sets without introducing corrections that are commonly applied by other methods to score conservation of natural sequences (such as phylogenetic tree construction or amino acid

background frequencies). We found a moderate correlation in property entropy (R^2 = 0.43) between naturally occurring and engineered sequences (Figure 2.5B). Specifically, for most residues, the higher the evolutionary conservation, the higher the functional conservation was. As found for the logo plots of four of the β strand sequences, most positions could tolerate more mutations than would be expected by their evolutionary conservation.

Color-coding the ratio between the functional conservation score and the evolutionary conservation score on the human RRM2 structure allowed us to identify regions whose scores do not match (Figure 2.5C). This comparison reveals that the RNA-binding residues N127, F129, K131, S154, F170, H172, and Y197, as well as some of the adjacent residues, G126, I128, I130, L153, S155, V171, V196, and V198, are functionally more conserved than suggested by their evolutionary conservation. A second region in which function shows greater conservation than evolutionary conservation encompasses certain residues that face the RRM1 interface. Contacts between helix α2 of RRM1 and helix α1 of RRM2, and between strand β4 of RRM1 and strand β2 of RRM2, stabilize the RNA-binding trough formed by the two tandem β sheets and facilitate the binding to eIF4G and poly(A) (Deo *et al.*, 1999; Safaee *et al.*, 2012). Strand β2 residues that mediate these interdomain interactions (K156 and L153) are functionally more important than evolutionary conservation would suggest. A third region that shows that divergence between function and evolutionary conservation comprises the eIF4G-binding site (Otero, 1999) including residues E181 and A185.

For these three regions, the high ratio of functional conservation to evolutionary conservation may reflect a sensitized activity of Pab1 in this assay. For example, decreased RNA binding due to lack of RRM4 may result in oversensitivity to mutations that further degrade this activity, either directly (such as sensitivity to mutations to RNA-binding residues) or indirectly (such as sensitivity to mutations that destabilize the trough formation between the RRM1 and RRM2 RNA-binding surfaces). Alternatively, the high functional conservation to evolutionary conservation ratio may suggest a specialized function for these residues in yeast that cannot be complemented by equivalent residues from other species. The failure of human PABP-1 segments to complement for eIF4G binding (Otero, 1999) supports this possibility.

2.3.6 *Epistatic interactions between two mutations*

An epistatic interaction between two mutations describes an observed gain or loss of function of a double mutant that exceeds predictions based on the functional consequences of each of the constituent single mutations alone (Horovitz, 1996). To study epistasis in Pab1 RRM2, we used the enrichment scores of 39,609 trios for which the scores of the two single mutants and the corresponding double mutant

were available. We used a product interaction model previously applied to large-scale mutational data (Fowler *et al.*, 2010b; Araya *et al.*, 2012) such that in the absence of epistasis, the observed enrichment score of the double mutant should equal the product of the two single mutants' enrichment scores. The observed enrichment scores correlated well with the products of the single mutants ($R^2 = 0.76$), suggesting that, in general, no substantial epistasis occurs in double mutants (Figure S2.9A). We used LOESS function to correct for input read counts effects (Figure S2.9B) and selected those double mutants whose epistasis scores exceeded two standard deviations from the mean as candidates for displaying strong epistatic interactions. Of the 39,609 double mutants, 411 showed positive epistasis (i.e., performance higher than expected) and 1444 negative epistasis (i.e., performance lower than expected).

The distribution of the spacing along the primary sequence between the mutations in double mutants revealed enrichment for short distances in the case of mutation pairs that showed either positive or negative epistatic interactions (Figure 2.6A). Variants with positive epistasis showed a preference for the inclusion of interacting mutations that lie no more than three residues apart. Variants with negative epistasis also showed a preference for short distances between interacting mutations, with zero to five residues apart being the most significant range (wilcoxon *P*-value = 0.0024). From these distributions (Figure 2.6A), we estimate that primary sequence proximity is responsible for ~8.6% of the variants showing positive epistasis and 7.4% of those showing negative epistasis.

We also examined epistasis with respect to the distribution of physical distance using the structure of the human PABP-1 RRM2 domain as a proxy for the yeast domain structure. Residues can be in close physical distance either because they reside nearby in the primary sequence or because they are distant in the primary sequence and come together due to protein folding. To eliminate effects due to sequence proximity, we followed the distribution of physical distances between mutations that are five residues or more apart. This distribution revealed a similar association between short distance and epistasis. In particular, positively interacting mutations showed enrichment for distances shorter than ~12 Å between the centers of mass of the two residues, and negatively interacting mutations showed enrichment for distances between ~10 and 15 Å between these centers of mass (Fig. 6B). Based on these distributions, we estimate the upper limit of 12 Å accounts for ~17% of variants displaying positive epistatic interaction, and the range of 10–15 Å accounts for ~7% of variants displaying negative epistatic interaction. Although the two physical distances slightly overlap, no pair of residues was shared between the two groups, suggesting that physical distance acts on different sets of residues with respect to positive and negative epistasis.

Specific residues and mutations serve as hot spots for epistatic interaction (Hinkley *et al.*, 2011; Araya *et al.*, 2012). In particular, substitutions N139S and N139T were responsible for >30% of the

positive epistasis interactions in segment 1. Though N139S and N139T as single mutations had only a slightly negative effect on function (enrichment scores 0.88 and 0.86, respectively), these mutations partly rescued deleterious mutations at RNA-binding sites (e.g., F129I and K131N), core residues (e.g., I128N and I130S), and G150 (e.g., G150A and G150S). Similarly, I157L (enrichment score 0.96) was present in >13% of positive epistatic interactions in segment 2, and suppressed deleterious mutations in an RNA-binding residue (F168S, F168I, and F168C) and in RRM1–RRM2 interface residues (K156R, K156M, and L153V).

In the human RRM2 structure, the residues corresponding to N139 and I157 form a hydrogen bond between the carbonyl group of an asparagine side chain and the backbone amino group of a valine that connect helix α1 and strand β2 (Figure 2.6D). N139S and N139T may slightly destabilize the association between the two elements by weakening the presumed hydrogen bond between N139 and I157. I157L may cause a similar outcome by destabilizing the hydrophobic interactions between loops L1 and L3. Although as single mutations these had only a minor effect on RRM2 function, they may confer flexibility to the RRM2 structure to allow this domain to adjust its structure to accommodate other mutations that interfere with RNA binding or protein stability.

While the enrichment score of the G150T mutation, which destabilizes RRM2 (Figure 2.4), was too low to detect negatively interacting mutations, we found G150A and G150S substitutions comprised ∼15% of all negative epistatic interactions in segment 1. These two substitutions had only a moderate effect on Pab1 function (enrichment scores 0.62 and 0.66, respectively) relative to other substitutions at this position, probably due to the small size of their side chains. G150A and G150S negatively interacted with a similar set of mutations (*P*-value of Fisher exact test = $1.9e^{-23}$) listed in Supplemental Table S1.5. This set of mutations may enhance the destabilizing effect of G150 mutations, pointing to loop 1, which carries most of the G150A and G150S interacting mutations, as an important element for RRM stability. Moreover, mutations at N132, an RNA-binding residue in loop 1, negatively interacted with G150A and G150S, suggesting an additional role for N132 in supporting the integrity of RRM2 structure.

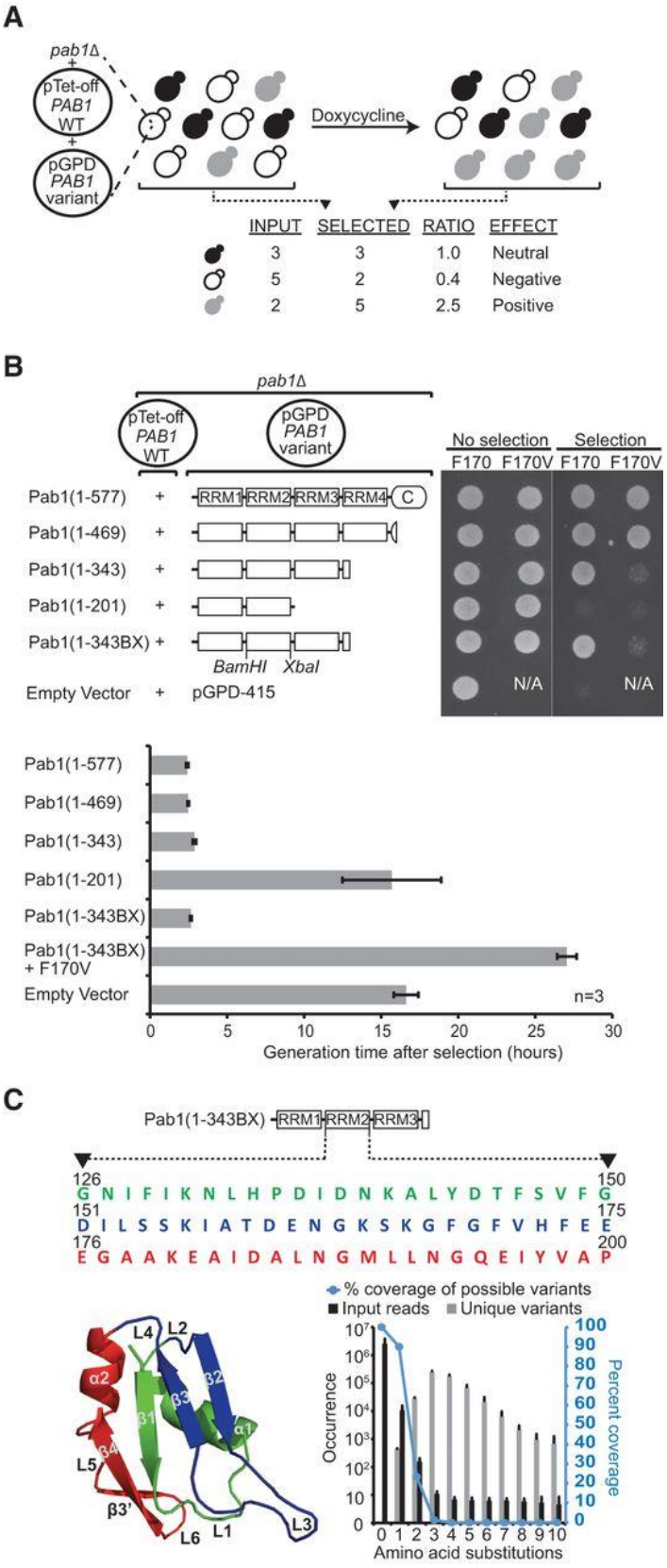


Figure 2.1 Experimental design of the deep mutational scan for the Pab1 RRM2 domain.

(*A*) Protocol to assess the effects of RRM2 mutations on the in vivo function of Pab1. *pab1*Δ cells carry two plasmids, one expressing the full-length Pab1 protein under a tetracycline-off promoter (pTet-*PAB1* WT) and the other expressing one of many variants of Pab1 from a constitutively active promoter (pGPD-*PAB1* variant). The cells are grown to logarithmic phase in liquid culture, and a tetracycline analog (doxycycline) is added to the media. Cells expressing variants of Pab1 that cannot fully complement the loss of the wild-type protein grow slower than cells expressing neutral variants of Pab1. Sequencing the mutated fragments of the variant population before (input) and after selection (output) can be used to quantify these effects as the ratio of frequencies in each pool for each variant. (*B*) Selection of a Pab1 fragment that displays growth-rate sensitivity to a single point mutation. *pab1*Δ cells carrying the two plasmids specified in *A* with the variant plasmid expressing truncated forms of Pab1 with or without F170V substitution were tested for growth in the presence of doxycycline (20 μg/mL) on plates (*top*) and in liquid culture (*bottom*). Generation time was calculated starting from 8 h after doxycycline addition to eliminate cell divisions due to residual Pab1 activity. Note that cells carrying an empty vector grow at a low rate, probably due to leaky expression of the wild-type protein. (*C*) RRM2 mutagenesis. Shown are the Pab1(1–343BX) construct and the RRM2 sequence that was mutagneized. Each colored sequence and the corresponding elements on the structure of the human RRM2 domain (PDB_ID 1CVJ) represents a 25-amino acid-long RRM2 sequence that was doped with an average of three DNA base substitutions per variant. The graph at the *bottom right* depicts averages values of the listed properties of the three input libraries with respect to variants carrying a specified number of amino acid substitutions.

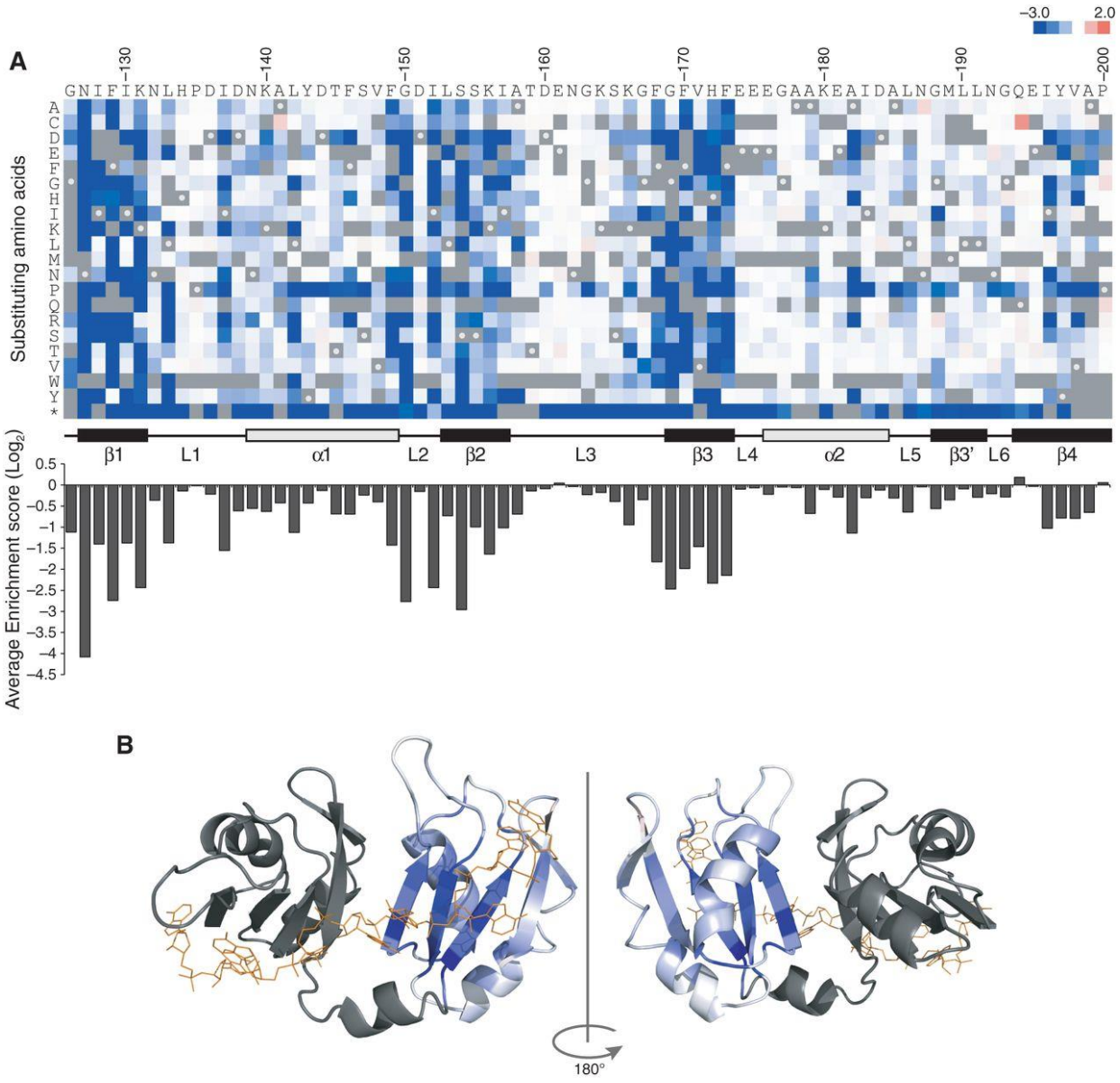


Figure 2.2 Effect of single amino acid substitutions on the in vivo function of the Pab1 RRM2 domain.

(*A*) A heat map displaying the enrichment scores (log_2 transformed) of single amino acid substitutions in the RRM2 domain. Each column represents an RRM2 sequence position and each row a substitution to a specific amino acid. An asterisk designates the row of nonsense mutations. Color ranges from blue for the most deleterious mutations to red for the most beneficial ones. Substitutions that were not sequenced in the input or selected pools or that were eliminated by subsequent quality filtration steps are shown in gray; wild-type residues are marked with white dots. The secondary structure of the RRM2 domain aligned to the sequence is shown *below* the heat map as well as the average enrichment score for each position. (*B*) The average enrichment scores are projected on the crystal structure of the human RRM2 (PDB_ID 1CVJ). RRM1 and the connecting linker are shown in black and the poly(A) RNA in orange.

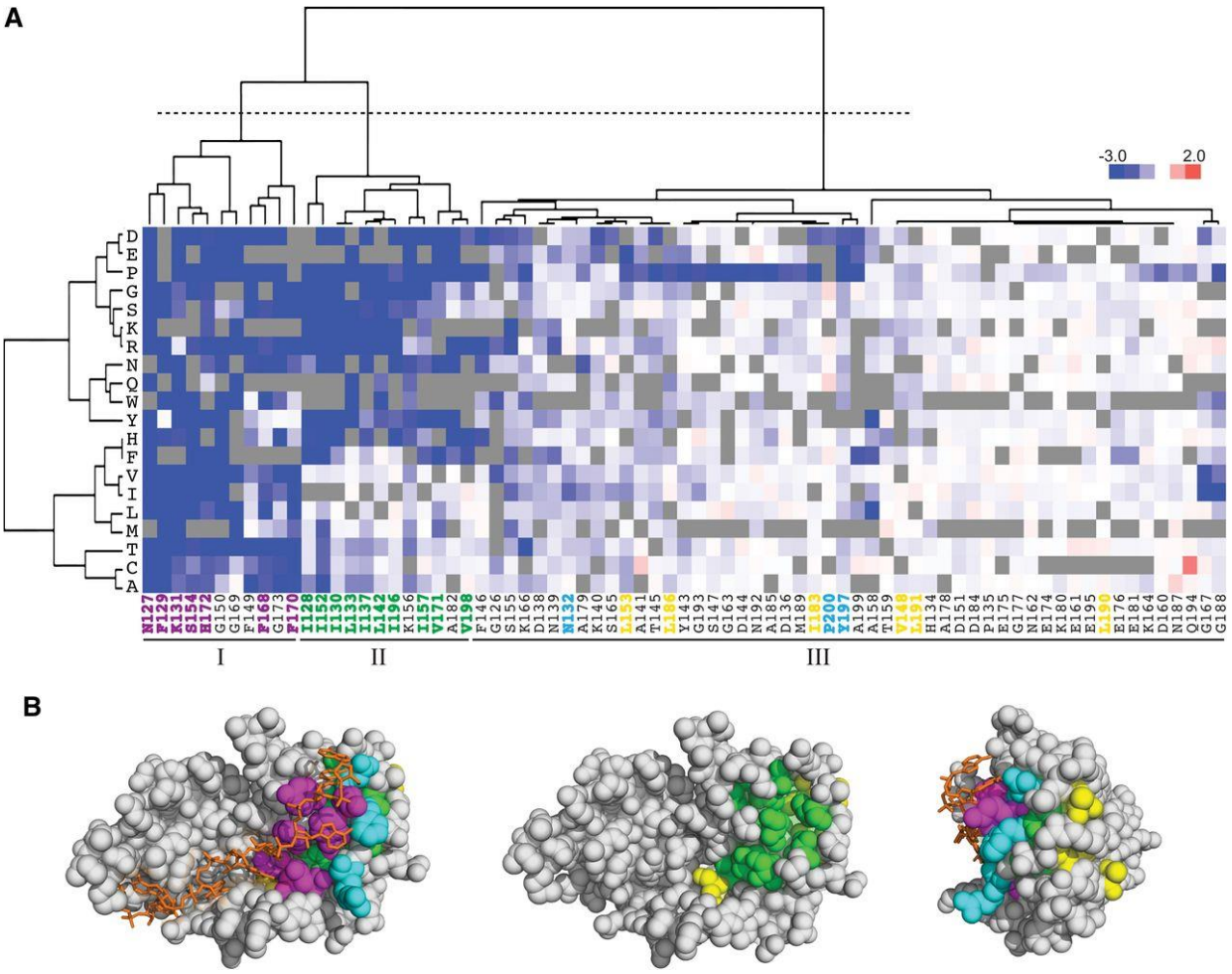


Figure 2.3 Clustering the effects of single amino acid substitutions groups structurally related residues.

(*A*) Pab1 RRM2 positions and substituting amino acids were clustered based on enrichment score values and color coded as shown in Figure 2.2. The dotted line creates three clusters of RRM2 residues. Positions corresponding to RNA-binding residues in the human RRM2 are colored for their clustering (magenta) or lack of clustering (cyan) to group I. Positions corresponding to aliphatic residues are colored for their clustering (green) or lack of clustering (yellow) to group II. (*B*) Clustered residues displayed on the structure of the human RRM2 domain (PDB_ID 1CVJ) and color coded as in *A*. (*Left*) RNA-binding surface of RRM1–RRM2 with poly(A) shown in orange; (*middle*) RNA-binding residues with the poly(A) and the RNA-binding residues removed to observe the RRM2 core residues; (*right*) image as at *left* rotated 90° at the horizontal axis.

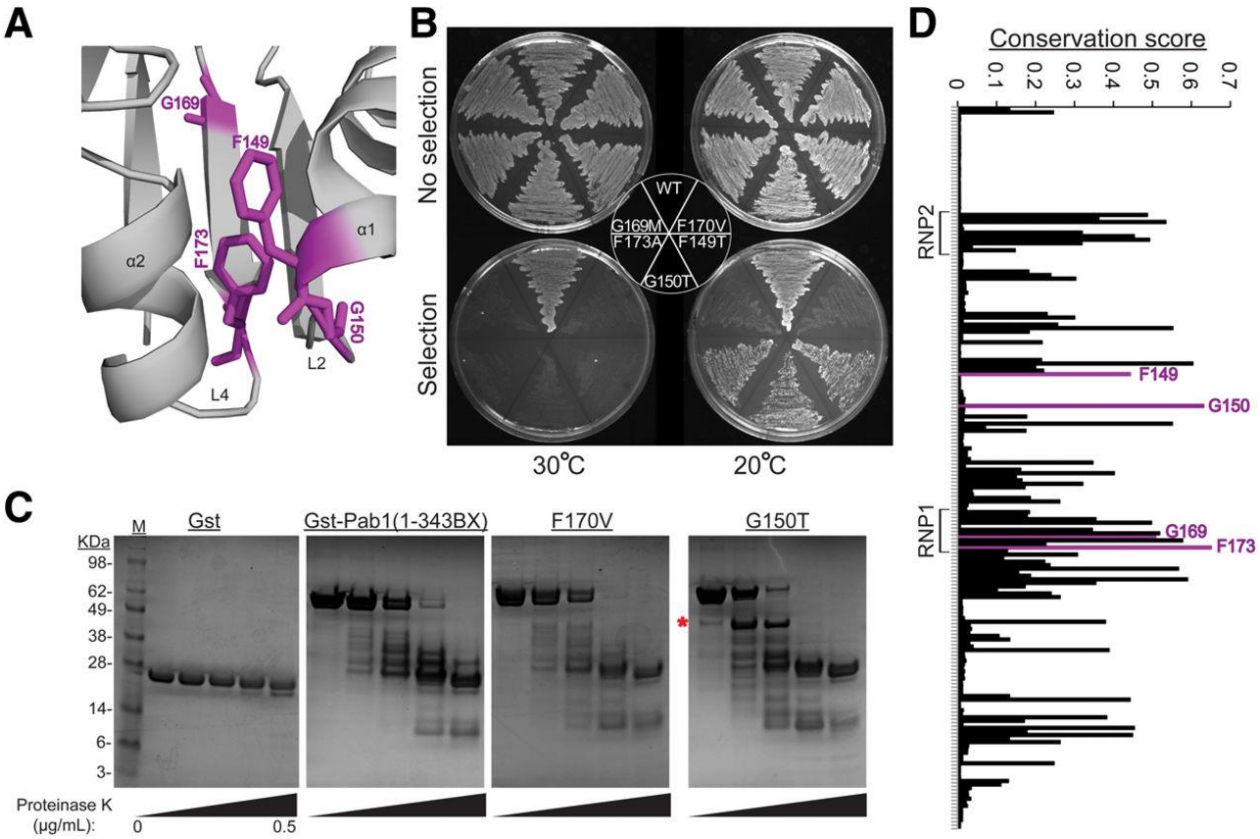


Figure 2.4 Mutational sensitivity of residues in the helices α1 and α2 interface suggests a role for these residues in Pab1 stability.

(*A*) Three of the four residues highly sensitive to mutation but not RNA binding (F149, G150, and F173) are found in close proximity in the RRM2 opening between the two helices. (*B*) Cold-suppressible phenotype of mutants carrying G150T, F149T, and F173A. *pab1*Δ cells carrying the two plasmids shown in Figure 2.1A with the variant plasmid expressing the specified mutations from Pab1(1–343BX) were grown in the absence or in the presence of 5-fluoro-orotic acid (5FOA) to follow the survival of the mutants upon wild-type plasmid loss. (*C*) Protease sensitivity of a G150T mutant. Western blot showing GST fusions of Pab1(1–343BX) constructs following treatment with increasing concentrations of proteinase K. (*D*) Conservation scores for multiple sequence alignment positions created from 119 RRM sequences in the protein data bank. The RNP elements and the four mutation sensitive residues are shown.

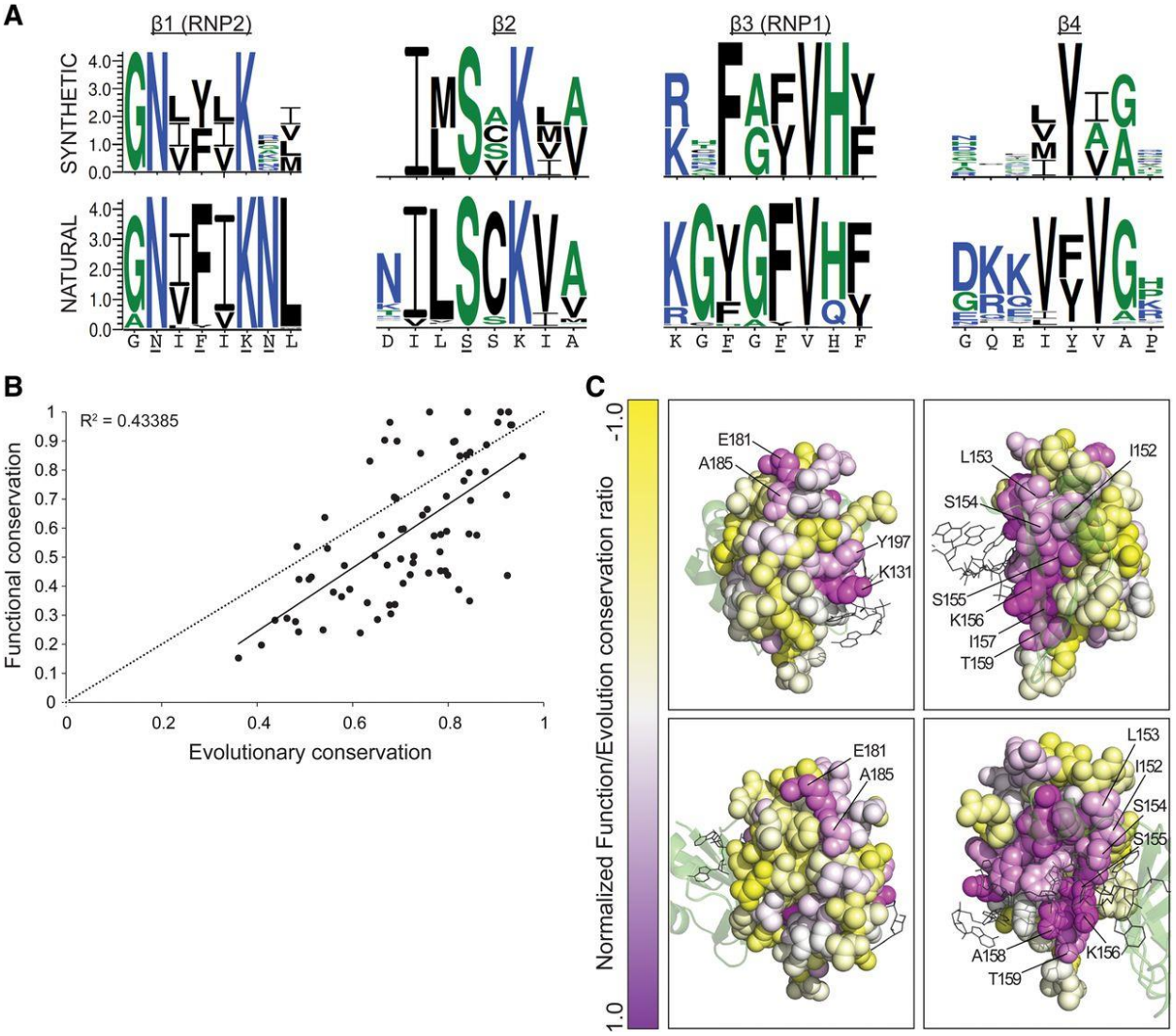


Figure 2.5 Discrepancy between mutation sensitivity data and evolutionary conservation provides functional insights.

(*A*, *top*) Logo plots generated for all amino acid substitutions that resulted in no more than a 5% reduction in performance compared with the wild type. Presented are only the four β-strands and flanking residues. (*Bottom*) Logo plots generated for the same RRM2 elements from a multiple sequence alignment of 306 Pab1 homologous sequences. The yeast RRM2 sequence corresponding to these logo plots is shown *below*. Residues shown to bind RNA in the human RRM2 domain (Deo *et al.*, 1999) are underlined. (*B*) Comparison of the property entropy of each Pab1 RRM2 position in the multiple sequence alignment created from homologous sequences to the property entropies that were derived from all amino acid substitutions that showed no more than a 5% reduction in performance. The trendline is shown in a solid line, together with the Pearson's R^2 value. Dotted line represents perfect correlation. (*C*) The ratio between the functional conservation score to the evolutionary conservation score is color coded on the structure of the human RRM2 (PDB ID 1CVJ). From *top*, *left* in clockwise direction: lateral (facing RRM3), lateral (facing RRM1), dorsal, and ventral views of RRM2. RRM1 is shown in transparent green and the poly(A) in gray.

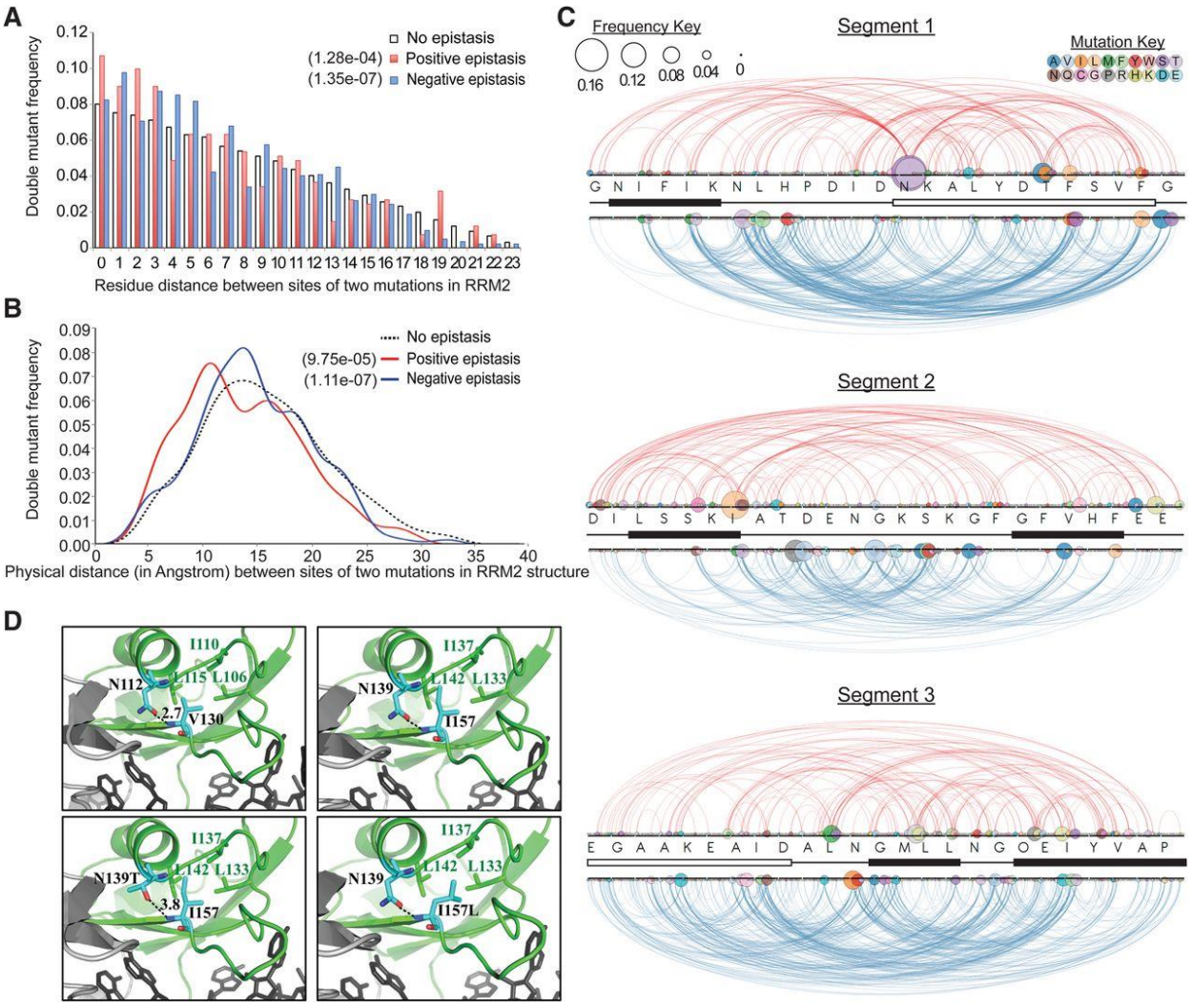


Figure 2.6 Epistatic interactions in double mutants.

(*A*) Contribution of spacing between residues in the primary sequence corresponding to the two mutations in each double mutant to epistatic interactions. A sequence distance of 0 corresponds to adjacent residues. (*B*) Contribution of physical distance between the two mutations forming each double mutant to epistatic interactions. Shown is the distribution of distances between the center of the masses of the two mutated residues based on the human RRM2 structure (PDB ID 1CVJ). Only variants with five or more residues separating the two mutated residues in the RRM2 sequence were included. For *A* and *B*, the *P*-values of wilcoxon-rank tests for the differences between the positive and negative epistasis groups to the no-epistasis group are specified. (*C*) Arc diagrams displaying the interactions between mutation pairs in variants with positive (red) or negative (blue) epistatic interactions. The sequence and the secondary structure of each segment are shown. The color of each node represents substitution to a specific amino acid and the size represents the fraction of variants with that particular mutation. A color map describing the identity of substituting amino acids is shown. (*D*) Presumed effects of N139 and I157 substitutions on strand β2 and helix α1 association may account for suppression of deleterious mutations. (*Top*, *left*) Structure of the human RRM2 in green with N112 and V130 residues shown in color-coded sticks (carbon, light blue; nitrogen, blue; oxygen, red). The hydrogen bond between the carbonyl group of the asparagine side chain and the amino group of the valine backbone is shown by a black dotted line, along with the distance. Hydrophobic residues found in close proximity to Val130 are also shown. (*Top*, *right*)

Replacing the specified residues with the yeast residues supports a similar association between I157 and N139 and between I157 to the conserved aliphatic residues. (*Bottom, left*) Substitution N139T (as well as N139S) may weaken the hydrogen bonding with I157 by increasing the distance between the hydrogen donor and acceptor groups. (*Bottom, right*) Substitution I157L may cause a similar effect by destabilizing Loop L1 and L3 association.

2.4 DISCUSSION

By assaying variants of the RRM2 domain of the yeast Pab1 in high throughput, we scored most (83%) of the possible 1500 single amino acid substitutions (including stop codons), and more than 100,000 variants with multiple substitution events, in a 75-residue-long sequence. The results highlight the RNA-binding surface of RRM2 as the most important element for its function, although each position in the RRM2 shows a nearly unique pattern of mutational sensitivity. We clustered the data to reveal other residues highly sensitive to mutation, as well as core hydrophobic residues that tolerated substitution only by other hydrophobic amino acids. By comparing the evolutionary conservation of RRM residues with their ability to function in the context of the yeast Pab1 protein, we could implicate some residues in yeast-specific functions. Finally, we used epistasis analysis to identify interacting residues in Pab1.

Beside the two RNP motifs, the deep mutational scan suggests that residue G150, present in the loop L2 between helix α1 and strand β2, is an additional signature of the RRM domain family. This residue was one of the few non-RNA-binding residues to display extreme sensitivity to mutations and to be highly conserved within the RRM family. The cold-suppressible phenotype of yeast carrying Pab1 with the G150T mutation, along with the proteinase sensitivity of the mutant protein, suggests a critical role for this residue in stabilizing the RRM structure. G150, which is in the L2 β-turn, may be essential to maintain the gap between the two RRM helices inaccessible to solvent. In agreement with this general function, a mutation at the corresponding residue (G53S) in the RRM1 domain of the *C. elegans* UNC-75 protein eliminated activity (Kuroyanagi *et al.*, 2013). Given the similar mutational sensitivity, cold-suppressible phenotypes of mutants, and the close proximity of their side chains to the gap between the two helices, residues F149 and F173 may act with G150 in the same structural role. Indeed, a temperature-sensitive mutant (F87A) in the residue corresponding to F173 in the RRM1 domain of the splicing factor Prp24 (Kwan and Brow, 2005) suggests that the presumed role for this RNP1 consensus residue in RRM2 stabilization might be general in other RRM sequences. Another loop L2 residue, I152, may contribute to the solvent inaccessibility of the gap between helix α1 and helix α2. This proposed function is supported both by I152 having the highest sensitivity to mutations of the RRM2 aliphatic residues and by its partial solvent accessibility, which distinguish it from other core residues.

While the consensus residues of the RNP motifs are the most commonly found in nature, it is not known how well RRM consensus residues can substitute for wild-type residues in a single, specified RRM domain. The Pab1 RRM2 mutational data provide evidence both for the functional redundancy of some of these consensus residues and for the inability of other consensus residues to support Pab1 activity. In particular, the RRM consensus residues of the two RNP elements that were tolerated in the

yeast RNP motifs appear in some PABP sequences, while consensus residues that were not tolerated are absent from all PABP sequences. However, for some RNP positions (such as H172), the yeast RRM2 tolerated neither the general consensus residue nor a PABP consensus residue, suggesting a highly specific function in yeast for these residues.

We found that clustering RRM2 positions based on their mutational sensitivity could distinguish between core and non-core aliphatic residues and could identify other residues, such as K156, that function within the hydrophobic core. These data are consistent with a long history of mutagenesis (e.g., Lim and Sauer, 1989) that has found hydrophobicity to be the most essential feature of residues in a protein's core. Although a structure of the RRM2 domain is available only for the human PABP-1 protein (Deo *et al.*, 1999), the match between the mutational sensitivity of the yeast residues to their positions in the human protein suggests that the in vivo structure of the yeast domain resembles the in vitro structure of the human one. Given the striking signature in the mutational profile of core residues, these profiles can serve as a general approach to evaluate structures that have been defined in vitro, as previously shown (Adkar *et al.*, 2012) or to refine folding predictions for proteins whose structure is unknown.

The deep mutational scanning approach has been used to study the in vivo effect of all possible 171 single amino acid substitutions across a 9-amino acid-long stretch of the yeast Hsp90 (Hietpas *et al.*, 2011) and many of the possible 1425 single amino acid substitutions in the 75-amino acid-long ubiquitin sequence (Roscoe *et al.*, 2013). However, we found that current yeast methods and sequencing technology allow an in vivo assessment of nearly two orders of magnitude more variants, which enables an analysis of variants with multiple mutations while still maintaining high coverage for variants with a single-point mutation. Using these data to study the epistatic interactions between two mutations in ∼40,000 double mutants revealed ∼2000 double mutants with extreme epistasis scores. We identified a small but significant preference for sequence proximity, up to three residues (positive epistasis) and five residues (negative epistasis), and for short physical distance (up to 12 Å for positive epistasis and 10–15 Å for negative epistasis). Overall, we found that short sequence and physical distance play a role in ∼25% of the positive epistasis events and 14% of the negative epistasis events.

Though further study will be needed to determine the extent to which sequence and physical proximity govern positive and negative epistasis in other proteins, in a comprehensive analysis of drosophilid genomes (Callahan *et al.*, 2011) found a correlation between amino acid substitutions undergoing positive selection and their separation within primary protein sequences, with a second substitution strongly enhanced within ∼10 residues of the first. Among the explanations provided for this correlation was epistasis. Thus, both the evolutionary study and our mutational one point to a role in positive epistasis of residues nearby in the primary sequence.

Epistasis analysis revealed mutations N139T, N139S, and I157L as suppressors of multiple deleterious mutations in residues associated with RNA binding, protein stability, and interdomain interactions. Residues corresponding to N139 and I157 interact in the human RRM2 domain, implying that the association between these two residues is critical for suppression. We suggest that weakening of the hydrogen bonding between these two residues by the three substitutions slightly interferes with the association of helix α1 and strand β2. However, the flexibility that results from the substitutions may allow RRM2 to adjust to mutations that damage its structure. Given the structural conservation of the RRM domain family, the contact site between helix α1 and strand β2 may serve as a general target for mutations that can suppress the effects of disruptive mutations.

This mutational analysis of Pab1 RRM2 has allowed us to assess the in vivo function of residues in this domain. A similar approach with the three other Pab1 RRM domains could highlight both common and unique properties of the sequence–function relationship of each Pab1 RRM domain.

2.5 MATERIALS AND METHODS

2.5.1 *Plasmids*

To create a tetracycline-regulated Pab1 expression system, we cloned the complete coding sequence of *PAB1* into the BamHI and NotI sites of pCM188 (*URA3*, *tetO2* promoter, *CEN*) (Gari *et al.*, 1997). Truncated variants Pab1(1–469) and Pab1(1–343) were generated by cloning the complete coding sequence of *PAB1*, Pab1(1–577), into the XmaI and XhoI sites of p415GPD (*LEU2*, *GPD1* promoter, *CEN*) (Mumberg *et al.*, 1995) and removing the 3′ terminal sequences by treating the plasmid with either SphI and XhoI or with NdeI and XhoI, respectively. The fragment encoding Pab1(1–201) was cloned into the XmaI and XhoI sites of p415GPD. For bacterial expression, the Pab1(1–343BX) fragment carrying either wild-type, F170V, or G150T mutation was cloned into the XmaI and XhoI sites of pGEX4T2 (Addgene).

2.5.2 *Yeast strains and growth conditions*

The *pab1* knockout strain (*MATa ura3Δ0 leu2Δ0 met15Δ0 his3Δ1 pab1*Δ::NatMX) was created by replacing the endogenous *PAB1* gene in strain BY4741 with a NatMX cassette from a PUG6 plasmid (Güldener *et al.*, 1996) and selecting for clonNAT-resistant transformants. To maintain cell viability, we expressed the complete coding sequence of *PAB1* from the Tet-off vector pCM188 prior to gene disruption. We refer to *the pab1*Δ strain that expresses *PAB1* from pCM188 as *pab1*Δ *[PAB1]*. Truncated and mutated *PAB1* variants were cloned into p415GPD and transformed into *pab1*Δ*[PAB1]*. Cells were

grown at 30°C in synthetic complete (SC) media lacking leucine and uracil and supplemented with 2% glucose. The effect of the *PAB1* mutations on growth was tested by adding the tetracycline analog, doxycycline (Sigma, D-9891), to a final concentration of 20 μg/mL, unless otherwise indicated.

2.5.3 *Construction of PAB1 RRM2 libraries in yeast*

The DNA encoding the complete Pab1 protein followed by two stop codons was cloned into the XmaI and XhoI sites of p415GPD. After disrupting the BamHI and XbaI sites in the multiple cloning site, we introduced a series of synonymous mutations at eight codons on either side of RRM2 (codons 118–125 [CC**T**TC**C**C**T**A**C**G**T**AAA**A**AA**A**GG**A**TC**C**] and 203–210 [TC**T**AGAAA**A**GA**GA**G**G**GA**T**TC**C**CA**G**] with synonymous mutations shown in bold and restriction sites underlined) to create silent and unique BamHI and XbaI restriction sites. The changes also allowed specific amplification of the *PAB1* RRM2 insert for high-throughput sequencing. Three oligonucleotides covering codons 126–150, 151–175, and 176–200 in the *PAB1* coding sequence were synthesized with a 4% error rate by TriLink Biotechnologies, filled in, and cloned into the BamHI and XbaI sites that flanked the RRM2 domain.

Following propagation in bacteria, the *pab1*Δ*[PAB1]* strain was transformed by each library by a modified version of the LiAc-PEG method (Daniel Gietz and Woods, 2002). Specifically, an overnight culture was diluted into 50 mL of fresh YPD medium to an OD_{600} of 0.45 and cultured at 30°C for two cell divisions. Cells were washed and resuspended in 2 mL of LiSORB solution (100 mM LiAc, 1 M Sorbitol in TE) and incubated for 30 min at room temperature with constant shaking; 1 mL of LiPEG solution (100 mM LiAc, 40% PEG 3350 in TE) was mixed with 1 μg of plasmid and 50 μg of salmon sperm DNA (Sigma, D1626) and added to 200 μL of cell suspension. After a 30-min incubation at room temperature with constant shaking, 100 μL of DMSO was added to the sample, followed by heat shock at 42°C for 15 min. Cells were recovered in 10 mL of YPD supplemented with 0.5 M Sorbitol at 30°C for 1 h, providing transformation efficiency of 3×10^5 transformants per 1 μg of plasmid DNA.

2.5.4 *Yeast selection*

Yeast carrying one of the three libraries were grown to log phase in SC medium lacking leucine and uracil, supplemented with 2% glucose, and diluted into fresh medium containing 20 μg/mL of doxycycline to a final concentration of 4×10^4 cells/mL. Selection was carried out for 22 h with the culture growing to a density of 5×10^6–1×10^7 cells/mL. Next, 2.5×10^8 cells from each culture were collected before (“input”) and after selection (“selected”).

2.5.5 *Library preparation for high throughput sequencing*

Cells were resuspended in miniprep buffer P1 (Qiagen, 27106) and treated with 100 µg of Zymolase 20T (ImmunO, 320921) in the presence of 50 mM DTT for 2 h at 37°C to digest yeast cell walls. After one freeze and thaw cycle from −80°C to 30 sec at 42°C, yeast DNA was recovered in 50 µL of 10 mM Tris-HCl (pH 8.0) by the standard Qiagen miniprep protocol (Qiagen, 27106). DNA was treated with 60 units of Exonuclease I (USB, 70073X) and with 15 units of Lambda exonuclease (NEB, M0262S) for 2 h at 37°C to remove excess of yeast genomic DNA, and plasmid DNA was purified and concentrated using a Zymo Research kit (D4004). Library fragments were amplified by 18 PCR cycles using primers specific to the synonymous changes that flank the RRM2 domain, and sequenced by an Illumina GAIIx sequencer by pair-end reads.

2.5.6 *Scoring the performance of library variants*

We used the Enrich software package (Fowler *et al.*, 2011) to remove low-quality reads (discarding reads with base Q score <20); to determine the location and identity of mutations, while filtering out variants with insertions or deletions; to calculate the frequency of each variant appearing in the input and selected pools; and to provide an enrichment score for each variant appearing in both pools by calculating the ratio between the two frequencies (selected/input). Enrichment scores were further normalized to the wild-type score.

2.5.7 *Use of synonymous mutations to set input read cutoff and enrichment score distribution of neutral variants*

Enrichment scores for variants carrying missense mutations were arranged from low to high. At each enrichment score X, the proportion of synonymous variants with a score at least as extreme as X was multiplied by the total number of missense variants and divided by the number of missense variants with a score at least as extreme as X to yield an estimate of the False Discovery Rate. Missense variants with less extreme enrichment scores but lower estimated FDRs have their FDRs set to the highest FDR among the set of variants with more extreme enrichment scores as used for the calculation of *Q*-values. Final estimated FDR values were multiplied by two to account for the two extremes.

2.5.8 *Clustering of enrichment scores*

Enrichment scores of single amino acid substitutions were log2 transformed and visualized using Matrix2png (Pavlidis and Noble, 2003). Complete linkage hierarchical clustering with a Euclidean

distance similarity metric for both RRM2 residues and substituting amino acids was performed using Gene Cluster 3.0 (Hoon *et al.*, 2004) and visualized by Java TreeView 1.1.6r2 (Saldanha, 2004). Accessible Surface Areas (ASA) of hydrophobic residues from human RRM2 structure (1CVJ) were obtained from PDBePISA (Krissinel and Henrick, 2007).

2.5.9 *GST–Pab1 purification and Proteinase K sensitivity assay*

Glutathione S-transferase (GST) fusions of Pab1(1–346) were overexpressed in *Escherichia coli* strain DE3. Cells were collected and lysed by sonication in lysis buffer (20 mM Tris-HCl at pH 7.6, 200 mM NaCl, 0.2 mM EDTA, 1 mM DTT) in the presence of protease inhibitor cocktail (Roche, *05056489001*). Proteins were bound to Glutathione-Sepharose beads (GE Healthcare, 17-5132-01), washed (20 mM Tris-HCl at pH 7.6, 1 M NaCl, 0.2 mM EDTA, 1 mM DTT), and eluted (20 mM Tris-HCl at pH 7.6, 1 M NaCl, 0.2 mM EDTA, 10 mM glutathione) according to the manufacturer's instructions. Proteins were dialyzed overnight at 4°C in PBS (25 mM $NaPO_4$ at pH 7.0, 150 mM NaCl) containing 20% glycerol. To assess Proteinase K sensitivity, 10 μg of GST and GST fusion proteins were treated with 0, 0.004, 0.02, 0.1, and 0.5 ng/μL of Proteinase K (NEB, P8102S) in a 20-μL reaction buffer (25 mM Tris-HCl at pH 7.5, 2.5 mM $CaCl_2$) for 1 h at 37°C. Digestion was stopped by adding PMSF to a final concentration of 5 mM.

2.5.10 *Calculating RRM conservation*

To evaluate the general conservation of residues in RRMs by an unbiased approach, we searched the protein databank (PDB) for RRM domains using the terms “RRM” and “RNA Recognition Motif” and collected the PDB-ID of all proteins with a known RRM structure. Using these IDs, we extracted all of the sequences of the structurally defined RRMs from the UniProt Knowledge Base with the exception of proteins with multiple structurally resolved RRM domains, where we randomly selected a single domain for the analysis. Taking this approach provided us with 119 RRM sequences, all from unique proteins (see Supplemental Table S2.3 for the list of sequences). Multiple sequence alignment was performed using the MAFFT program (Katoh and Toh, 2008), and a conservation score for each site was determined by the Protein Residue Conservation Prediction program using the Jensen-Shannon divergence (JSD) scoring method (Capra and Singh, 2007).

2.5.11 *Comparing functional conservation to evolutionary conservation*

To create a consensus sequence that represents every mutation tolerated in the β-sheet structure, all mutations were unlinked from the input and the selected sequence pools and their frequencies were

determined. For each position in each pool, the frequency of every mutation was normalized to the frequency of the wild-type residue, which was set to 1.0. Hence, for every mutation the ratio of frequencies (selected/input) indicates its enrichment relative to the enrichment score of the wild-type residue at the same position, which is equal to 1.0. Amino acid substitutions with an enrichment ratio below 0.95 were assumed to be deleterious for Pab1 RRM2 function and were removed from the analysis, while the other mutations and wild-type residues were used to create 1000 arbitrary sequences that represent their relative enrichment scores. Logo plots from these sequences were created using WebLogo 3.0 (Crooks *et al.*, 2004). Logo plots for natural Pab1 homologs were created by providing WebLogo a MAFFT-based multiple sequence alignment of 306 sequences selected by ConSurf server (Ashkenazy *et al.*, 2010) from the UniRef90 database (Pruitt *et al.*, 2012) showing a maximal 95% identity between sequences and a minimum of 35% identity with Pab1 (Supplemental Table S1.4). To calculate functional and evolutionary conservations, we measured the property entropy for each site in the library sequences and UniRef90-based multiple sequence alignment that were used to create the logo plots by the Protein Residue Conservation Prediction tool (Capra and Singh, 2007). We used a window size of two residues, which incorporates the estimated conservation of adjacent residues into the score of each site.

2.5.12 *Epistasis analysis*

Epistasis scores were calculated using the product model formula (($\varepsilon = W_{AB} – W_A \times W_B$), where W symbolizes a variant's enrichment score, and A and B represent two different amino acid substitutions). Variants carrying stop codons as one of the single amino acid substitutions and others with predicted read counts lower than 1 ($W_A \times W_B \times$ input_read_counts of variant$_{AB}$) were eliminated. To correct for input read effect on epistasis data, the R package *locfit* was used to fit a local regression to the graph of epistasis versus input reads. We used the standard local polynomial model with cubic decay and a nearest neighbor fraction of 0.7, which provides an estimate of the mean epistasis score as a function of input reads. An additional local regression was fitted to the squared residuals of the epistasis scores from their estimated mean and double mutants, with a local estimated z-score greater than 2 or less than −2 were collected as highly positively and highly negatively epistatic mutants, respectively.

2.5.13 *Structure visualization*

PyMol visualization software (v1.5.0.5) was used to create all figures of PABP-1 structure.

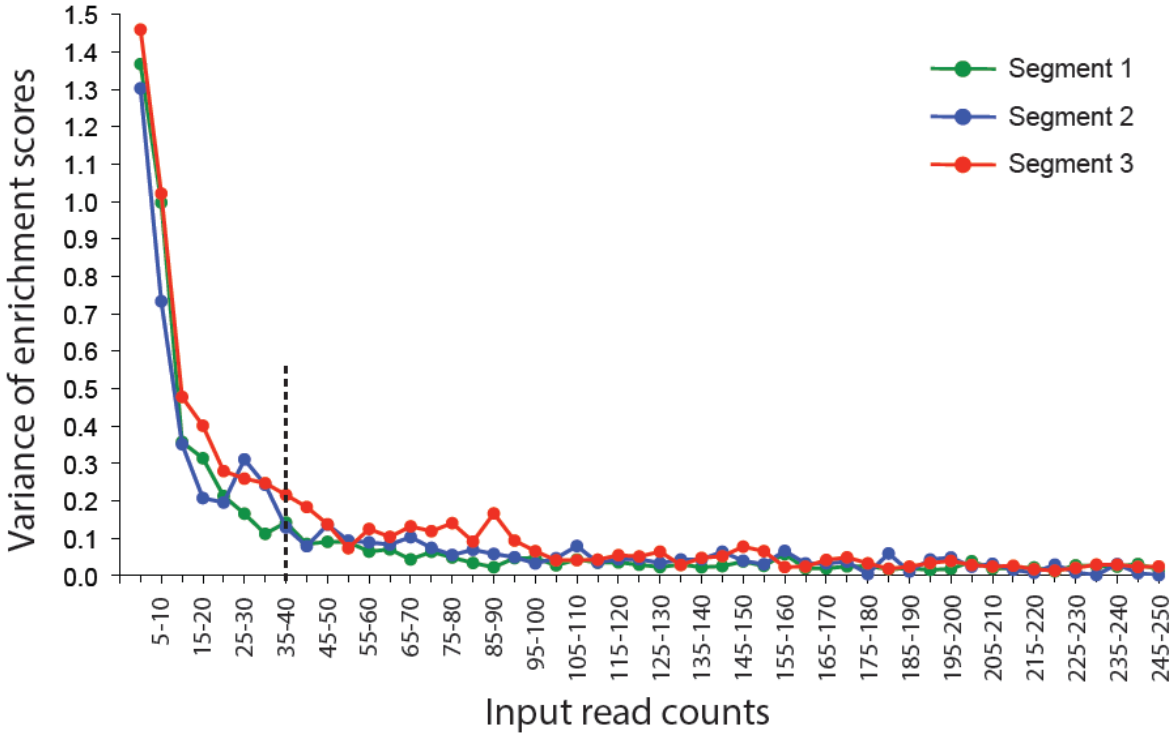


Figure S2.7. Use of RRM2 synonymous variants to set an input read cutoff.

For each library, enrichment scores of synonymous variants carrying single or multiple DNA base substitutions were binned by a 5 input read count window and their variance was calculated. Based on the high variance of enrichment scores from variants with low input read counts, a 40 input read count cutoff was selected (dashed line) and applied to the three libraries. This step removed ~77% of unique protein sequences from the data set and allowed the analysis of the remaining 110,745 variants.

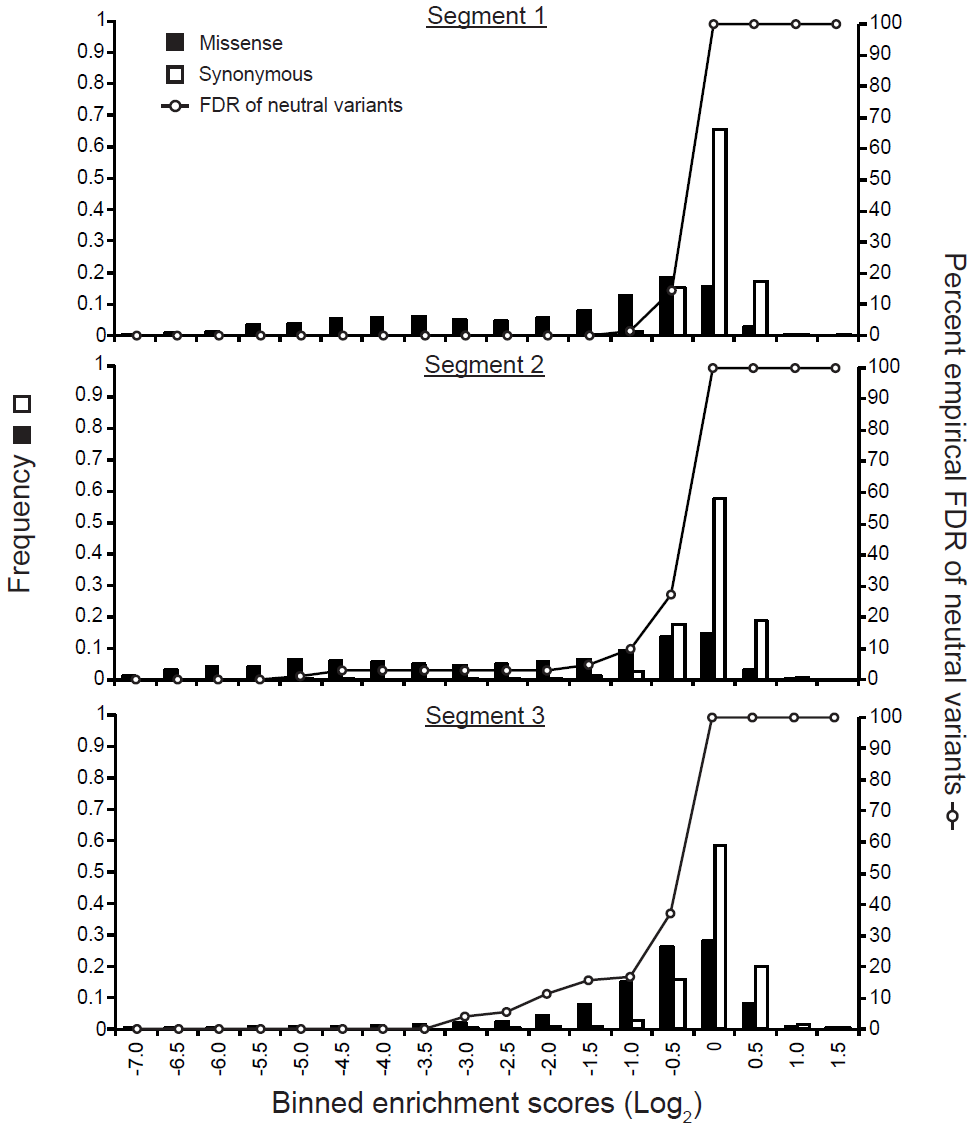


Figure S2.8. Empirical estimate of False Discovery Rate (FDR) of neutral variants.

Shown are enrichment score distributions of variants carrying synonymous or missense mutations for each of the three library segments. Plotted FDR depicts the discovery rate of neutral variants for the enrichment scores specified on the x-axis.

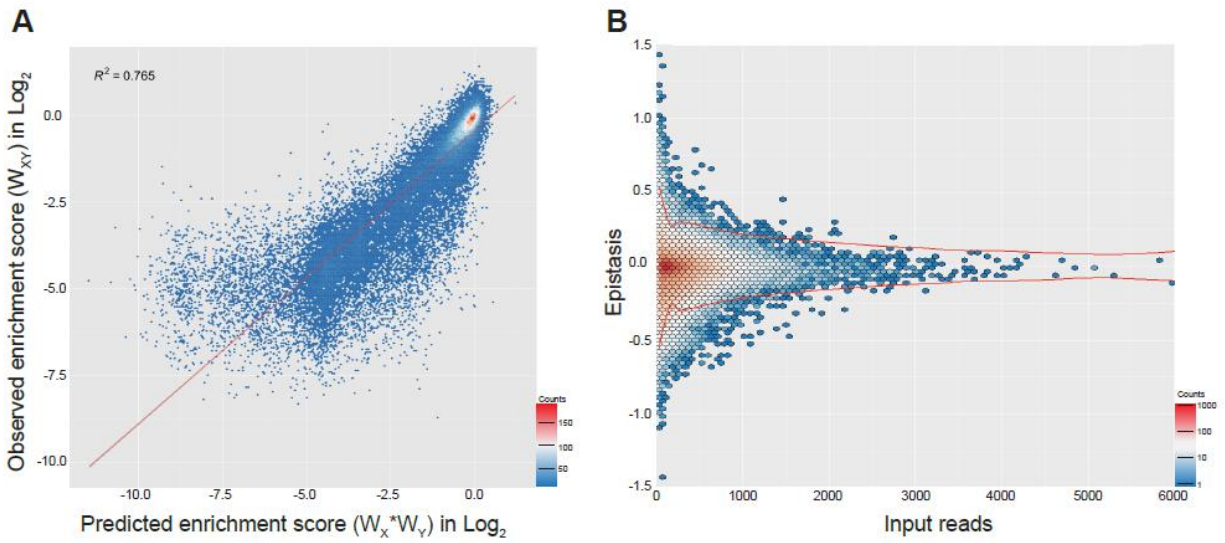


Figure S2.9. A threshold for significant epistatic interactions in double mutants.

(A) Observed enrichment scores of 39,609 double mutants *vs*. predicted enrichment scores as calculated by the product of the two enrichment scores of the single amino acid substitutions that comprise each double mutant. Trend-line is in red. (B) Epistasis scores of the 39,609 double mutants vs. their input read counts. The red lines describe 2 standard deviations cutoff, calculated locally using the loess normalization method. Variants with epistasis scores greater than +2 standard deviations or lower than -2 standard deviations were selected as candidates for displaying positive or negative epistasis, respectively. To cope with stochastic noise due to neutrality (see Figure S2.8), double and single mutants with enrichment scores >1 were eliminated from these lists if reducing their enrichment score to 1.0 resulted in non-significant epistasis score (see materials and methods).

Supplemental Table S2.1. Sequencing statistics

See (Melamed *et al.*, 2013)

Supplemental Table S2.2. Enrichment scores of single amino acid substitutions

See (Melamed *et al.*, 2013)

Supplemental Table S2.3. List of RRM domains in PDB

See (Melamed *et al.*, 2013)

Supplemental Table S1.4. PAB1 homologous proteins

See (Melamed *et al.*, 2013)

Supplemental Table S1.5. PAB1 epistasis scores for all double mutants

See (Melamed *et al.*, 2013)

Chapter 3. IDENTIFYING PROTEIN INTERACTION SITES ON PAB1 USING DMS AND PAB1 HOMOLOGS

Chapter 3 appeared in this form in the journal *PLoS Genetics* (Melamed *et al.*, 2015). My contributions were in the processing, filtering, and scoring of the variants from the sequence data, and finding the epistatic interactions in PAB1 homologs for use in humanizing the PAB/eIF4G interaction. I contributed to the creation of Figure 3.1, Figure 3.2, Figure 3.6, Figure S3.8.

3.1 ABSTRACT

Many protein interactions are conserved among organisms despite changes in the amino acid sequences that comprise their contact sites, a property that has been used to infer the location of these sites from protein homology. In an inter-species complementation experiment, a sequence present in a homologue is substituted into a protein and tested for its ability to support function. Therefore, substitutions that inhibit function can identify interaction sites that changed over evolution. However, most of the sequence differences within a protein family remain unexplored because of the small-scale nature of these complementation approaches. Here we use existing high throughput mutational data on the *in vivo* function of the RRM2 domain of the *Saccharomyces cerevisiae* poly(A)-binding protein, Pab1, to analyze its sites of interaction. Of 197 single amino acid differences in 52 Pab1 homologues, 17 reduce the function of Pab1 when substituted into the yeast protein. The majority of these deleterious mutations interfere with the binding of the RRM2 domain to eIF4G1 and eIF4G2, isoforms of a translation initiation factor. A large-scale mutational analysis of the RRM2 domain in a two-hybrid assay for eIF4G1 binding supports these findings and identifies peripheral residues that make a smaller contribution to eIF4G1 binding. Three single amino acid substitutions in yeast Pab1 corresponding to residues from the human orthologue are deleterious and eliminate binding to the yeast eIF4G isoforms. We create a triple mutant that carries these substitutions and other humanizing substitutions that collectively support a switch in binding specificity of RRM2 from the yeast eIF4G1 to its human orthologue. Finally, we map other deleterious substitutions in Pab1 to inter-domain (RRM2–RRM1) or protein-RNA (RRM2–poly(A)) interaction sites. Thus, the combined approach of large-scale mutational data and evolutionary conservation can be used to characterize interaction sites at single amino acid resolution.

3.2 AUTHOR SUMMARY

The interactions of proteins with each other are essential for almost all biological processes.

Many of the sites of protein contact have evolved to maintain these interactions, but use different sets of amino acid residues. As a result, the residues at a contact site in a protein from one species might not allow a protein interaction when they are tested in a second species. This property underlies the idea of inter-species complementation assays, which test the effect of replacing protein segments from one species by their equivalents from another species. However, this approach has been highly limited in the number of changes that could be analyzed in a single study. Here, we present a novel approach that combines a high-throughput analysis of mutations in a single protein with the set of natural sequences corresponding to evolutionarily divergent variants of this protein. This integration step allows us to map at high resolution both sites of inter-protein interaction as well as intra-protein interaction. Our approach can be used with proteins that have limited functional and structural data, and it can be applied to improve the performance of computational tools that use sequence homology to predict function.

3.3 INTRODUCTION

Protein activity, folding and stability are regulated by the interactions of proteins with other macromolecules. Thus, the identification of sites on a protein where these interactions occur is a critical but difficult undertaking. In some cases, structural analyses provide these sites at high resolution. In other cases, combinations of biochemical, biophysical and genetic methods with mutagenesis strategies have delineated specific residues that contribute to physical interactions. However, the vast number of protein-protein interactions and the low throughput and robustness of approaches to identify interaction sites have led to the limited and often imprecise characterization of only a tiny fraction of the contact sites. Sequence-based computational methods offer an alternative and cost-effective approach that can predict interacting positions by making use of homologous sequences. For example, the evolutionary trace method (Lichtarge *et al.*, 1996) assumes that the locations of interaction sites are conserved over evolution, and that sequence variation within these sites occurs in response to changes in evolutionary constraints to allow the protein to maintain its activity. Other computational methods are based on the idea that physical interaction between two proteins leads to linked evolutionary changes between their contact sites (Lovell and Robertson, 2010; de Juan *et al.*, 2013; Ovchinnikov *et al.*, 2014). Thus, the correlated changes between pairs of positions in multiple sequence alignments of two interacting proteins can identify binding sites (de Juan *et al.*, 2013).

However, despite improvements in the construction of multiple sequence alignments and phylogenetic trees, and the huge increase in the number of homologous sequences, the accuracy of these methods remains challenged by fundamental problems (Cheng *et al.*, 2005; Chakrabarti and Lanczycki, 2007). For example, transient interactions often yield poor evolutionary signals due to increased rates of

substitutions at contact sites (Mintseris and Weng, 2005). In consequence, these contact sites resemble other, less critical residues in the protein that also tolerate multiple substitutions.

We begin with the idea that substitutions tolerated in nature usually cause only minor changes in structure (Chothia and Lesk, 1986). Thus, if a position in a protein is substituted with an amino acid that is found at that position in homologous proteins, the resulting protein is likely still to function in its native organism. However, when such a substitution has a detrimental effect, it may have affected a functional site that has changed over evolution (Marini *et al.*, 2010). For a protein contact site, such a detrimental effect is likely due to the lack of other compensating substitutions also present in the homologous protein that have co-evolved to support its binding to a partner protein. Alternatively, compensatory substitutions might be present in the homologue of the protein partner. Complementation assays using a protein with such natural substitutions have been used to characterize binding site residues (Otero, 1999; Sowa *et al.*, 2001; Harrison and Burton, 2006; Mody *et al.*, 2009). However, the utility of this approach has been limited by the lack of large-scale assays that can test a protein's activity when it carries all the possible substitutions that occur in homologous sequences.

Recently, a method known as deep mutational scanning was developed to assess the functional consequences of up to hundreds of thousands of variants of a protein in a single experiment (Fowler *et al.*, 2010b; Araya and Fowler, 2011). This method combines next generation sequencing with a functional selection, using the change in frequency for each variant over the course of the selection as a proxy for the variant's activity. We previously applied this method to study the *in vivo* function of an RNA recognition motif (RRM) of the *Saccharomyces cerevisiae* poly(A)-binding protein, Pab1 (Melamed *et al.*, 2013).

The eukaryotic poly(A)-binding protein regulates mRNA translation and decay (Adam *et al.*, 1986; Sachs *et al.*, 1986; Mangus *et al.*, 2003) by binding to the poly(A) tail of an mRNA via its four RRMs (Sachs *et al.*, 1987; Burd *et al.*, 1991). This binding leads to an interaction between RRM2 and the translation initiation factor eIF4G, a constituent of the mRNA cap-binding complex, eIF4F (Kessler and Sachs, 1998), which is assumed to enhance the rate of translation by supporting the establishment of a closed loop structure of the mRNA (Jacobson and Favreau, 1983; Wells *et al.*, 1998; Amrani *et al.*, 2008). Yeast encode two eIF4G paralogues, eIF4G1 and eIF4G2 (Goyer *et al.*, 1993), which both interact with Pab1 (Otero, 1999). Complementation assays by Otero et al. (Otero, 1999) with yeast Pab1 containing residues from the human orthologue mapped the binding site for the two eIF4G isoforms to five amino acids on the surface of Pab1 RRM2 (Otero, 1999). However, this study addressed only the 25 Pab1 residues in the RRM2 domain that vary between human and yeast, and thus the contribution of the other 50 RRM2 residues and the precise Pab1 contact sites for the two isoforms of eIF4G were not determined.

We analyzed deep mutational scanning data for the RRM2 domain of yeast Pab1 to examine the functional consequences in yeast of single amino acid substitutions that differentiate the yeast domain from its homologues. This large-scale inter-species complementation data allowed us to characterize the eIF4G1 and eIF4G2 binding sites on the RRM2 surface at single amino acid resolution and to identify residues associated with the RRM2–poly(A) and RRM2–RRM1 interactions. By combining epistasis data for double mutants with natural variation data, we identify a humanizing substitution that promotes a change in binding specificity of the yeast Pab1 RRM2 from the yeast to the human eIF4G1 protein. Taken together, *in vivo* deep mutational scanning data integrated with evolutionary variation can be used to characterize interaction sites with high resolution and to predict epistatically interacting residues in natural homologues of a protein.

3.4 RESULTS

3.4.1 *Effects of substituting amino acids in Pab1 with those of Pab1 homologues*

We recently scored the *in vivo* function of more than 100,000 variants of the RRM2 domain of the yeast Pab1 (Melamed *et al.*, 2013). The assay was based on turning off the expression of a wild-type copy of the *PAB1* coding sequence and assaying growth of yeast dependent on mutated versions of a C-terminally truncated form (Pab1-343) that includes the first three RRM domains and a small portion of RRM4. For each variant, we assigned an enrichment score that represents the ratio between the fractions of its sequence read counts after and before selection, normalized to the wild-type enrichment score. Hence, enrichment scores serve as indirect readouts for the effects of mutations on growth rate. We obtained scores for 1246 single amino acid substitutions, including 1190 missense mutations and 56 nonsense mutations (~83% of all possible single amino acid substitutions in the 75 amino-acid long sequence that covers most of this domain) (Melamed *et al.*, 2013).

We realized that the scores of variants with amino acid substitutions present in Pab1 homologues might provide insight into functional sites that diverged in sequence throughout the evolution of this protein. To this end, we collected sequences of 52 poly(A)-binding proteins that represent all Pab1 homologues in the UniProtKB/Swiss-Prot database. The 52 homologous sequences include both orthologues and paralogues of the poly(A)-binding protein and are derived from eukaryotic species including fungi, plants and mammals. All 52 proteins carry four tandem RRM domains, allowing us to align the Pab1 RRM2 against all its corresponding domains. The multiple sequence alignment showed conservation between the homologous RRM2 sequences and the yeast Pab1 RRM2 ranging from 88%

identity for *Candida glabrata* to 55% identity for *Encephalitozoon cuniculi*. The alignment revealed 210 single amino acid differences ("natural substitutions") with respect to the yeast Pab1 RRM2 sequence. The *in vivo* deep mutational scanning data from our previous study (Melamed *et al.*, 2013) provide functional scores for 197 of these 210 substitutions (Figure 3.1A).

Most of these natural substitutions resulted in small effects (Figure 3.1B), with a median score of −0.07 relative to the wild-type (the score, in $\log_2$ scale, is comparable to ∼5% reduction from the wild-type score) and narrow upper and lower quartiles. On the contrary, substitutions that do not appear in Pab1 homologues ("non-natural substitutions") displayed a much larger range and more negative effects, with a median score of −0.53 (comparable to ∼30% reduction from the wild-type score). That most natural changes result in small effects suggests that the functional constraints on the poly(A)-binding protein remained largely constant throughout its evolution. However, a few natural substitutions showed low enrichment scores that correspond to poor Pab1 performance in *S. cerevisiae*. In particular, enrichment scores of 45 natural substitutions ranged between −0.15 and −0.5 (a range that we term mildly deleterious, comparable to ∼10–30% reduction from the wild-type score) and enrichment scores of 17 other natural substitutions were lower than −0.5 (a range that we term strongly deleterious, comparable to more than 30% reduction from the wild-type score) (Figure 3.1A). We further compared the score distribution of natural variants to the score distribution of synonymous variants which serve as a proxy for non-deleterious variants, as previously described (Melamed *et al.*, 2013). This comparison allowed us to assess the contamination of the mildly and the strongly deleterious groups by variants that carry non-deleterious mutations (Figure S3.7). Based on this analysis, we estimated that the natural substitutions in the mildly deleterious group are contaminated by 35% non-deleterious variants, while the natural substitutions in the strongly deleterious group are contaminated by only 8% non-deleterious variants. Given these results, we further analyzed only mutations classified as strongly deleterious.

The solvent accessibility of residues in the structure of a human orthologue of Pab1 reveals that both natural non-deleterious and natural strongly deleterious substitutions, compared to all other non-natural substitutions, occur preferentially at solvent-exposed areas (Figure 3.1C). However, an evaluation of the conservation of each substitution using its Blosum62 score revealed a significant difference between the natural non-deleterious and the natural strongly deleterious groups (Figure 3.1D). Though both groups showed high conservation compared to non-natural substitutions, the natural strongly deleterious substitutions displayed a lower conservation score (median of −1) than the natural non-deleterious substitutions (median of 0). The differences in Blosum62 score distributions of the two groups suggests that natural deleterious effects in general are due to substitutions to amino acids that display physicochemical properties that are neither as disruptive as non-natural substitutions nor as subtle as

natural non-deleterious ones. Nonetheless, a few natural-deleterious substitutions resulted from replacements by highly similar amino acids (*e.g.* L186M and L153V), indicating that sometimes the exact identity of the Pab1 residue is of crucial importance.

3.4.2 *Delineation of the eIF4G-binding site in Pab1*

Of the 25 single amino acid substitutions that differentiate the yeast Pab1 RRM2 domain from its human orthologue (Figure 3.2), 24 have enrichment scores in our dataset. Three of these mutations (E181R, A185K and L186M) are strongly deleterious (Figure 3.2B). These three substitutions occur in two short stretches of the yeast Pab1, 180-KE-181 and 184-DAL-186, that when replaced with the corresponding human stretches to create 180-ER-181 and 184-EKM-186 interfere with *in vitro* binding to ~100 amino acid fragments of yeast eIF4G1 and eIF4G2 (Kessler and Sachs, 1998; Otero, 1999). The large-scale mutational data indicate that the other two mutations in these short stretches, K180E and D184E, cause no measurable effect on function (Figure 3.2B).

To test whether the *in vivo* effects on Pab1 performance correlate with eIF4G1 and eIF4G2 binding, we established a two-hybrid assay between yeast Pab1 and the N-terminal 341 amino acids of yeast eIF4G1 or eIF4G2, which contain the binding sites for Pab1 (Kessler and Sachs, 1998; Otero, 1999). The full-length Pab1 tested with the eIF4G1 or eIF4G2 fragment did not activate *HIS3* reporter gene expression (Figure 3.2C). However, as some protein-protein interactions can be detected by the yeast two-hybrid system only when parts of the proteins are removed (Walhout *et al.*, 2000), we tested various truncation products of Pab1 for eIF4G1 and eIF4G2 association. Indeed, RRM2 alone produced a positive interaction signal with both isoforms (Figure 3.2C).

In agreement with Otero et al. (Otero, 1999), the replacement of residues 184–186 with those from human resulted in complete loss of binding to both eIF4G1 and eIF4G2 (Figure 3.2D). When tested individually, A185K and L186M did not bind eIF4G1 or eIF4G2, while D184E showed wild-type binding. The replacement of residues 180–181 with those from human abolished eIF4G1 binding and reduced eIF4G2 binding. This residual binding to eIF4G2 may reflect the greater sensitivity of the two-hybrid assay compared to the *in vitro* assay (Otero, 1999). When tested individually, E181R resulted in loss of eIF4G1 and eIF4G2 binding, while K180E had no effect (Figure 3.2D). Since the E181R effect on eIF4G2 binding was more severe in the absence of the K180E substitution, K180E might suppress the negative effect of the E181R mutation on eIF4G2 binding by decreasing the local positive charge. Overall, the *in vivo* function of Pab1 carrying any of the five single amino acid substitutions correlates with the ability of Pab1 to support eIF4G1 and eIF4G2 binding.

We hypothesized that the deleterious effects of some of the other natural substitutions might be due to a loss of eIF4G1 and eIF4G2 binding. We therefore tested in the two-hybrid assay the 17 substitutions that cause a strongly deleterious effect, as well as A185K and D184W, which score similarly but had lower sequence read coverage in the original experiments (Melamed *et al.*, 2013). Of these 19 mutations, 10 (occurring in 8 different residues) impaired the ability of RRM2 to bind eIF4G1, with I137F, T145H, T145L, V148K, E181R, A185H, A185K and L186M showing the most severe effects (Figure 3.3A, left). D138T and A141D resulted in modest effects on eIF4G1 binding (Figure 3.3A, left). The same Pab1 variants assayed against eIF4G2 revealed similar effects (Figure 3.3A, right), suggesting that eIF4G1 and eIF4G2 use the same set of Pab1 RRM2 residues for binding. However, eIF4G1 binding was more sensitive to A141D and T145L compared to eIF4G2. Based on the effects of the natural amino acid substitutions on binding, we set the boundaries of eIF4G recognition site to the upper surface of RRM2 (Figure 3.3B), a region much wider than previously identified (Otero, 1999).

3.4.3 *A large-scale mutational analysis of the Pab1 RRM2-eIF4G1 interaction*

While combining natural variation with *in vivo* deep mutational scanning highlights the contribution to protein-protein interactions of residues that change over evolutionary time, it overlooks highly conserved residues and ignores the effects of substitutions to amino acids that do not appear in homologues. We therefore sought to study the effects of mutations on Pab1 RRM2–eIF4G1 association by an alternative approach. To this end, we performed a large-scale two-hybrid analysis. We expressed each of three libraries of RRM2 as a DNA-binding domain hybrid, with mutations covering Pab1 positions 131–150, 151–175 or 176–197, and tested for the binding of these hybrids to the yeast eIF4G1 expressed as an activation domain hybrid. Samples were collected before (input) and after (selected) two-hybrid selection, and the library segments were recovered and sequenced. For each variant, the change in its frequency from input to selected pool (i.e. its enrichment score) was determined as previously described (Melamed *et al.*, 2013). We were able to extract enrichment scores for 802 single amino acid substitutions across the three library segments, which comprise 60% of all possible substitutions (Supplementary Table 3.1). While mutations that disrupt RRM2 structure caused fortuitous activation of the yeast two-hybrid reporter gene, positions that were shown to be sensitive to natural substitutions when tested individually showed similar sensitivities to mutation in this large-scale assay, suggesting that the enrichment scores for mutations that specifically affect the contact site for eIF4G1 were valid (Figure S3.8). In particular, of the 44 mutations that reduced the enrichment score by more than 50% ($\log_2$ enrichment score < -1), 22 mutations occur at the eight positions that were found by our natural variation analysis to be involved in eIF4G1 binding (I137, D138, A141, T145, V148, E181, A185 and L186); eight

mutations occur at the immediate sequence neighbors of these positions (D136, S147, F149 and D184); and 11 mutations occur at residues that show physical but not immediate sequence proximity to these contact site residues (G150, G188, M189, L190 and N192). Overall, in addition to identifying eIF4G1 contact site residues that were elucidated by the combined approach of the *in vivo* mutational data and the natural variation data, the large-scale two-hybrid results highlighted the contribution of residues at the periphery of this site (Figure 3.4A). To understand why mutations at these positions were not discovered using our combined approach, we examined the level of natural variation at these sites. While F149 and G150 are fully conserved, the other residues show some degree of variation in Pab1 homologues. Though some of these natural changes interfered with eIF4G1 binding in the two-hybrid assay, none of them showed a strongly deleterious effect *in vivo* (Figure 3.4B), suggesting that the central residues of the eIF4G1 binding site are more sensitive to natural variation substitutions *in vivo* than the peripheral ones.

3.4.4 *Evolutionary paths of deleterious substitutions*

To understand how incompatible Pab1 variants have evolved in different lineages, we constructed a maximum likelihood tree from the 52 Pab1 homologues. In agreement with theoretical expectations (Orr, 1995), we found that the number of substitutions in Pab1 that were strongly deleterious in *S. cerevisiae* increases with evolutionary distance (Figure 3.5A). Specifically, while closely related fungi provide zero or one strongly deleterious substitution, the microsporidian *Encephalitozoon cuniculi*, which carries the most diverse PABP sequence, contributes six deleterious substitutions. The deep divergence of *E. cuniculi* PABP, likely due to rapid evolution of microsporidia after branching off the fungal lineage (Gribaldo and Philippe, 2002), provides a unique set of mutations (I137F, D138T and A141D) that interfered with eIF4G1 binding. However, unlike the metazoan substitutions that interfered with this binding, the *E. cuniculi* substitutions localize to helix α1 (Figure 3.5B), which suggests two alternative paths of eIF4G-binding site evolution. In addition, the deleterious effects of substitutions T145L and T145H, from the non-yeast paralogues of the poly(A) binding protein (PABP5 and PABP4L), reveal the critical function of T145 in eIF4G binding. Taken together, these results highlight the need to analyze evolutionarily remote sequences in order to obtain a detailed map of functional sites in proteins.

The functional scores of the natural substitutions that occurred throughout evolution suggest ancestral states that were likely to promote the divergence of the eIF4G1-binding site. In particular, for position 185, we observe a stepwise decrease in charge in the *S. cerevisiae* lineage, from lysine through histidine and asparagine to alanine (Figure 3.5B, middle). Both A185K and A185H were strongly deleterious in yeast, suggesting that the lack of positive charge in yeast was accompanied by other changes in eIF4G or in Pab1 orthologues that are no longer compatible with the ancestral state of this

position. At positions 181 and 186, substitutions matching variation within the *S. cerevisiae* lineage were mildly deleterious or non-deleterious, while substitutions matching variation that occurred after the fungal–metazoan divergence were strongly deleterious. Therefore, changes in eIF4G or in Pab1 orthologues that compensate for the otherwise detrimental effects of these mutations are likely to be conserved along the metazoan branch of the tree.

3.4.5 *Changing Pab1 RRM2 binding specificity from yeast eIF4G1 to human eIF4G1*

We asked whether the yeast Pab1 and eIF4G protein sequences might enable us to infer the compensatory changes that allowed the establishment of the strongly deleterious substitutions E181R, A185K and L186M in the human orthologue of Pab1. For instance, a pair of mutations comprising one humanizing substitution in yeast Pab1 that interferes with yeast eIF4G1 association and a compensating, second humanizing mutation in the yeast eIF4G1 might restore binding. However, the identification of candidate humanizing substitutions in the yeast eIF4G1 that may form deleterious–compensatory clusters with humanizing mutations in Pab1 is challenging due to the extreme diversification of eIF4G1 and its contact site residues throughout evolution (Figure 3.6A). Thus, we decided to explore the inter-protein interactions in Pab1 that underpin the binding of the RRM2 domain to either the yeast or human eIF4G1. While the human and yeast RRM2 domains interacted with their cognate eIF4G1 fragment, neither bound to its non-cognate eIF4G1 fragment (Figure 3.6B), suggesting that eIF4G1 binding specificity is dependent on the 25 positions that differ between the yeast and the human RRM2 domains.

We tested a few humanizing mutations in Pab1 RRM2 for their ability to change the binding specificity towards human eIF4G1. Though there are many possible combinations of humanizing substitutions, we used the deep mutational scanning results to narrow down the list of candidate residues. We first evaluated the ability of Pab1 RRM2 fragments that carry each of the three humanizing substitutions (E181R, A185K and L186M) that abolished binding to the yeast eIF4G1 to bind the human eIF4G1 fragment. The E181R variant activated the two-hybrid reporter gene (Figure 3.6B), indicating that despite other sequence differences, elements within the yeast Pab1 RRM2 domain support this change in binding specificity. Unlike E181R, A185K and L186M did not bind to human eIF4G1, suggesting that these two substitutions require other humanizing changes in Pab1 RRM2 to function. Combining A185K and L186M with E181R to form a triple mutant did not enable binding of yeast Pab1 to human eIF4G1 (Figure 3.6B). Because this triple mutant carries all of the strongly deleterious substitutions that differ between the human and the yeast Pab1 RRM2 domain, this finding suggests that some of the remaining mildly deleterious or non-deleterious substitutions are necessary to compensate for the detrimental effects

of A185K and L186M on eIF4G1 binding.

Because the deep mutational scanning of Pab1 RRM2 provided functional scores for multiple variants that change two amino acids (Melamed *et al.*, 2013), we realized that the contribution of other humanizing substitutions to the function of contact site residues might be inferred from the epistasis scores of such variants. We calculated epistasis by taking the enrichment score of a double mutant and subtracting the product of the scores of the component single mutants. Humanizing substitutions that compensate for the deleterious effects of E181R, A185K or L186M are likely to show positive epistasis (*i.e.* the double variant functions better than predicted) while humanizing substitutions that do not should display no epistasis. We extracted the epistasis scores for 866 double mutants ((Melamed *et al.*, 2013), Supplementary Table 3.2), each carrying two substitutions that are found in one of the 52 homologues of Pab1. Comparing the epistasis score distribution of these variants to that of 38,742 double mutants that carry pairs of mutations that do not occur in any of the individual homologues of Pab1 that were sampled in our analysis revealed a small yet significant increase (Wilcoxon rank sum test p-value = 3.712e-10) in epistatic interactions between substitutions that are present together in natural variants (Figure 3.6C). Thus, two mutations found in a natural protein variant are more likely to interact positively, either by synergistic or compensatory mechanisms.

Of the 866 double mutants with two substitutions found in Pab1 homologues, eight carry one of the strongly deleterious humanizing substitutions together with a second humanizing mutation (Supplementary Table 3.2). Of these, a double mutant carrying the deleterious substitution L186M together with the non-deleterious substitution G177E had a high epistasis score (Figure 3.6C). Specifically, while L186M alone resulted in ~30% loss of *in vivo* function, addition of the non-deleterious G177E substitution restored Pab1 function to the wild-type level (Supplementary Table 3.2). G177E was able to partly restore eIF4G2 binding of an RRM2 mutant that carries the L186M substitution (Figure 3.6D, suggesting that the positive epistasis of G177E and L186M is at least in part due to an improved association of the double mutant with eIF4G2. While adding G177E to the triple mutant did not shift the binding specificity towards the human eIF4G1, humanizing its adjacent residue by E176Q substitution supported this switch (Figure 3.6D), suggesting that the local humanized environment of G177E is important for its function. The contribution of E176Q and G177E to human eIF4G1 binding is specific, as other groups of humanizing substitutions, found either at a distance or in close physical proximity to the three deleterious substitutions, were not able to promote this shift in binding specificity (Figure S3.9A). Thus, despite the lack of measurable effects of single amino acid substitutions at position 177 of yeast Pab1 (Figure 3.1A), the amino acid at this position is important for Pab1 binding to the human eIF4G1. The ancestral state of position 177 in the Pab1 lineage was glutamic acid, which was

replaced by glycine in the recent ancestor of *S. cerevisiae* (Figure S3.9B). Therefore, it is likely that the pre-establishment of glutamic acid at position 177 compensated in human for the detrimental effects of at least one of the three deleterious substitutions, while becoming dispensable in the evolutionary path that was taken by *S. cerevisiae*.

3.4.6 *Deleterious mutations that map to other functional sites*

Of the other nine natural and strongly deleterious substitutions in Pab1, five (K140A, L153V, S155V, K156N and A158E) map to the interface between RRM1 and RRM2. In particular, L153 and K156, present in the human orthologue, are key residues in the interaction between RRM1 and RRM2 that allow for efficient poly(A) binding (Deo *et al.*, 1999). In addition, an allosteric change in the RRM1 and RRM2 interface upon poly(A) binding is suggested to facilitate the association of RRM2 with eIF4G (Safaee *et al.*, 2012). Therefore, deleterious substitutions in the RRM1–RRM2 contact site are likely to result from loss of either poly(A) or eIF4G binding activity, or both.

Three other substitutions (Y197N, Y197V and A199E) map to the poly(A)-binding site (Deo *et al.*, 1999). Residue 197 is the only RNA-binding residue that is highly divergent, as all the other residues that bind RNA are either identical across the 52 homologues or display a small variation that is highly tolerated by the yeast protein. It is likely that the structure of the poly(A) forces extreme conservation on the RNA-binding residues, similar to enzyme-substrate binding sites (Caffrey *et al.*, 2004), in a way that prevents useful characterization by natural substitutions.

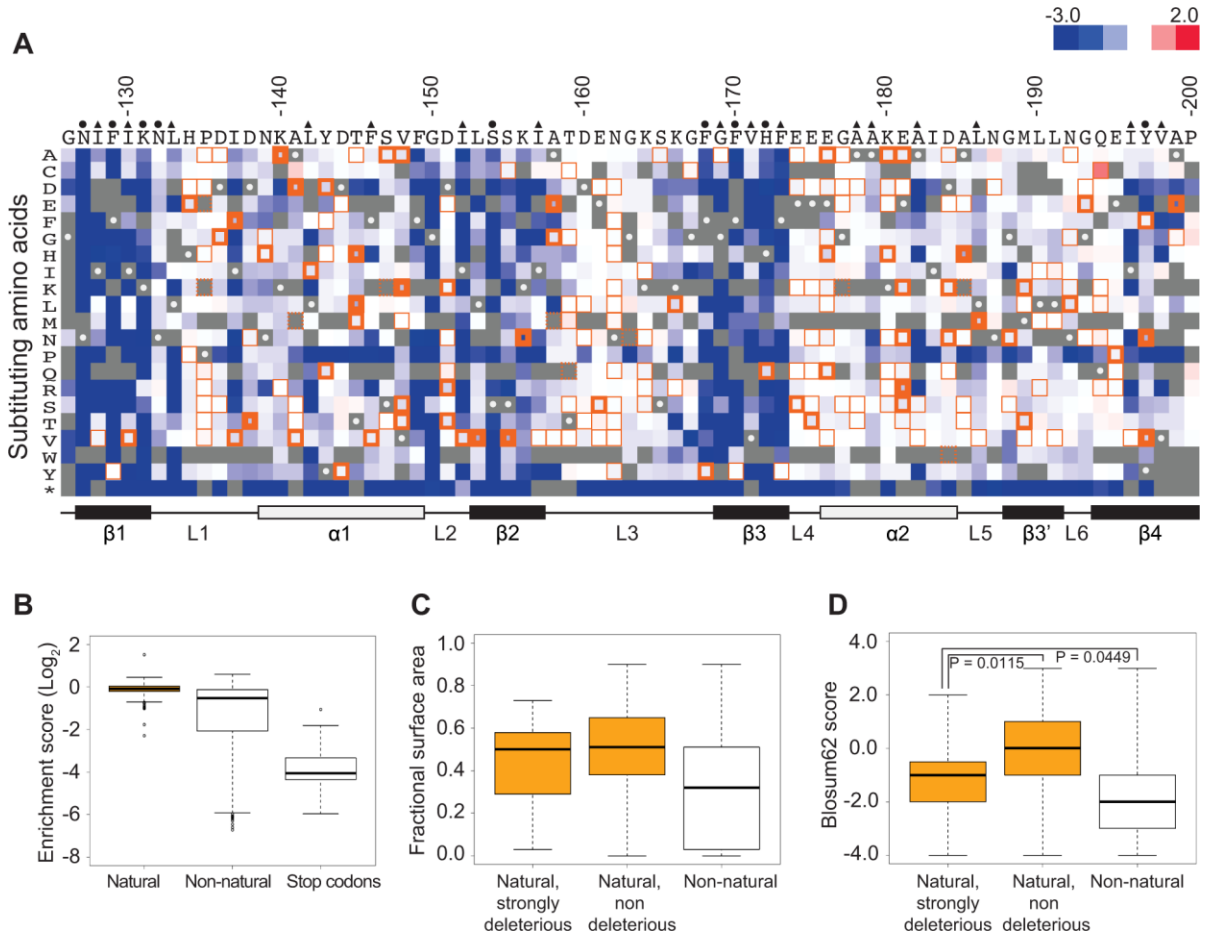


Figure 3.1 Functional characterization of single amino acid substitutions occurring in 52 Pab1 RRM2 homologues.

A) A heat map displaying the enrichment scores (log_2 transformed) for single amino acid substitutions in the Pab1 RRM2 sequence (Melamed *et al.*, 2013). Each column represents a site in the RRM2 sequence and each row a substitution to a specific amino acid. An asterisk designates nonsense mutations. Color ranges from blue for the most detrimental mutations to red for the most beneficial. Orange outlines depict substitutions found in at least one of the 52 Pab1 homologues that were analyzed, with effects that are non-deleterious indicated by thin lines, effects that are mildly deleterious by intermediate lines, and effects that are strongly deleterious by thick lines. Gray stands for missing data and substitutions that did not pass the 40 input read counts cutoff or other quality filtration steps. Wild-type residues are marked by gray with white central dots. Residues known to interact with poly(A) in the human RRM2 domain (Deo *et al.*, 1999) are marked with green and residues showing low solvent accessibility (PDB ID 2K8G, fractional accessible surface area ≤ 0.1) are marked with magenta. The secondary structure of the RRM2 domain aligned to the sequence is shown below. B) The distribution of enrichment scores for amino acid substitutions found in Pab1 RRM2 homologues (natural) and for all other substitutions that do not occur in those sequences (non-natural). Shown also is the distribution of stop codons to highlight enrichment scores of null mutations. C) Box plots displaying the fractional surface area for the equivalent yeast residues in the human RRM2 structure in complex with poly(A) (PDB ID 1CVJ). Shown are the fractional surface area for the 17 amino acid substitutions in Pab1 RRM2 homologues that displayed enrichment scores lower than −0.5, the other naturally occurring substitutions, and the non-natural substitutions. D) Box plots depicting the Blosum62 score of the same substitution groups as in C. p-values from Wilcoxon rank sum tests for the differences between the score distributions are shown.

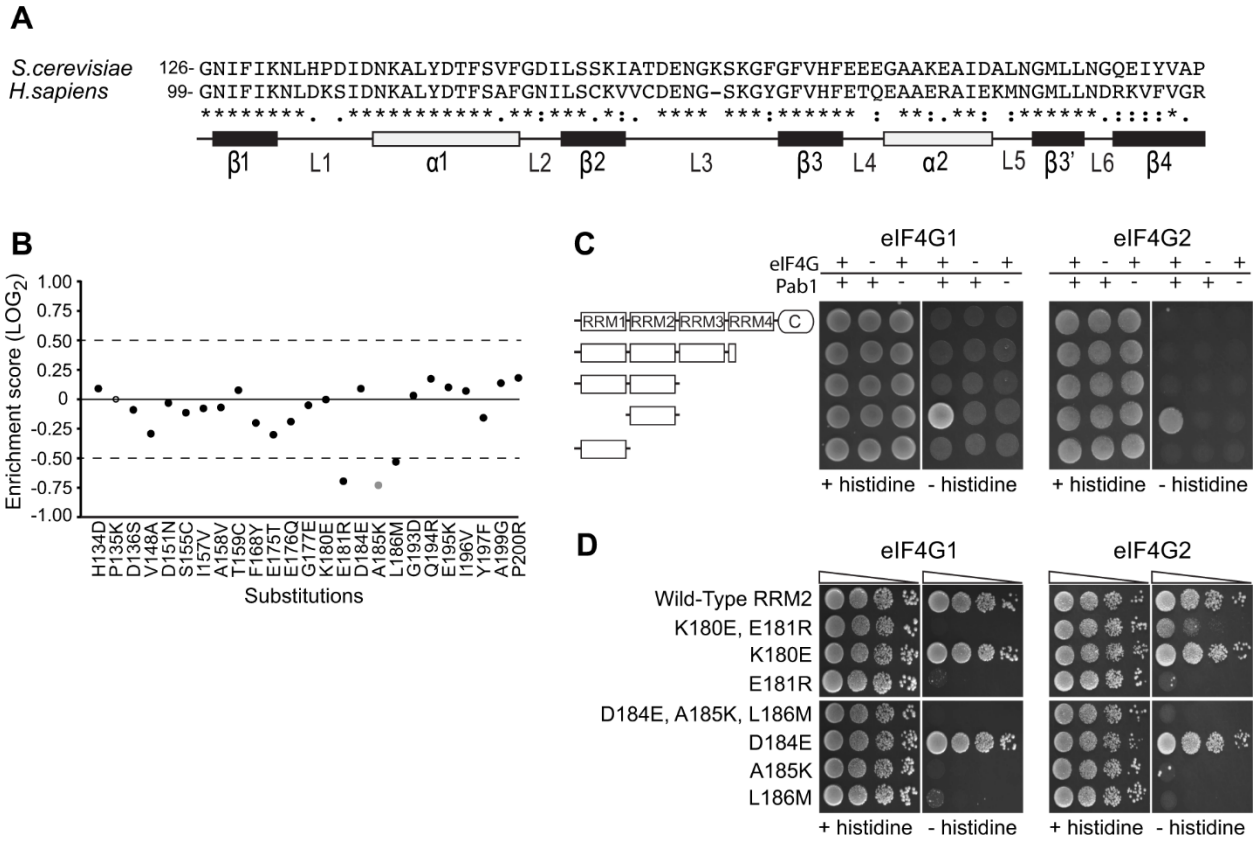


Figure 3.2 Functional consequences of single amino acid substitutions corresponding to residues found in the human RRM2 domain.

A) Sequence alignment of the *S. cerevisiae* Pab1 RRM2 region that was subjected to deep mutational scanning (Melamed *et al.*, 2013) and its human homologue. The illustration below depicts the secondary structure of the human RRM2 domain (based on PDB ID 1CVJ). Black and white rectangles represent β-strands and α-helices, respectively. Lines represent loop structures. B) Enrichment scores for 24 of the 25 single amino acid substitutions that differentiate the yeast RRM2 from its human orthologue. Scores were normalized to the enrichment score of the wild-type protein (black solid line). Black circles represent scores obtained for variants with more than 40 sequencing reads in the input library, which served as a cutoff for reliable measurements (Melamed *et al.*, 2013). The gray circle represents a less reliable score due to lower input read count. The empty circle stands for missing data. C) A yeast two-hybrid assay testing for the interaction between full length or C-terminal truncations of Pab1 (fused to the Gal4 DNA-binding domain) and the N-termini (amino acids 1–341) of the two eIF4G isoforms (fused to the Gal4 activation domain). For *HIS3* expression, $2*10^4$ cells were seeded on synthetic complete plates selecting only for the two plasmid markers (no histidine selection) or for the two-hybrid interaction (histidine selection), and grown for three days. D) The same eIF4G isoforms were tested against Pab1 RRM2 variants that carry the specified mutations, as described in C. Triangles represent 10-fold dilutions in the number of cells at each spot, beginning with $2*10^4$ cells at the leftmost spots

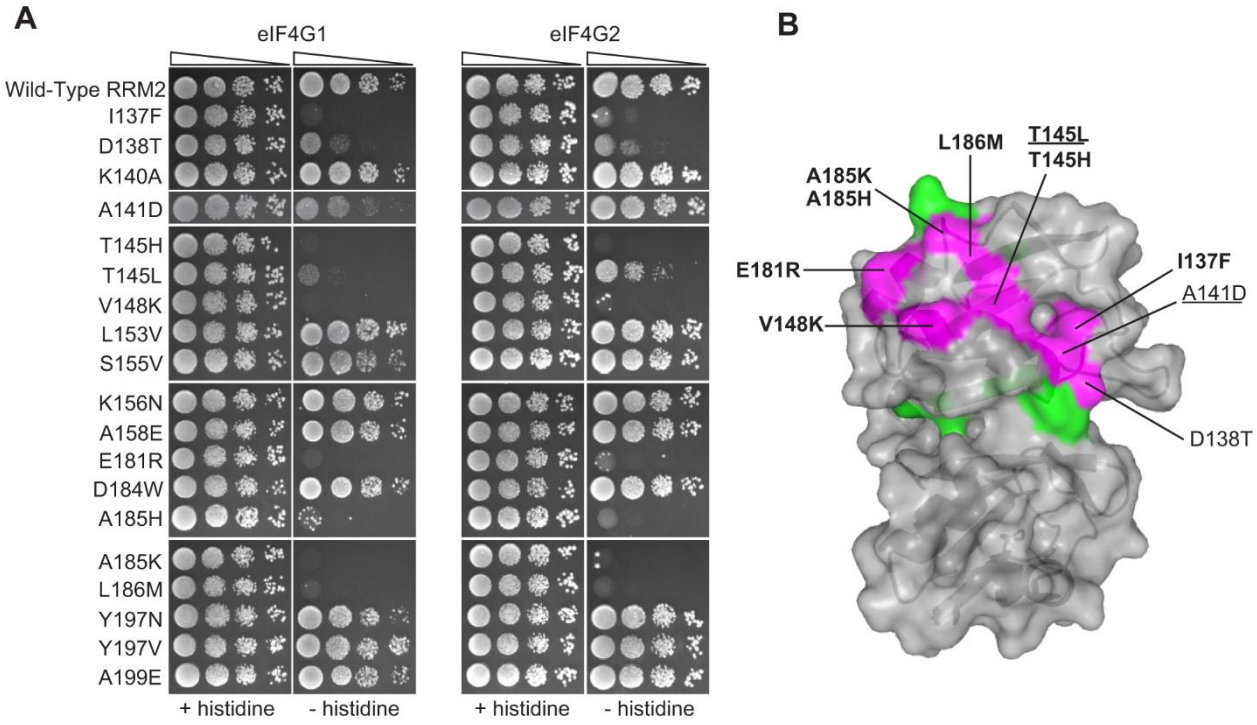
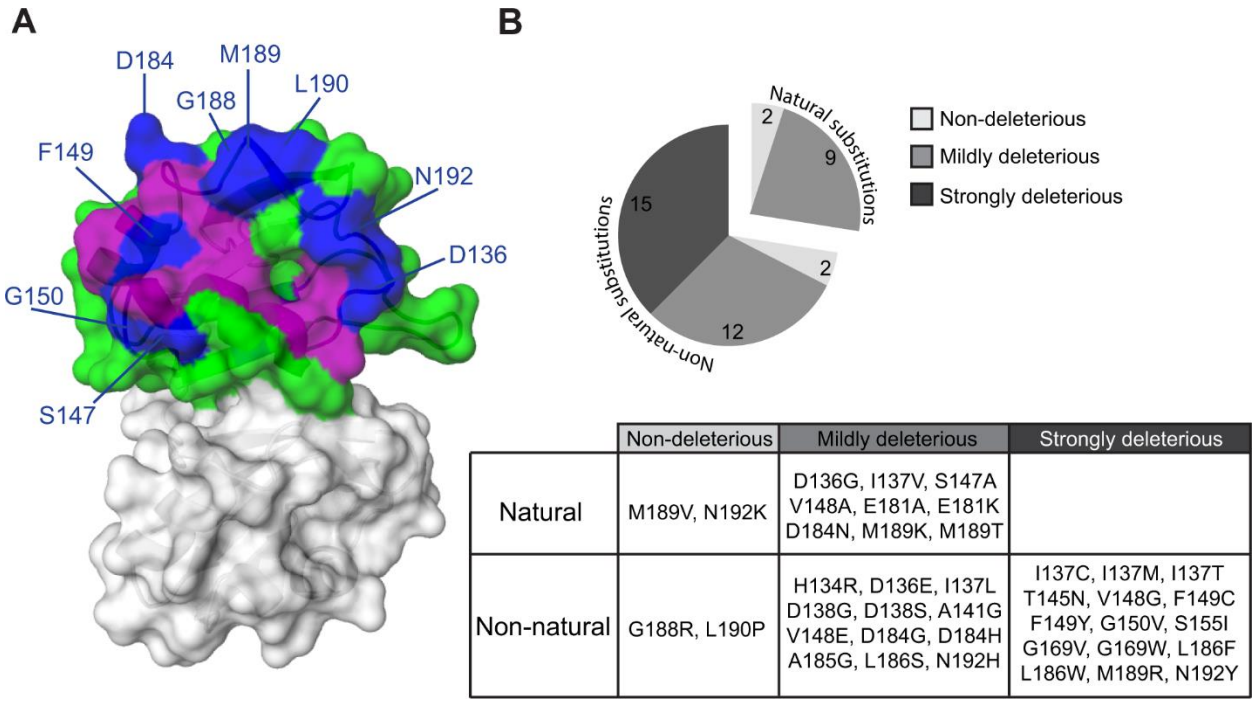


Figure 3.3 Effects of deleterious substitutions found in Pab1 RRM2 homologues on Pab1 binding to eIF4G.

A) A yeast two-hybrid assay testing 19 deleterious substitutions found in Pab1 RRM2 homologues for the ability of Pab1 to bind to the two eIF4G isoforms. The *HIS3* assays were performed as described in Figure 3.2. The two substitutions marked with gray did not pass the input read cutoff and therefore their low enrichment score is less reliable. B) Structure of the human orthologue of the Pab1 RRM1-RRM2 fragment bound to poly(A) (PDB ID 1CVJ). Shown is the surface at the opposite side of the RNA-binding site and the secondary structure depicted by a cartoon. Residues with mutations that impaired eIF4G1 or eIF4G2 binding are shown in purple with the substitution names specified. Bold and non-bold markings represent no growth and intermediate growth of mutants under selection. Substitutions that are underlined depict stronger effect on eIF4G1 binding over eIF4G2. Residues with mutations that did not interfere with eIF4G1 or eIF4G2 binding are shown in green.



	Non-deleterious	Mildly deleterious	Strongly deleterious
Natural	M189V, N192K	D136G, I137V, S147A V148A, E181A, E181K D184N, M189K, M189T	
Non-natural	G188R, L190P	H134R, D136E, I137L D138G, D138S, A141G V148E, D184G, D184H A185G, L186S, N192H	I137C, I137M, I137T T145N, V148G, F149C F149Y, G150V, S155I G169V, G169W, L186F L186W, M189R, N192Y

Figure 3.4 Effects of single amino acid substitutions in the Pab1 RRM2 domain on its interaction with eIF4G1.

(A) The crystal structure of the human RRM2 (PDB_ID 1CVJ) is shown. RRM2 positions with at least one amino acid substitution resulting in an enrichment score lower than 50% of the wild-type are colored in purple if the position was discovered by the combined analysis or in blue if the position was identified solely by the two-hybrid assay. All other RRM2 positions are shown in green. (B) Distribution of enrichment scores for amino acid substitutions that were not identified by the combined approach.

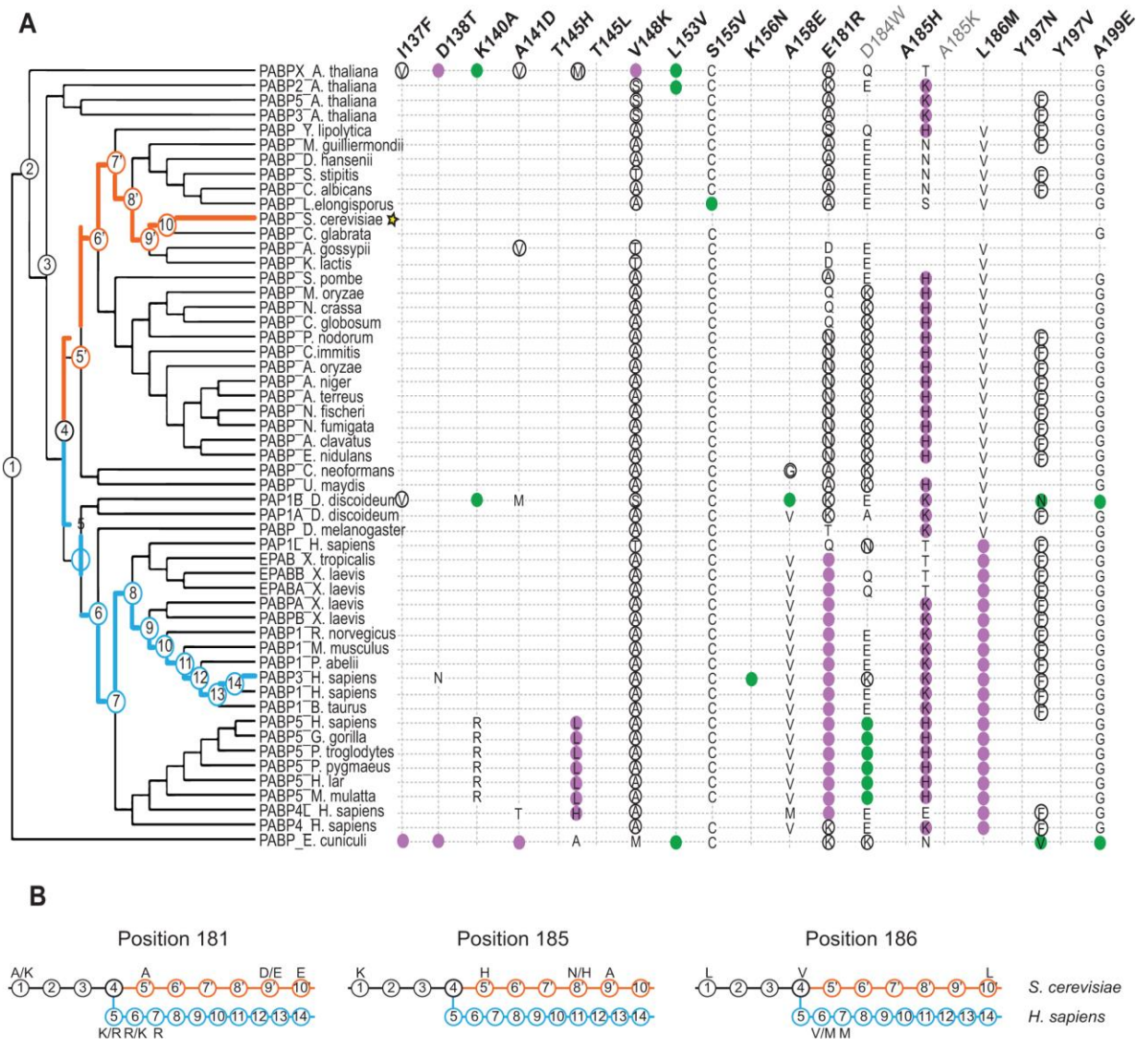
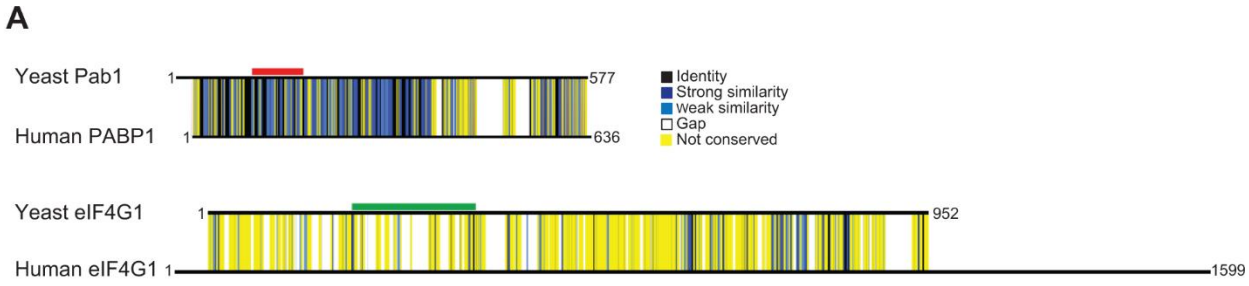
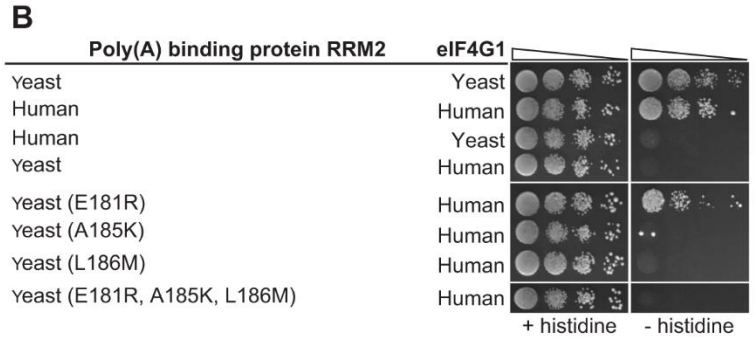


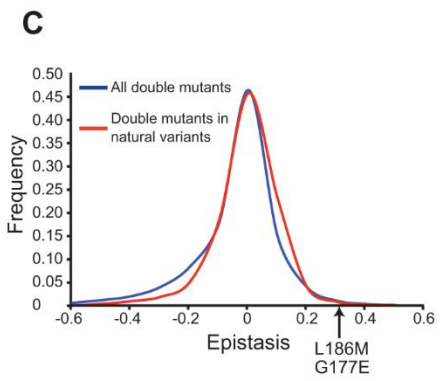
Figure 3.5 Distribution of single amino acid substitution effects across the Pab1 phylogenetic tree.

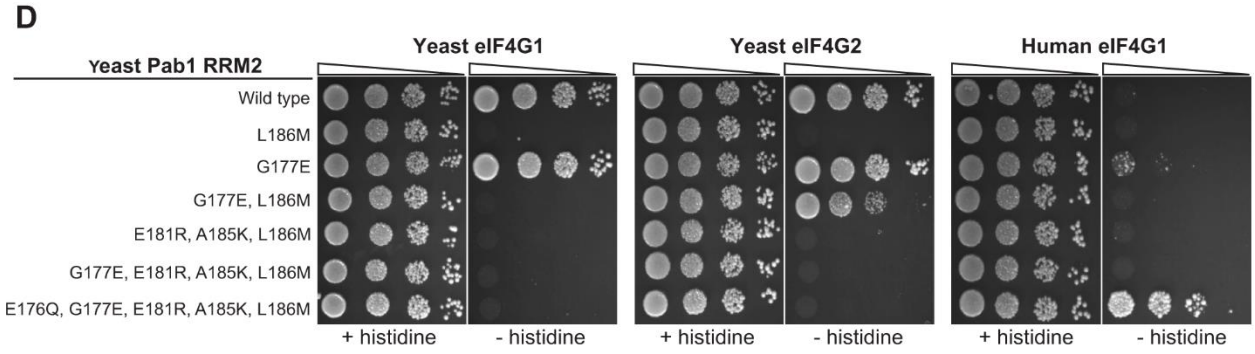
A) A cladogram describing the phylogenetic lineage of 52 Pab1 homologues. The numbered nodes along the tree represent the common ancestors shared between different branches of the tree that lead to the speciation of PABP-1 orthologues in *S. cerevisiae* and *H. sapiens*. The light blue and the orange paths represent the evolutionary lineages of the human and the yeast, respectively, since the fungal and metazoan poly(A)-binding protein divergence. The position of *S. cerevisiae* in the tree is designated by a star. Each column represents one of the 19 single amino acid substitutions from Pab1 RRM2 homologues that were found to be deleterious. Effects by substitution names depicted in gray are less reliable due to lower input read count. The rows are the Pab1 homologues shown in the phylogenetic tree. A filled circle depicts the presence of the specified substitution in a particular sequence. Purple circles depict mutations that interfere with eIF4G binding and black circles represent mutations that had no effect on eIF4G binding. For two deleterious mutations that occur at the same position, the identities are specified within the colored circle. Also specified are the identities of other amino acids found at the same sites as the deleterious mutations that are either mildly deleterious (open black circle) or non-deleterious (all others). The absence of any marking for a given homologue indicates that the same residue is present in *S. cerevisiae* at that position. B) The common ancestors from the *S. cerevisiae* and the *H. sapiens* lineages

shown in (A) are displayed. For each of the designated positions the most probable amino acid is shown (see Material and Methods for reconstruction probability cutoff). Two amino acids are shown with the most likely amino acid designated on the right, if both have high probability score but neither is above the reconstruction probability cutoff. Ancestral sequences with no designated amino acid share the same amino acid identity with the former ancestral sequence.


A
Yeast Pab1
1
577
Human PABP1
1
636
Identity
Strong similarity
weak similarity
Gap
Not conserved
Yeast eIF4G1
1
952
Human eIF4G1
1
1599


B
Poly(A) binding protein RRM2
eIF4G1
Yeast
Yeast
Human
Human
Human
Yeast
Yeast
Human
Yeast (E181R)
Human
Yeast (A185K)
Human
Yeast (L186M)
Human
Yeast (E181R, A185K, L186M)
Human
+ histidine
- histidine


C
All double mutants
Double mutants in natural variants
Frequency
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
Epistasis
L186M
G177E


D
Yeast Pab1 RRM2
Yeast eIF4G1
Yeast eIF4G2
Human eIF4G1
Wild type
L186M
G177E
G177E, L186M
E181R, A185K, L186M
G177E, E181R, A185K, L186M
E176Q, G177E, E181R, A185K, L186M
+ histidine
- histidine
+ histidine
- histidine
+ histidine
- histidine

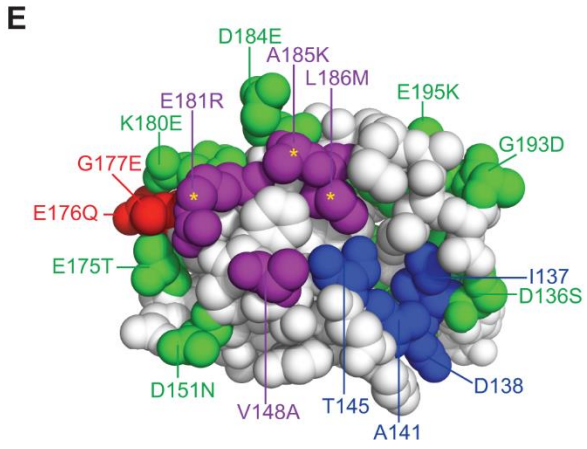

E
D184E
A185K
L186M
E195K
E181R
K180E
G193D
G177E
E176Q
E175T
I137
D136S
D151N
V148A
T145
D138
A141

Figure 3.6 Testing humanizing substitutions for their ability to change the binding specificity of the yeast Pab1 RRM2 to the human eIF4G1.

A) Pairwise alignments of the yeast and the human orthologues of the poly(A)-binding protein-1 and eIF4G1 were generated using Clustal Omega (Sievers *et al.*, 2011). Black horizontal lines designate the protein sequences. Degrees of similarity between aligned positions as well as gaps are color coded with the color key depicted on the left. The yeast Pab1 sequence that is associated with eIF4G1 binding (Kessler and Sachs, 1998) is highlighted with a red line and the yeast eIF4G1 region associated with Pab1 binding (Tarun *et al.*, 1997) is shown with green. B) Yeast two-hybrid assays testing Pab1 RRM2 variants that carry the specified mutations for their ability to interact with the N-terminus (amino acids 1–260) of the human eIF4G1. Cells were serially diluted and seeded as described in Figure 3.2C) Epistasis score distributions of Pab1 variants carrying two mutations that are either found together in natural homologues of the poly(A)-binding protein (red line) or not (blue line). A positive epistasis score indicates a growth rate of the double mutant that was better than expected based on the effects of each of the constituent single mutants alone. An arrow shows the epistasis score of the G177E and L186M double mutant. D) Yeast two-hybrid assays testing Pab1 RRM2 variants that carry the specified mutations for their ability to interact with the N-termini of the yeast eIF4G1 and eIF4G1 and with the human eIF4G1. E) An upper view of the eIF4G1 contact surface of the human PABP1 RRM2 domain (PDB ID 1CVJ). Differences between the yeast and the human sequences are specified by the amino acid substitution nomenclature. eIF4G1 contact site residues that are conserved between yeast and human are shown in blue and contact site residues that differ between the two species are in purple. Deleterious substitutions are marked with yellow asterisks and non-deleterious substitutions are colored in green. E176Q and G177E are shown in red. All other conserved residues between yeast and human are white.

3.5 DISCUSSION

We used the deep mutational scanning data on the yeast Pab1 RRM2 domain to delineate the functional consequences of 197 single amino acid substitutions to residues that occur in Pab1 homologues. As expected (Kondrashov *et al.*, 2002; Soylemez and Kondrashov, 2012), the great majority of these natural substitutions had a minor effect on Pab1 activity, indicating that the primary constraints on poly(A)-binding protein function remain the same among different organisms. Of the 17 strongly deleterious substitutions, nearly all mapped to either protein-protein (RRM2–eIF4G), inter-domain (RRM2–RRM1) or protein-RNA (RRM2–poly(A)) interaction sites, suggesting that all known ligand-binding sites in Pab1 RRM2 experienced some degree of divergence over evolutionary time.

We characterized the eIF4G-binding site in Pab1 at single amino acid resolution, demonstrating that integrating results from mutagenesis with natural variation data provides a compact list of mutations that are likely to interfere with protein-ligand interaction. The rapid generation of this list overcomes limitations of other mutagenesis approaches. In particular, deletion experiments are unlikely to provide an accurate map of the eIF4G contact site in Pab1, as its critical residues span most of the primary sequence of RRM2 and are brought together by the three-dimensional structure. An alanine scan, which tests the effects of substituting single amino acids to alanine, would prove insufficient to identify the involvement of T145, V148 and E181 in eIF4G binding, as judged by the minor effects of these alanine changes on the *in vivo* function of Pab1 (Melamed *et al.*, 2013).

Although our combined approach delineated eight Pab1 RRM2 surface residues that are associated with eIF4G1 binding, the large-scale two-hybrid assay identified nine additional residues, located mostly at the periphery of the contact site. The higher sensitivity of the RRM2 domain to mutations at the eIF4G1 contact site in the two-hybrid assay is likely due to the higher selection pressure in this assay, than *in vivo*, for the RRM2 eIF4G1 interaction, as mutations that disrupt this interaction are not lethal (Kessler and Sachs, 1998; Otero, 1999; Melamed *et al.*, 2013). Nonetheless, the differences between the regions that were highlighted by the two approaches point to the central residues, discovered by our combined approach, as more important to the *in vivo* function of the eIF4G1 binding site than the peripheral residues, added by the yeast two-hybrid assay. This difference in the importance of residues to the interaction is likely to mirror the higher evolutionary conservation of the central binding site residues compared to the peripheral ones (Bordner and Abagyan, 2005).

Despite the greater ability of the two-hybrid assay to uncover positions on the RRM2 surface that associate with eIF4G1 binding, our combined approach of using natural variation to filter the deep mutational scanning results on the *in vivo* function of RRM2 yields an increased fraction of mutations that

interfere with eIF4G1 binding, Because large-scale mutational data are usually not available for a protein interaction, these results emphasize the advantage of this combined approach to identify the effects of mutations on protein interactions. In addition, while mutations that damage the structure of a protein can affect the two-hybrid readout, the short list of candidate mutations created by the *in vivo* approach is likely to exclude such indirect effects (Chothia and Lesk, 1986)

3.5.1 *Implications for Pab1 and eIF4G1 activity*

Elucidating contact sites with high resolution is important to clarify how proteins exert their functions. With respect to Pab1, we found that the binding sites for eIF4G1 and eIF4G2 extend beyond the helix α2 element (Otero, 1999) to include part of helix α1. The inclusion of this helix provides a plausible explanation for the molecular mechanism that couples poly(A) and eIF4G binding by Pab1. In yeast, binding of eIF4G to Pab1 requires the prior association of Pab1 with poly(A) in order to promote translation (Kessler and Sachs, 1998). In human, these sequential steps are separated by inter-domain allostery of RRM2 and RRM1, allowing PABP1 to adopt a more extended conformation in the presence of RNA (Safaee *et al.*, 2012). Since the association of RRM2 and RRM1 involves direct interactions between helix α1 of RRM2 and helix α2 of RRM1 (Deo *et al.*, 1999; Safaee *et al.*, 2012), conformational changes of the two domains might make helix α1 of RRM2 and its surrounding residues available for eIF4G association upon poly(A) binding. Our finding that a Pab1 fragment consisting only of RRM1–RRM2 was unable to bind eIF4G supports the regulatory role of this inter-domain interaction in this function.

eIF4G1 and eIF4G2 are functionally interchangeable under optimal growth conditions (Clarkson *et al.*, 2010). However, differences in eIF4E co-purification and *in vitro* translation efficiencies suggest that each of the two isoforms possesses unique roles in translation under non-optimal conditions (Tarun and Sachs, 1996; Tarun *et al.*, 1997). Despite the overlap in location and similar mutational sensitivity of the binding sites for eIF4G1 and eIF4G2, a few Pab1 RRM2 substitutions resulted in differential sensitivities to binding. Whether this difference in Pab1 RRM2 binding points to altered mechanisms of action is a matter of further studies. T145L, which bound only to eIF4G2, might be useful in clarifying specific roles for the isoforms in translation.

3.5.2 *Implications for the evolution of binding sites*

We identified three substitutions (E181R, A185K and L186M), corresponding to the residues present in the human PABP1, each of which when introduced into the yeast Pab1 eliminated its binding to yeast eIF4G1. We tested whether these substitutions might switch the specificity of Pab1 to bind human

eIF4G1. The single humanizing substitution, E181R, allowed the yeast Pab1 RRM2 to bind to human eIF4G1, demonstrating that in spite of sequence diversification, the human and yeast orthologues of eIF4G1 and Pab1 share similarities with respect to their physical association. However, Pab1 carrying A185K and L186M did not bind to the human eIF4G1, even after humanizing the contact site by other substitutions. Thus, this shared similarity in binding activity is likely to be maintained by other intra-protein interactions in Pab1 that compensate for the otherwise deleterious effect of A185K and L186M.

Our finding that two substitutions that are both present in an individual homologue are more likely to display positive epistatic interactions suggests that compensating mutations reconstruct functional modules that are conserved between organisms despite changes in the amino acid sequence that comprise these modules. Indeed, that the addition of the G177E substitution repairs the binding of an RRM2 L186M mutant to the yeast eIF4G2 suggests that the two humanizing substitutions restore a functional binding site for the yeast eIF4G2. Additional studies will be required to determine whether the tendency for positive epistasis of two substitutions present in a homologous sequence is a universal property of proteins or a specific feature of Pab1. Nonetheless, it is likely that substitutions from paralogues of closely related species are more prone to this type of epistasis than substitutions from other homologous sequences, given the functional conservation and the small number of amino acid changes in these paralogues.

Additionally, G177E together with E176Q combined with the three deleterious substitutions E181R, A185K and L186M to allow yeast Pab1 binding to the human eIF4G1. This finding supports the use of epistatic interactions between two natural substitutions tested in a model organism to infer similar interactions between those residues in their natural context. We suggest that systematic integration of large-scale epistasis data with bioinformatic tools that use sequence homology might improve prediction accuracies of co-evolutionary relationships and functional association between residues.

3.6 CONCLUSIONS

Approximately 20% of *S. cerevisiae* genes are essential for growth on rich glucose medium (Giaever *et al.*, 2002), with many of the remaining genes required upon environmental or genetic perturbations. Therefore, growth selections compatible with deep mutational scanning can be used to study the *in vivo* function of a large fraction of yeast proteins. This experimental strategy can also be applied to cross-species complementation assays to analyze human proteins in yeast (Zhang *et al.*, 2003; Marini *et al.*, 2010; Dunham and Fowler, 2013). However, the score assigned to each protein variant reflects the consequence of mutation only on growth rate. Therefore, inferring the direct impact of mutations on an *in vivo* activity such as ligand binding remains challenging. Here we show that

integrating deep mutational scanning results with natural variation data provides a high throughput inter-species complementation assay that can be used to identify and characterize functional regions in proteins, including protein-protein contact sites. In addition, the large-scale analysis of natural amino acid substitutions can provide an experimental platform to evaluate the performance of computational tools that use protein homology to predict function and co-evolutionary relationships.

3.7 MATERIALS AND METHODS

3.7.1 *Deep mutational scanning*

The procedures for Pab1 RRM2 deep mutational scanning, including establishment of the experimental platform, construction of mutant libraries, sequencing of RRM2 DNA fragments and data analysis were previously described (Melamed *et al.*, 2013). Unless otherwise indicated, only variants with input-read counts greater than 40 were used for the analysis.

3.7.2 *Plasmids*

pOBD2 and pOAD were used to test the interactions between Pab1 and eIF4G isoforms in the yeast two-hybrid system. Full length *PAB1* encoding amino acids 1–578 (DMP87) was PCR amplified from pCM188-Pab1 (Melamed *et al.*, 2013) and cloned into the NcoI and SalI sites of pOBD2. The following *PAB1* truncations, encoding amino acids 1–343 (DMP88), 1–204 (DMP183), 123–204 (DMP180) and 1–120 (DMP179) were PCR amplified from p415GPD-Pab1-343BX (Melamed *et al.*, 2013) and cloned into the NcoI and SalI sites of pOBD2. *PAB1* fragments encoding amino acids 123–204 (RRM2) with the point mutations I137F (DMP201), D138T (DMP202), K140A (DMP203), A141D (DMP230), T145H (DMP204), T145L (DMP189), V148K (DMP205), L153V (DMP206), S155V (DMP207), K156N (DMP208), A158E (DMP209), K180E (DMP197), E181R (DMP193), D184E (DMP188), D184W (DMP210), A185H (DMP211), A185K (DMP186), L186M (DMP185), Y197N (DMP191), Y197V (DMP194), A199E (DMP192), [K180E, E181R] (DMP198), [D184E, A185K, L186M] (DMP190), [E181R, A185K, L186M] (DMP235), [E181R, A185K, L186M, A158V, T159C] (DMP286), [E181R, A185K, L186M, E176Q, G177E] (DMP287), [E181R, A185K, L186M, V148A, K180E, D184E] (DMP291), [E181R, A185K, L186M, P135K, Q194R] (DMP292), G177E (DMP293), [G177E, L186M] (DMP297) and [E181R, A185K, L186M, G177E] (DMP298) were created by PCR using the same p415GPD-Pab1-343BX plasmid as a template and cloned into the NcoI and SalI sites of pOBD2, C-terminal and in-frame with the Gal4 DNA binding domain. eIF4G1 and eIF4G2 fragments encoding amino acids 1–341 were amplified from yeast genomic DNA (strain W-303) and cloned into the

EcoRI and SalI sites of pOAD, C-terminal and in-frame with the Gal4 activation domain (DMP92 and DMP212, respectively). The human PABP1 fragment encoding amino acids 95–176 was amplified from HsCD00042197 (PlasmidID) and cloned into the NcoI and SalI sites of pOBD2 (DMP264). The human eIF4G1 fragment encoding amino acids 1–260 was amplified from HsCD00342900 (PlasmidID) and cloned into the NcoI and SalI sites of pOAD (DMP265)

3.7.3 *Individual yeast two-hybrid assays*

Yeast strain PJ694a (*MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4Δ gal80Δ LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ*) carrying pOBD2- and pOAD-based vectors were grown overnight, at 30°C in synthetic complete media lacking leucine and tryptophan. To test for activation of the *HIS3* reporter gene, cells were spotted in a dilution series on synthetic complete plates lacking leucine and tryptophan, with or without histidine and grown at 30°C for three days.

3.7.4 *Data for natural single amino acid variations*

We collected 52 Pab1 homologues (see Supplementary File 3.3 for sequences and accession numbers), representing sequences of all poly(A)-binding proteins that carry four consecutive RRM domains that can be found in the UniProtKB/SwissProt database (Magrane and Consortium, 2011), which contains high quality annotations of protein sequences. Multiple sequence alignment (MSA) was performed using Clustal Omega (Sievers *et al.*, 2011) with default parameters (Supplementary File 3.4). Enrichment scores for natural and non-natural single amino acid substitutions were obtained from Supplementary Table 2 of Melamed et al (Melamed *et al.*, 2013).

3.7.5 *Properties of natural and non-natural amino acid substitutions*

To assess the fraction of natural substitutions that result in impaired function, enrichment score distributions of 160 natural single amino acid substitutions, 539 non-natural single amino acid substitutions and 229 synonymous variants with input read counts greater than 500 were determined. The stringent input read count threshold was set to minimize fluctuations of enrichment scores due to low representation of variants in the library pools. The enrichment scores distribution of the synonymous variants was used as a proxy for the enrichment scores distribution of non-deleterious variants in the dataset. To estimate the fraction of deleterious substitutions within the natural substitutions, for each enrichment score bin shown in Figure S3.7B, we subtracted the estimated fraction of non-deleterious variants from the fraction of the natural variants.

For each single amino acid substitution, the fractional accessible surface area (ASA) was obtained for the side chains of the wild-type residue in the human PABP1 RRM2 structure (PDB ID 2K8G) using VADAR server, version 1.8 (Willard *et al.*, 2003). Data for K164 residue was omitted, as this residue is absent from the human RRM2 (see Figure 3.2A). The Blosum62 matrix was used to score each substitution to determine the degree of conservation. Box plots were generated using R-studio software.

3.7.6 *Large-scale yeast two-hybrid assays*

RRM2 sequences containing one of the three library segments were PCR amplified from the library plasmids that were previously described (Melamed *et al.*, 2013) and cloned into the NcoI and SalI sites of pOBD2. Yeast expressing the RRM2 hybrid containing one of the three libraries were grown to log phase in SC medium lacking leucine and tryptophan, supplemented with 2% glucose, and diluted into fresh medium lacking leucine, tryptophan and histidine to a final concentration of 4×10^4 cells/mL. Selection was carried out for 21 h with the culture growing to a density of 5×10^6–1×10^7 cells/mL. 2.5×10^8 cells from each culture were collected before ("input") and after selection ("selected"). Library preparation for high throughput sequencing was carried as previously described (Melamed *et al.*, 2013). Amplicons were created with internal primers that flanked each library segment and carried at their 5' end common sequencing targets for Illumina read1, read2 and index primers (11 PCR cycles) and with external primers that added Illumina adapter sequences (8 PCR cycles). Amplicons were sequenced by an Illumina NextSeq500 using paired-end reads. We used the Enrich software package (Fowler *et al.*, 2011) to filter for high quality reads (base Q score >20). Based on the variance of enrichment scores of 2423 synonymous variants (i.e. variants that encode the wild-type Pab1 RRM2 protein sequence and carry at least one synonymous mutation), we selected variants with at least 20 input read counts for further analysis (synonymous variance <0.4 for all three libraries). Enrichment scores of single amino acid substitutions were log2 transformed and visualized using Java TreeView 1.1.6r2 (Saldanha, 2004). Average linkage hierarchical clustering with a Euclidean distance similarity metric for both RRM2 residues and substituting amino acids was performed using Gene Cluster 3.0 (Hoon *et al.*, 2004).

3.7.7 *Construction of a phylogenetic tree and determination of ancestral states*

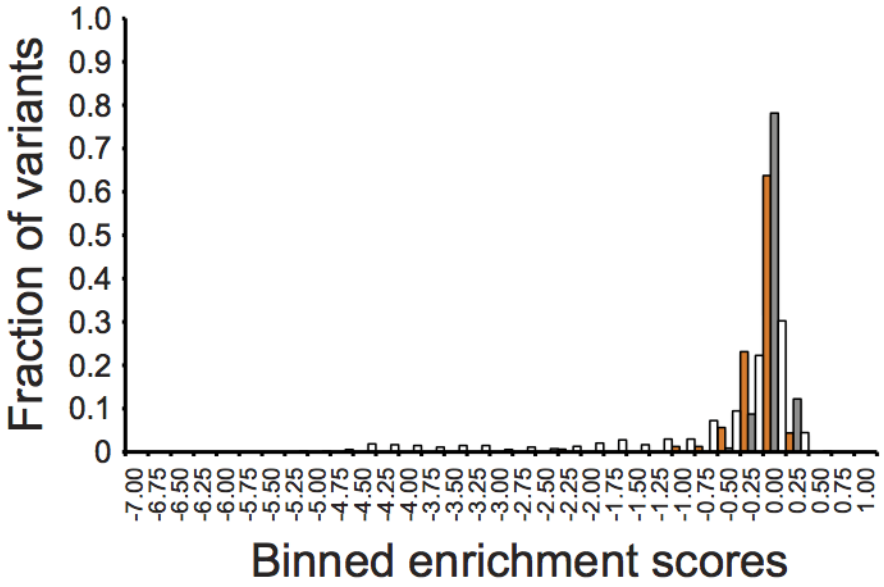
A maximum likelihood tree was constructed using the Phylogeny.fr tool (Dereeper *et al.*, 2008) using default parameters. Probabilities for ancestral states were calculated using the JTT model of substitution by the FastML tool (Ashkenazy *et al.*, 2010). Ancestral amino acids were considered "true" if their reconstruction probabilities were greater than 0.7 (the sum of probabilities for all amino acids equals

1.0). Otherwise, the two most probable amino acids with a minimal probability of 0.3 for each, and sum of probabilities greater than 0.75 were considered.

3.7.8 *Structure visualization*

The human PABP1 RRM1-RRM2 structure (PDB ID 1CVJ) was visualized using PyMol software (version 1.5.0.5).

A



B

	Total variants	Non deleterious Score > -0.15	Mildly deleterious -0.15>Score>-0.5	Strongly deleterious -0.15>Score>-0.5
Synonymous Observed	229 (100%)	209 (91.3%)	19 (8.3%)	1 (0.4%)
Natural Observed	160 (100%)	115 (71.25%)	38 (23.75%)	8 (5%)
Natural Expected	-	146.1 (91.3%)	13.3 (8.3%)	0.64 (0.4%)
Estimated fraction of non deleterious variants	-	1.00	0.35	0.08

Figure S3.7 Enrichment scores distribution of synonymous mutations to assess the contamination of the mildly and strongly deleterious groups by non-deleterious mutants.

A) A bar graph shows the enrichment score distributions of natural (orange), non-natural (white), and synonymous variants (gray) with input read counts greater than 500. B) Natural and synonymous variants falling into the non-deleterious, mildly deleterious and strongly deleterious categories were counted (Synonymous Observed and Natural Observed rows). The expected number of natural variants in each category based on the distribution of synonymous variants, which serve as a proxy for non-deleterious variants, is shown (Natural Expected). For the natural substitution variants, the ratio between the expected and the observed numbers in each category represents the estimated fraction of contamination by non-deleterious mutants.

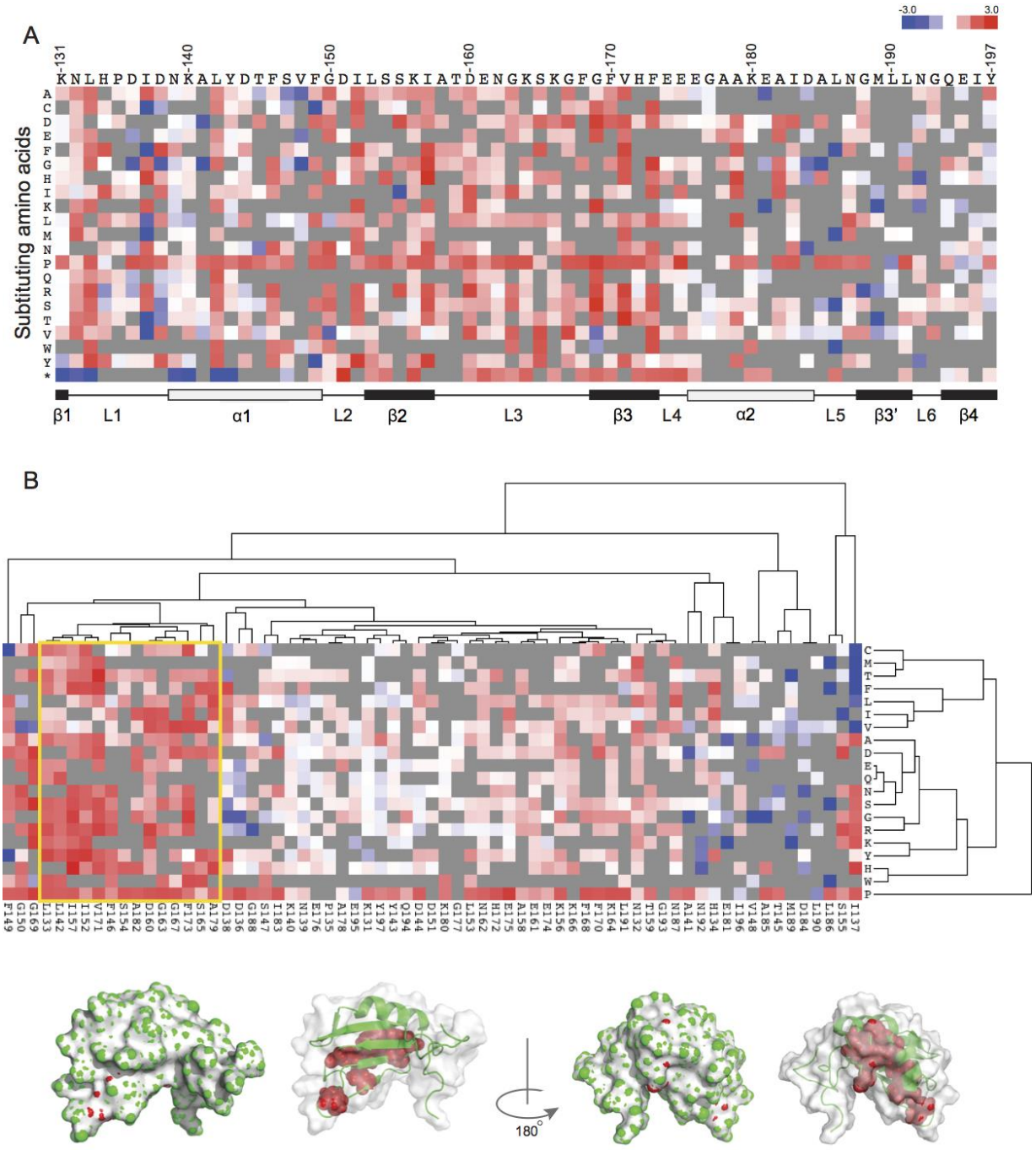


Figure S3.8 Effects of single amino acid substitutions in the RRM2 domain on the two-hybrid interaction for RRM2–eIF4G1 binding.

A) A heat map showing the enrichment scores (log_2 transformed) for single amino acid substitutions in the Pab1 RRM2 sequence. Each column represents a site in the RRM2 sequence and each row a substitution to a specific amino acid. An asterisk designates nonsense mutations. Color ranges from blue for the most detrimental mutations to red for the most beneficial. Effects of stop codons are highly detrimental through most of the first library segment, but become beneficial if they occur C-terminal to helix-α1. In addition, substitutions to proline, which were the most deleterious *in vivo* due to their presumed interference with α-helix and β-sheet structures (Melamed *et al.*, 2013), were highly enriched in

this assay. B) Clustering the effects of single amino acid substitutions based on enrichment scores. The yellow frame outlines a cluster of positions that displayed a beneficial effect in response to most substitutions. Below, the locations of these positions in the protein's core are highlighted in red spheres in the RRM2 structure (PDB_ID 2K8G). All other positions are shown in green. While these observations point to fortuitous activation of two-hybrid transcription due to mutations that disrupt the RRM2 structure, the assay was highly sensitive to mutations in other positions known to be involved in eIF4G1 binding.

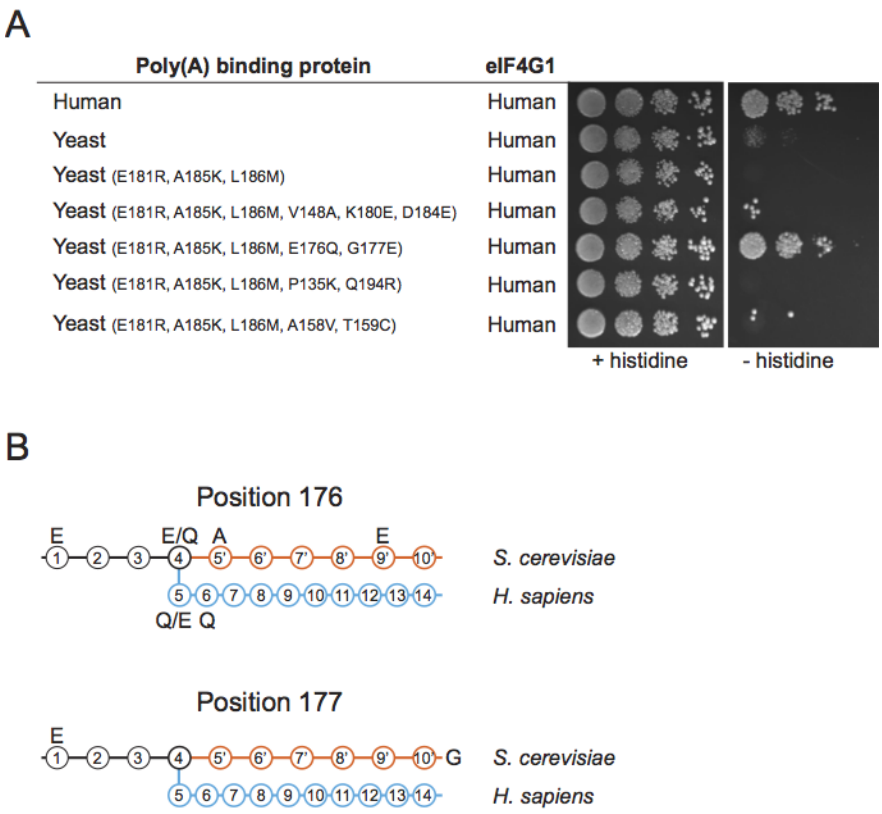


Figure S3.9 A test of humanizing substitutions for their ability to change the binding specificity of the yeast Pab1 RRM2 to the human eIF4G1.

A) Yeast two-hybrid assays testing Pab1 RRM2 variants that carry the specified mutations for their ability to interact with the N-terminus (amino acids 1–260) of the human eIF4G1. Cells were serially diluted and seeded as described in Fig. 2. B) The common ancestors of positions 176 and 177 from the *S. cerevisiae* and the *H. sapiens* lineages shown in Fig. 4A are displayed. For each of the designated positions the most probable amino acid is shown (see Materials and Methods for reconstruction probability cutoff).

Supplementary Table 3.1 See (Melamed *et al.*, 2015)

Supplementary Table 3.2 See (Melamed *et al.*, 2015)

Supplementary File 3.3 See (Melamed *et al.*, 2015)

Supplementary File 3.4 See (Melamed *et al.*, 2015)

Chapter 4. A DEEP MUTATIONAL SCAN OF A TRNA

Chapter 4 appeared in this form in the journal *Genes & Development* (Guy/Young *et al.*, 2014). My contributions were in the sequencing, processing, filtering, and scoring of the variants from the sequence data, as well as in the analyses and graphing of the epistasis data. I also contributed to the comparison of functional data with folding free energy and ensemble defect. I contributed to the creation of all figures.

4.1 ABSTRACT

Sequence variation in tRNA genes influences the structure, modification, and stability of tRNA, affects translation fidelity, impacts the activity of numerous isodecoders in metazoans, and leads to human diseases. To comprehensively define the effects of sequence variation on tRNA function, we developed a high-throughput in vivo screen to quantify the activity of a model tRNA, the nonsense suppressor $SUP4_{oc}$ of *Saccharomyces cerevisiae.* Using a highly sensitive fluorescent reporter gene with an ochre mutation, fluorescence-activated cell sorting of a library of $SUP4_{oc}$ mutant yeast strains, and deep sequencing, we scored 25,491 variants. Unexpectedly, $SUP4_{oc}$ tolerates numerous sequence variations, accommodates slippage in tertiary and secondary interactions, and exhibits genetic interactions that suggest an alternative functional tRNA conformation. Furthermore, we used this methodology to define tRNA variants subject to rapid tRNA decay (RTD). Even though RTD normally degrades tRNAs with exposed 5′ ends, mutations that sensitize $SUP4_{oc}$ to RTD were found to be located throughout the sequence, including the anti-codon stem. Thus, the integrity of the entire tRNA molecule is under surveillance by cellular quality control machinery. This approach to assess activity at high throughput is widely applicable to many problems in tRNA biology.

4.2 INTRODUCTION

tRNAs allow the genetic code to be correctly interpreted for protein synthesis. In consequence, their sequence is under three strong functional constraints. First, tRNAs require similar overall structures to participate equivalently in translation (Kim *et al.*, 1974; Westhof *et al.*, 1985; Basavappa and Sigler, 1991; Giegé *et al.*, 2012). Second, tRNAs require unique features to ensure specific charging by cognate synthetases and accurately decode mRNA, both of which often require specific modifications (Pütz *et al.*, 1994; Giegé *et al.*, 1998; Murphy *et al.*, 2004; Agris *et al.*, 2007) and elements outside the anti-codon (Musier-Forsyth *et al.*, 1991; Cochella and Green, 2005; Ledoux *et al.*, 2009; Ling *et al.*, 2009; Shepotinovskaya and Uhlenbeck, 2013). Third, tRNAs must be stable enough to survive for multiple

generations (Gudipati *et al.*, 2012) and avoid turnover (Whipple *et al.*, 2011) yet flexible enough to accommodate conformational changes during ribosome passage (Valle *et al.*, 2003; Schmeing *et al.*, 2009; Zhou *et al.*, 2013).

These sequence constraints on tRNA function suggest that tRNAs would be largely intolerant to mutation. Indeed, numerous tRNA mutations in yeast adversely affect function (Kurjan *et al.*, 1980). Over 230 mitochondrial tRNA mutations have been associated with human diseases (Ruiz-Pesini *et al.*, 2007), including encephalopathy, hearing loss, ataxia, myopathy, diabetes, epilepsy, neuropathy, and gastrointestinal dysfunction (Yarham *et al.*, 2010); these mutations occur in all stems and loops.

Nonetheless, there are also data demonstrating that tRNAs can tolerate variant sequences in the stems and loops. For example, the yeast *Saccharomyces cerevisiae* $tRNA^{Arg(CCG)}$ gene retains function with any of several D-loop or anti-codon loop mutations (Geslain *et al.*, 2003), and the yeast $tRNA^{Ser(CGA)}$ gene is fully functional with any of seven different base-pair swaps in the acceptor or T stems that retain secondary structure (Whipple *et al.*, 2011). Similarly, several *Escherichia coli* variants of a partially impaired $tRNA^{Ala(CUA)}$ amber suppressor tRNA retain activity with individual mutations in the acceptor stem, the anti-codon stem, or the T stem (Hou and Schimmel, 1992).

These seemingly conflicting data make it difficult to predict the effects on tRNA function of the numerous naturally occurring sequence variations in the metazoan tRNA isodecoders, which have the same anti-codon but altered tRNA bodies (Goodenbour and Pan, 2006). In addition, most of the numerous disease-associated mitochondrial tRNA variants are poorly understood (Suzuki *et al.*, 2011). Prediction of the function of variants is further complicated by the multiple modifications and quality control pathways that influence tRNA activity (Kadaba *et al.*, 2004; Chernyakov *et al.*, 2008; Hopper, 2013; Kramer and Hopper, 2013) and by the tRNA internal promoter, which is not quantitatively understood (Koski *et al.*, 1980; Pearson *et al.*, 1985; Kaiser and Brow, 1995; Marck *et al.*, 2006; Orioli *et al.*, 2012).

Although there is a wealth of information on the effects of mutating individual tRNA residues on specific steps of tRNA processing and function (Normanly *et al.*, 1986; Schultz and Yarus, 1994; Yan and Francklyn, 1994; Fechter *et al.*, 2000; Schrader *et al.*, 2009), there has been no quantitative analysis at a large scale of the effects of mutations on tRNA biology. Here we describe the use of a sensitive fluorescent reporter and deep sequencing to quantify the in vivo function of thousands of variants of a tRNA suppressor in the yeast *S. cerevisiae* and the use of this system to comprehensively define the biological substrates of a prominent tRNA decay pathway (Alexandrov *et al.*, 2006). We identified a large number of mutated tRNAs that are functional, suggesting that tRNA structure is much more flexible than anticipated, and found that the tRNA decay pathway unexpectedly acts on many more classes of variants than previously known

or predicted.

4.3 RESULTS

4.3.1 *Quantification of tRNA function by cell sorting of yeast carrying a library of tRNA variants*

To analyze the effect of mutations on tRNA function in vivo, we sought a model system in which we could assay tRNA activity quantitatively, with high sensitivity and on a large scale. In yeast, suppression of a stop codon in the green fluorescent protein gene (*GFP*) allows fluorescence-activated cell sorting (FACS) of millions of cells based on their level of suppression by a nonsense suppressor tRNA. To test the feasibility of this approach, we integrated the nonsense suppressor $SUP4_{oc}$ ($tRNA^{Tyr}$-G34U) into a yeast strain bearing GFP_{oc} in the RNA-ID reporter (Dean and Grayhack, 2012); this strain allows a comparison of the expression of GFP_{oc} to the control red fluorescent protein gene (*RFP*) (Figure 4.1A). GFP_{oc}/RFP was minimal without suppression (0.004 of *GFP/RFP*) but was nearly normal (0.94 of *GFP/RFP*) with $SUP4_{oc}$ (Fig. 1B), as anticipated for this stop codon because of its poor termination context (Bonetti *et al.*, 1995; Dean and Grayhack, 2012). Based on these data, this GFP_{oc} expression assay discriminates with high resolution among tRNA variants, with a 235-fold dynamic range of expression and limited variation in GFP/RFP values for individual cells of a variant (Dean and Grayhack, 2012). Moreover, the assay measures the net contribution of all steps of tRNA biogenesis and translation except fidelity.

We constructed a library of ~220,000 $SUP4_{oc}$ variants, each integrated into the yeast RNA-ID strain and bearing ~3% random mutations in nucleotides 1–33 and 38–73 (conventional numbering) (Supplemental Figure S4.7A). We grew this library at 28°C, sorted cells into four bins by FACS (Figure 4.1C), PCR-amplified the $SUP4_{oc}$ allele from the pooled genomic DNA from each bin, and evaluated the bin distribution of individual variants by sequencing (Supplemental Table S4.1), similar to an approach used to measure gene expression from thousands of designed promoters (Sharon *et al.*, 2012). The fractional representation of reads for each variant in each bin was converted to a GFP/RFP ratio, which was normalized to the $SUP4_{oc}$ ratio to define relative function (termed GFP^{SEQ}). Filters were then applied to score only those variants with ≥100 reads and enough reads to measure the distribution of ≥30 cells (Supplemental Figure S4.7B).

Overall, we scored 25,491 variants (Supplemental Table S4.2), including all 213 single variants. GFP^{SEQ} was highly reproducible for single mutants of a biological replicate (Supplemental Figure S4.7C), with $R^2 = 0.99$. We also confirmed tRNA activity of 60 variants by reconstruction and flow cytometry analysis of the variants; each activity, normalized to the $SUP4_{oc}$ ratio, yielded a GFP/RFP ratio termed

GFP^{FLOW}, which correlated highly with the corresponding GFP^{SEQ} up to GFP^{FLOW} of 0.4 (R^2 = 0.90) (Supplemental Figure S4.7D). To further enhance resolution of highly active variants, we used FACS to subdivide bin 1 into three fractions, extending the linear range of GFP^{SEQ} values to GFP^{FLOW} of 0.55. It is not clear why GFP^{SEQ} is systematically approximately twofold higher than GFP^{FLOW}, resulting in a correlation between GFP^{SEQ} and GFP^{FLOW} that only extends up to 0.55 (Supplemental Figure S4.7D). Some of the discrepancy is likely due to the limited resolution of high-fluorescence variants, even in the bin 1 subdivision data set. In addition, PCR chimerism in low-fluorescence bins can lead to spurious wild-type reads, thereby underestimating the function of the wild-type tRNA by the sequencing approach, which in turn leads to overestimation of variant function by GFP^{SEQ}. Finally, minor systematic errors may be introduced by the use of different instruments for GFP^{SEQ} and GFP^{FLOW} measurements and the steps of bin collection and plating, PCR amplification, and sequencing.

4.3.2 *SUP4oc is highly tolerant of mutations*

To characterize the mutational consequences in $SUP4_{oc}$, we initially analyzed the 213 single variants, given both their relative simplicity and the previous studies that examined single mutations in this tRNA. $SUP4_{oc}$ is remarkably tolerant of single mutations, with 44 highly functional variants ($GFP^{SEQ} \geq 0.9$) (Figure 4.1D, dark blue) and 27 substantially functional variants (0.18–0.9) (Figure 4.1D, blue), along with nine marginally functional variants (0.026–0.18) (Figure 4.1D, light blue). We note that there are minimal consequences due to the higher values of GFP^{SEQ} relative to GFP^{FLOW}. Thus, 26 of 32 nonfunctional or marginally active variants by GFP^{SEQ} were correctly annotated based on reconstruction and GFP^{FLOW} analysis, and six nonfunctional variants had trace amounts of GFP^{FLOW} activity (Supplemental Figure S4.7D). Similarly, 13 of 16 of highly functional variants that were tested by reconstruction and flow cytometry had GFP^{FLOW} values >0.7.

The highly or substantially active variants were heavily clustered in specific residues. These included each of the three possible mutations of all five D-loop uridine residues; U4 of the acceptor stem; A9, A13, and A22 of the D loop; residues 44, 45, and 47 of the variable loop; C59 of the T loop; and G62 of the T stem. These results are consistent with, and substantially extend, previous analyses of functional variants of $SUP4_{oc}$ (Kurjan and Hall, 1982; Kohalmi and Kunz, 1992), yeast $tRNA^{Arg(CCG)}$ (Geslain *et al.*, 2003), and an *E. coli* alanine amber suppressor tRNA (Hou and Schimmel, 1992). In contrast, residues that did not tolerate any single mutations included those in conserved tertiary pairs (U8–A14, R15–Y48, G18–U55, G19–C56, and U54–A58), emphasizing the requirement of the L-shaped tertiary fold of the tRNA for activity.

Although our data emphasize that the integrity of the four stems must be intact for tRNA to have

full function, flexibility is observed at two locations. Single- and double-mutant variants that preserve canonical pairing were often functional (Figure 4.2A), with the notable exceptions of the G53–C61 pair, which is highly conserved as part of the B-box of the internal promoter (Marck *et al.*, 2006); C1–G72, which is a determinant for tyrosine charging (Fechter *et al.*, 2000); and the G10–C25 and C11–G24 pairs of the D stem, which is comprised of only 3 base pairs (bp). In contrast, only eight of 140 stem variants with noncanonical pairing had a $GFP^{SEQ} > 0.5$, and these eight included four variants of U4–G69 and two variants of C52–G62. Although functional variants with mismatches at U4–G69 might be anticipated because of the weak U–G pair and the known mismatches that occasionally occur among stem base pairs in tRNAs, it is unclear why $SUP4_{oc}$ tolerated mismatches at C52–G62, since this position is rarely occupied by a mismatched pair, G–U, or U–G (Marck and Grosjean, 2002).

Our data also indicate that the tertiary fold must be intact, since little sequence variation is observed in the conserved tertiary pairs. Indeed, 45 of 47 variants with mutations in these pairs resulted in a completely nonfunctional tRNA, and the remaining two had only marginal activity (Figure 4.2A).

Among the 9349 double-mutant variants, 1499 were active, including 685 substantially or highly functional variants. One important requirement for activity is a low ensemble defect (ED), which is a parameter that estimates the propensity of a tRNA to misfold (Zadeh *et al.*, 2011a). According to our data, almost all functional variants had an estimated per nucleotide ED < 0.21 (Figure 4.2B, 95% cutoff, yellow), which is well within the range of native eukaryotic tRNAs (Supplemental Figure S4.8A).

4.3.3 *Unexpected positive interactions between residues*

To identify previously unappreciated parameters important for tRNA function, we examined double-mutant variants that displayed positive (or negative) epistasis, indicating that they functioned substantially better (or worse) than anticipated from the scores of the corresponding single-mutant variants. Epistasis within a protein or RNA can reveal interactions between residues when the phenotype caused by one mutation is dependent on mutation at another residue. Based on a multiplicative model, we calculated an epistasis score by subtracting the product of the GFP^{SEQ} scores of two single variants from that of the corresponding double variant (Supplemental Table S4.3, Supplemental Table S4.4). Most double variants scored close to their predicted values (Figure 4.2C), but 6.9% had substantial negative epistasis (defined as a score ≤0.18) (Supplemental Figure S4.8B), and 1.5% had positive epistasis (>0.18) (Figure 4.2D). As might be expected for a molecule with severe sequence constraints, there was a large excess of negative epistasis over positive epistasis. Indeed, of the double variants that had GFP^{SEQ} scores that allowed the possibility of negative or positive epistasis, 62% were negatively epistatic, whereas only 1.4% were positively epistatic, and this excess was not dependent on the epistasis cutoff score used (Supplemental

Figure S4.8C). However, the 6.9% of total double variants with negative epistasis includes a remarkably large number (202 of 648, 31%) of completely nonfunctional doubles in which both singles were highly functional, suggesting that while the tRNA tolerates single mutations at multiple locations with little loss of function, it is extremely sensitive to a second mutation.

Many of the 1.5% of double variants displaying positive epistasis can be explained simply, such as by restoration of a base pair that was lost in both of the corresponding single variants. However, there were several striking examples of unexpected positive epistasis, four classes of which are highlighted below because they suggest structural rearrangements.

First, an alternative tRNA conformation appears to form in variants with mutations in the 26- to 44-nucleotide (nt) pair. The nucleotides at residues 26 and 44 are mismatched ~65% of the time in eukaryotes and, in known structures, often form a propeller-twisted noncanonical base pair in a Watson-Crick-like orientation (Figure 4.3A; (Kim *et al.*, 1974)); however, these nucleotides are also frequently canonically paired, with Watson-Crick (17%) or G–U (18%) pairings (Marck and Grosjean, 2002). We found that the A44U mutation (opposite G26) had nearly opposite effects on the function of double variants, dependent on the identity of the other mutation. Thus, the A44U mutation substantially rescued the function of variants with the destabilizing anti-codon stem mutations A29C, A29U, and A28U; in contrast, the A44U mutation had large negative epistatic effects with A9U, A9C, A22U, and G57A (Figure 4.3B,C; Supplemental Table S4.4; Supplemental Figure S4.9A), all of which often participate in the tertiary fold (Giegé *et al.*, 2012). One likely interpretation of these results is that A44U alters tRNA conformation by pairing with G26, strengthening the anti-codon stem and thereby countering other destabilizing anti-codon stem mutations while simultaneously causing structural shifts that impair the function of variants with otherwise benign mutations affecting the tRNA fold. It is notable that the 26–44 pair is in the "hinge" region of tRNA, which undergoes substantial conformational changes during ribosome passage in the A/T state with EF-Tu (Valle *et al.*, 2003; Schmeing *et al.*, 2009) and in the pe*/E state during translocation (Zhou *et al.*, 2013). Flexibility in this region of the tRNA may also explain why the inactive G26U variant (opposite A44) was substantially rescued by mutation of G45 (Supplemental Table S4.4; Supplemental Figure S4.9B–D). G45 sometimes interacts with the 10–25 pair (Westhof *et al.*, 1985; Gautheret *et al.*, 1995; Giegé *et al.*, 2012), suggesting that mutating G45 could alter or break this tertiary interaction, thus adding more flexibility to the hinge region and allowing for a Watson-Crick pair at 26–44.

Second, the virtually universally conserved U8–A14 pair (Randau *et al.*, 2009) could be replaced by A8–G14, resulting in substantial function (Supplemental Table S4.4; Supplemental Figure S4.10A–C), whereas none of eight other substitutions of this pair resulted in a tRNA that was functional (Supplemental Table S4.3). Since U8–A14 forms a critical reverse Hoogsteen pair to help position the D stem, it seems

plausible that A8–G14 is functional in part because it maintains this geometry (Sterner *et al.*, 1995), perhaps with N1 of A8 protonated (Supplemental Figure S4.10D;(Leontis *et al.*, 2002)). However, it is not clear why only the A8–G14 variant had function, since five of the other eight 8–14 pairs that we scored are predicted to accommodate this geometry, albeit with slightly differing spacing (Leontis *et al.*, 2002).

Third, tertiary interactions involving the D loop appear to shift to adjacent residues. For example, although G18 and G19 in the D loop are virtually universally conserved and interact with U55 and C56, respectively, the inactive G19U variant was completely rescued by U17G but not by U16G, U20aG, or other mutations (Figure 4.4A; Supplemental Table S4.4; Supplemental Figure S4.11A–C), and the G18A variant was rescued only by U20G (Supplemental Table S4.4). A plausible explanation for these data is that the positions of the guanosines can be altered while retaining critical tertiary interactions with Ψ55 and C56, presumably by physical displacement of D-loop residues. This mechanism is consistent with the known variability of D-loop size (Giegé *et al.*, 2012), but since the crystal structure of $tRNA^{Tyr}$ is not known, epistasis at these residues may be due to another mechanism. We note that $tRNA^{Asp}$ lacks the G19–C56 interaction (Westhof *et al.*, 1985).

Fourth, flexibility in the anti-codon stem and V loop may accommodate a bulged base. Although it was puzzling that a destabilizing A28C variant (opposite U42) was substantially rescued by a destabilizing C27U mutation (opposite G43) (Figure 4.4B; Supplemental Table S4.4; Supplemental Figure S4.11D,E), a plausible explanation is that U42 bulges out of the anti-codon stem helix, allowing U27 to form a Watson-Crick pair with A44 of the V loop and allowing C28 to pair with G43 while reducing the V-loop size by 1 nt (Figure 4.4C). Consistent with this interpretation, activity was retained if the putative bulged U42 was deleted from the C27U A28C variant (Figure 4.4B; Supplemental Table S4.4).

4.3.4 *The rapid tRNA decay (RTD) pathway monitors the integrity of the entire tRNA molecule*

Our high-throughput screening approach to quantify tRNA function allows us to vary the parameters of the assay to define how mutations affect many distinct aspects of tRNA biology. One critical process modulating tRNA turnover is the RTD pathway, which targets specific mature tRNAs for degradation due to lack of one or more body modifications or to a destabilized acceptor or T stem, resulting in attack by the 5′–3′ exonucleases Rat1 and Xrn1 (Figure 4.5A; (Alexandrov *et al.*, 2006; Chernyakov *et al.*, 2008; Whipple *et al.*, 2011; Dewe *et al.*, 2012). However, the full scope of sequence variants subject to RTD is not clear, as only the $tRNA^{Ser}$ family has been examined in any detail; the roles of the anti-codon stem–loop, the D stem–loop, and the T loop have been only minimally examined; $tRNA^{Ser}$ family members are in the minority class II of tRNAs that have a long variable stem; and acceptor stem/T-stem stability

estimates do not always accurately predict RTD susceptibility for other tRNA species (Whipple *et al.*, 2011).

We applied this library-based approach to comprehensively define $SUP4_{oc}$ variants that are substrates for RTD. RTD is readily detected with the RNA-ID reporter, since the known substrate $SUP4_{oc}$-G62C (Whipple *et al.*, 2011) had reduced GFP^{FLOW} in $MET22^{+}$ (wild-type) cells compared with that in *met22*Δ cells (Figure 4.5B, Supplemental Table S4.5), in which RTD is inactivated (Chernyakov *et al.*, 2008). We made a S$UP4_{oc}$ library in the *met22Δ* strain, analyzed variants by FACS and sequencing (Supplemental Figure S4.12A), and compared GFP^{SEQ} of variants with that from wild-type cells. GFP^{SEQ} from the *met22*Δ strain was highly reproducible and correlated with GFP^{FLOW} (Supplemental Figure S4.12B,C).

This analysis revealed many single variants that were putative RTD substrates, with mutations surprisingly located throughout the tRNA body. In the *met22Δ* strain, 70 single variants were highly functional, including all 44 that we identified in the wild type (Supplemental Figure S4.12D). Overall, 38 single variants were more than twofold more active in *met22Δ* cells than in wild-type cells (GFP^{SEQ} RTD ratio >2), suggesting that they are RTD substrates (Figure 4.5C, green; Supplemental Table S4.6), and for 16 of these, the increase in activity was >0.3 (Figure 4.5C, dark wedge outlines). Eleven of these 38 RTD candidates have mutations in the acceptor or T stem, as expected for RTD substrates (Whipple *et al.*, 2011). Remarkably, the other 27 RTD candidates have mutations in regions not previously associated with RTD, including 17 in the anti-codon stem and loop, six in the D stem, and one each in the D loop, V loop, and T loop and at N8 (Figure 4.5C).

We determined that a number of these variants are RTD substrates by two approaches. First, we reconstructed individual RTD candidate variants with mutations in different regions of the tRNA and tested them by flow cytometry when integrated into *met22Δ* and wild-type reporter strains. Nineteen of 21 putative RTD substrates with GFP^{SEQ} RTD ratios ranging from 24.4 to 2.3 had GFP^{FLOW} RTD ratios >2.0 (Supplemental Figure S4.13A; Supplemental Table S4.5), whereas 26 of 30 putative non-RTD substrates with GFP^{SEQ} RTD ratios ranging from 1.4 to 0.9 had GFP^{FLOW} RTD ratios <2.0. We therefore conclude that RTD ratios determined by GFP^{SEQ} scores have high predictive value for potential RTD substrates as measured by GFP^{FLOW}. Second, a primer extension assay with ddCTP instead of dCTP (which results in a G34 stop for $tRNA^{Tyr}$ and a G30 stop for $SUP4_{oc}$) showed that tRNA levels of RTD candidates were increased in the *met22Δ* strain relative to the wild-type strain (Figure 4.5D). As expected, $SUP4_{oc}$ levels did not change in *met22Δ* compared with wild-type cells; however, tRNA levels were substantially increased in the *met22Δ* mutant for the acceptor stem U2C variant, the D-stem C25U variant, the T-stem G62C variant, and the anti-codon stem C27A, A29U, and A31U variants, providing strong evidence that these are

all RTD substrates. Based on these data, we estimate that the vast majority of the 38 single-mutant and 605 double-mutant variants that are candidate RTD substrates are authentic (Supplemental Table S4.6), suggesting that RTD places a significant constraint on tRNA sequences.

Previous analysis of determinants for RTD in the $tRNA^{Ser}$ family demonstrated that the predicted folding stability of the combined acceptor and T stem correlated inversely with susceptibility to RTD (Whipple *et al.*, 2011). However, since mutations throughout the tRNA elicited RTD, we examined the relationship of RTD to the predicted stability of the entire molecule, as quantified by $\Delta\Delta G°_{28}$ (Reuter and Mathews, 2010). Consistent with the importance of stability in RTD, a threshold of $\Delta\Delta G°_{28}$ of 2.65 kcal/mol has good predictive value, since 28 of 38 qualified single-mutant variants with $\Delta\Delta G°_{28}$ >2.65 kcal/mol are RTD substrates (Supplemental Table S4.6). Since these 28 variants occur in all of the stems of *SUP4*$_{oc}$, we conclude that the influence of stability on RTD extends to the entire molecule. In contrast, only 10 of 65 variants with $\Delta\Delta G°_{28}$ <2.65 kcal/mol were RTD substrates. Since a number of these 10 variants have mutations in loop residues that participate in tertiary interactions, we presume that stability is affected here, too, but is not captured by calculated $\Delta\Delta G°_{28}$, which only measures secondary structure contributions. Overall, a $\Delta\Delta G°_{28}$ cutoff of 2.65 kcal/mol results in a true positive rate of 0.74 and a false positive rate of 0.15 (Figure 4.5E). We also found that $\Delta\Delta G°_{28}$ is predictive of RTD for double-mutant variants (Supplemental Figure S4.13B).

The numerous examples of positive epistasis involving the stabilizing U4C acceptor stem mutation (opposite G69) may be due to protection from RTD. A large number of variants that were rescued by the U4C mutation are themselves RTD substrates, including several variants with mutations in the anti-codon stem (Figure 4.6A,B; Supplemental Figure S4.14A,B). Moreover, for each of three variants examined, the U4C double variants had similar tRNA levels in wild-type and *met22*Δ cells (Figure 4.6C; Supplemental Figure S4.14C). This result suggests that U4C protects the 5′ end of variants subject to RTD from exonucleolytic attack, presumably by stabilizing the 4–69 base pair.

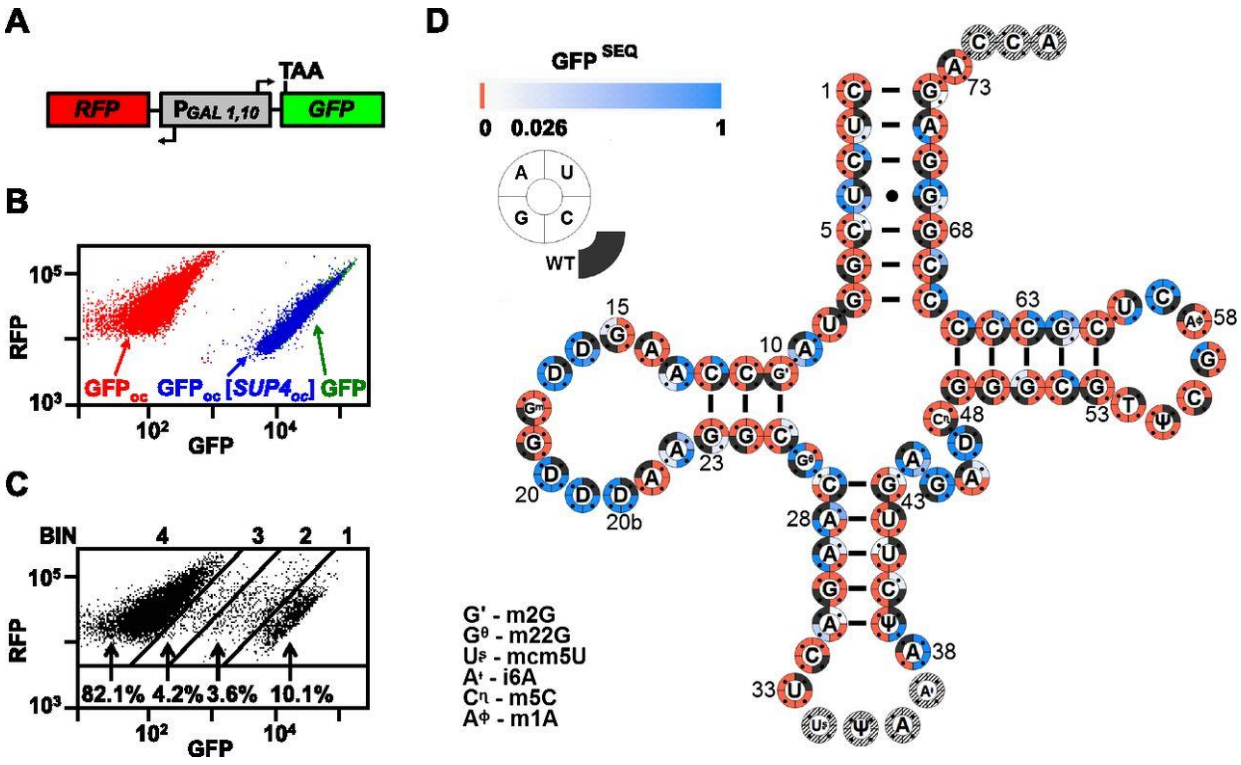


Figure 4.1 High-throughput quantification of tRNA function of *SUP4oc* variants.

(*A*) Schematic of the RNA-ID reporter used to quantify tRNA function. (*B*) $SUP4_{oc}$ efficiently suppresses GFP_{oc}. Scatter plot of flow cytometry of cells with integrated RNA-ID reporter expressing GFP (green), GFP_{oc} (red), and GFP_{oc} and $SUP4_{oc}$ (blue). (*C*) FACS of $SUP4_{oc}$ variant library. Cells were grown in YP galactose medium and sorted. (*D*) *SUP4oc* tolerates numerous mutations. Cloverleaf heat map showing GFP^{SEQ} of single-mutant variants. Quadrant color around residues indicates variant activity. Active variants are white (GFP^{SEQ} of 0.026) to blue (GFP^{SEQ} of 1) gradient, and inactive variants are red. Modified bases are indicated in the figure.

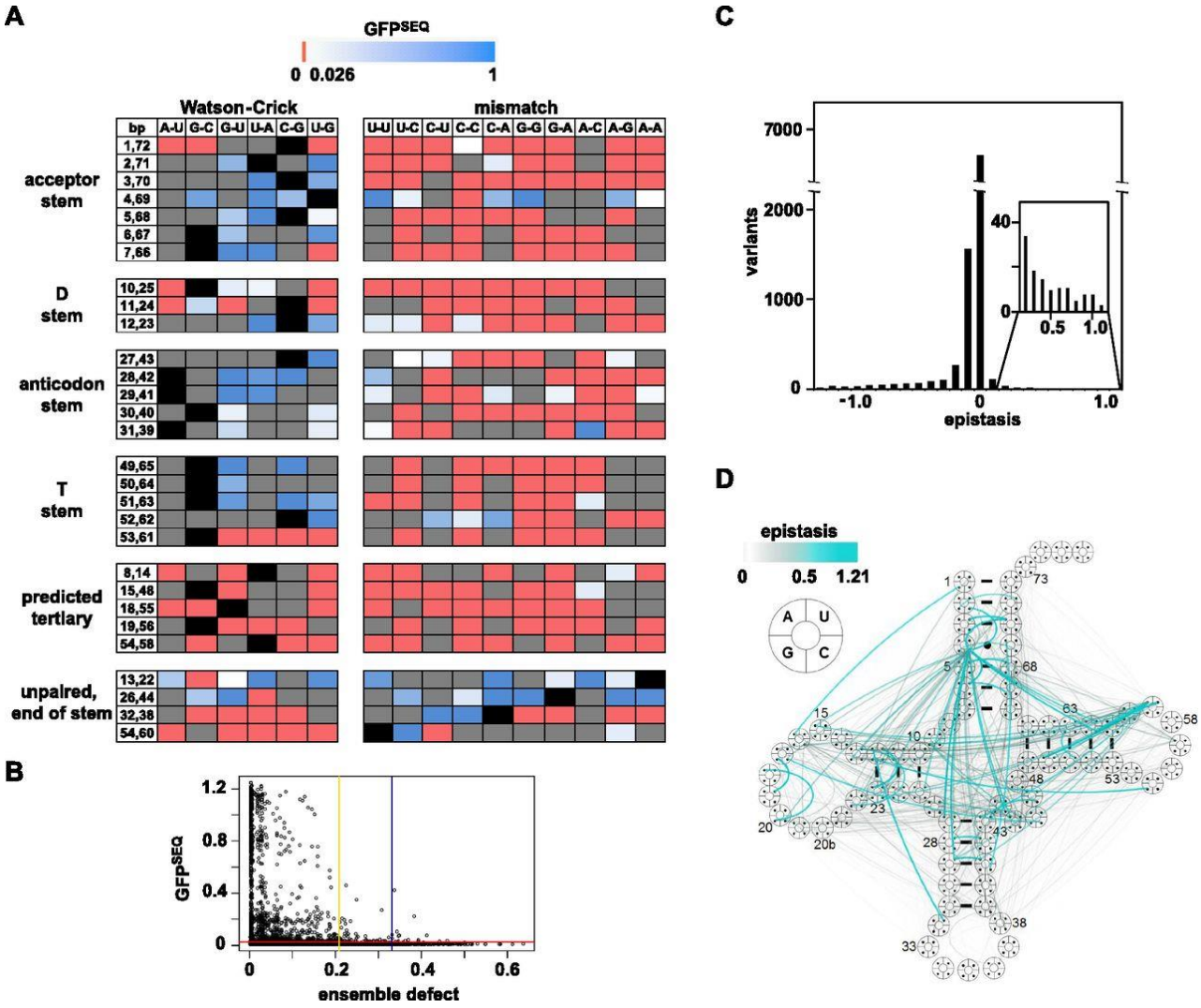


Figure 4.2 Analysis of single- and double-mutant *SUP4*$_{oc}$ variants.

(*A*) GFP^{SEQ} of stem base pair and tertiary pair variants. Color-coding as in Figure 4.1D. (Gray boxes) Variant not scored. (*B*) Plot of GFP^{SEQ} versus ensemble defect (ED) for all single and double variants. (Red) Undetectable GFP^{SEQ} activity cutoff; (yellow) 95% ED cutoff; (blue) 99% ED cutoff. (*C*) Epistasis of double-mutant variants. (*D*) Cloverleaf schematic map of positive epistatic interactions between residues in *SUP4*$_{oc}$. Color and width of lines correspond to the strength of the interactions.

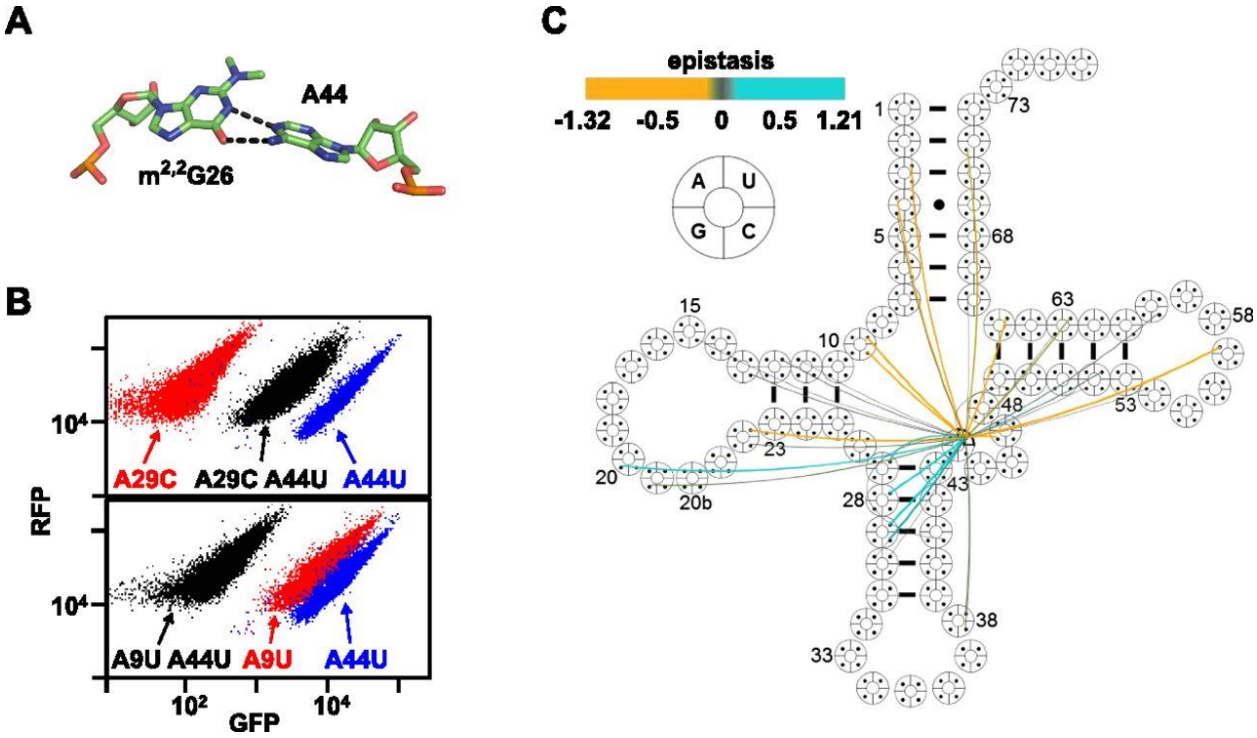


Figure 4.3 Evidence for an alternative conformation in *SUP4*$_{oc}$ 26–44 variants.

(*A*) G26–A44 structure in tRNA^{Phe}. Data from Protein Data Bank (PDB) ID 1EHZ. (*B*) An A44U mutation confers both positive and negative epistasis on variants. Flow cytometry of cells expressing A44U and/or A29C (*top*) and/or A9U (*bottom*). (*C*) Cloverleaf map of epistatic interactions involving A44U. (Cyan) Positive epistasis; (amber) negative epistasis.

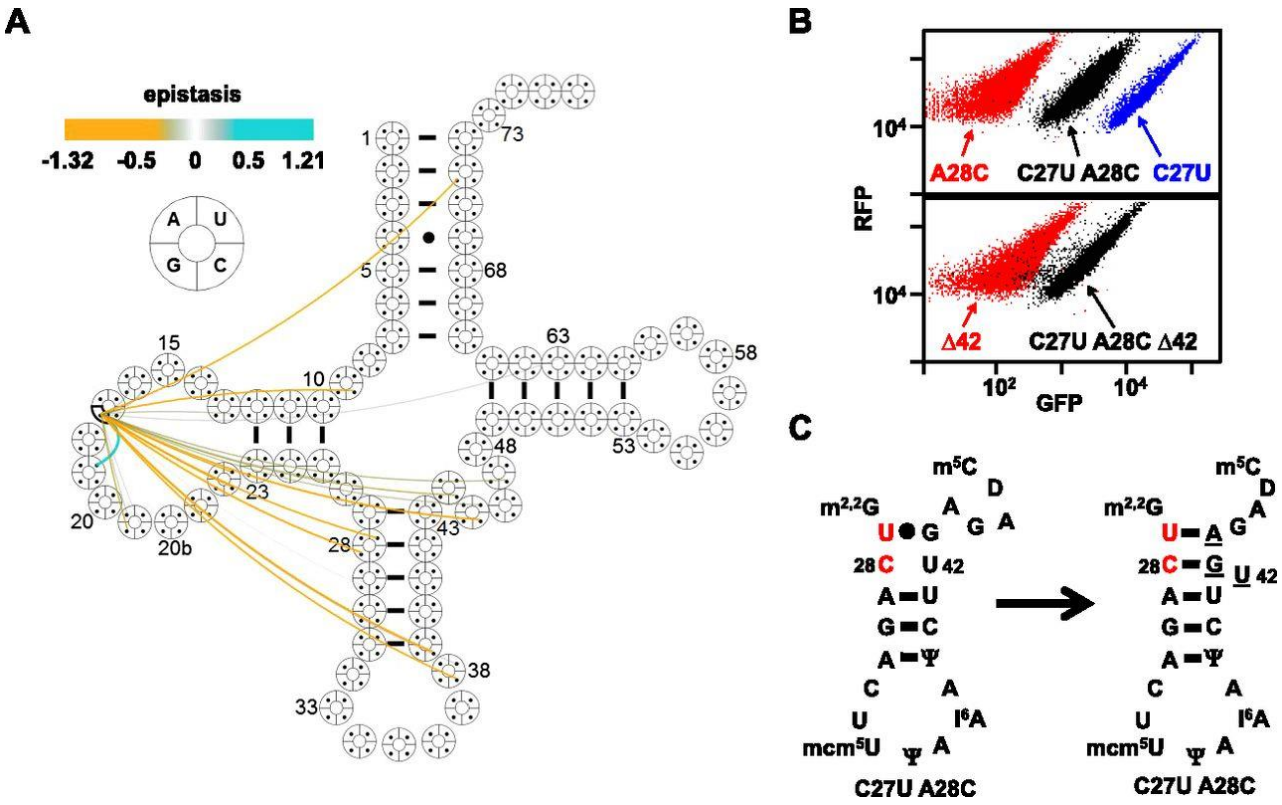


Figure 4.4 Positive epistasis due to shift of interactions to neighboring residues in *SUP4*$_{oc}$ variants.

(*A*) Cloverleaf map of epistatic interactions involving U17G. (Cyan) Positive epistasis; (amber) negative epistasis. (*B*) A C27U mutation restores activity to an A28C variant. Flow cytometry of cells expressing C27U and/or A28C (*top*) and Δ42 derivatives of C27U A28C and *SUP4*$_{oc}$ (*bottom*). (*C*) Predicted base pair rearrangement of the C27A U28A variant. (Red) Mutations; (underlined) proposed rearranged bases.

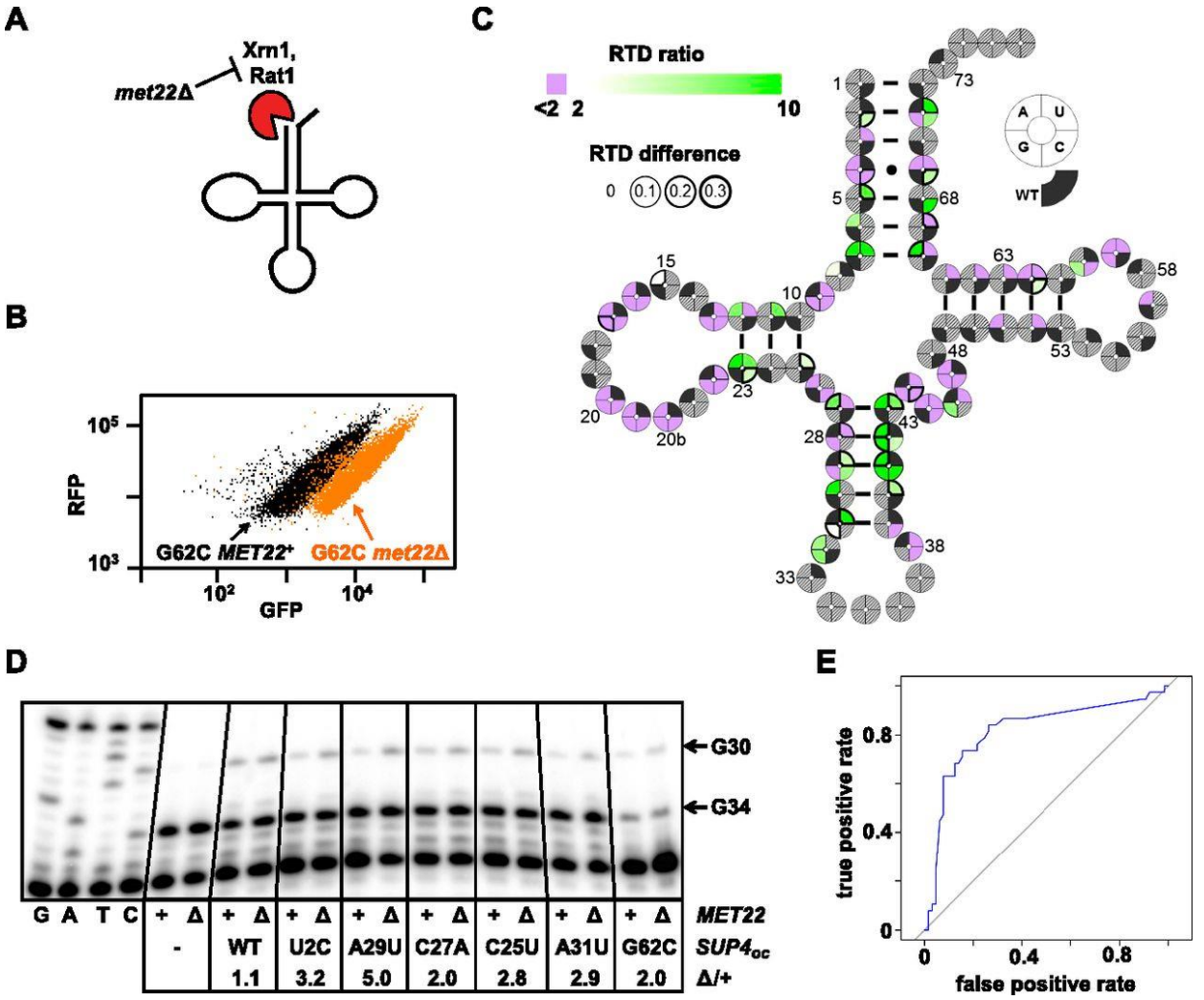


Figure 4.5 Analysis of *SUP4oc* RTD substrates.

(*A*) Schematic of RTD. (*B*) The RTD substrate *SUP4-3*$_{oc}$ (*SUP4-G62C*) has increased GFP^{FLOW} in the *met22*Δ strain. (*C*) Mutations throughout the tRNA appear to trigger RTD. Cloverleaf heat map of *SUP4*$_{oc}$ single variants analyzed for RTD based on GFP^{SEQ} RTD ratios [GFP^{SEQ}(*met22*Δ)/GFP^{SEQ}(wild type)]. (Green shades) RTD substrate; (purple) nonsubstrate; (wedge border thickness) GFP^{SEQ}(*met22*Δ) − GFP^{SEQ}(wild type). Gray hatches indicate variants not scored by sequencing or those with a GFP^{SEQ} < 0.052 in *met22Δ* cells (the minimum score in *met22Δ* cells to observe an RTD ratio >2.0). (*D*) Analysis of *SUP4*$_{oc}$ levels of putative RTD variants in *met22*Δ and wild-type strains. Bulk RNA from the indicated strains was analyzed using poison primer extension with ddCTP. (*E*) ROC (receiver operating characteristic) curves for RTD prediction based on estimated $\Delta\Delta G^{\circ}_{28}$ for single variants for which the RTD ratio could be scored (see the Materials and Methods).

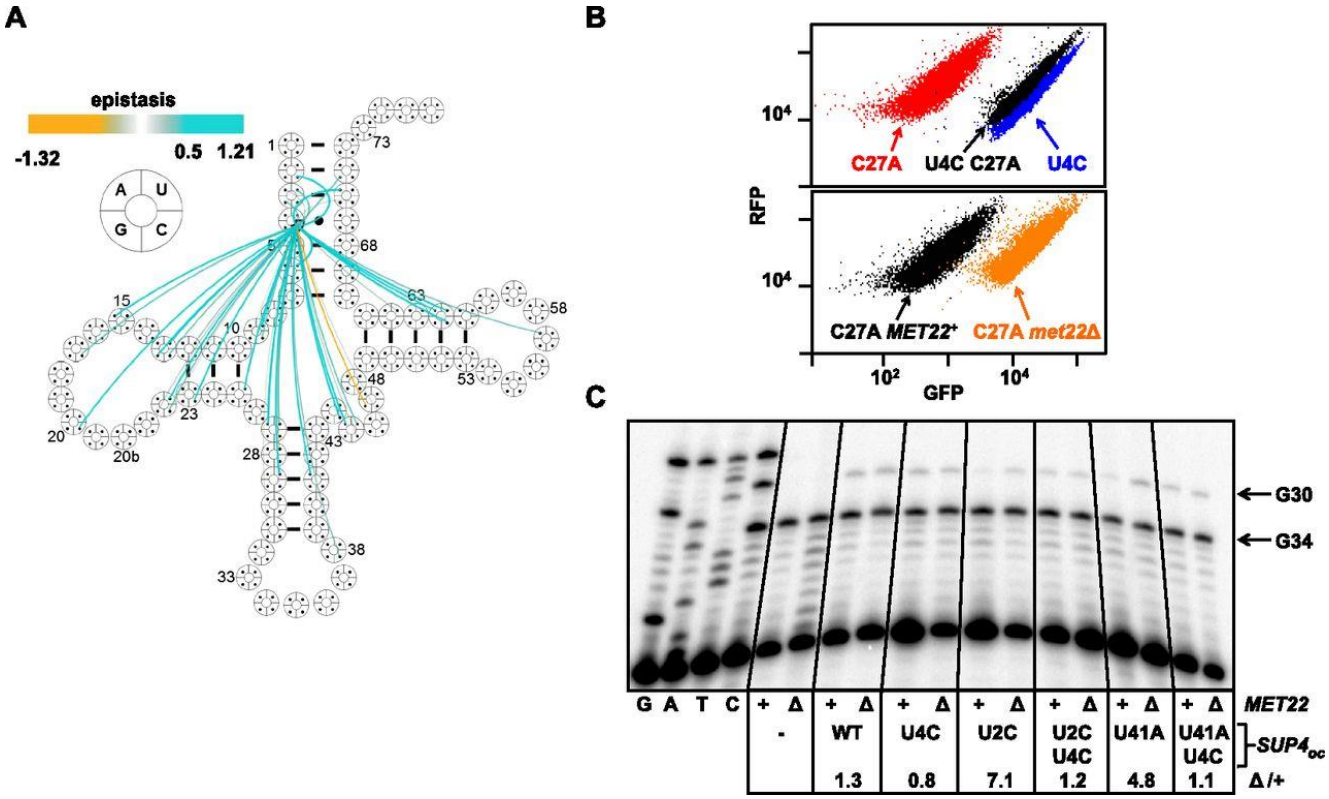


Figure 4.6 A U4C stabilizing mutation rescues variants that are RTD substrates.

(*A*) Cloverleaf map of epistatic interactions involving U4C. (*B*, *top*) Flow cytometry of cells expressing the indicated *SUP4*$_{oc}$ variants. (*Bottom*) Flow cytometry of cells expressing the indicated *SUP4*$_{oc}$ variant in wild-type and *met22*Δ cells. (*C*) A U4C mutation restores *SUP4*$_{oc}$ levels of RTD substrate variants in wild-type cells to those in the *met22*Δ strain. Levels of *SUP4*$_{oc}$ were determined as described in Figure 4.5D.

4.4 DISCUSSION

Although tRNAs have evolved for their efficient transcription, processing, and modification, high structural stability, and accurate and efficient usage in the translation cycle, the analysis of >25,000 variants of the model yeast tRNA $SUP4_{oc}$ demonstrates that it is highly robust to mutation. This robustness was unexpected based on the numerous constraints on tRNA sequences but was consistent with models of RNA evolution in which sequences converge to those that are robust to mutation (Nimwegen *et al.*, 1999). Nonetheless, although many single-base changes are tolerated in $SUP4_{oc}$, a second mutation is much more likely to abolish function than to rescue it (Supplemental Figure S4.8C). Since this excess of negative epistasis was present also in *met22*Δ cells (Supplemental Table S4.3), RTD is not the primary cause of negative epistasis, although it is likely one contributing factor. Part of the reason that $SUP4_{oc}$ has a preponderance of negative epistasis may be that the multiple constraints on structure and function are too great to accommodate most double mutations.

The analysis of positive epistasis suggests a remarkable amount of flexibility allowed in the sequence of $SUP4_{oc}$. The suggestion that an alternative tRNA conformation is provoked by mutation of the 26–44 pair in the hinge region (Figure 4.3; Supplemental Figure S4.9A) may be compatible with conformational changes during translation (Valle *et al.*, 2003; Schmeing *et al.*, 2009; Zhou *et al.*, 2013), as previously proposed for a D-stem variant (Cochella and Green, 2005). In the *met22Δ* mutant, there is even more pronounced evidence for this alternative conformation based on additional examples of negative and positive epistasis for mutations affecting the 26–44 pair (Supplemental Table S4.3) and more extreme epistasis values. These data suggest that the large fraction (35%) of 1984 surveyed eukaryotic tRNAs with canonical 26–44 pairing (Marck and Grosjean, 2002) may have common compensatory features that distinguish them from tRNAs with unmatched residues at this position. The epistasis data derived from the *met22Δ* mutant underscore the flexibility of tRNA, since increased positive epistasis was observed for the C27U A28C variant and the U8A A14G variant. In addition, there are several examples of epistasis to preserve adjacent guanosine residues in the D loop (normally located at positions 18 and 19) (see Figure 4.4A), including one variant that was not scored in wild-type cells (Supplemental Table S4.3).

Several other positive epistatic interactions cannot be explained easily by structural alterations (Supplemental Table S4.3). Since our approach measures the overall function of the tRNA, which includes all steps from its transcription by RNA polymerase III through its role in translation, such positive epistasis could arise because of altered function in the double mutant due to any combination of steps during the biogenesis of tRNA or its deployment in translation.

Our results suggest that RTD monitors the integrity of the entire tRNA molecule, greatly expanding

the scope of variants subject to this pathway. Indeed, since 446 of 838 substantially functional double variants are likely subject to RTD, RTD is a major factor in determining the sequence limits to tRNA function (Supplemental Table S4.6). Since our analysis suggests that increased $\Delta\Delta G°_{28}$ correlates with susceptibility to RTD regardless of the stem that is affected (Figure 4.5C; Supplemental Table S4.5, Supplemental Table S4.6; Supplemental Figure S4.13), this brings up the question of how overall tRNA stability influences RTD. Our suggestion that degradation occurs through the 5′ end is supported by widespread U4C suppression of RTD (Figure 4.6; Supplemental Fig. S8). Mutations to $SUP4_{oc}$ in the D stem, D loop, and T loop may trigger RTD by altering the tertiary fold, allowing increased 5′ end attack. However, it is difficult to rationalize how mutations in the anti-codon stem expose the 5′ end, since residues in this stem do not interact with other regions. These anti-codon stem mutations might trigger RTD due to cooperative unfolding of the acceptor stem, altered stacking with the D stem and consequent destabilization, or another sensing mechanism, perhaps an element of the translation machinery.

Our results provide a framework for understanding how sequence variation influences many aspects of tRNA biology, including the role and function of tRNA isodecoders in metazoans and the molecular basis of diseases caused by mitochondrial tRNA mutations (Yarham *et al.*, 2010). An analysis of ED shows that tRNA structure prediction software may be useful for giving an upper bound on tolerated defects, but other parameters need to be incorporated for these programs to predict function successfully. Application of the high-throughput approach described here to define functional determinants of other tRNA species should lead to large improvements in our ability to predict function of variants.

The approach described here is generally applicable to many problems in tRNA biology. FACS followed by deep sequencing of tRNA genes can be used to score the effect of mutations that affect tRNA processing, modification, or translation by comparing the scores with those in a wild-type background. A prerequisite for this approach is that the tRNA can be made into a suppressor or that cells carrying the tRNA can be scored for growth or another activity. By the use of appropriate screens, this approach can also measure tRNA charging fidelity (Kramer *et al.*, 2010). Furthermore, this overall approach can be adapted to study many problems in the biology of noncoding RNAs.

4.5 MATERIALS AND METHODS

4.5.1 *Yeast strains*

The BY4741 *can1::*P_{GAL1}*-GFP*$_{oc}$*-*P_{GAL10}*-RFP* strain (YK380-1) was constructed by PCR amplification of the P_{GAL1}*-GFP*$_{oc}$*-*P_{GAL10}*-RFP* reporter and its adjacent *MET15* marker from plasmid EKD1302 (Dean and Grayhack, 2012), using primers with sequence complementary to the 5′ and 3′ ends

of *CAN1*, followed by linear transformation of the DNA into BY4741. A *met22*Δ derivative of the YK380-1 (YK391-1) was generated by PCR amplification of the *met22::kanMX* strain (Open Biosystems) followed by linear transformation. $SUP4_{oc}$ and $SUP4_{oc}$ variant derivatives of strains YK380-1 and YK391-1 were generated by linear transformation to integrate the StuI fragment of the plasmid containing $SUP4_{oc}$ (derived from AB230-1) into the *ADE2* locus, followed by selection on S-His dropout medium. Since the StuI fragment has different sequences of *ADE2* DNA at each end flanking the DNA containing $SUP4_{oc}$ and *Schizosaccharomyces pombe* $his5^+$, linear transformation should not generate multiple integrants at this locus. For each variant strain analyzed, three individual transformants were constructed and used.

4.5.2 *Plasmids*

AB230-1was constructed by replacement of the *MET15* marker of JW132 (Whipple *et al.*, 2011) with a fragment of DNA from pUG27 expressing *S. pombe* $his5^+$ (which complements *S. cerevisiae his3*) (Gueldener *et al.*, 2002) followed by insertion of a 1508-base-pair (bp) fragment of *Fluc* DNA into the BglII and XhoI restriction sites to facilitate detection of inserts when $SUP4_{oc}$ variants were inserted into these sites.

Variant tRNAs were constructed by insertion of the appropriate $SUP4_{oc}$ tRNA sequence, flanked by the 22 bp 5′ of the +1 site and the 7 bp 3′ of residue 73 of mature $tRNA^{His(GUG)}$ [tH(GUG)G2], into the BglII XhoI site of AB230-1, essentially as described previously (Whipple *et al.*, 2011). The final sequence inserted was as follows: 5′-AACAAAGTTCATAAAGAAATTA<u>CTCTCGGTAGCCAAGTTGGTTTAAGGCGCAAGACT</u>**<u>TTA</u>**<u>A</u>*T TTATCACTACGAA*<u>ATCTTGAGATCGGGCGTTCGACTCGCCCCCGGGAGA</u>TTTTTTCCTCGAG-3′, with the $SUP4_{oc}$ exon sequence underlined, the anti-codon in bold, and the intron in italics.

4.5.3 *Analytical flow cytometry*

Strains were grown overnight at 28°C in S-His liquid dropout medium containing 2% raffinose and 2% galactose supplemented with 80 mg/L adenine, followed by growth for 24 h in YP medium containing 2% raffinose and 2% galactose supplemented with 80 mg/L adenine. Dilutions were made as necessary to maintain log phase growth. Cells were then diluted in the same medium to an OD_{600} of 0.3 and grown to an OD_{600} between 0.8 and 1.2. Samples were prepared and analyzed essentially as described previously. Briefly, 10,000 events were recorded after analysis on an LSR-II flow cytometer (BD Biosciences) using laser and fluorescence detection filter parameters as described previously, with filter voltages set so that both GFP and RFP fluorescence intensities were ~26,000 and with only those cells that passed an RFP cutoff of 5×10^3 analyzed (Dean and Grayhack, 2012). Data analysis was performed using FlowJo software

(Tree Star). GFP^{FLOW} scores for a culture of a given *SUP4*$_{oc}$ variant represent the ratio of the median GFP divided by the median RFP, normalized to the median GFP/median RFP for wild-type *SUP4*$_{oc}$. Biological triplicates were used to obtain standard deviations.

4.5.4 *SUP4*$_{oc}$ *library construction and analysis*

The *SUP4*$_{oc}$ library was generated by annealing two partially complementary oligonucleotides (MPG P346 and MPG P347) with 3% random mutations in *SUP4*$_{oc}$ residues 1–33 and 39–73 (IDT) followed by filling in the unpaired overhangs (Supplemental Figure S4.7) and cloning. The sequence of P346 was 5′-TTTTGAGATCTAACAAAGTTCATAAAGAAATTACTCTCGGTAGCCAAGTTGGTTTAAGGCGCAAGACT**TTAATTTATCACTACGAA**-3′, and that of P347 was 5′-AGTTGCTCGAGGAAAAAATCTCCCGGGGGCGAGTCGAACGCCCGATCTCAAGAT**TTCGTAGTGATAAATTAA**-3′, with the residues containing mutations underlined and the complementary sequence (comprising residues 34–38 and the intron) in bold. Annealing of the two oligonucleotides was done by heating for 5 min to 100°C followed by slow cooling to 30°C and then immediate placement on ice. The unpaired overhangs were then filled using the Klenow fragment of DNA polymerase at 37°C, and the reaction product was digested with BglII and XhoI, purified by gel extraction, and ligated into AB230-1, giving ~325,000 *E. coli* transformants. An aliquot of these transformants containing $\sim 2.7 \times 10^{10}$ cells was amplified by ~4.3 generations of growth, and then plasmid DNA was extracted and digested with StuI for integration into yeast as described above. The yeast transformants were scraped, pooled, and frozen in aliquots for subsequent use.

To grow yeast *SUP4*$_{oc}$ libraries, ~4.9 million cells were thawed, inoculated into S-His medium containing 2% raffinose supplemented with 80 mg/L adenine, and grown for 24 h at 28°C followed by dilution to OD_{600} of 0.04 and growth for 24 h in YP medium containing 2% raffinose and 2% galactose supplemented with 80 mg/L adenine. Dilutions were made as necessary to maintain log phase and ensure that at least 4.9 million cells were propagated at each step. Cells were then diluted in the same medium to an OD_{600} of 0.4 and grown to an OD_{600} of 1.1 prior to FACS into four bins on an Aria-ll cell sorter (BD Biosciences) at the University of Rochester Medical Center Flow Cytometry Core facility. Laser and fluorescence detection filter parameters were set as previously described, and only those cells with a RFP $> 5 \times 10^3$ were collected (Dean and Grayhack, 2012). Bin borders were set at a GFP^{FLOW} of 0.007 (the lowest activity readily distinguished for a strain containing *SUP4*$_{oc}$ variants as compared with strains with no *SUP4*$_{oc}$, corresponding to GFP^{SEQ} of 0.026), with successive borders at 0.038 and 0.384. At least 2 million cells were collected (Supplemental Table S4.1) and then plated on YPD medium. After incubation

for 3 d at 25°C, cells were scraped, pooled, and stored at −80°C. Genomic DNA was then directly isolated from frozen aliquots of the stored cells in each bin. Libraries WT1 and WT2 are replicates of the *SUP4*$_{oc}$ library analyzed in wild-type cells, and libraries Δ1 and Δ2 are replicates of the *SUP4*$_{oc}$ library analyzed in *met22*Δ cells. The GFP^{SEQ} RTD analysis was based on comparing the WT2 and Δ2 libraries.

To enhance resolution of highly functional variants, aliquots of stored bin 1 cells (cells with high GFP expression) collected from the first sorting of the WT2 library were thawed, grown, and further sorted by FACS into four bins, three of which were subdivisions of the original bin1, and one of which was the original bin 2. Pooled cells were treated as described above prior to sequence analysis. Data from this analysis are referred to as WT2 6 bins and are the data set used for single, double, and epistasis analysis of *SUP4*$_{oc}$ in wild-type cells.

4.5.5 *Sequencing*

The *SUP4*$_{oc}$ construct, including 27 5′ and 16 3′ nucleotides, was amplified for 20 cycles (10 sec at 98°C, 30 sec at 52°C, and 30 sec at 72°C) from 1–3 μg of genomic DNA using Phusion polymerase and one of four sets of primers. Y19 (5′-AATGATACGGCGACCACCGAGATCTACACCTCCGCCTAACCCGAGTCCACCCGTCCCNNNNGATCTAACAAAGTTCATAAAGAAATTA-3′), used in all four primer sets, contains the Illumina adaptor sequence (1–29), a sequencing primer (30–56), and four Ns to mitigate low-complexity library cluster registration problems on the MiSeq. Y22 (5′-CAAGCAGAAGACGGCATACGAGATATTCCTTTCTTCCCTGCCCACCACCAGCTCCGTTGCTCGAGGAAAAAA-3′), Y23 (5′-CAAGCAGAAGACGGCATACGAGATCGGGTAAACTTCCCTGCCCACCACCAGCTCCGTTGCTCGAGGAAAAAA-3′), Y24 (5′-CAAGCAGAAGACGGCATACGAGATGGATATAGCTTCCCTGCCCACCACCAGCTCCGTTGCTCGAGGAAAAAA-3′), and Y25 (5′-CAAGCAGAAGACGGCATACGAGATTCCAGCCCCTTCCCTGCCCACCACCAGCTCCGTTGCTCGAGGAAAAAA-3′) each contain the Illumina adaptor, a sequencing primer, and a different index to discriminate between the bins (ATTCCTTT, CGGGTAAA, GGATATAG, and TCCAGCCC). Amplicons were gel-purified, quantitated with the qbit and Kapa quantitative PCR (qPCR) quantification kits, and sequenced on either the MiSeq V2 (using the 2x150 kit) or the HiSeq 2500 (using the 2x101 kit) using the manufacturer's instructions.

4.5.6 *Sequence assembly and quality filtering*

Sequences were trimmed and split into bins using a custom python script. Forward and reverse reads were combined and filtered for quality using Enrich version 0.2 with a minimum *phred* score of 30 for any given cycle (Fowler *et al.*, 2011).

4.5.7 *Calculation of GFP^{SEQ}*

The number of reads corresponding to each unique variant for each bin was tabulated using Enrich version 0.2 (Fowler *et al.*, 2011). The reads were then normalized to the sequencing depth of each bin, and the frequency of a given variant in a given bin was converted to an estimated number of collected cells by multiplying by the total number of cells collected for that bin during the FACS analysis. To calculate an approximate cellular fluorescence score for a given variant, the fraction of that variant's cells in each bin was multiplied by the median GFP/median RFP value for each bin, and the results were summed, resulting in a weighted average fluorescence for that variant. This score was then normalized to the wild-type score (labeled NA-NA in the ID column of Supplemental Table S4.2) to give the normalized weighted average fluorescence referred to as GFP^{SEQ}.

4.5.8 *Data quality control*

To focus analysis on validated sequences for which there were sufficient data, two sets of filters were applied: read counts per variant from sequencing and estimated cell counts per variant. A comparison of GFP^{SEQ} values between WT1 and WT2 was taken as the gauge of quality in order to select the thresholds for filtering. Fifty-six combinations of both filtering parameters derived from seven different values from zero to 60 cell counts and eight different values from zero to 200 sequencing reads were applied to the data set, and a manual selection was made to reduce spurious correlations while not reducing the number of sequences needlessly. Scatter plots of comparisons of GFP^{SEQ} values between the WT1 and WT2 libraries are shown in Supplemental Figure S4.7B, for nine representative combinations. The thresholds chosen were total read counts ≥100 and total cell counts ≥30.

4.5.9 *Prediction of ED*

The EDcalculator from RNAstructure (Reuter and Mathews, 2010) was applied to calculate each mutant's ED (Zadeh *et al.*, 2011b) using the secondary structure of $SUP4_{oc}$ as reference. Modified bases dihydrouridine (D), N^2,N^2-dimethylguanosine ($m^{2,2}G$), and 1-methyladenosine (m^1A) were forced to be unpaired in the ED prediction, since these bases are known to block canonical base-pairing, and the

assumption was also made that only mutation of the modified base would remove the modification. For the assessment of the ED of natural tRNAs, we used all eukaryotic tRNA sequences for which the modification status is known from the following species: *Bombyx mori*, *Bos taurus*, *Candida cylindracea*, *Drosophila melanogaster*, *Homo sapiens*, *Leishmania tarentolae*, *Lupinus albus*, *Lupinus luteus*, *Mus musculus*, *Nicotiana rustica*, *Nicotiana tabacum*, *Oryctolagus cuniculus*, *Pichia jadinii*, *Rattus norvegicus*, *S. cerevisiae*, *S. pombe*, *Tetrahymena thermophila*, *Triticum aestivum*, and *Xenopus laevis*. Sequences were obtained from the Modomics database (http://modomics.genesilico.pl).

4.5.10 *Calculation of $\Delta\Delta G°_{28}$*

The $\Delta\Delta G°_{28}$ is the folding free energy change difference between the mutant and wild-type tRNA structure at 28°C. The predicted $\Delta\Delta G°_{28}$ for each tRNA variant was computed with a custom program using the C^{2+} classes from the *RNAstructure* package (Reuter and Mathews, 2010). $\Delta\Delta G°$s were calculated using nearest neighbor rules (Mathews *et al.*, 2004), where base pairs were disrupted if a nucleotide mutation prevented canonical pairing.

4.5.11 *Epistasis analysis*

The formula to determine epistasis for double mutants that passed the read count filter was

$$Epistasis = GFP^{SEQ}_{double\ mutant} - \left(GFP^{SEQ}_{single\ mutant1} * GFP^{SEQ}_{single\ mutant2}\right)$$

The product of GFP^{SEQ} of the two single mutants was correlated with the actual double-mutant GFP^{SEQ} with an R^2 of 0.53 for Δ1, 0.56 for Δ2, 0.545 for WT1, and 0.534 for WT2. Epistasis can be >1 or less than −1, since some double-mutant variants have GFP^{SEQ} values >1, because all sequencing reads for that variant were in bin 1, whereas some wild-type $SUP4_{oc}$ sequence reads occurred outside of bin 1.

Parameters that can be manipulated on the interactive $SUP4_{oc}$ Web site (http://depts.washington.edu/sfields/tRNA_supplemental/tRNA_interactive.html) are as follows: predicted fitness cutoff, only display epistatic interactions (links) in which the product of the GFP^{SEQ} for the constituent single mutant variants is greater than this number; cell count cutoff, only display interactions for double mutants with more total estimated cell counts than this number; read count cutoff, only display interactions for double mutants with more total sequencing reads than this number; epistasis color control, controls the point at which the negative and positive epistasis values switch from gray to color for links; opacity cutoff, the epistasis value below which link opacity is set to "minimum opacity"; and minimum opacity, a value between 0 (transparent) and 1 (opaque) that defines the transparency of epistatic links with values below the "opacity cutoff." The default is 0.3.

4.5.12 *ROC (receiver operating characteristic) analysis*

To assess the extent to which $\Delta\Delta G°_{28}$ correlates with RTD, we performed ROC analysis. All of the single and double variants were filtered with the criterion requiring *met22*Δ2 $GFP^{SEQ} \geq 0.052$ (twice the cutoff for a variant to be considered active). With that, all of the variants with a ratio of *met22*Δ2 GFP^{SEQ} to WT2 $GFP^{SEQ} \geq 2.0$ were classified as RTD substrates. $\Delta\Delta G°_{28}$ was taken as a predictor to compute the false positive rate and the true positive rate for ROC analysis. The plot was generated, and thresholds were computed by the R package pROC (http://www.R-project.org) (Robin *et al.*, 2011).

4.5.13 *Isolation of bulk RNA and tRNA purification*

Wild-type and *met22*Δ cells containing integrated $SUP4_{oc}$ variants were grown as described for analytical flow cytometry. Bulk low-molecular-weight RNA was extracted from 300 OD-mL pellets by hot phenol extraction followed by two ethanol precipitations and resuspension in ddH_2O, as previously described (Jackman *et al.*, 2003). $tRNA^{Tyr}$ was purified using biotinylated oligomer MP129, which is complementary to residues 76–52 of endogenous $tRNA^{Tyr}$, *SUP4oc*, and the variants analyzed (Jackman *et al.*, 2003).

4.5.14 *Primer extension of SUP4oc variants*

Bulk low-molecular-weight RNA was subjected to a poison primer extension assay using a primer from nucleotides 57–37 or 62–43 of mature $tRNA^{Tyr}$ that was 5′ end-labeled with T4 polynucleotide kinase and [γ-^{32}P]ATP. Two-hundred nanograms of bulk low-molecular-weight RNA (or 7 ng of purified tRNA) was annealed to ~1 pmol of 5′ radiolabeled primer after incubation for 3 min at 95°C before slow cooling and incubation for 30 at 50°C. Annealed RNA was then incubated for 1 h at 50°C in the presence of 1 mM each ddCTP, dATP, dGTP, and dTTP and 2 U of AMV reverse transcriptase (Promega). After completion, the reaction was resolved on a 7 M urea and 15% polyacrylamide gel for ~4 h. The resulting gel was then dried and exposed to a phosphorimager plate for analysis, as previously described (Jackman *et al.*, 2003).

An interactive Web site for analysis of GFP^{SEQ}, epistasis, and RTD on tRNA cloverleaf maps is also available at http://depts.washington.edu/sfields/tRNA_supplemental/tRNA_interactive.html.

Table S4.1 Summary of *SUP4oc* library FACS and sequencing.

library[a]	FACS, sequencing parameters	total	Bin 1	Bin 2	Bin 3	Bin 4
WT1	cells sorted	28,979,311	2,630,557	1,000,630	1,348,124	24,000,000
WT1	percentage of total	100.0	9.1	3.5	4.7	82.8
WT1	sequencing reads[b]	18,526,477	3,776,058	2,899,542	3,140,476	8,710,401
WT2	cells sorted	24,259,904	2,456,401	868,278	1,011,676	19,923,458
WT2	percentage of total	100.0	10.1	3.6	4.2	82.1
WT2	sequencing reads[b]	145,683,620	14,552,206	10,910,754	6,547,510	113,673,150
Δ1	cells sorted	2,212,947	309,776	101,315	111,619	1,690,237
Δ1	percentage of total	100.0	14.0	4.6	5.0	76.4
Δ1	sequencing reads[b]	9,421,650	1,782,280	2,073,087	2,804,735	2,761,548
Δ2	cells sorted	29,019,400	3,303,947	1,389,328	2,326,125	22,000,000
Δ2	percentage of total	100.0	11.4	4.8	8.0	75.8
Δ2	sequencing reads[b]	14,351,445	4,856,469	3,542,948	3,203,677	2,748,351

[a]WT1: replicate 1 of *SUP4oc* library analyzed in wild type cells; WT2: replicate 2 of *SUP4oc* library analyzed in wild type cells; Δ1: replicate 1 of *SUP4oc* library analyzed in *met22*Δ cells; Δ2: replicate 2 of *SUP4oc* library analyzed in *met22*Δ cells

[b]Sequencing reads with a minimum phred score of 30 for any given cycle

Table S4.2 FACS and sequencing data for all scored *SUP4oc* variants.

See (Guy/Young *et al.*, 2014)

Table S4.3 Epistasis analysis of *SUP4oc* double mutant variants

See (Guy/Young *et al.*, 2014)

Table S4.4 GFP^{SEQ} and GFP^{FLOW} values for reconstructed *SUP4oc* double mutant variants.

See (Guy/Young *et al.*, 2014)

Table S4.5 GFP^{SEQ} and GFP^{FLOW} values for *SUP4oc* RTD candidates

See (Guy/Young *et al.*, 2014)

Table S4.6. GFP^{SEQ} and GFP^{FLOW} values for reconstructed *SUP4oc* double mutant variants

variant[a]	GFP^{FLOW}	GFP^{SEQ}	epistasis
A29C	0.005 ± 0.0002	0.01	n/a
A44U	0.89 ± 0.006	0.98	n/a
A29C A44U	0.10 ± 0.008	0.22	0.21
A9U	0.37 ± 0.004	0.8	n/a
A9U A44U	0.005 ± 0.002	0.065	-0.72
G26U	0.007 ± 0.0001	0.01	n/a
G45C	0.67 ± 0.02	1.14	n/a
G26U,G45C	0.04 ± 0.005	0.32	0.31
U8A	0.001 ± 0.0002	0.02	n/a
A14G	0.001 ± 0.0001	0.02	n/a
U8A A14G	0.05 ± 0.002	0.18	0.18
U17G	0.96 ± 0.01	1.00	n/a
G19U	0.006 ± 0.0004	0.01	n/a
U17G G19U	0.94 ± 0.01	1.05	1.04
G18A	0.006 ± 0.00005	0.01	n/a
U20G	0.94 ± 0.01	0.99	n/a
G18A U20G	0.98 ± 0.01	0.91	0.89
C27U	0.78 ± 0.003	1.07	n/a
A28C	0.005 ± 0.0002	0.01	n/a
C27U A28C	0.08 ± 0.007	0.20	0.18
Δ42	0.007 ± 0.001	n/a	n/a
C27U A28C Δ42	0.075 ± 0.002	n/a	n/a

[a]Mean and standard deviation for triplicate isolates

Table S4.7 Strains used in this study.

See (Guy/Young *et al.*, 2014)

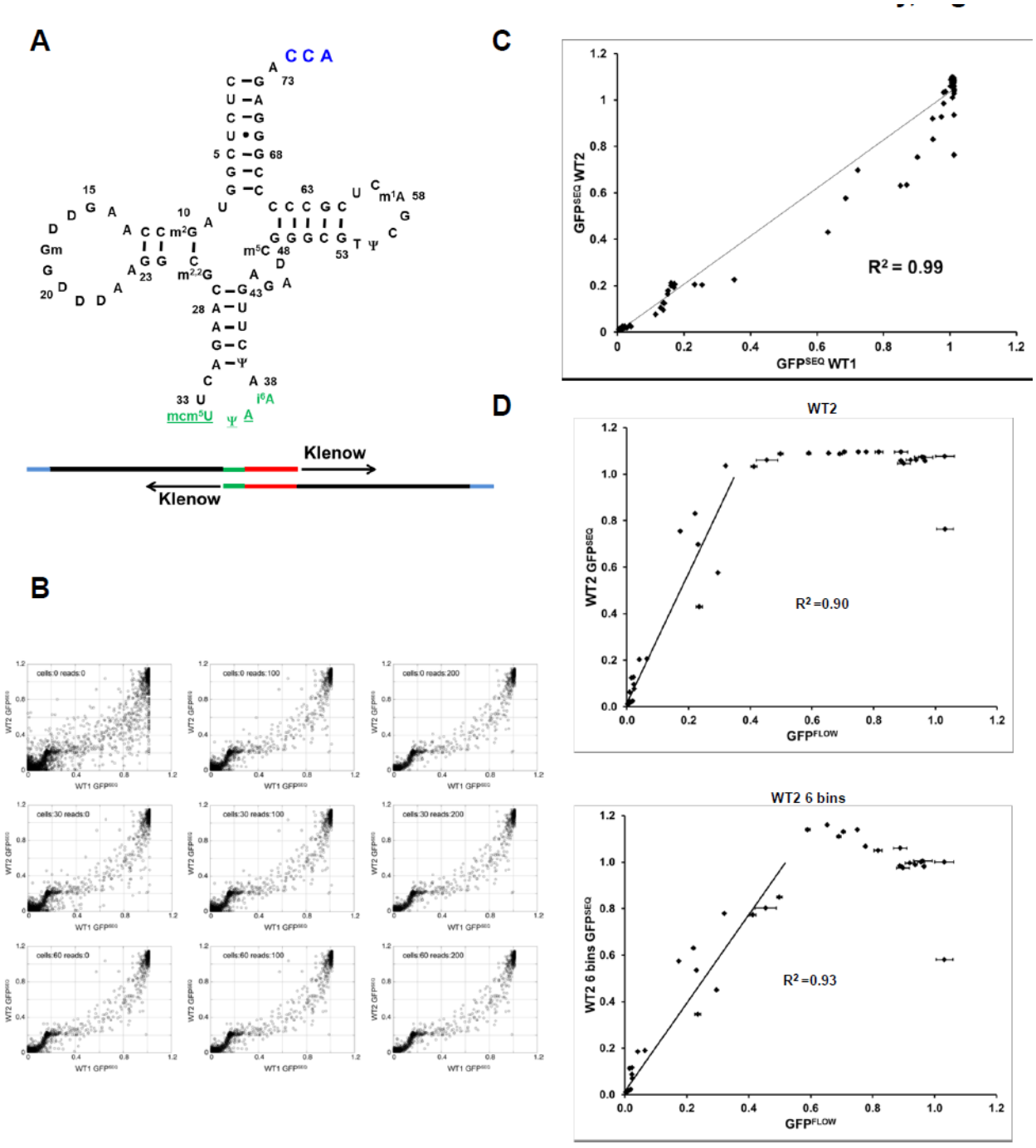


Figure S4.7 Generation and analysis of a *SUP4oc* randomly mutated tRNA library.

(A) Construction of a randomly mutated *SUP4oc* tRNA library. DNA oligonucleotides with 3% random mutations (black), and no mutations in positions 34-38 (green) and the intron (red) were annealed, and the Klenow fragment of DNA polymerase I was used to fill in the gaps, prior to restriction digestion and ligation. The CCA end (blue) is represented for completeness, and is added post-transcriptionally. (B) Determination of filtering parameters for maximal correlation of GFP_{SEQ} from replicate analyses of function of *SUP4oc* variants. To select the thresholds for filtering reads, we compared GFP_{SEQ} values of biological replicates (the WT1 and WT2 libraries; see materials and methods). Selected scatterplots

comparing GFP_{SEQ} from WT1 and WT2 libraries after filtering *SUP4oc* variants by total sequencing reads and total estimated cells are depicted (indicated in top left of plots). Thresholds were manually selected as total read counts ≥ 100 and total cell counts ≥ 30, since these thresholds appeared to reduce spurious correlations without overly reducing the number of variant sequences. (C) GFP_{SEQ} scores determined by sequencing are highly correlated between biological replicates of the *SUP4oc* library. Plot of the correlation between GFP_{SEQ} for singly mutated variants from two independent analyses of the *SUP4oc* library in wild type cells, by FACS and sequencing. (D) tRNA function determined by GFPFLOW correlates with GFP_{SEQ} determined by sequencing. Plot of GFP_{SEQ} as determined by FACS and sequencing, compared to GFP_{FLOW} as determined after reconstruction and analytical flow cytometry. GFP_{FLOW} values are mean and standard deviation from measurement of triplicate isolates of each strain. top, GFP_{SEQ} of WT2; bottom, GFP_{SEQ} of WT2 6 bins.

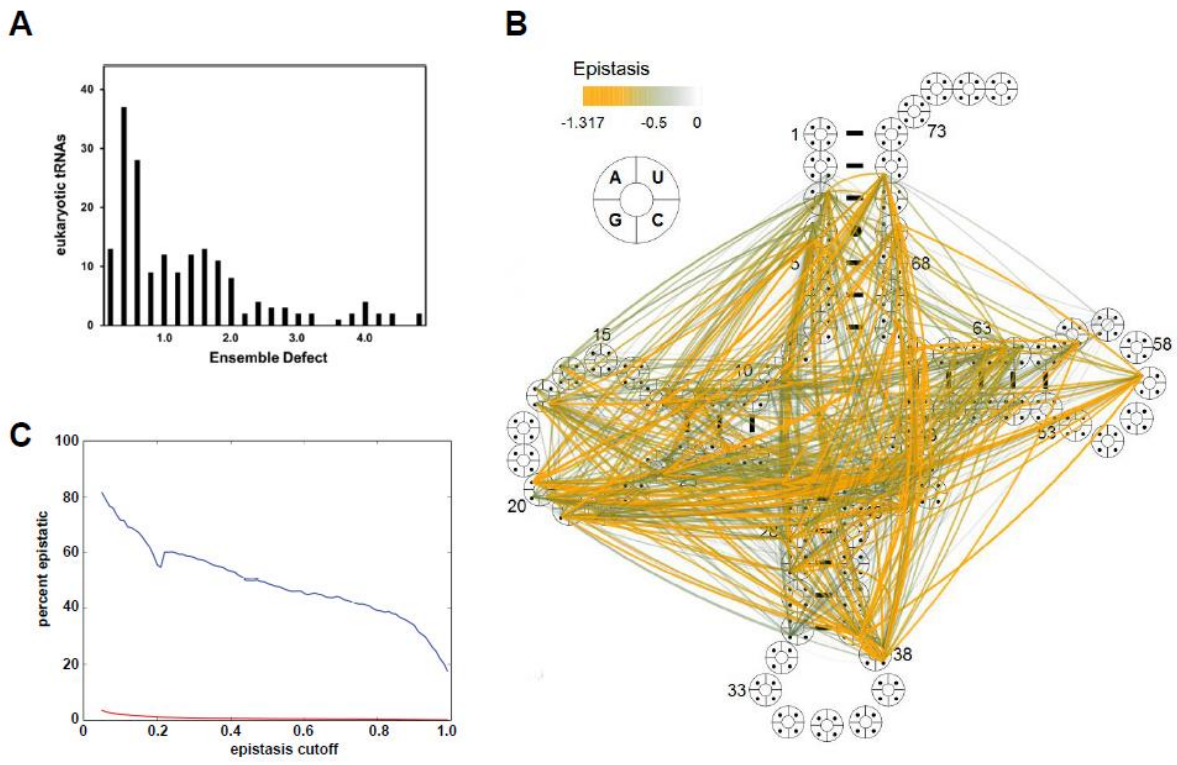


Figure S4.8 Analysis of *SUP4oc* double mutant variants.

(A) Calculated ensemble defect of eukaryotic tRNAs. Bar graph depicting the distribution of calculated ensemble defect for representative eukaryotic tRNAs for which the modification status is known (see Materials and Methods). (B) Cloverleaf schematic map of negative epistatic interactions between bases in *SUP4oc*. Color and width of lines corresponds to strength of the interactions. (C) Negative epistasis is more prevalent than positive epistasis. Negative epistasis (blue); Positive epistasis (red).

A

A9U epistasis

ID	S1	S2	D	Epistasis
4,9-C,T	0.59	0.80	1.07	0.60
9,68-T,A	0.80	0.02	0.14	0.12
9,58-T,T	0.80	0.01	0.11	0.10
9,58-T,G	0.80	0.01	0.10	0.09
9,68-T,T	0.80	0.01	0.07	0.06
+73; epistasis of 0.06 to -0.14				
9,31-T,G	0.80	0.35	0.03	-0.24
9,45-T,T	0.80	1.03	0.57	-0.25
9,12-T,T	0.80	0.77	0.29	-0.33
4,9-A,T	0.69	0.80	0.20	-0.36
9,44-T,C	0.80	0.49	0.02	-0.37
9,45-T,A	0.80	0.85	0.20	-0.49
3,9-T,T	0.74	0.80	0.01	-0.58
9,62-T,A	0.80	0.78	0.04	-0.59
9,16-T,G	0.80	1.01	0.20	-0.60
9,38-T,T	0.80	1.04	0.22	-0.62
9,20-T,G	0.80	0.99	0.12	-0.67
9,16-T,A	0.80	1.00	0.13	-0.67
9,29-T,G	0.80	1.08	0.19	-0.67
9,17-T,C	0.80	1.05	0.13	-0.71
9,45-T,C	0.80	1.14	0.20	-0.72
9,44-T,T	0.80	0.98	0.07	-0.72
9,28-T,G	0.80	1.11	0.10	-0.79
9,47-T,C	0.80	1.05	0.05	-0.79
9,57-T,A	0.80	1.12	0.08	-0.81
9,38-T,C	0.80	1.16	0.05	-0.88
9,44-T,G	0.80	1.14	0.02	-0.90
9,65-T,T	0.80	1.16	0.03	-0.90

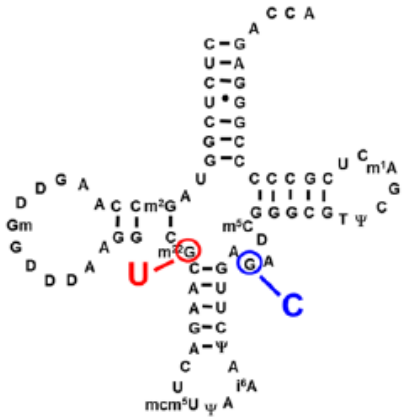
A29C epistasis

ID	S1	S2	D	Epistasis
29,41-C,A	0.01	0.09	0.22	0.22
29,44-C,T	0.01	0.98	0.22	0.21
24,29-C,C	0.01	0.01	0.02	0.02
29,63-C,G	0.01	0.01	0.02	0.02
29,72-C,A	0.01	0.01	0.01	0.01
29,73-C,G	0.01	0.01	0.01	0.01
5,29-G,C	0.01	0.01	0.01	0.01
29,71-C,C	0.01	0.01	0.01	0.01
29,56-C,G	0.01	0.01	0.01	0.01
29,60-C,G	0.01	0.01	0.01	0.01
14,29-T,C	0.02	0.01	0.01	0.01
29,73-C,T	0.01	0.01	0.01	0.01
29,67-C,G	0.01	0.01	0.01	0.01
2,29-A,C	0.01	0.01	0.01	0.01
29,50-C,C	0.01	0.01	0.01	0.01
25,29-G,C	0.01	0.01	0.01	0.01
6,29-C,C	0.01	0.01	0.01	0.01
23,29-C,C	0.18	0.01	0.01	0.01
29,64-C,G	0.01	0.01	0.01	0.01
7,29-T,C	0.01	0.01	0.01	0.01
29,30-C,T	0.01	0.01	0.01	0.01
29,46-C,C	0.01	0.01	0.01	0.01
29,32-C,T	0.01	0.01	0.01	0.01
29,41-C,C	0.01	0.01	0.01	0.01
25,29-A,C	0.01	0.01	0.01	0.01
5,29-A,C	0.01	0.01	0.01	0.01
24,29-A,C	0.01	0.01	0.01	0.01
+66; epistasis of 0.01 to 0				

A44U epistasis

ID	S1	S2	D	Epistasis
29,44-T,T	0.18	0.98	0.65	0.47
29,44-C,T	0.01	0.98	0.22	0.21
28,44-T,T	0.57	0.98	0.77	0.20
20,44-G,T	0.99	0.98	1.12	0.14
27,44-T,T	1.07	0.98	1.18	0.13
31,44-T,T	0.07	0.98	0.20	0.12
44,46-T,G	0.98	0.02	0.11	0.09
44,52-T,T	0.98	1.05	1.11	0.07
44,60-T,C	0.98	1.01	1.05	0.06
41,44-C,T	0.01	0.98	0.03	0.01
2,44-C,T	0.19	0.98	0.19	0.01
44,50-T,T	0.98	0.01	0.02	0.01
7,44-C,T	0.01	0.98	0.02	0.01
7,44-C,T	0.01	0.98	0.02	0.01
+55; epistasis of 0.01 to -0.1				
44,71-T,G	0.98	1.16	1.02	-0.12
4,44-A,T	0.69	0.98	0.52	-0.16
44,69-T,C	0.98	0.16	0.02	-0.17
44,51-T,A	0.98	0.19	0.05	-0.17
3,44-T,T	0.74	0.98	0.55	-0.18
15,44-A,T	0.19	0.98	0.02	-0.20
22,44-G,T	0.19	0.98	0.02	-0.20
44,65-T,T	0.98	1.16	0.82	-0.32
44,47-T,G	0.98	1.14	0.79	-0.33
22,44-T,T	0.49	0.98	0.06	-0.42
9,44-C,T	1.05	0.98	0.58	-0.45
9,44-T,T	0.80	0.98	0.07	-0.72
44,57-T,A	0.98	1.12	0.22	-0.88

B



C

G26U epistasis

ID	S1	S2	D	Epistasis
26,44-T,C	0.01	0.49	0.67	0.66
26,45-T,C	0.01	1.14	0.32	0.31
26,45-T,T	0.01	1.03	0.19	0.18
12,26-T,T	0.77	0.01	0.07	0.06
26,62-T,T	0.01	0.63	0.04	0.04
26,62-T,A	0.01	0.78	0.03	0.02
26,31-T,T	0.01	0.07	0.02	0.01
13,26-C,T	1.13	0.01	0.03	0.01
21,26-C,T	0.01	0.01	0.01	0.01
2,26-G,T	0.01	0.01	0.01	0.01
9,26-T,T	0.80	0.01	0.02	0.01
5,26-A,T	0.01	0.01	0.01	0.01
14,26-T,T	0.02	0.01	0.01	0.01
3,26-T,T	0.74	0.01	0.02	0.01
26,64-T,G	0.01	0.01	0.01	0.01
21,26-G,T	0.02	0.01	0.01	0.01
26,61-T,A	0.01	0.01	0.01	0.01
26,67-T,A	0.01	0.01	0.01	0.01
26,43-T,A	0.01	0.01	0.01	0.01
8,26-A,T	0.02	0.01	0.01	0.01
26,71-T,T	0.01	0.02	0.01	0.01
18,26-A,T	0.01	0.01	0.01	0.01
10,26-T,T	0.01	0.01	0.01	0.01
26,50-T,T	0.01	0.01	0.01	0.01
26,33-T,C	0.01	0.01	0.01	0.01
26,66-T,G	0.01	0.01	0.01	0.01
19,26-C,T	0.01	0.01	0.01	0.01
+55; epistasis of 0.01 to 0				

G45C epistasis

ID	S1	S2	D	Epistasis
9,45-G,C	0.45	1.14	1.06	0.55
26,45-T,C	0.01	1.14	0.32	0.31
45,71-C,T	1.14	0.02	0.30	0.28
45,65-C,G	1.14	0.01	0.25	0.23
+79; epistasis of 0.19 to -0.19 +6; epistasis of -0.2 to -0.43				
20,45-G,C	0.99	1.14	0.61	-0.51
22,45-T,C	0.49	1.14	0.02	-0.54
45,52-C,T	1.14	1.05	0.63	-0.56
45,62-C,T	1.14	0.63	0.13	-0.59
28,45-T,C	0.57	1.14	0.02	-0.64
3,45-T,C	0.74	1.14	0.16	-0.68
9,45-T,C	0.80	1.14	0.20	-0.72
17,45-G,C	1.00	1.14	0.37	-0.77
45,63-C,T	1.14	0.88	0.20	-0.80
45,60-C,C	1.14	1.01	0.23	-0.92
26,45-C,C	1.17	1.14	0.39	-0.94
29,45-G,C	1.08	1.14	0.27	-0.96
13,45-T,C	1.03	1.14	0.19	-0.99
45,47-C,C	1.14	1.05	0.20	-1.00
28,45-G,C	1.11	1.14	0.25	-1.01
45,57-C,A	1.14	1.12	0.21	-1.06
45,71-C,G	1.14	1.16	0.25	-1.08
20a,45-G,C	1.04	1.14	0.11	-1.08
20b,45-A,C	1.13	1.14	0.20	-1.08
38,45-C,C	1.16	1.14	0.22	-1.10
45,47-C,G	1.14	1.14	0.12	-1.18
27,45-T,C	1.07	1.14	0.03	-1.19
20,45-A,C	1.18	1.14	0.07	-1.27

D

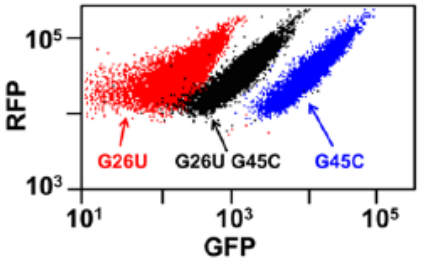
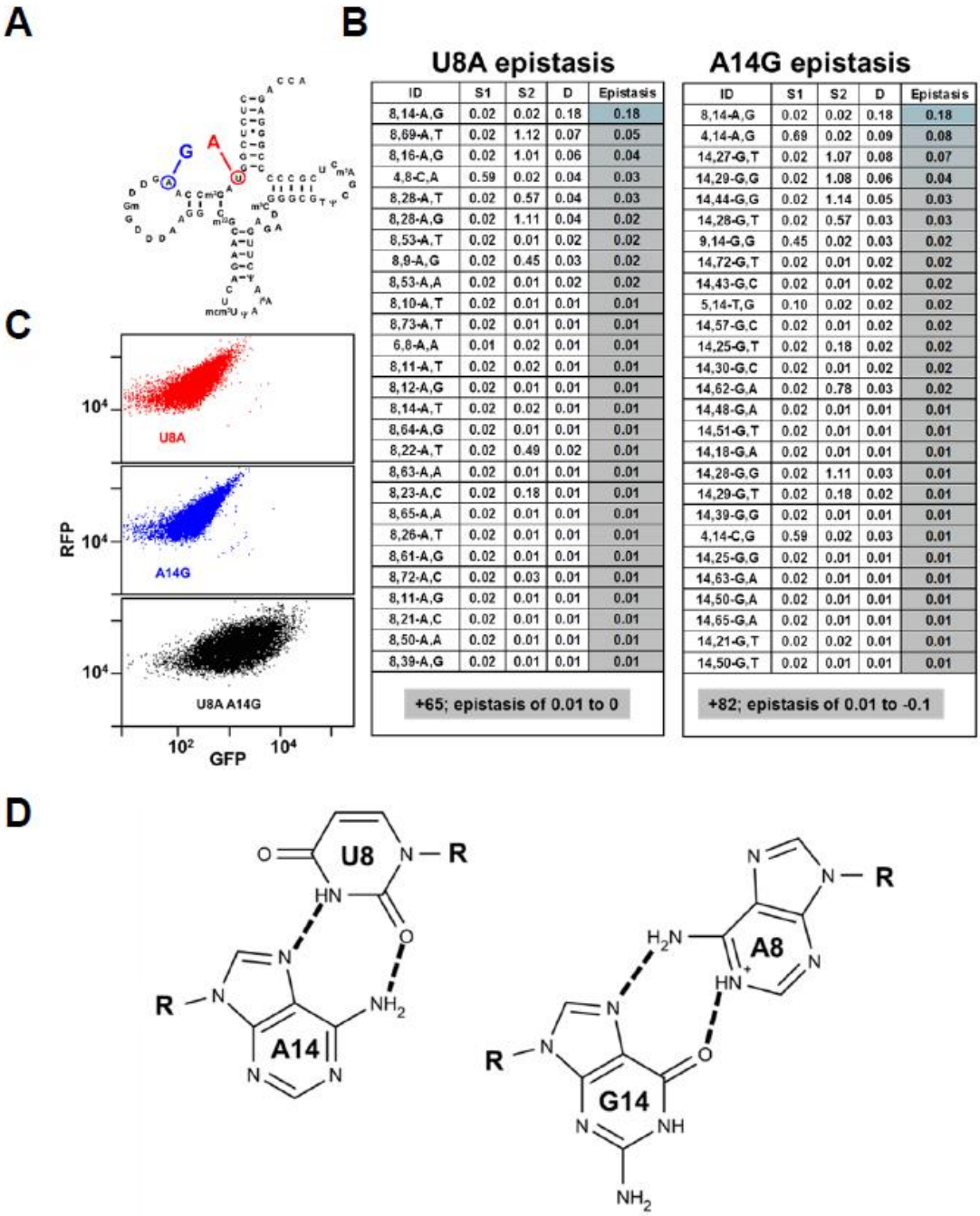


Figure S4.9 Evidence for an alternative conformation at the 26-44 hinge.

(A) A44U has both positive and negative epistatic interactions. Tables of epistasis values for *SUP4oc* double mutants containing the indicated mutations are shown. (ID) Identity of the mutation using standard tRNA numbering, (S1) and (S2) GFPSEQ of constituent single mutant variants, (D) GFPSEQ of double mutant variant. (B) Cloverleaf schematic of *SUP4oc* with the location of the analyzed mutations. (C) Tables of epistasis values for double mutants containing the indicated mutations. (D) Scatter plots of RFP and GFP of cells expressing the indicated *SUP4oc* variants.



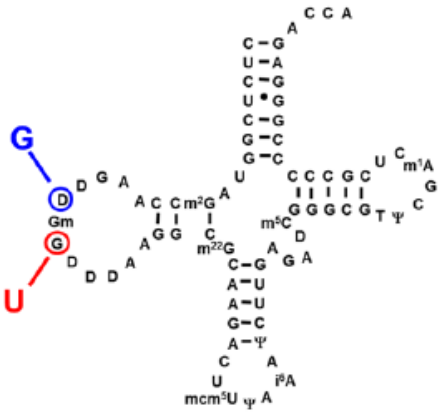
ID	S1	S2	D	Epistasis
8,14-A,G	0.02	0.02	0.18	0.18
8,69-A,T	0.02	1.12	0.07	0.05
8,16-A,G	0.02	1.01	0.06	0.04
4,8-C,A	0.59	0.02	0.04	0.03
8,28-A,T	0.02	0.57	0.04	0.03
8,28-A,G	0.02	1.11	0.04	0.02
8,53-A,T	0.02	0.01	0.02	0.02
8,9-A,G	0.02	0.45	0.03	0.02
8,53-A,A	0.02	0.01	0.02	0.02
8,10-A,T	0.02	0.01	0.01	0.01
8,73-A,T	0.02	0.01	0.01	0.01
6,8-A,A	0.01	0.02	0.01	0.01
8,11-A,T	0.02	0.02	0.01	0.01
8,12-A,G	0.02	0.01	0.01	0.01
8,14-A,T	0.02	0.02	0.01	0.01
8,64-A,G	0.02	0.01	0.01	0.01
8,22-A,T	0.02	0.49	0.02	0.01
8,63-A,A	0.02	0.01	0.01	0.01
8,23-A,C	0.02	0.18	0.01	0.01
8,65-A,A	0.02	0.01	0.01	0.01
8,26-A,T	0.02	0.01	0.01	0.01
8,61-A,G	0.02	0.01	0.01	0.01
8,72-A,C	0.02	0.03	0.01	0.01
8,11-A,G	0.02	0.01	0.01	0.01
8,21-A,C	0.02	0.01	0.01	0.01
8,50-A,A	0.02	0.01	0.01	0.01
8,39-A,G	0.02	0.01	0.01	0.01

ID	S1	S2	D	Epistasis
8,14-A,G	0.02	0.02	0.18	0.18
4,14-A,G	0.69	0.02	0.09	0.08
14,27-G,T	0.02	1.07	0.08	0.07
14,29-G,G	0.02	1.08	0.06	0.04
14,44-G,G	0.02	1.14	0.05	0.03
14,28-G,T	0.02	0.57	0.03	0.03
9,14-G,G	0.45	0.02	0.03	0.02
14,72-G,T	0.02	0.01	0.02	0.02
14,43-G,C	0.02	0.01	0.02	0.02
5,14-T,G	0.10	0.02	0.02	0.02
14,57-G,C	0.02	0.01	0.02	0.02
14,25-G,T	0.02	0.18	0.02	0.02
14,30-G,C	0.02	0.01	0.02	0.02
14,62-G,A	0.02	0.78	0.03	0.02
14,48-G,A	0.02	0.01	0.01	0.01
14,51-G,T	0.02	0.01	0.01	0.01
14,18-G,A	0.02	0.01	0.01	0.01
14,28-G,G	0.02	1.11	0.03	0.01
14,29-G,T	0.02	0.18	0.02	0.01
14,39-G,G	0.02	0.01	0.01	0.01
4,14-C,G	0.59	0.02	0.03	0.01
14,25-G,G	0.02	0.01	0.01	0.01
14,63-G,A	0.02	0.01	0.01	0.01
14,50-G,A	0.02	0.01	0.01	0.01
14,65-G,A	0.02	0.01	0.01	0.01
14,21-G,T	0.02	0.02	0.01	0.01
14,50-G,T	0.02	0.01	0.01	0.01

Figure S4.10 A U8A A14G *SUP4oc* variant has activity.

(A) Cloverleaf schematic of *SUP4oc* with the location of the analyzed mutations. (B) Tables of epistasis values for double mutants containing the indicated mutations. Column headings as in Supplemental Figure S4.9A. (C) Scatter plots of RFP and GFP of cells expressing the indicated *SUP4oc* variants. (D) Proposed model of the 8-14 reverse Hoogsteen tertiary interaction restored in the *SUP4oc* U8A A14G variant. Left, Reverse Hoogsteen U8-A14 tertiary interaction as seen in the structures of tRNA. Right, proposed reverse Hoogsteen interaction for A8-G14 (Leontis et al. 2002).

A



B

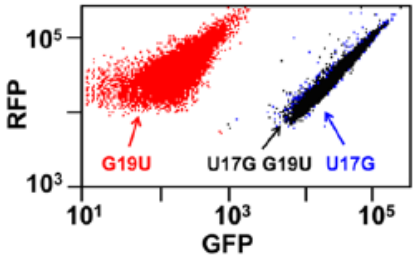
U17G epistasis

ID	S1	S2	D	Epistasis
17,19-G,T	1.00	0.01	1.05	1.04
16,17-C,G	0.86	1.00	1.07	0.20
17,19-G,A	1.00	0.01	0.20	0.19
16,17-A,G	1.00	1.00	1.16	0.16
17,26-G,A	1.00	0.97	1.13	0.15
17,20a-G,A	1.00	1.02	1.16	0.14
17,38-G,T	1.00	1.04	1.16	0.11
17,20a-G,C	1.00	0.95	1.05	0.10
+57 epistasis of -0.16 to 0.07				
13,17-C,G	1.13	1.00	0.96	-0.17
17,69-G,C	1.00	0.16	0.01	-0.18
17,29-G,T	1.00	0.18	0.02	-0.19
17,45-G,T	1.00	1.03	0.83	-0.20
17,25-G,T	1.00	0.18	0.01	-0.20
17,62-G,C	1.00	0.20	0.01	-0.23
17,47-G,C	1.00	1.05	0.80	-0.25
17,44-G,C	1.00	0.49	0.19	-0.30
17,20a-G,G	1.00	1.04	0.72	-0.32
17,22-G,T	1.00	0.49	0.16	-0.33
17,71-G,G	1.00	1.16	0.75	-0.41
9,17-C,G	1.05	1.00	0.60	-0.45
17,22-G,C	1.00	1.13	0.65	-0.48
17,28-G,T	1.00	0.57	0.06	-0.51
17,38-G,C	1.00	1.16	0.64	-0.52
17,28-G,G	1.00	1.11	0.40	-0.71
17,45-G,C	1.00	1.14	0.37	-0.77
17,26-G,C	1.00	1.17	0.21	-0.97
17,39-G,C	1.00	1.09	0.08	-1.01

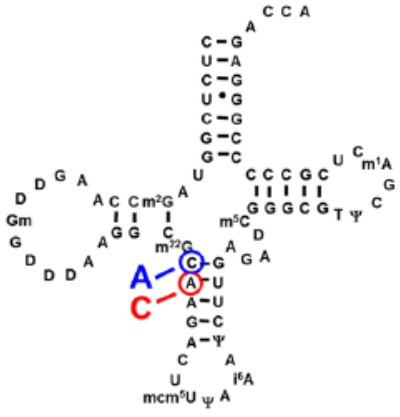
G19U epistasis

ID	S1	S2	D	Epistasis
17,19-G,T	1.00	0.01	1.05	1.04
19,26-T,A	0.01	0.97	0.03	0.02
19,64-T,A	0.01	0.01	0.02	0.02
19,60-T,G	0.01	0.01	0.02	0.02
19,62-T,C	0.01	0.20	0.02	0.01
19,51-T,A	0.01	0.19	0.02	0.01
10,19-T,T	0.01	0.01	0.01	0.01
19,25-T,A	0.01	0.01	0.01	0.01
19,21-T,T	0.01	0.02	0.01	0.01
19,33-T,G	0.01	0.01	0.01	0.01
19,21-T,C	0.01	0.01	0.01	0.01
19,56-T,A	0.01	0.01	0.01	0.01
18,19-T,T	0.01	0.01	0.01	0.01
19,50-T,C	0.01	0.01	0.01	0.01
12,19-A,T	0.01	0.01	0.01	0.01
19,41-T,G	0.01	0.01	0.01	0.01
19,40-T,A	0.01	0.01	0.01	0.01
11,19-G,T	0.01	0.01	0.01	0.01
19,58-T,C	0.01	0.01	0.01	0.01
2,19-G,T	0.01	0.01	0.01	0.01
14,19-G,T	0.02	0.01	0.01	0.01
19,54-T,G	0.01	0.01	0.01	0.01
14,19-T,T	0.02	0.01	0.01	0.01
19,24-T,A	0.01	0.01	0.01	0.01
19,39-T,G	0.01	0.01	0.01	0.01
12,19-T,T	0.77	0.01	0.02	0.01
19,68-T,A	0.01	0.02	0.01	0.01
+65 epistasis of 0 to 0.01				

C



D



E

C27U epistasis

ID	S1	S2	D	Epistasis
4,27-C,T	0.59	1.07	1.03	0.40
27,60-T,A	1.07	0.01	0.22	0.21
16,27-C,T	0.86	1.07	1.11	0.19
27,28-T,C	1.07	0.01	0.20	0.18
17,27-A,T	0.94	1.07	1.15	0.15
+72 epistasis of -0.17 to 0.13 +6 epistasis of -0.3 to -0.2				
27,60-T,C	1.07	1.01	0.77	-0.30
27,45-T,T	1.07	1.03	0.75	-0.35
13,27-T,T	1.03	1.07	0.75	-0.35
27,31-T,G	1.07	0.35	0.01	-0.36
9,27-C,T	1.05	1.07	0.75	-0.37
13,27-C,T	1.13	1.07	0.81	-0.40
22,27-C,T	1.13	1.07	0.75	-0.46
27,28-T,T	1.07	0.57	0.12	-0.49
27,67-T,T	1.07	0.54	0.04	-0.53
12,27-T,T	0.77	1.07	0.21	-0.62
27,62-T,T	1.07	0.63	0.02	-0.65
27,71-T,G	1.07	1.16	0.44	-0.80
27,52-T,T	1.07	1.05	0.27	-0.85
20,27-C,T	0.85	1.07	0.03	-0.88
16,27-A,T	1.00	1.07	0.19	-0.88
27,66-T,T	1.07	1.06	0.22	-0.90
27,57-T,A	1.07	1.12	0.24	-0.95
27,44-T,G	1.07	1.14	0.11	-1.10
27,29-T,G	1.07	1.08	0.04	-1.11
27,69-T,T	1.07	1.12	0.03	-1.17
27,45-T,C	1.07	1.14	0.03	-1.19
27,38-T,C	1.07	1.16	0.02	-1.21

A28C epistasis

ID	S1	S2	D	Epistasis
28,42-C,G	0.01	0.02	1.02	1.02
27,28-T,C	1.07	0.01	0.20	0.18
28,41-C,C	0.01	0.01	0.01	0.01
28,50-C,T	0.01	0.01	0.01	0.01
28,57-C,C	0.01	0.01	0.01	0.01
28,72-C,T	0.01	0.01	0.01	0.01
24,28-T,C	0.01	0.01	0.01	0.01
28,69-C,C	0.01	0.16	0.01	0.01
5,28-A,C	0.01	0.01	0.01	0.01
28,62-C,C	0.01	0.20	0.01	0.01
28,67-C,A	0.01	0.01	0.01	0.01
12,28-A,C	0.01	0.01	0.01	0.01
27,28-G,C	0.01	0.01	0.01	0.01
28,56-C,T	0.01	0.01	0.01	0.01
27,28-A,C	0.11	0.01	0.01	0.01
1,28-A,C	0.01	0.01	0.01	0.01
28,60-C,G	0.01	0.01	0.01	0.01
28,39-C,A	0.01	0.01	0.01	0.01
23,28-A,C	0.02	0.01	0.01	0.01
28,51-C,A	0.01	0.19	0.01	0.01
28,73-C,G	0.01	0.01	0.01	0.01
24,28-C,C	0.01	0.01	0.01	0.01
3,28-A,C	0.01	0.01	0.01	0.01
24,28-A,C	0.01	0.01	0.01	0.01
28,53-C,C	0.01	0.01	0.01	0.01
28,70-C,T	0.01	0.01	0.01	0.01
12,28-G,C	0.01	0.01	0.01	0.01
+62 epistasis of 0 to 0.01				

Figure S4.11 Positive epistasis due to shift of interactions to neighboring residues in *SUP4oc* variants.

(A) Cloverleaf schematic of *SUP4oc* with the location of the analyzed mutations. (B) Tables of epistasis values for double mutants containing the indicated mutations. Column headings as in Supplemental Figure S4.9A. (C) Scatter plots of RFP and GFP of cells expressing the indicated *SUP4oc* variants. (D) Cloverleaf schematic of *SUP4oc* with the location 9 of the analyzed mutations. (E) Tables of epistasis values for double mutants containing the indicated mutations. Column headings as in Supplemental Figure S4.9A.

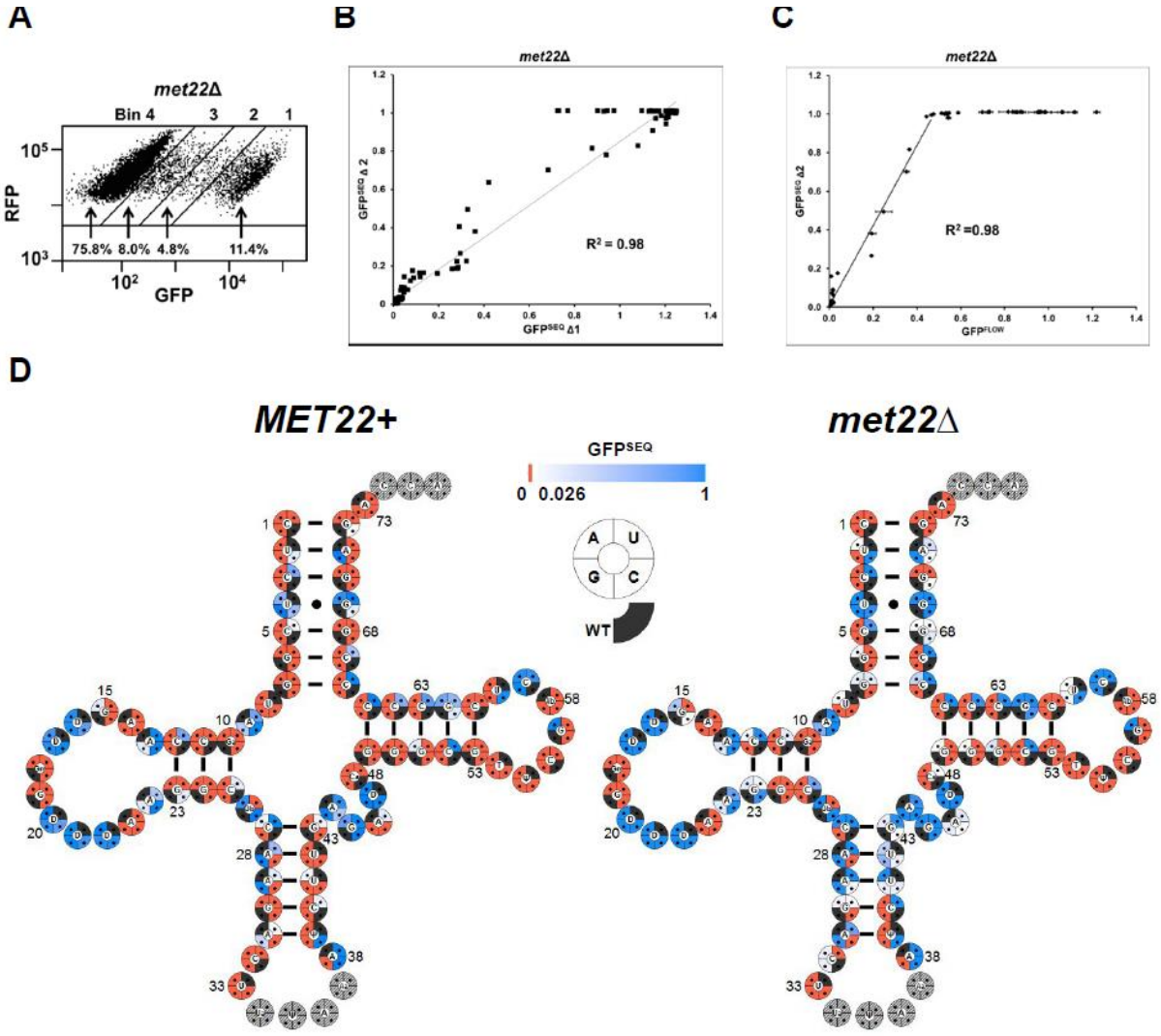
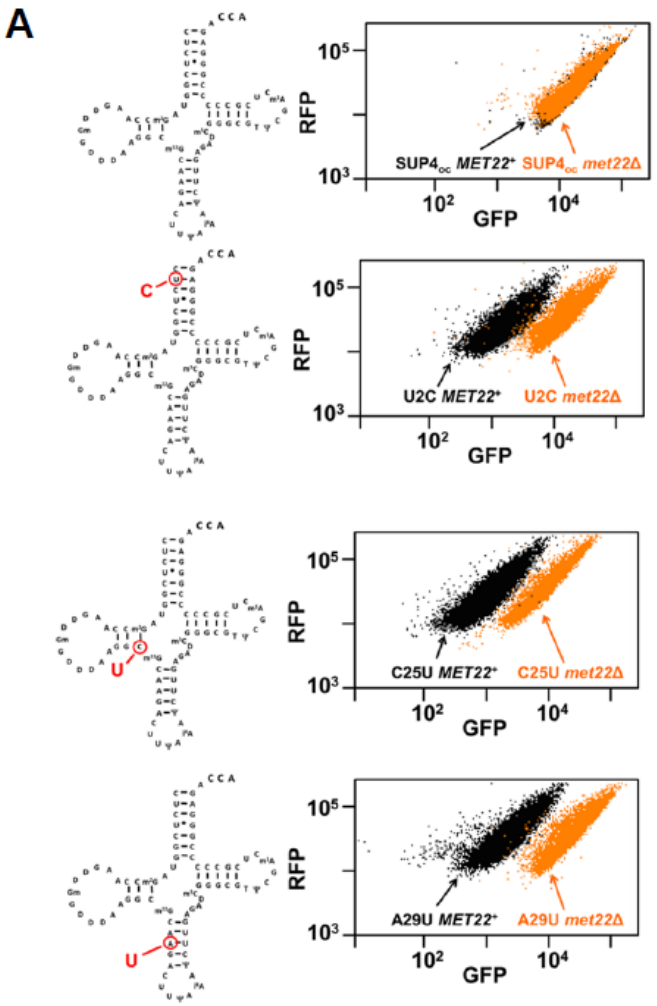


Figure S4.12 FACS and sequencing analysis of the *SUP4oc* library in *met22Δ* cells.

(A) FACS of the *SUP4oc* library in *met22Δ* cells. *met22Δ* cells with the integrated *SUP4oc* library and the integrated *GFPoc/RFP* reporter were analyzed by FACS, followed by sequencing. (B)GFP^{SEQ} scores determined by sequencing are highly correlated between biological replicates of the *SUP4oc* library in the *met22Δ* strain. Plot of the correlation between GFP^{SEQ} for singly mutated variants from two independent analyses of the *SUP4oc* library in the *met22Δ* strain, by FACS and sequencing. (C) tRNA function determined by GFP^{FLOW} correlates with GFP^{SEQ} determined by sequencing in the *met22Δ* strain. Plot of GFP^{SEQ} as determined by FACS and sequencing, compared to GFP^{FLOW} as determined after reconstruction and analytical flow cytometry. GFP^{FLOW} values are mean and standard deviation from measurement of triplicate isolates of each strain. (D) Heat map illustrating the activity of *SUP4oc* singly mutated variants in wild type and *met22Δ* cells. (Red), no function. Intensity of blue corresponds to activity for detectably functional variants ranging from GFP^{SEQ} 0.026 to 1.



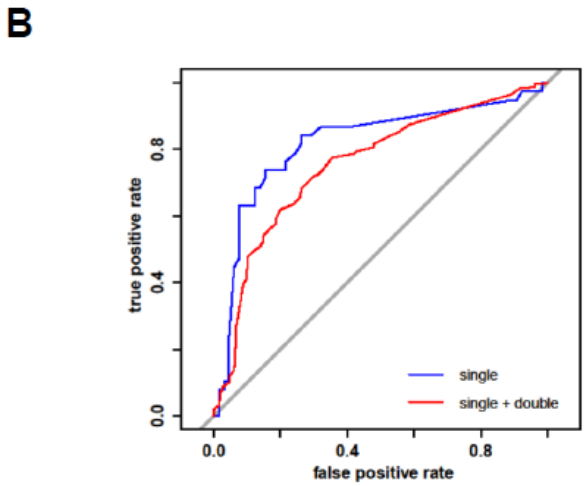


Figure S4.13 RTD substrate candidates are found in the acceptor stem, D-stem, and anticodon stem.

(A) Cloverleaf schematic of indicated *SUP4oc* variants with the location of the analyzed mutation. Scatter plots of RFP and GFP in wild type (black) and *met22Δ* (orange) cells expressing the indicated *SUP4oc* variant. (B) ROC (Receiver Operating Characteristic) curves for RTD prediction based on estimated ΔΔG°28. Blue curve represents the ROC curve for the single variants, and red curve is ROC curve for both the single and double variants for which RTD ratio could be scored (see materials and methods).

A

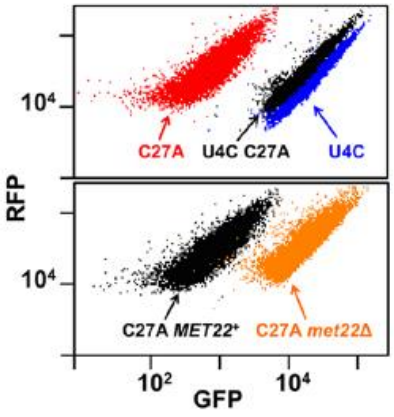
U4C epistasis

ID	S1	S2	D	Epistasis
4,70-C,A	0.59	0.01	1.09	1.09
4,27-C,A	0.59	0.11	1.11	1.05
4,6-C,T	0.59	0.01	0.96	0.95
4,29-C,T	0.59	0.18	1.04	0.94
4,62-C,C	0.59	0.20	1.06	0.94
2,4-C,C	0.19	0.59	1.01	0.90
4,67-C,T	0.59	0.54	1.17	0.86
4,44-C,C	0.59	0.49	1.05	0.76
4,62-C,T	0.59	0.63	1.12	0.75
4,45-C,A	0.59	0.85	1.25	0.75
4,25-C,T	0.59	0.18	0.79	0.69
4,13-C,G	0.59	0.19	0.77	0.66
4,15-C,A	0.59	0.19	0.75	0.64
4,23-C,C	0.59	0.18	0.72	0.61
4,9-C,T	0.59	0.80	1.07	0.60
4,41-C,A	0.59	0.09	0.61	0.56
+18, epistasis of 0.56 to 0.2 +62, epistasis of 0.19 to 0				
4,58-C,G	0.59	0.01	0.01	0.00
4,49-C,T	0.59	0.01	0.01	0.00
4,57-C,T	0.59	0.01	0.01	0.00
4,56-C,A	0.59	0.01	0.01	0.00
4,19-C,A	0.59	0.01	0.01	0.00
4,46-C,G	0.59	0.02	0.02	0.00
4,42-C,G	0.59	0.02	0.01	0.00
4,69-C,C	0.59	0.16	0.03	-0.11
4,31-C,G	0.59	0.35	0.04	-0.16
4,28-C,T	0.59	0.57	0.08	-0.26
4,47-C,G	0.59	1.14	0.09	-0.57

C27A epistasis

ID	S1	S2	D	Epistasis
4,27-C,A	0.59	0.11	1.11	1.05
27,56-A,A	0.11	0.01	0.02	0.02
1,27-A,A	0.01	0.11	0.02	0.01
27,60-A,G	0.11	0.01	0.02	0.01
14,27-T,A	0.02	0.11	0.02	0.01
27,73-A,C	0.11	0.01	0.02	0.01
11,27-T,A	0.02	0.11	0.02	0.01
27,39-A,G	0.11	0.01	0.01	0.01
19,27-C,A	0.01	0.11	0.01	0.01
10,27-T,A	0.01	0.11	0.01	0.01
11,27-G,A	0.01	0.11	0.01	0.01
27,65-A,G	0.11	0.01	0.01	0.01
+55, epistasis of 0.01 to -0.09				
27,44-A,G	0.11	1.14	0.03	-0.10
27,59-A,T	0.11	0.97	0.03	-0.10
27,39-A,C	0.11	1.09	0.02	-0.10
27,57-A,A	0.11	1.12	0.01	-0.11
20b,27-C,A	1.00	0.11	0.02	-0.11
27,60-A,C	0.11	1.01	0.02	-0.12
27,47-A,G	0.11	1.14	0.02	-0.12
22,27-C,A	1.13	0.11	0.02	-0.12
16,27-A,A	1.00	0.11	0.02	-0.12
20,27-C,A	0.85	0.11	0.01	-0.12
20,27-A,A	1.18	0.11	0.01	-0.12
27,38-A,T	0.11	1.04	0.01	-0.12
9,27-C,A	1.05	0.11	0.01	-0.12
27,45-A,T	0.11	1.03	0.01	-0.12
27,47-A,C	0.11	1.05	0.01	-0.12

B



C

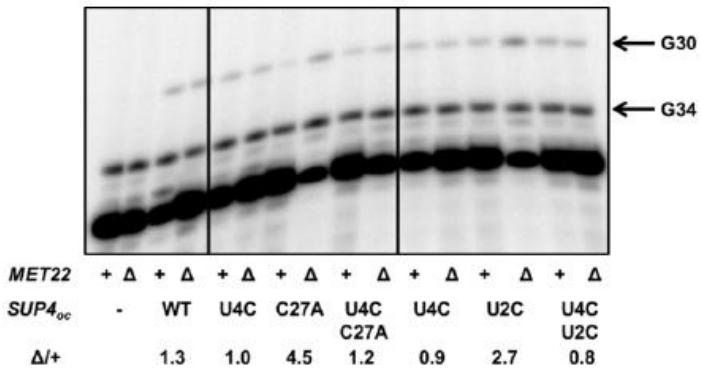


Figure S4.14 A U4C stabilizing mutation rescues variants that are RTD substrates.

(A) Tables of epistasis values for double mutants containing the indicated mutations. Column headings as in Supplemental Figure S4.9A. (B) Top, Flow cytometry of cells expressing the indicated *SUP4oc* variants. Bottom, Flow cytometry of cells expressing the indicated *SUP4oc* variant in wild type and *met22Δ* cells. (C) A U4C mutation restores *SUP4oc* levels of RTD substrate variants in wild type cells to those in the *met22*Δ strain. Levels of *SUP4oc* were determined in tRNATyr purified from indicated strains using poisoned primer extension assays with ddCTP, which results in a G34 stop for tRNATyr and a G30 stop for *SUP4oc*. Numbers at bottom correspond to the ratio of *SUP4oc* to total tRNATyr (wild type and *SUP4oc*) in *met22*Δ cells relative to that in wild type cells.

Chapter 5. PREDICTION OF THE DNA REPAIR ACTIVITY OF BRCA1 VARIANTS BY HIGH THROUGHPUT MUTAGENESIS

Chapter 5 appeared in this form in the journal *Genetics* (Starita *et al.*, 2015). My contributions were in the modeling and prediction of Homology-Directed DNA Repair (HDR) and in the processing, filtering, and scoring of the variants from the sequence data. I contributed to the creation of Figure 5.1, Figure 5.3, Figure 5.4, Figure S5.5, Figure S5.7, Figure S5.8, Figure S5.10, Figure S5.11.

5.1 ABSTRACT

Interpreting variants of uncertain significance (VUS) is a central challenge in medical genetics. One approach is to experimentally measure the functional consequences of VUS, but to date this approach has been *post hoc* and low throughput. Here we use massively parallel assays to measure the effects of nearly 2000 missense substitutions in the RING domain of BRCA1 on its E3 ubiquitin ligase activity and its binding to the BARD1 RING domain. From the resulting scores, we generate a model to predict the capacities of full-length BRCA1 variants to support homology-directed DNA repair, the essential role of BRCA1 in tumor suppression, and show that it outperforms widely used biological-effect prediction algorithms. We envision that massively parallel functional assays may facilitate the prospective interpretation of variants observed in clinical sequencing.

5.2 INTRODUCTION

In an era of increasingly widespread genetic testing, DNA sequencing identifies many missense substitutions with unknown effects on protein function and disease risk. In the absence of genetic evidence, experimental measurement is the most reliable way to determine the functional impact of a variant of uncertain significance (VUS). However, initiating an experiment for each new variant observed in a gene is often impractical. When experiments are done, they are nearly always performed in a retrospective manner (Bouwman *et al.*, 2013), such that the resulting data are not useful for the patient in whom the VUS was observed.

By prospectively measuring, in a high-throughput fashion, the consequences of all possible missense mutations on a gene's function, we can generate a look-up table for interpreting newly observed VUS. Although functional analysis at this scale is made possible by deep mutational scanning (Fowler

and Fields, 2014), a central challenge is that any single assay may not recapitulate all the activities of a given protein in human disease. To address this challenge, we hypothesized that integrating the results of assays of multiple biochemical functions would strengthen estimates of the effects of mutations on disease risk (strategy outlined in Figure 5.1A). As a proof-of-concept, we initiated massively parallel functional analysis of BRCA1, a protein for which there are multiple biochemical functions as well as known pathogenic and benign missense substitutions to benchmark results.

BRCA1 has been subject to intense study since its implication in hereditary, early onset breast and ovarian cancer (Miki *et al.*, 1994). All missense substitutions in BRCA1 that are known to be pathogenic occur in either the amino-terminal RING domain or the carboxy-terminal BRCT repeat (http://brca.iarc.fr/LOVD/home.php?select_db=BRCA1). Although the RING domain represents only 5% of the BRCA1 protein, 58% of the pathogenic missense substitutions occur within this domain. Sixty-two missense substitutions in the RING domain have been observed in patients, the general population, or tumor samples. Of these, only 22 have been classified—19 as pathogenic and 3 as benign (Supporting Information,

Table S)—by multifactorial models based on information from personal history, family history, and pathological profile and by A-GVGD (Tavtigian *et al.*, 2006), a conservation-based, biological-effect prediction algorithm (reviewed in Lindor *et al.*, 2012).

Although BRCA1 has multiple roles in the cell, its activity in homology-directed DNA repair (HDR) is most closely associated with cancer risk (Moynahan *et al.*, 1999; Towler *et al.*, 2013). Cell-based HDR rescue assays on the full-length BRCA1 protein have been performed for a small number of variants (Ransburgh *et al.*, 2010; Towler *et al.*, 2013). However, those assays are too laborious to be applied to each possible BRCA1 variant. We therefore sought to implement alternative BRCA1 functional assays that are more amenable to multiplexing.

The BRCA1 RING domain heterodimerizes with the RING domain of BARD1 to comprise an E3 ubiquitin ligase (Hashizume *et al.*, 2001). The structural stability of the heterodimer is required for the stability of full-length BRCA1 (Wu *et al.*, 2010). BRCA1 variants that cannot dimerize result in defects in HDR and loss-of-tumor suppression (Ransburgh *et al.*, 2010; Drost *et al.*, 2011). Assays for both BRCA1 E3 ligase activity and interaction with BARD1 are sensitive to amino acid substitutions that destabilize the structure of the heterodimer (Brzovic *et al.*, 2003; Morris *et al.*, 2006; Ransburgh *et al.*, 2010). We therefore developed massively parallel assays (Fowler *et al.*, 2010b) to measure the impact of thousands of missense substitutions on these two functions.

5.3 RESULTS

To assay E3 ligase activity, we subjected an allelic series (Kitzman *et al.*, 2015) (Figure S5.5) of the BRCA1 N terminus amino acids (2–304) to a phage display assay (Starita *et al.*, 2013) that selects for protein variants capable of autoubiquitination (Christensen *et al.*, 2007). We expressed BRCA1(2–304) variants on the surface of phage and selected for BRCA1 ubiquitination activity over five sequential rounds of selection in the presence of an E1, an E2, and Flag–ubiquitin by capturing phage with anti-Flag beads (Figure S5.6). Phages that encode active BRCA1 RING variants increase in abundance and those that encode inactive variants decrease in abundance over the multiple rounds of selection. We used deep sequencing to count each allele in the input phage population and after each round. We calculated E3 ligase scores by tracking the changes in the relative abundance of each allele during the selection (Araya *et al.*, 2012). The scores were normalized such that the wild type had a score of one and the mean score for variants with premature termination codons had a score of zero. We obtained scores for 5154 of the 5757 possible single-amino-acid substitutions (Table S5.2). Using an input frequency threshold (Figure S5.7A), we filtered these to a high-confidence set corresponding to 3881 amino acid substitutions, with the six replicates having Spearman's rank correlation values between 0.76 and 0.83 (Figure S5.7B).

E3 ligase activity for variants with missense substitutions ranged from completely nonfunctional (scores of zero) to nearly three times higher than wild type. Scores for residues in the RING domain (2–103) are shown in Figure 5.1B and for residues 104–304 in Figure S5.8; all scores are reported in Table S5.2. As expected, substitutions in the residues that coordinate zinc ions and the residues in loop 1 and the central helix that contact the E2 enzyme (Brzovic *et al.* 2003) were the most intolerant to mutation (Figure 5.1B; Wilcoxon rank sum test (WRST), $P = 0.0008$), with the exception of Phe46, where most substitutions were hyperactivating. We compared the E3 ligase scores to previous work by splitting the published activities of BRCA1 RING domain variants in *in vitro* ubiquitination assays (Brzovic *et al.*, 2003; Morris *et al.*, 2006) into three categories: completely nonfunctional, impaired, or wild-type like. E3 ligase scores corresponding to variants in the nonfunctional category were lower than those in the impaired category (WRST, $P = 1.4 \times 10–5$), which were in turn lower than those in the wild-type-like category (WRST, $P = 0.03$, Figure 5.1D).

In separate experiments, we used a multiplexed yeast two-hybrid assay to select for the ability of BRCA1 RING domain (2–103) (Brzovic *et al.*, 2001) variants to interact with the RING domain of BARD1. The DNA-binding domain of the yeast transcription factor Gal4 was fused to the BRCA1(2–304) allelic series and the Gal4 activation domain was fused to BARD1(26–126) (Figure S5.9). Here, BRCA1 binding to BARD1 drives the expression of a selectable reporter gene such that yeast expressing

BRCA1 variants that bind to BARD1 increase in abundance during the selection and those expressing nonfunctional variants decrease. We used deep sequencing to quantify the relative abundance of alleles after transformation into the yeast and at multiple time points during the selection (*Materials and Methods* and Table S5.2). We calculated a BARD1-binding score for 1855 of 1938 possible amino acid substitutions, excluding the carboxy-terminal 201 amino acids, which were required only for the autoubiquitination assay but not the BARD1-binding assay (Brzovic *et al.*, 2001). Using an input frequency threshold, we filtered these to a high-confidence subset corresponding to 1529 substitutions, whose scores were highly reproducible ($\rho = 0.82–0.95$, Figure S5.10 and Table S5.2).

Overall, BARD1-binding scores agreed with what is known about the RING–RING interaction. The residues that coordinate the zinc ions were the most intolerant to substitution with the exception of H41 (Brzovic *et al.*, 2001) (Figure 5.1C). The effect size for most other substitutions was small, which was expected given the large interface between the two RING domains (Brzovic *et al.*, 2001). We compared our results with those published for co-immunoprecipitation of BRCA1 RING domain variants with BARD1 (Brzovic *et al.*, 2003; Ransburgh *et al.*, 2010). While the scores from the yeast two-hybrid BARD1-binding assay were lower for BRCA1 variants reported not to bind to BARD1 (WRST, $P = 7.5 \times 10–7$), these scores spanned the entire range from zero to one (Figure 5.1E). Intermediate BARD1-binding scores for BRCA1 variants with weak or no BARD1 binding in co-immunoprecipitation assays may derive from differences in variant thermostability between the yeast assay (carried out at 30°) and the mammalian cell culture assay (carried out at 37°), and the *in vivo* transcriptional readout of the two-hybrid assay being more sensitive than co-immunoprecipitation.

We compared the E3 ligase scores to the BARD1-binding scores and observed that neither assay was sufficient alone to accurately discriminate BRCA1 variants with respect to their pathogenicity (Figure 5.1F, colored points). Because BARD1-binding is required for E3 ligase function, the scores from both assays were modestly correlated ($\rho = 0.386$; $P = 9.67 \times 10–56$), but many more positions were intolerant to substitutions in the E3 ligase assay (Figure 5.1F). Although the E3 ligase activity of BRCA1 may not be required for HDR and therefore tumor suppression (Reid *et al.*, 2008; Shakya *et al.*, 2011), the E3 ligase and BARD1-binding activities likely reflect the structural stability of the RING domain. Indeed both assays had some power to discriminate BRCA1 variants with respect to their pathogenicity (Figure 5.1F, colored points). We hypothesized that these two rich mutational data sets could be combined to accurately identify deleterious substitutions in the BRCA1 RING domain.

A test of whether the results from these high-throughput biochemical assays can be used to discriminate disease risk alleles needs “gold standards” as benchmarks. Since only 22 mutations in the BRCA1 RING domain have been classified for pathogenicity, we required a larger set of BRCA1 variants

with established, disease-relevant functional consequences. Therefore, we tested additional full-length BRCA1 variants in the assay that best correlates with tumor suppression: rescue of HDR at an induced double-strand break by expression of a BRCA1 variant following siRNA knockdown of endogenous BRCA1 (Figure 5.2A). We curated results from this assay (Ransburgh *et al.*, 2010; Towler *et al.*, 2013) for 17 missense substitutions in the BRCA1 RING domain and tested an additional 28 (Figure 5.2B) for a total of 45. Of the 19 known pathogenic mutants, 8 have now been tested for HDR rescue. As expected, after excluding R71G, a variant that affects BRCA1 splicing (Vega *et al.*, 2001), these pathogenic mutants all had low HDR rescue scores (mean = 0.19, max = 0.33) that separate them from the three known benign variants, which have much higher scores (mean = 0.88, min = 0.77; Figure 5.2B and Table S5.2). We defined a BRCA1 HDR rescue score of 0.53—the value midway between the average HDR rescue score for known pathogenic BRCA1 variants and the average score for known benign variants—as the inflection point for discriminating between functional and nonfunctional variants, as was done for BRCA2 (Guidugli *et al.*, 2014). With this inflection point, the HDR rescue assay has 100% sensitivity and 100% specificity.

We then asked whether models trained on the E3 ligase and BARD1-binding scores can predict the effects on HDR rescue of substitutions in the full-length protein. We evaluated the accuracy of several models using leave-one-out cross-validation (LOOCV), wherein we serially predicted HDR rescue scores for each of the 44 missense substitutions for which we had empirical HDR rescue and functional scores from models fit on the 43 remaining variants. We first compared the performance of models tested on scores from the E3 ligase and BARD1-binding assays alone or in combination. A linear model based on scores from both assays outperformed linear models based on scores from either assay alone (Figure 5.3, A and B). However, because we observed a nonlinear relationship between E3 ligase and BARD1-binding scores (Figure 5.1F), we tested whether nonlinear models would improve HDR prediction results. A support vector regression (SVR) model trained on E3 ligase scores and BARD1-binding scores yielded the best predictive power for HDR rescue (Figure 5.3C).

We next replaced our experimental data with computational predictions from several popular variant effect prediction algorithms (Grantham, 1974; Ng and Henikoff, 2003; Cooper *et al.*, 2005; Tavtigian *et al.*, 2006; Adzhubei *et al.*, 2010; Kircher *et al.*, 2014), which incorporate evolutionary constraints and/or chemical differences between amino acid side chains, and repeated the model training procedure. Individually, these prediction-based models performed poorly at predicting a substitution's effect on HDR (Figure 5.3D, white bars, and Figure S5.11). Although A-GVGD was the best performing algorithm, it yielded higher error and lower correlation than all experimentally-based models (Figure 5.4D and Figure S5.11F). Furthermore, when we added the A-GVGD predictions to the experimental data

and trained a hybrid model, performance was not enhanced over the experimentally based models (Figure 5.3D, gray bar, and Figure S5.11G). A plausible explanation for the comparatively poor performance of models trained on computational predictions is that they are largely based on features that are not specific to BRCA1 (*e.g.*, Grantham chemical difference scores) or on evolutionary constraint information that captures organismal fitness over evolutionary timescales, which may poorly discriminate subtle and strong molecular effects on BRCA1 function.

Because the SVR model based on combined functional data sets from the two assays was the most accurate, we used it to predict HDR scores for the 1287 BRCA1 RING domain missense variants with both high-confidence E3 ligase and BARD1-binding scores (Figure 5.4, A–C and Table S5.2), 1225 of which have not yet been reported in clinical sequencing. The distribution of predicted HDR scores is bimodal; 785 missense substitutions are predicted to have little effect on HDR, with predicted rescue scores >0.77 (Figure 5.4A). Conversely, 160 substitutions are predicted to be damaging to HDR, with scores <0.33; these variants would potentially increase the risk of hereditary breast and ovarian cancer. Based on this SVR modeling, only 342 variants have predicted scores in the indeterminate region between functional and nonfunctional.

As expected, predicted HDR scores for most of the 19 known pathogenic mutants in the BRCA1 RING domain are low (Figure 5.4B). Excluding pathogenic mutants known to affect splicing or used to train the model leaves 10 pathogenic mutants. All 10 have predicted HDR scores <0.53, the threshold for classifying a variant as functional. Nine have predicted HDR scores <0.33, the maximum empirical HDR rescue score for a known pathogenic mutant. Thus, our model demonstrates strong performance in predicting HDR activity of known variants (Figure 5.4B and Table S5.2). For 31 VUS identified in patients, predicted HDR scores range from near zero to wild-type-like, with 8 of 31 <0.53 and 5 <0.33, suggesting that a substantial fraction of individuals with VUS diagnoses may carry pathogenic BRCA1 alleles.

The data in Figure 5.4C represent a prospective map or look-up table for the effects of missense substitutions in the RING domain of BRCA1 on HDR function. This experimentally-derived map is more accurate than any map that could be generated using current computational tools. In terms of defining BRCA1 activity, these systematic mutational analyses uncovered positions in the four-helix bundle that show extreme sensitivity to substitution. For example, V11 does not tolerate substitutions with charged or amine-containing polar amino acids; M18 does not tolerate charged, polar, or aromatic substitutions; and F93 and D96 do not tolerate any substitutions. Our data support the idea some variants with defects in the E3 ligase activity are not compromised for HDR and tumor suppression (Reid *et al.*, 2008; Shakya *et al.*, 2011). The benign variants R7C and D67Y showed no binding defect with BARD1 and were able to

rescue HDR but they performed poorly in the E3 ligase selections. However, they may retain enough E3 ligase activity to satisfy a possibly low threshold of requisite activity.

Our results demonstrate the power of empirical measurements to assess the impact of missense variants on complex protein functions. Thus, we envision that massively parallel experiments to measure the effects of large numbers of substitutions will meet an urgent need in the clinical translation of genetic information.

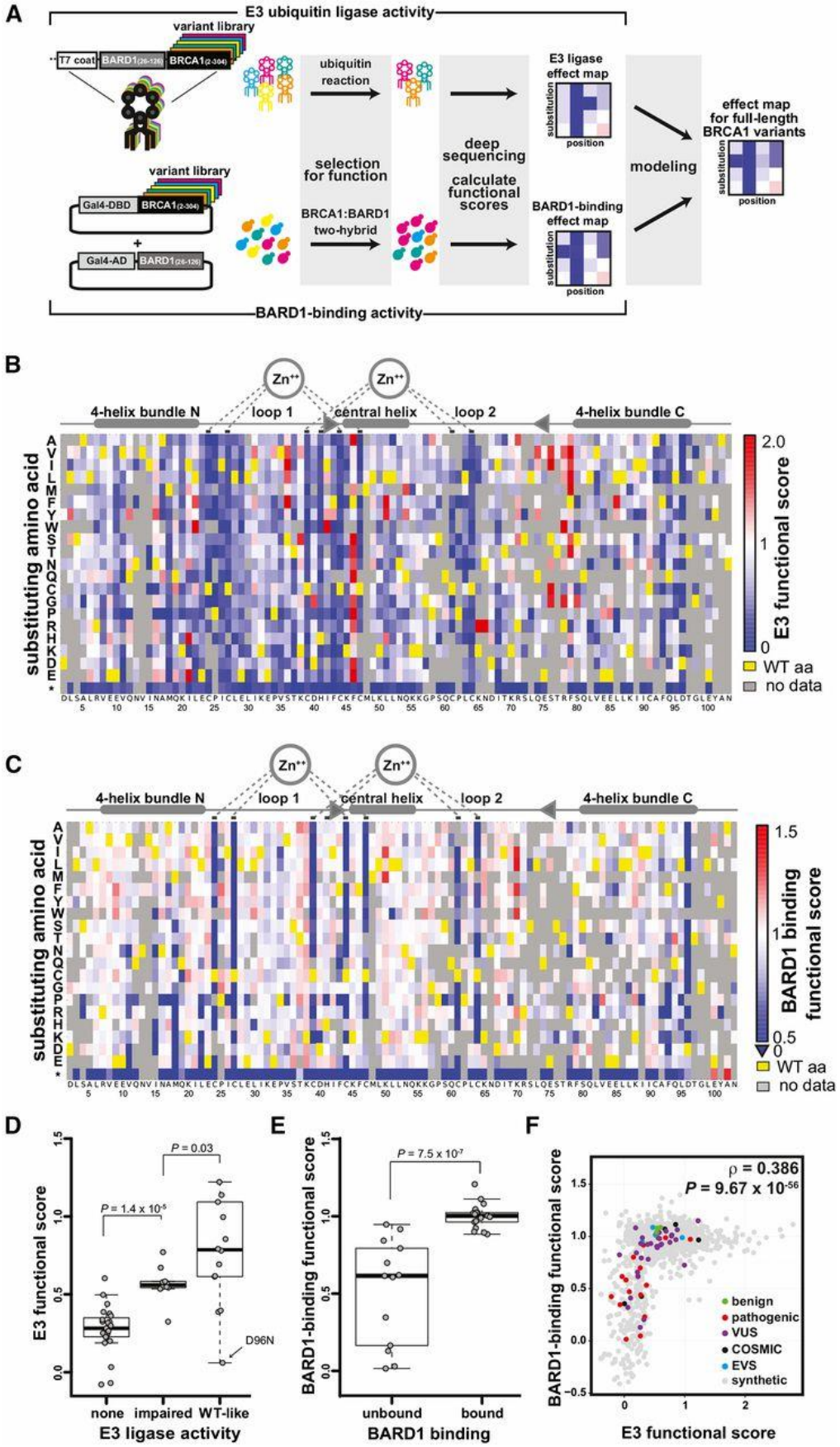

A
E3 ubiquitin ligase activity
variant library
T7 coat
BARD1(26-126)
BRCA1(2-304)
ubiquitin reaction
selection for function
deep sequencing
calculate functional scores
E3 ligase effect map
substitution
position
modeling
effect map for full-length BRCA1 variants
variant library
Gal4-DBD
BRCA1(2-304)
Gal4-AD
BARD1(26-126)
BRCA1:BARD1 two-hybrid
BARD1-binding effect map
BARD1-binding activity
B
Zn++
Zn++
4-helix bundle N
loop 1
central helix
loop 2
4-helix bundle C
substituting amino acid
E3 functional score
2.0
1
0
WT aa
no data
C
BARD1 binding functional score
1.5
1
0.5
0
D
E3 functional score
P = 1.4 x 10-5
P = 0.03
D96N
none
impaired
WT-like
E3 ligase activity
E
BARD1-binding functional score
P = 7.5 x 10-7
unbound
bound
BARD1 binding
F
ρ = 0.386
P = 9.67 x 10-56
benign
pathogenic
VUS
COSMIC
EVS
synthetic
E3 functional score

Figure 5.1 Deep mutational scans of BRCA1 for BARD-binding and Ubiquitin ligase activity

(A) Scheme for leveraging scores from parallelized assays for BRCA1 RING function into predictions for the function of the full-length BRCA1 protein in homology-directed DNA repair. (B-F) Scoring the E3 ligase and BARD1-binding activities of BRCA1 RING domain variants. (B) A sequence-function map of the effect of missense mutations in the BRCA1 RING domain on E3 ligase function. The functional score for each variant is the slope of the fit curve, normalized by setting stop codons to a score of 0 and the wild-type to a score of 1. Each position in BRCA1(2-103) is arranged along the x-axes, structural features of the RING domain are diagrammed above. The amino acid substitutions, grouped by side-chain properties, are on the y-axes. The E3 ligase scores range from improved activity versus wild-type (red), equivalent to wild-type (white), to less than wild-type (blue). Yellow represents the wild-type residue and gray missing or low confidence data. (C) A sequence-function map of the effect of missense mutations in the BRCA1 RING domain on BARD1-RING binding. Coloring as in panel B. (D) Comparison of the variant scores from the deep mutational scan for E3 ligase activity versus literature-reported E3 ligase activities for the same BRCA1 variants (Brzovic *et al.*, 2003; Morris *et al.*, 2006). The Wilcoxon rank sum test (WRST) was used to test for significant differences between the categories. The biggest outlier in the wild type-like category, D96N, not only performed poorly as an E3 ligase score but also failed to bind to BARD1 and to support homology-directed repair in cells (Table S5.2). (E) Comparison of BARD1-binding scores from the two-hybrid experiment versus literature-reported BARD1 binding by the same BRCA1 variants (Brzovic *et al.*, 2003; Ransburgh *et al.*, 2010). The WRST was used to test for significant differences between categories. (F) The relationship between the quality-filtered E3 ligase functional scores and the BARD1-binding scores. Colors indicate the clinical classification or database of origin for each variant.

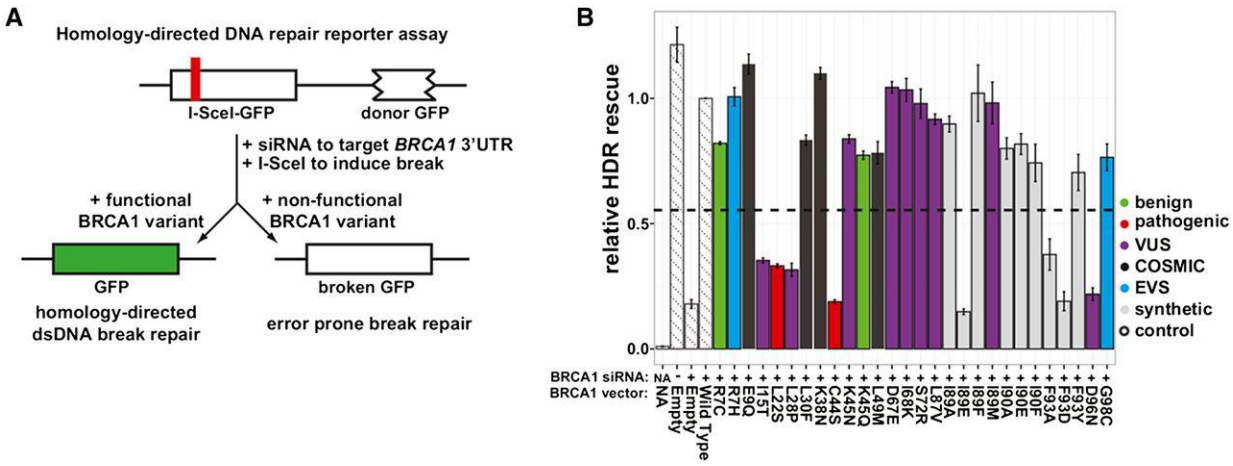


Figure 5.2 Testing BRCA1 variants for their ability to rescue homology-directed DNA repair.

(A) Integrated into the genome of the HDR reporter strain is one copy of the GFP gene containing an I-*Sce*I homing endonuclease site that introduces an in-frame stop codon, along with another copy of the GFP gene lacking both its start and stop codons that functions as a donor for DNA repair (Pierce *et al.* 2001). The cell line is depleted for BRCA1 by transfection with an siRNA that targets the 3′-UTR of the endogenous gene. I-*Sce*I and a full-length variant of BRCA1 are then transfected into the cells. After 3 days, the GFP+ population is assessed by flow cytometry. (B) The wild-type normalized percentage of cells that were GFP+ in the BRCA1 HDR rescue assay is shown. Experiments were performed in triplicate and error bars represent the standard error. siRNA, BRCA1 variant, and clinical classification or database of origin is indicated by color. Dashed horizontal line represents the midpoint between the average HDR scores for known pathogenic and benign variants.

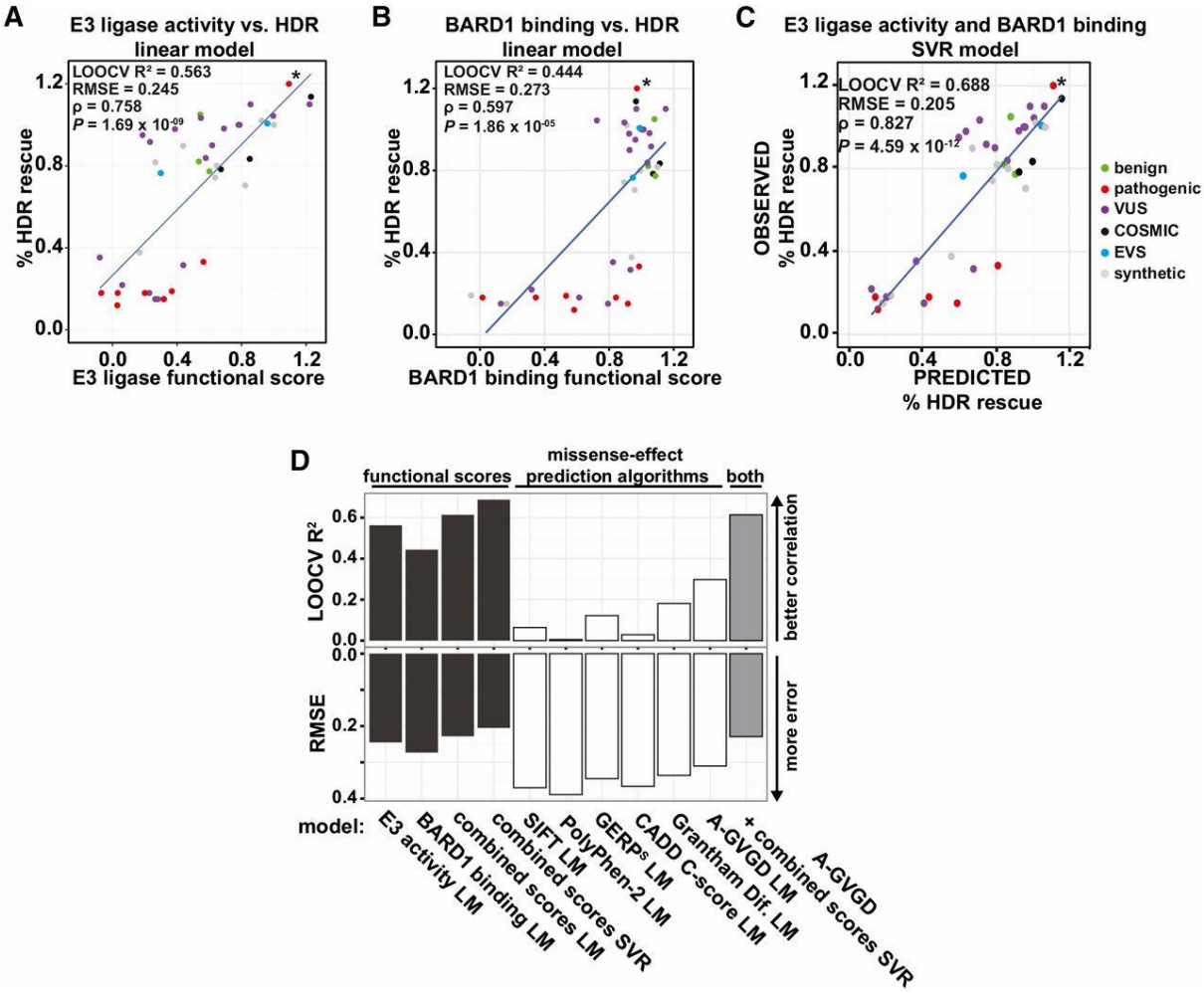


Figure 5.3 Scores from massively parallel E3 ligase and BARD1-binding assays on BRCA1 RING domain variants are better predictors of the HDR activity of the full-length protein.

The linear relationship of the E3 ligase scores (A), BARD1-binding scores (B), and HDR scores. (C) A support vector regression (SVR) model of HDR rescue scores from the combination of the E3 ligase and BARD1 binding functional scores. Variants are colored by database of origin. The blue line represents the least-squares fit of the displayed data. Known pathogenic splice variant R71G is marked with an asterisk. (D) Experimentally or computationally derived values for the effect of missense variants on protein function were used to predict the effect on HDR. The LOOCV *R*2 and RMSE for each model is indicated. The RMSE of LOOCV indicates the average distance between the HDR rescue predictions and the true HDR rescue scores, and the LOOCV *R*2 is the overall correlation between predicted and observed values; low RMSE and high *R*2 indicate better predictive power. For A-GVGD, the Grantham deviation value was used. Source of HDR predictions is indicated by color, linear model (LM), and SVR.

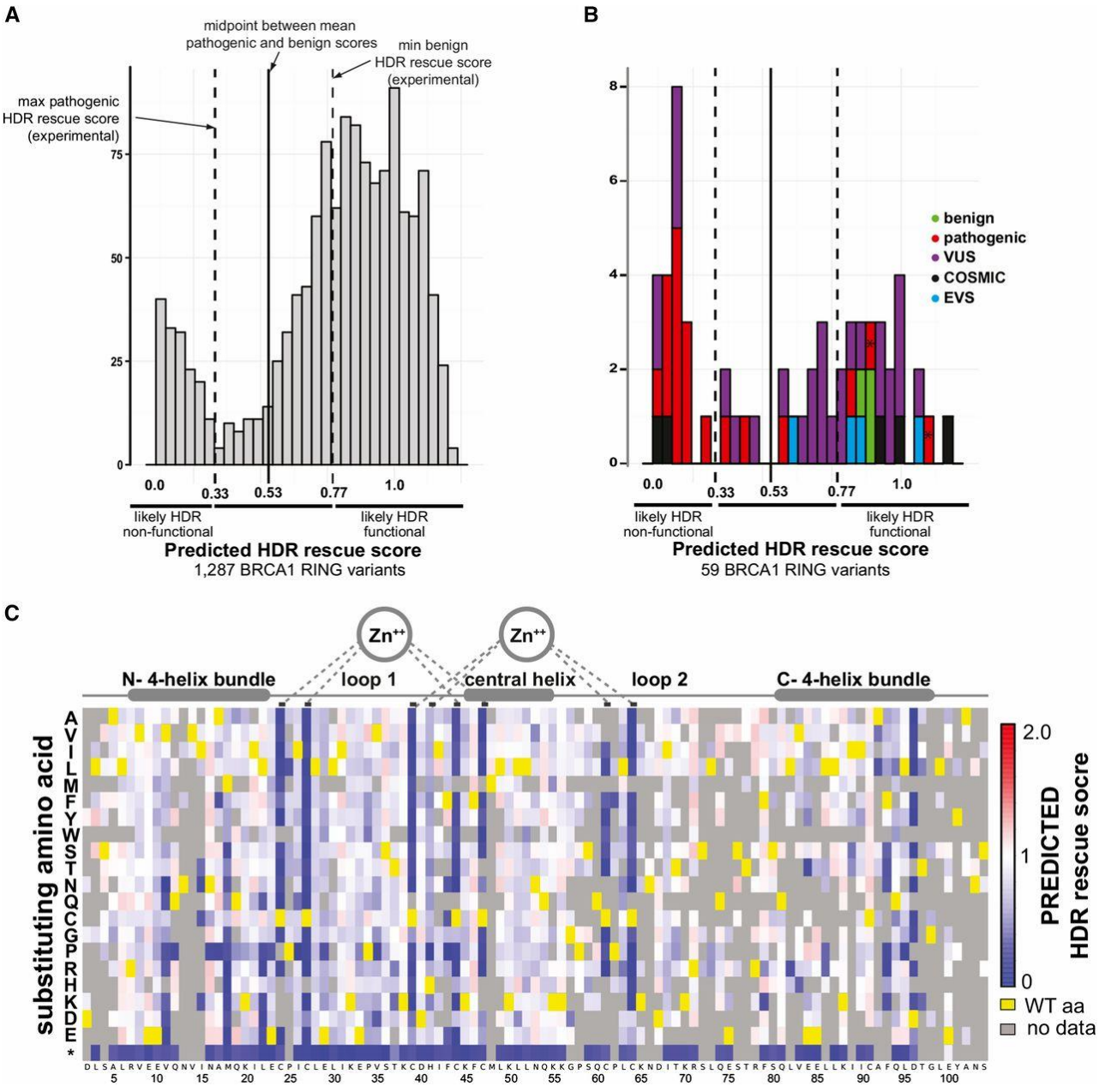


Figure 5.4 Predicted HDR rescue scores for 1287 BRCA1 RING variants create a prospective map of the effect of missense substitutions.

(A) A histogram of the predicted HDR scores for the 1287 BRCA1 RING variants with both high-confidence E3 ligase and BARD1-binding scores. (B) A histogram of the 59 of the 62 BRCA1 RING variants found in the human population, clinical classification, or database of origin is indicated by color. Known splice variants R71G and R71K are marked with an asterisk. (C) Sequence-function map of the predicted HDR rescue scores for BRCA1 RING variants. Colors are centered on 1.0 as wild type (white). Structural features of the BRCA1 RING domain are diagrammed above.

5.4 MATERIALS AND METHODS

5.4.1 *BRCA1(2-304) single codon substitution library construction by the Programmed Allelic Series method*

Oligonucleotides, 90-mers, to direct the single codon mutagenesis of BRCA1 were synthesized on and released from a 12,000-feature array by Custom Array (Bothell, WA) (example in Table S5.3, BRCA1_00284.0). The BARD1(26–126)–GS–BRCA1(2–304) fusion open reading frame (ORF) (Christensen *et al.*, 2007) was moved to the pGEM vector and the *Eco*RI site in BRCA1 was destroyed. This fragment of BRCA1 was used as a template for PALS mutagenesis (Kitzman *et al.*, 2015). Sixteen base random barcodes (16N) were added 3′ of the stop codon in the final PCR step. The ligation of the final mutagenized and barcoded amplicon was transformed into DH10B electromax cells (NEB) and yielded 250,000 unique transformants of the pGEM_BARD1_GS_BRCA1-var_barcode library.

5.4.2 *Subassembly to match 16N barcodes to BRCA1 variants*

Since BRCA1(2–304) is too long to be sequenced in one pass by current Illumina technology, we developed a method to create randomly shortened contigs that could be grouped by barcode to use in an assembly method call tag-directed read grouping or subassembly (Hiatt *et al.*, 2010). A total of 5 μg of the plasmid pGEM_BARD1_GS_BRCA1-var was cut at the 5′ end of the BRCA1 ORF with *Bam*HI and purified. The purified DNA was digested using the double-strand exonuclease Bal-31 (NEB), 1 unit Bal31 per 1.6 pmol DNA ends at 30°. Aliquots were taken at 0, 3, 7, 11, 13, and 15 min and placed in the DNA-binding buffer from the Zymo clean and concentrate kit to stop the reactions. One-fifth of the digested and cleaned DNA was cut with *Hin*dIII and examined by PAGE to determine the degree of digestion. DNA from all time points was pooled and treated with End-it (Epicentre) to blunt and phosphorylate the ends. Blunt-ended, cleaned DNA was A-tailed using goTaq (NEB) and cleaned again. A double-stranded linker containing the Tru-seq Illumina Read 2 primer was ligated onto the A-tailed DNA (W-E4B-subassembly-linker and phosphorylated C-E4B-subassembly linker). The linkered and cleaned DNA was cut with *Sac*I (NEB) to separate the ORF and barcode from the rest of the plasmid. Primers that annealed to the linker and plasmid DNA directly 3′ of the barcode that contain the p5 and p7 Illumina cluster generating sequences (newBRCA1-side_R_CG1 and BRCA1-side_R_CG2) were used to amplify the fragments and barcode for Illumina sequencing in reactions containing the high-fidelity polymerase KAPA HFHSRM and SYBR green [conditions: 95° 3:00 (98° 0:20, 63° 0:15, 72° 1:50) × 12–15]. The amplicons were sequenced on an Illumina HiSeq2000 in paired-end, 2 × 101 run mode and with an Illumina MiSeq paired-end, 2 × 250 kit.

Reads were filtered for quality and grouped by the sequence of the 16-base barcode. A Smith–Waterman algorithm was used to align the grouped reads to the wild-type BRCA1(2–304) sequence and a consensus sequence was determined for each barcode group as in (Hiatt *et al.*, 2010) and (Patwardhan *et al.*, 2012). A minimum quality score of 20 was required for a base to contribute to an assembly. Full-length BRCA1(2–304) sequences were filtered for quality by requiring that a given base in the assembly was observed at least twice and was present at an intra-tag group allele frequency of one for positions covered by two to four reads or a frequency of at least 0.8 for positions covered by five or more reads. If these conditions were not met the assembly was discarded. We assembled 128,237 barcoded variants, of which 60,256 corresponded to 5156 single-amino-acid changes out of the possible 5757 (89% of the 19 substitutions × 303 codons) in BRCA1(2–304) (see Figure S5.5). A database of barcodes and their associated full-length BRCA1(2–304) assembly was created to facilitate linking barcodes sequenced from the experimental samples to the full-length subassemblies.

5.4.3 *Phage-based E3 ligase assays*

The BARD1(26–126)_glycine–serine linker_BRCA1(2–304) library was subcloned from pGEM_BARD1_GS_BRCA1-var_barcode by cutting and gel purifying the *Eco*RI and *Hin*dIII fragment and ligating into the genome of T7–Select 10-3b bacteriophage. Phage genomic DNA was packaged into phage particles, the number of ligation/packaging events was estimated by titer as 2.56 × 107 plaque-forming units (PFU), and the phages were amplified in *E. coli* strain BLT5403 according to the T7Select Cloning Kit instructions (EMD Millipore). The selections for functional BRCA1(2–304) phages were performed as in (Starita *et al.*, 2013) with these differences: amplified phage were never stored more than 24 hr before a sequential round of selection and the 50-µl ubiquitination reactions contained 20 µl (~1 × 107 PFU) of amplified phage, 2 mM ATP, 5 mM MgCl2, 1 µM wheat E1 ubiquitin activating enzyme, 4 µM UBE2D3 (UbcH5c), and 8 µM Flag-tagged ubiquitin.

DNA from the initial amplified phage population and amplified phage from each replicate from each of five rounds of selection was purified from 200 µl of lysate by phenol chloroform extraction. Barcodes were PCR amplified in two sequential reactions. The first reaction contained primers jkA0390_BBcplxcheckF and E4B-index01-8_CG-R or T7_barcodes_common primer_R and 200 ng of phage DNA in reactions containing the high-fidelity polymerase Phusion (NEB), 2 mM added MgCl2, and SYBR green [conditions: 95° 3:00 (98° 0:20, 63° 0:15, 72° 1:50) × 10–13]. Reaction products were monitored by qPCR and removed during exponential amplification. The first reactions were purified using the Zymo clean and concentrate kit. One-tenth of that product was amplified with JK19 and one of the index containing primers E4B-index01-8_CG-R or common_index_primers such as NexV2ad2_A1

[conditions: 95° 3:00 (98° 0:20, 63° 0:15, 72° 1:50) × 4–6]. Again, reaction products were monitored by qPCR and removed during exponential amplification. Reaction products were treated with exonuclease I (Affymetrix) for 15 min at 37° then purified using the Zymo clean and concentrate kit. Samples were multiplexed and sequenced using primer jkA0390_BBcplxcheckF on an Illumina GAIIx or HiSeq2000 in single end mode.

5.4.4 *Yeast two-hybrid-based deep mutational scan for BRCA1-BARD1 binding*

The Gal4 DNA-binding domain (Gal4DBD) was amplified from pOBD2 (Cagney *et al.*, 2000) using primers *Spe*I_Gal4DBD_F and *Spe*I_Gal4DBD_R and cloned into p414–ADH (Mumberg *et al.*, 1995). The BRCA1(2–304) variant library was excised from pGEM_BARD1_GS_BRCA1-var_barcode library by digestion with *Bam*HI (NEB) and *Pst*I (NEB) and ligated into p414_Gal4DBD to create p414_Gal4DBD_BRCA1_var_barcode, yielding ∼1.16 × 105 total transformants. BARD1(26–126) was amplified from pGEM_BARD1_GS_BRCA1 using primers *Eco*RI_BARD1_Ln_F and *Nco*I_BARD1_Stop_R and cloned into pOAD (Cagney *et al.*, 2000) containing the Gal4 transcriptional activation domain.

The *S. cerevisiae* strain, PJ69a (James *et al.*, 1996), containing pOAD_BARD1 was transformed (Melamed *et al.*, 2013) with the p414_Gal4DBD_BRCA1_var_barcode library with a yield of ∼1.26 × 106 total transformants. Transformed yeast were transferred to 40 ml SD–Leu–Trp, cultured overnight and stored in 6.7 optical density units (ODU) aliquots at −80°.

Two independent experiments (A and B) were performed, each consisting of three independent selections: 12.5 ODU (A) or four ODU (B) of cells were collected from each culture at each time point for sequencing. Each experiment began by culturing one frozen aliquot of PJ69a transformed with pOAD_BARD1 and p414_Gal4DBD_BRCA1_var_barcode in SD–Leu–Trp to logarithmic phase (A, 1.01 OD/ml; B, 0.83 OD/ml). Cells from this input population were collected for sequencing and for back dilution into the selection medium (SD–His–Leu–Trp + 10 mM 3-amino-1,2,4 triazole (Sigma), A, 5 OD to 200 ml; B, 2 OD to 100 ml) in triplicate. Each replicate was cultured to logarithmic phase (A, 18 hr, 1.1 OD/ml; B, 16 hr, 0.7 OD/ml) after which cells were collected for sequencing and back diluted into fresh selection medium (A, 1 OD to 100 ml; B, 0.6 OD to 100 ml). Each replicate was again cultured to logarithmic phase (A, 37 hr, 0.62 OD/ml; B, 40.5 hr, 0.67 OD/ml), after which cells were collected for sequencing and back diluted into fresh selection medium (A, 12.5 OD to 125 ml; B, 1.1 OD to 100 ml). Each replicate was again cultured to logarithmic phase (A, 45 hr, 0.43 OD/ml; B, 64 hr, 1.4 OD/ml) and the final time point was collected.

Plasmid DNA was isolated from the input and samples collected during the selection for growth in the -histidine media using a Zymoprep Yeast Plasmid Miniprep II kit (Zymo Research). Sequencing amplicons were prepared individually for each sample by two successive PCR reactions using Phusion high-fidelity DNA polymerase. In the first reaction, primers jkA0390_BBcplxcheckF and BRCA1-Y2H_commonLinker_R were used to amplify the barcoded region of half of the prepared p414_Gal4DBD_BRCA1_var_barcode plasmid. Of the first reaction, 4% was amplified with primers JK19 and NexV2ad2_XX to append Illumina cluster generating sequences and a unique sample index sequence. Reaction products of all PCR reactions were monitored on a mini-opticon qPCR machine (Bio-Rad) and removed during exponential amplification. Samples were purified, multiplexed, and sequenced using primer jkA0390_BBcplxcheckF on a HiSeq2000 in single-end mode.

5.4.5 *Slope calculations and normalization*

The Illumina reads that matched to barcodes associated with full-length assemblies were retained and unmatched barcodes were discarded. The matched barcodes were converted to the sequence of the full length BRCA1(2–304) assembly and the Enrich software package (Fowler *et al.*, 2011) was used to determine locations and identity of substitutions and to tally the number of times each variant appeared in each population (Table S5.2).

Sequencing read counts corresponding to a given variant were equal to the sum of read counts from all barcodes matching that variant. For each time point, frequencies were calculated for all variants as the variant's read count divided by the sum of all read counts at that time point. Variants that dropped out (cannot be found in the selected populations) had their frequencies set to the lowest possible frequency at that time point. Ratios were calculated as the variant's frequency in the selected time point divided by its frequency in the input library. For each variant, a linear model was fit by least squares to the log ratios over time using numpy.polyfit. The inverse log of this slope corresponds to the percentage change in frequency per unit time. To obtain a normalized score, the average inverse log of the slope for all stop codons was subtracted from all inverse-log-slopes so that stop codons, on average, have a score of 0. These 0 centered values were then divided by the wild-type (WT) inverse-log-slope so that a score of 1 corresponds to WT function.

The normalized score for variant *i* is

$$\frac{2^{Slope_i} + \sum_0^m 2^{Slope_m}}{2^{Slope_{WT}}} \tag{5.1}$$

where *m* is the number of stop codons from positions 2–103 and all slopes were fit to the log ratios at each time point. A conservative estimate of the standard deviation of the slopes was generated using a Loess curve to model the relationship between frequency in the input population and variance across all replicates (based on the assumption that the variance is related to the frequency in the input population; see Figure S5.7 and Figure S5.10). For each variant, the conservative variance was set to whichever was larger: the variance across all replicates or the value of the Loess curve evaluated at the number of input reads for that variant. This estimate was used to generate the reported confidence intervals (Table S5.2). Additionally, we used cutoff based on the number of input reads to determine the high-confidence data set that would be used for the final HDR predictions. The heuristic to determine the cutoff is described in Figure S5.7 and Figure S5.10. HDR predictions were made only for variants with high-confidence scores in both the E3 ligase and BARD1-binding assays. Finally, a permutation test was used to compare each variant's slopes to the WT slopes across all replicates. The average difference between paired slopes was used as the test statistic and 10,000 permutations were performed for each variant (Table S5.2).

5.4.6 *Full-length BRCA1 variant construction and HDR assays*

Mutations in the BRCA1 RING domain were made by overlap-extension PCR and subcloned into the *Hin*dIII and *Eco*RI sites of pcDNA3–HA–BRCA1 (plasmid described in (Chiba and Parvin, 2001)). All constructs were verified by Sanger sequencing. BRCA1 rescue of HDR assays were performed in triplicate as in (Ransburgh *et al.*, 2010). All BRCA1 HDR rescue scores are normalized to that of the wild-type protein at HDR rescue = 1. The maximum HDR score for known pathogenic variant of BRCA1 is 0.33. Seven pathogenic variants (excluding splice variant R71G) have been tested for HDR rescue with a mean score of 0.19. Of the only three known benign BRCA1 RING domain variants, all have been tested for HDR rescue and have a minimum score of 0.77 with an average score 0.88. We defined a BRCA1 HDR rescue score of 0.53—the value midway between the average HDR rescue scores for known pathogenic BRCA1 variants and the average scores for known benign variants—as the inflection point for discriminating between functional and nonfunctional variants, as was done for BRCA2 (Guidugli *et al.*, 2014).

5.4.7 *HDR prediction model building and testing*

We obtained SIFT, Polyphen-2, GERP++, and CADD values from the CADD database (Kircher

et al., 2014) and references therein (http://cadd.gs.washington.edu/download). For every possible amino acid substitution in BRCA1 (2-103), we obtained Grantham chemical difference values from (Grantham, 1974), and GVGD values from the A-GVGD BRCA1 web-tool (http://agvgd.iarc.fr/agvgd_input.php). Grantham deviation (GD) values were used to predict HDR rescue scores.

All models were fit and cross-validated using the *R* package caret (Kuhn, 2008). Linear models were fit by least squares. Support vector regression models used the radial basis function kernel and were validated using a nested cross validation scheme (Cawley and Talbot, 2010). Briefly, for each step of the LOOCV, an inner LOOCV loop was used to determine model performance on each *C* and sigma pair in the tested parameter space and the best performing model (based on root mean square error, RMSE) was used to predict the holdout in the outer loop. The range of sigma values tested in the inner loop was determined using the sigest function from the *R* package kernlab and the *C* values tested were 0.1, 1, 2, 5, 10, 100, and 1000. The final model used for HDR predictions was chosen by picking the parameter pair with the lowest average RMSE across all iterations of the outer loop (Y2H and E3 model—$C = 5$ and sigma = 0.1633448, Y2H, E3; GVGD model—$C = 5$ and sigma= 0.08220825).

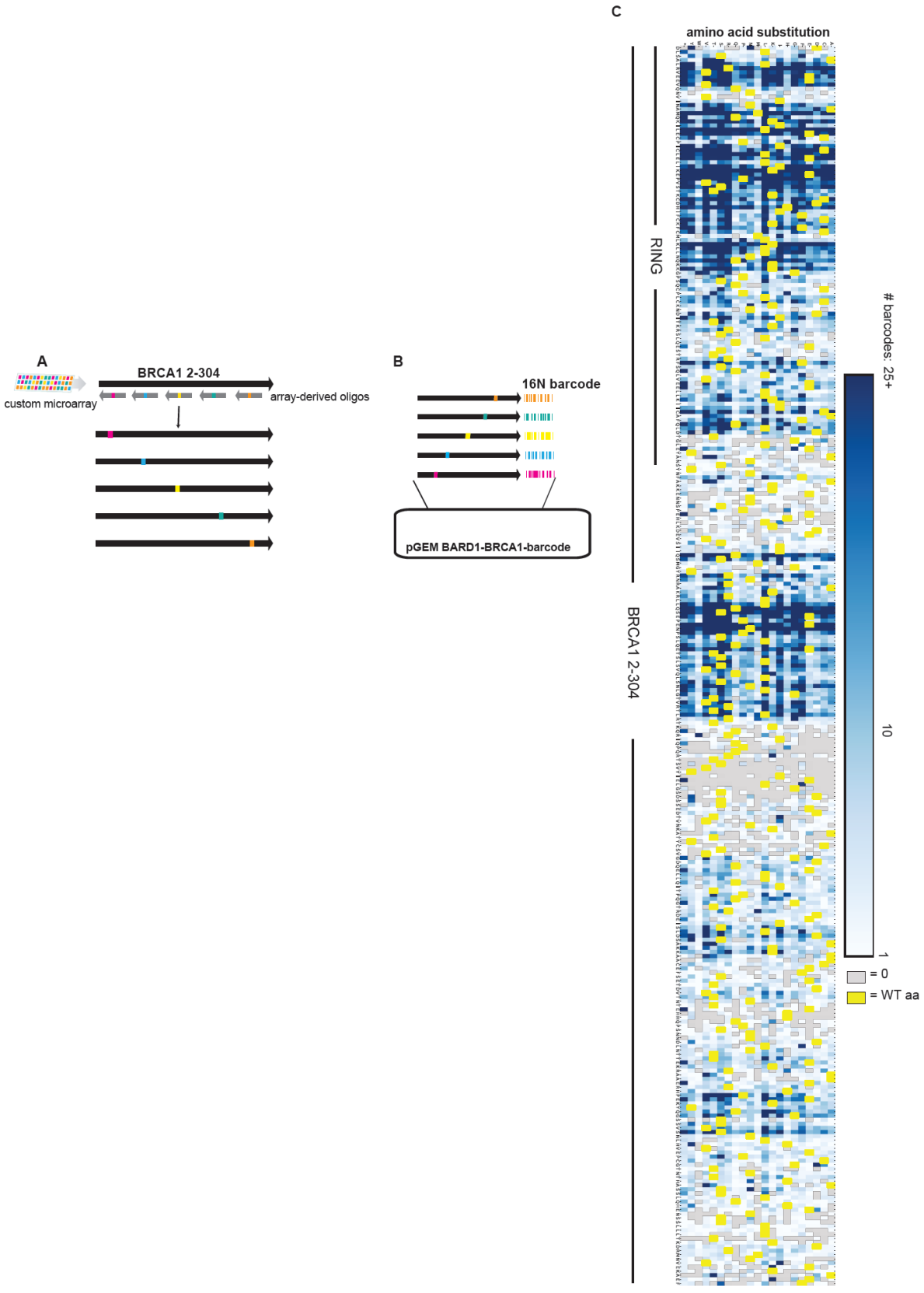

A
BRCA1 2-304
custom microarray
array-derived oligos
B
16N barcode
pGEM BARD1-BRCA1-barcode
C
amino acid substitution
RING
BRCA1 2-304
barcodes: 25+
10
1
= 0
= WT aa

Figure S5.5 Construction of the BRCA1(2-304) allelic series.

(A) We created an allelic series of variants within BRCA1(2-304) with single amino acid substitutions by the method known as Programmed Allelic Series (Kitzman et al. 2015), which uses mutagenic oligonucleotides synthesized on a programmed microarray to create a pool of variants with single codon changes by overlap-extension PCR. (B) Each variant was barcoded with a random 16-nucleotide tag that we associated with the mutation present in the BRCA1 domain. We assembled 128,237 barcoded variants, of which 60,256 corresponded to 5,156 single amino acid changes out of the possible 5,757 (89% of the 19 substitutions x 303 codons) in BRCA1(2-304). (C) The number of barcodes per assembled BRCA1(2-304) variant is represented in a heatmap. Shades of blue represent the number of barcodes per variant with the maximum color fill set to 25 barcodes. There were many variants that had more than 25 barcodes. Yellow represents wild-type residues and gray potential variants for which there was not a full length BRCA1(2-304) assembly.

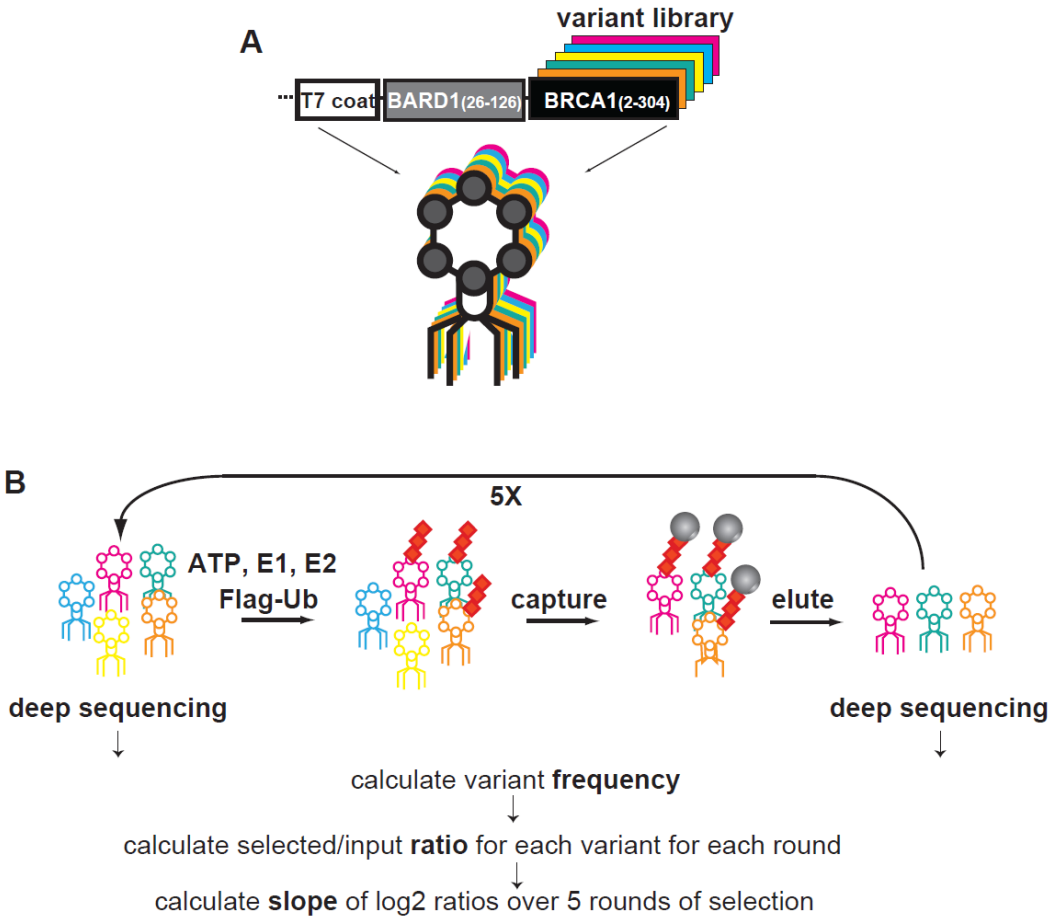


Figure S5.6 Scoring the effects of missense mutation on the E3 ligase activity of the BRCA1 RING domain.

(A) A fusion protein of BARD1(26-126) and BRCA1(2-304) is an active E3 ligase and capable of autoubiquitination in vitro. The allelic series of BARD1(26-126) - BRCA1(2-304) was expressed at the carboxy-terminus of the coat protein of bacteriophage T7. Residues 2-103 are the structured RING domain and lysine residues within 104-304 are required for autoubiquitination. (B) A phage population displaying the library of BRCA1 variants was incubated in ubiquitination reactions (purified E1, E2 (UbcH5c), Flag-tagged ubiquitin and ATP), in triplicate in two separate experiments. Phages encoding active variants of BRCA1 became ubiquitinated and were collected on anti-Flag beads. After washing, elution by competition with Flag peptide and re-amplification in E. coli, phages were used in the next round of selection. Phage DNA was extracted after each of five sequential rounds of selection and the barcodes were amplified by PCR and sequenced. Barcodes were tallied by single end Illumina sequencing. After converting the barcodes to BRCA1(2-304) variants, we calculated the frequency of each variant in the input and selected populations. For each of the five rounds of selection, we fit a linear curve to the log ratio of the frequency of each variant divided by its frequency in the input population for each of the six replicates. The functional score for each variant is the slope of the fit curve, normalized by setting stop codons to a score of 0 and the wild-type to a score of 1.

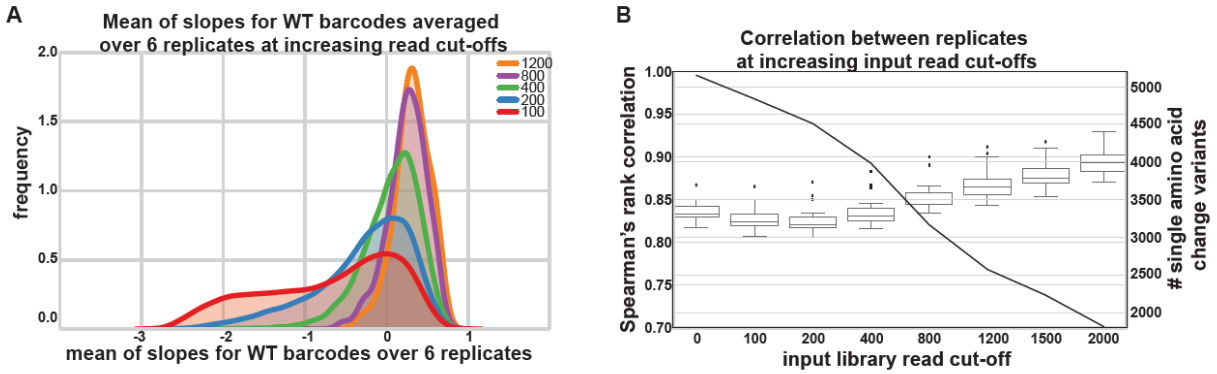


Figure S5.7 Heuristic for filtering high-confidence data set.

(A) The distribution of the log transformed slopes of the nearly 30,000 barcodes (Figure S5.5) associated with wild-type BRCA1(2-304) sequences (input read cut-offs represented by color). The poor scoring wild-type variants are thought to be due to loss of individual barcodes that follows a Poisson distribution due to experimental bottlenecks. (B) The 800 input read count cut-off maximizes the number of variants contributing to the analysis (black line) while maintaining the maximum Spearman's rank correlation between the six experimental replicates and minimizing barcode dropout due to bottlenecks (A). Estimates of variance and 95% confidence intervals can be found for each measurement in Table S5.2.

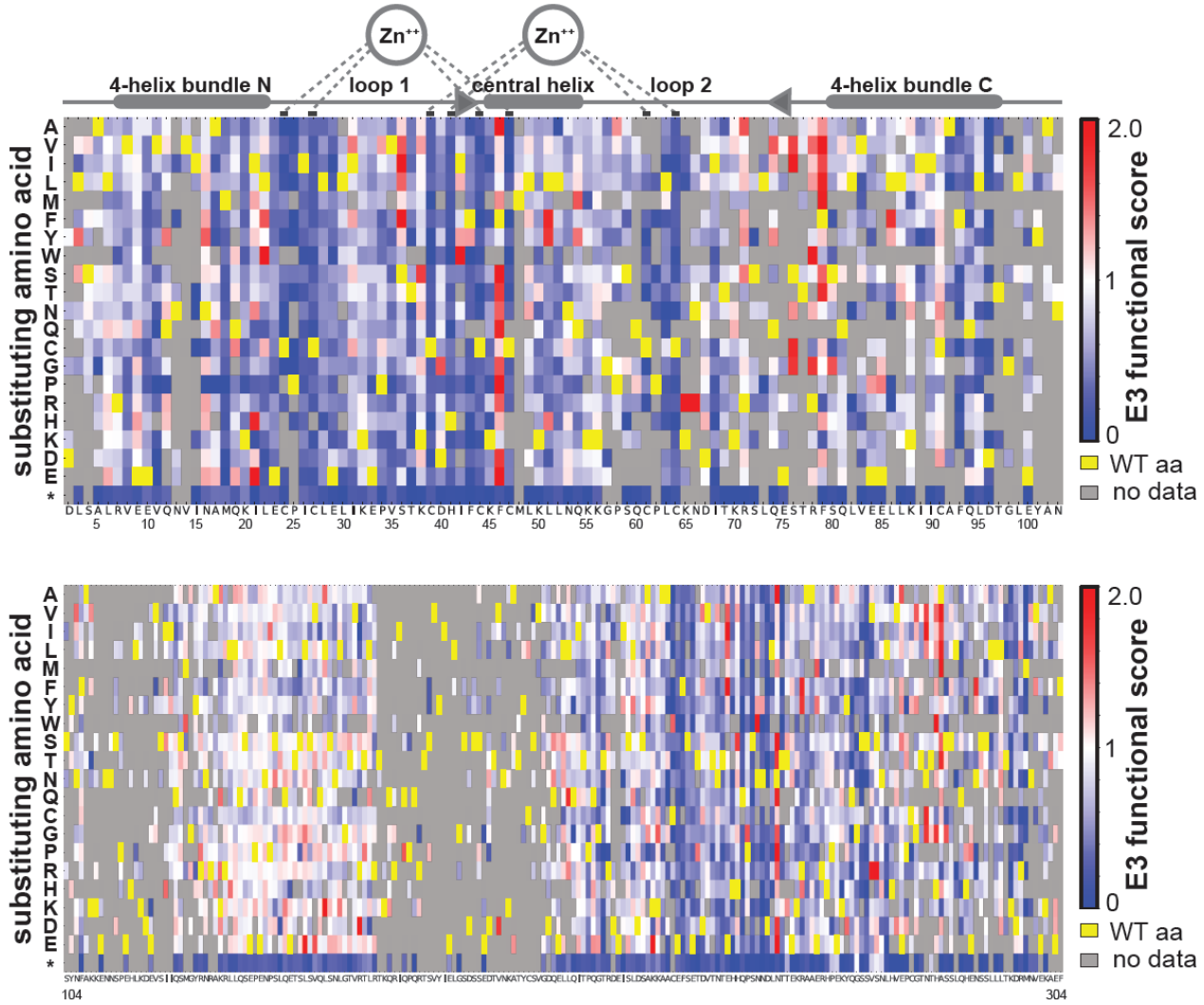


Figure S5.8 Sequence - function map of the effect of missense substitutions on E3 ligase function.

The functional score for each variant is the slope of the fit curve, normalized by setting stop codons to a score of 0 and the wild-type to a score of 1. Each position in BRCA1(2-304) is arranged along the x-axes, structural features of the RING domain are diagrammed above. The amino acid substitutions, grouped by side-chain properties, are on the y-axes. The E3 ligase scores range from improved activity versus wild-type (red), equivalent to wild-type (white), to less than wild-type (blue). Yellow represents the wild-type residue and gray missing or low confidence data.

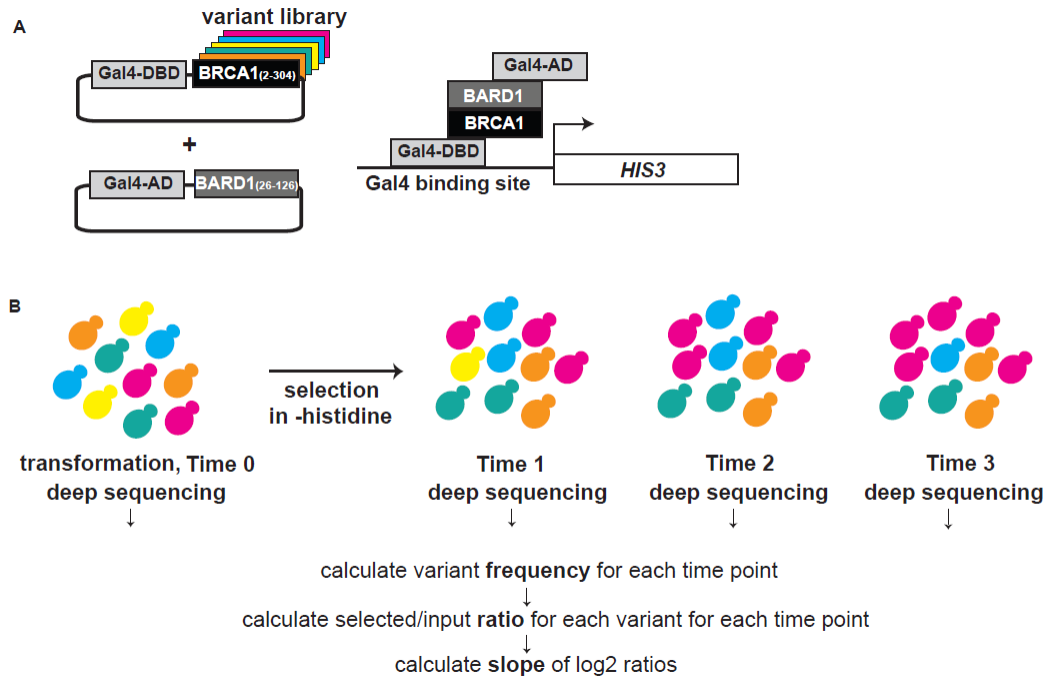


Figure S5.9 Diagram of the yeast-two-hybrid selection scheme to measure BRCA1-BARD1 binding.

(A) The barcoded allelic series of BRCA1(2-304) was fused to the carboxy-terminus of the Gal4 DNA-binding domain, and the BARD1(26-126) domain was fused to the carboxy-terminus of the Gal4 activation domain. Yeast harboring BRCA1 variants that bind to BARD1 drive the expression of the HIS3 reporter gene and therefore grow in media lacking histidine. (B) The two-hybrid reporter strain transformed with the plasmids encoding the BRCA1 allelic series and BARD1 was selected in triplicate in two separate experiments in media lacking histidine and containing 10 mM 3-amino-1,2,4-triazole (3-AT), a competitive inhibitor of the His3 enzyme. At mid-log phase, aliquots of the cultures were sampled and then back-diluted into fresh selective media and grown to mid-log phase two additional times. The BRCA1 plasmids were extracted at each of three time points and their barcodes were PCR amplified and sequenced. The barcodes associated with BRCA1(2-304) plasmids prepped from the yeast after each of the three time points of selection and the input population were deeply sequenced. We fit a linear curve to the log ratio of the frequency of each variant in the selected populations divided by its frequency in the input population and calculated the slope of that line, normalized again to stop codons (set to 0) and wild-type (set to 1).

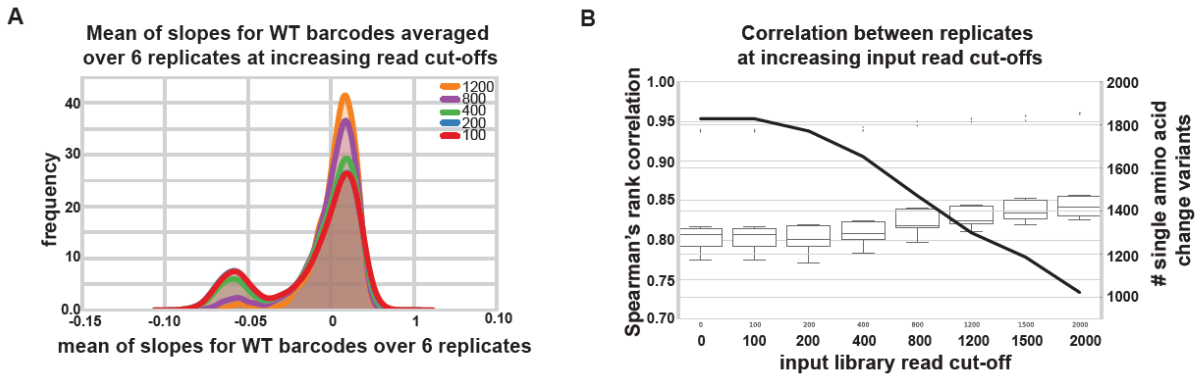


Figure S5.10 Heuristic for filtering high-confidence data set.

(A) The distribution of the log transformed slopes of the nearly 30,000 barcodes (Figure S5.5) associated with wild-type BRCA1(2-304) sequences (input read cut-offs represented by color). The poor scoring wild-type variants are thought to be due to loss of individual barcodes that follows a Poisson distribution due to experimental bottlenecks. (B) The 800 input read count cut-off maximizes the number of variants contributing to the analysis (black line) while maintaining the maximum Spearman's rank correlation between the six experimental replicates and minimizing barcode dropout due to bottlenecks (A). Estimates of variance and 95% confidence intervals can be found for each measurement in Table S5.2.

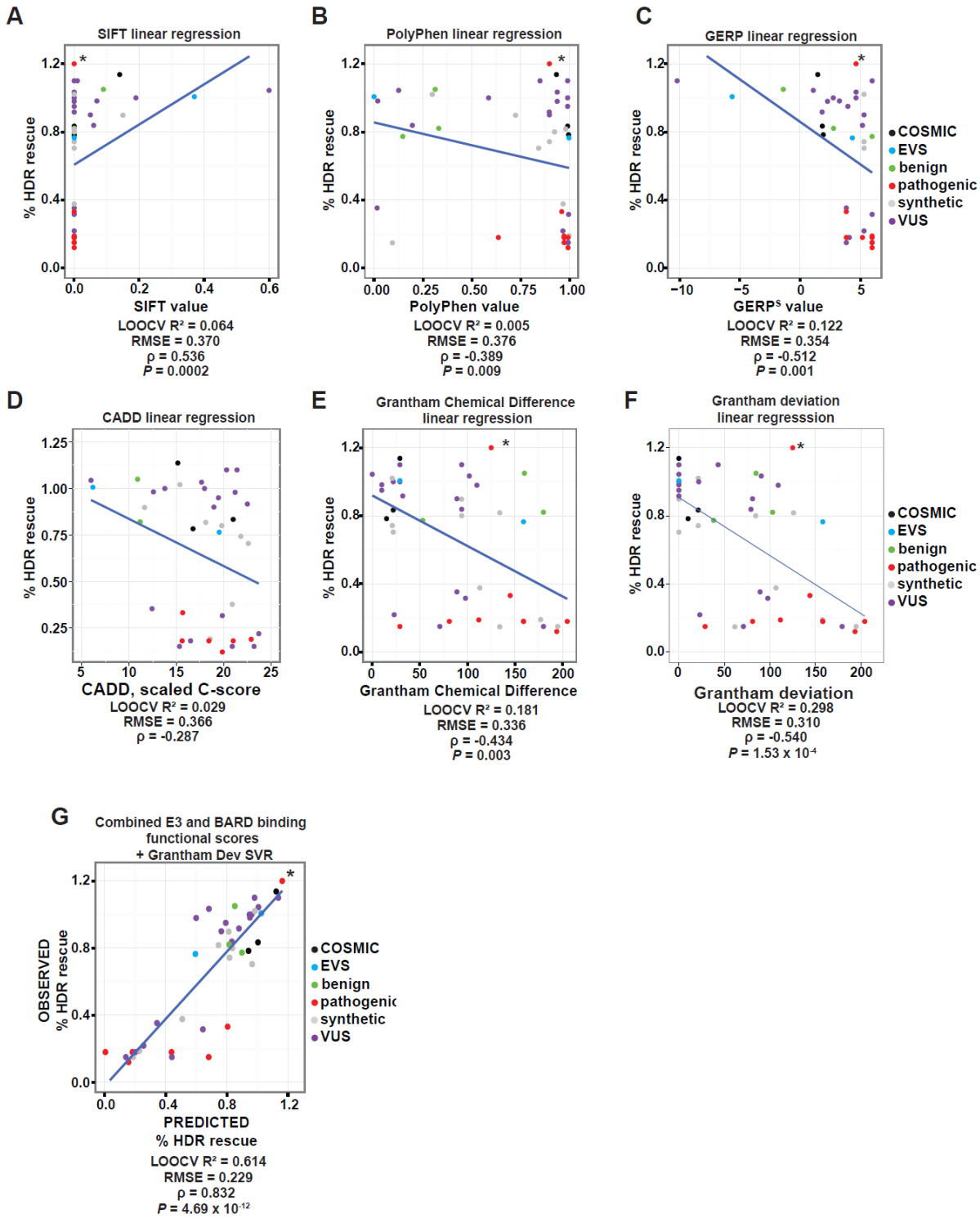


Figure S5.11 Scatter plots of regressions (models) of HDR rescue scores.

Model type and data source are indicted. Points are colored by database source or clinical classification of variant. LOOCV R2, RMSE, Spearman's Rank correlation (ρ), and P-value are reported.

Table S5.1 BRCA1 Variants

See (Starita *et al.*, 2015)

Table S5.2 BRCA1 DMS data

See (Starita *et al.*, 2015)

Table S5.3 BRCA1 Primers

See (Starita *et al.*, 2015)

Chapter 6. UNCOVERING THE STRUCTURAL BASIS OF GPCR FUNCTIONAL SELECTIVITY THROUGH DEEP MUTATIONAL SCANNING

6.1 ABSTRACT

Over the past two decades, interest in functional selectivity in G-protein coupled receptors has grown due to the potential of biased agonists to cause narrower ranges of biological effects, yielding new and more focused treatments while binding the same repertoire of GPCRs. We sought to help dissect the structural determinants of functional selectivity by comprehensive assessment of the effects of all single mutations in the Mu Opioid Receptor on surface expression and internalization in response to both morphine and fentanyl, two opiate agonists with well characterized biases in signaling. As a first step towards this goal, we have generated a barcoded library of nearly all single mutations of the hMOR and have developed an assay to assess the effects of these mutations on surface expression by porting a high throughput mutagenesis technique called Deep Mutational Scanning to mammalian cells. In the process, we have identified library delivery as the biggest barrier to high throughput mutagenesis in mammalian cells and made progress towards overcoming this barrier.

6.2 INTRODUCTION

GPCRs are fundamental to the way in which cells communicate with their environment. They mediate diverse processes in a wide range of species, from carbon fixation in halobacteria, to mating in yeast, to cognition in humans. The GPCR superfamily, with more than 5000 eukaryotic members, is the largest family mediating cellular responses to the environment(Unal and Karnik, 2012). Despite their central role in cellular biology, our understanding of the way in which they translate physical and chemical messages into internal signals is far from complete.

Classically, GPCRs have been thought to exist in two distinct conformations, the active and inactive states, which were respectively stabilized by agonists and inverse agonists. In this active state, GPCRs can catalyze the exchange of a GDP for a GTP in the active site of a G-protein, initiating a chain of downstream signals. This activation and signaling may, in turn, cause the GPCR to be phosphorylated on its C terminus by either G protein-coupled receptor kinases (GRK), which recognize and phosphorylate the C terminal tails of GPCRs in their active states, or by downstream signaling molecules such as PKC and PKA. A phosphorylated C terminus will attract arrestins, which will sterically inhibit

further G protein interaction and which will serve as mediators, through AP2 and directly through clathrin, of the formation of clathrin coated pits, causing the internalization of the receptor(Lefkowitz and Shenoy, 2005).

Over the past several decades, numerous examples of biased agonism have been reported across a range of GPCRs, providing more and more evidence that a two state receptor is likely not the norm (Spongier *et al.*, 1993; Gurwitz *et al.*, 1994; Berg *et al.*, 1998; Whistler and Zastrow, 1998; Kurrasch-Orbaugh *et al.*, 2003; Wei *et al.*, 2003; Kohout *et al.*, 2004; Gesty-Palmer *et al.*, 2006; Raote *et al.*, 2013). Instead, GPCRs are now thought to exist in a larger number of functionally distinct active and inactive conformations, allowing a single receptor to transmit a wider array of signals. For example, at the serotonin receptor HT2a, serotonin will cause an equal activation of PLC and PLA mediated signaling, whereas tryptamine has approximately double the efficacy in activating the PLC pathway when compared to the PLA pathway (Kurrasch-Orbaugh *et al.*, 2003). In addition, treatment of HT2a with serotonin leads to a PKC dependent internalization event, whereas internalization after the addition of dopamine is independent of PKC (Raote *et al.*, 2013). There is also some evidence that these signaling differences are caused by differing receptor conformations. Kahsai et al. used mass spectrometry to record the time course of incorporation of heavy (deuterated) or light (protiated) labels at cysteines and lysines of the beta2 adrenergic receptor in response to various agonists and antagonists(Kahsai *et al.*, 2011). As the label incorporation rate is dependent on the charge and solvent exposure of the microenvironment, the authors were able to use 9 separate sites to demonstrate both incorporation rates that correlate with levels of G-protein activation as well as ligand specific rates that point to the existence of ligand specific conformations.

Though this and other studies point to the existence of multiple, distinct active states, it is still unclear how many conformations exist, and structures of active states are limited. GPCRs are difficult to crystallize due to their dynamics and to their tendency to denature when removed from the membrane. Successful strategies, therefore, have relied on stabilizing the receptor, either through replacement of the dynamic third intracellular loop with T4 lysozyme(Cherezov *et al.*, 2007; Rosenbaum *et al.*, 2007), through alanine scans for stable variants(Warne *et al.*, 2011), or through stabilization of the inactive state with inverse agonists. Active states have proved even more difficult to crystallize, though the active state of the Beta2 adrenergic receptor has been crystallized by treatment with an active state stabilizing nanobody (a recombinant, camel-derived antibody) which acts as a surrogate G-protein(Rasmussen *et al.*, 2011b). Later, the active state was crystallized in complex with the G_s protein complex, which required both a change in detergent and the stabilization of the N terminus with a T4 lysozyme fusion(Rasmussen *et al.*, 2011a). The active state of the Neurotensin Receptor 1 (NTSR1) in complex with neurotensin was also crystallized a year later by using a variant with six stabilizing mutations as well as a T4 lysozyme

domain in place of intracellular loop 3 (ICL3)(White *et al.*, 2012). In all active crystal structures, a relatively large outward shift (6-14 Angstroms) of the cytoplasmic end of TM6 is observed, as well as smaller movements in TM3, TM4, and TM7. This outward shift occurs in part due to the disruption of a salt bridge between Asp (D) on TM3 and Tyr (Y) on TM6, which are both part of a highly conserved motif called the D(E)/RY motif. Though these structures do help in understanding how G-protein activation occurs in response to ligand, they don't provide as many insights into how a single GPCR can generate a variety of intracellular signals.

In many cases, some of this signaling diversity in individual GPCRs has been shown to be the result of β arrestin mediated signaling. β arrestin was originally identified for its role in desensitizing GPCRs, specifically in reducing rhodopsin mediated activation of transducin in response to light (Wilden *et al.*, 1986). It was later discovered that β arrestin, in addition to augmenting receptor desensitization and causing receptor internalization, can also transmit intracellular signals through activation of the MAP kinases ERK1 and ERK2, a function which is dependent in some cases on receptor internalization(Luttrell *et al.*, 1999). Since then, β arrestins have been shown to signal through many different pathways (Lefkowitz and Shenoy, 2005), affecting a range of phenotypes including cellular motility and apoptosis. The recently solved structure of a constitutively active rhodopsin mutant in complex with β arrestin showed several major differences between the G protein bound and β arrestin bound states including a smaller outward deviation of helix 6, the formation of an alpha helix in ECL2, and larger movements in helix 7 and helix 8. To date, one group has successfully crystalized a GPCR (the serotonin receptor HT_{2b}) bound to a biased agonist (Wacker *et al.*, 2013). They found that the receptor adopted a conformation intermediate between active and inactive states, with a smaller outward shift of TM6 and an intact D(E)/RY motif salt bridge. The later feature interferes with G protein binding, since the GPCR/G-protein interface is usually stabilized by an interaction involving the Asp (R) on TM3. However, it is still unclear how this conformation biases the receptor towards β arrestin bound state as opposed to the inactive state. These crystal structures have been extraordinarily informative, but the requirement that the receptor adopts a single, stabilized conformation limits the study of dynamic behaviors by means of x-ray crystallography, and it is unclear whether the full range of active conformations can be faithfully sampled through stabilizing strategies, especially without an idea of which binding partners stabilize those states. For example, relaxation dispersion NMR experiments have demonstrated that unbound receptors spend some small amount of time in the active conformation and have additionally shown the existence of other rarely populated conformational states which are likely important for GPCR function(Hansen *et al.*, 2008). It is therefore necessary that alternative strategies also be used to probe the determinants of functional selectivity.

Mutagenesis studies have been used before to characterize GPCR expression and function, and a recent study analyzed the complete set of rat neurotensin receptor single mutants for expression and stability (Schlinkmann *et al.*, 2012). In this study, mutant neurotensin receptors were labeled with a fluorescent agonist, and mutants were compared based on their frequency in the top 1% most fluorescent cells. It is clear from this study that the wild type sequence is not optimized for expression and that evolutionary constraints on class A GPCRs are not necessarily the constraints that arise when selecting for optimal expression. This is not unexpected, as correct signaling and the correct regulation of signaling likely involve constraints that are at odds with optimized surface expression. This is a good example of how this type of data can be used to tease apart the structural basis of the various functions of a GPCR. Mutagenesis may also yield variants that are useful in research or drug development. For example, stabilizing mutations may aid in crystallizing the target (Xie *et al.*, 2003) and constitutively active mutants can aid in understanding active states (Han *et al.*, 1998).

Several studies have demonstrated that the Mu Opioid Receptor (MOR) in particular displays functional selectivity in response to different classes of opiates. One of the earliest examples of biased agonism in the MOR was the demonstration that morphine causes very little internalization, whereas other agonists, like etorphine, cause robust, β-arrestin mediated internalization(Whistler and Zastrow, 1998). The authors were able to cause morphine induced internalization by overexpressing either β-arrestin or GRK2, which phosphorylates the intracellular loops of the MOR and is necessary for desensitization and β-arrestin binding, though it was unclear whether beta arrestin mediated internalization in response to morphine occurred in vivo. Functional selectivity in the MOR has the potential to be useful clinically, as it has been shown that morphine induced analgesia is potentiated and tolerance is profoundly reduced in β-arrestin 2 KO mice(Bohn *et al.*, 1999, 2000). In addition, these mice are partially protected from the side effects of morphine administration, including constipation and respiratory suppression(Raehal *et al.*, 2005). It should be noted here that although Bohn et al. did show increased efficacy of DAMGO (an opiate agonist) stimulated GTP-y-S binding in these mice, their measurements of response latency to a hotplate were capped at 30 seconds, rendering them unable to measure differences in the efficacy of morphine in these mice. Instead, they demonstrate increased potency and no decrease in potency in response to chronic morphine exposure. Despite this limitation, an agonist that activated $G_{\alpha i}$ protein signaling without causing β arrestin 2 recruitment still might be expected to have several clinical benefits.

Several biased MOR ligands have been developed in the hopes of achieving an improved therapeutic index. The first was a compound derived from salvinorin A, called herkinorin (Groer *et al.*, 2007). Unfortunately, herkinorin is also a Kappa Opioid Receptor agonist, and only produces Mu Opioid Receptor dependent analgesia peripherally, possibly due to degradation by circulating esterases, which

also degrade Salvinorin A. However, it was later found that chronic herkinorin administration over 5 days resulted in a higher level of bilateral antinociception with low levels of tolerance(Lamb *et al.*, 2012). More recently, structure activity studies led to the development of a compound called TRV130 that is strongly biased away from β-arrestin signaling and towards $G_{\alpha i}$ signaling, displaying 14% of the efficacy of morphine for the recruitment of β arrestin II (Chen *et al.*, 2013). This compound caused reduced tolerance as well as reduced gastrointestinal dysmotility and respiratory depression when compared to equi-analgesic doses of morphine in preclinical studies (DeWire *et al.*, 2013). The recently concluded phase II clinical trial for TRV130 also showed significantly smaller drops in oxygen saturation at equi-analgesic doses of TRV130 and morphine (Viscusi *et al.*, 2016).

The increased potency of morphine in β-arrestin KO mice is slightly paradoxical when you consider the failure of morphine to induce robust β-arrestin mediated internalization. One explanation might be that this internalization is dependent on cell type and location. One study has shown that morphine induced β-arrestin mediated internalization occurs in striatal neurons but not elsewhere(Haberstock-Debic *et al.*, 2005). The MOR has also shown functional selectivity with respect to receptor internalization in vivo(Arttamangkul *et al.*, 2008). It is unclear what the difference is between striatal neurons and HEK 293 cells which leads to the different response to morphine, but this result does seem to indicate that morphine can cause internalization in some contexts, leaving open the possibility that a mutation may exist which could allow morphine-induced β-arrestin interactions in HEK 293 cells. More recently, it was found that inhibiting JNK reverses many of the effects of β-arrestin II KO on morphine signaling but had no effect on fentanyl signaling (Mittal *et al.*, 2012). Nuclear localization of phosphoJNK had previously been shown to be a specific effect of morphine treatment, but not fentanyl treatment (Melief *et al.*, 2010). Taken together, it may be that a bias towards JNK signaling, rather than a bias away from β-arrestin, is the more direct cause of reduced tolerance and respiratory depression in herkinorin and TRV130 treatments.

The activation of GPCRs has also been studied from an evolutionary perspective. One study showed that, in an alignment of 940 GPCRs, placing a constraint on the amino acid identity of 10 conserved positions significantly constrained the possible identities of 47 other positions, indicating that these positions interacted, or changed together, throughout evolution(Süel *et al.*, 2003). These residues form a network that extends from the extracellular domains, to the ligand binding pocket, to the cytoplasmic surface. This, as well as other structural and mutagenic data, has led to the hypothesis that differences in signaling are largely due to differences in the coupling between GPCR domains(Unal and Karnik, 2012). These include the transmembrane domain, extracellular domain, intracellular domain, and, in some GPCRs, the long N-terminal and long C-terminal domains. Mutations that affect signaling exist in all domains, but the validity of the hypothesis that a specific network of residues transmits the

signal from the extracellular to intracellular domains might be tested by looking for clusters of amino acids that correlate functionally across multiple high throughput assays.

In order to study the relationship between structure and function, including biased function, in a GPCR, we developed a system to make pooled measurements of the effects of a large number of single amino acids on MOR expression and internalization in the human cell line HEK293. The system makes use of a translation marker, a myc-tagged MOR, flow cytometry, and high throughput sequencing to generate normalized estimates of cell surface expression.

6.3 RESULTS

Our assay (Figure 6.1A) requires that the measured phenotype in each cell is due to the expression of only one MOR variant. In order to ensure single copy expression, we first integrated a myc-tagged rat MOR into HEK293 cells using a lentiviral delivery vector pLVX _mCherry_2A in which rMOR is expressed off of a tetracycline activated promoter, Tet-On3G, as a fusion with mCherry via the self-cleaving peptide F2A (Kim *et al.*, 2011) (Figure 6.1B). The red fluorescent mCherry gene acts as a translation marker for the MOR. Before cloning into lentivirus, surface expression of the rMOR was verified by transient transfection of HEK293 cells followed by labeling with a mouse anti-myc antibody and an AlexaFluor 488 (AF488)-conjugated secondary (Figure S6.9).

To see if we could detect differences in surface expression via flow cytometry, we first mutagenized the rMOR to create several variants with previously identified effects on surface expression: N38D, N150D (Ballesteros and Weinstein, 1995 (BW) 3.35), and N188K (BW 4.46). N38D is equivalent to N40D, a polymorphism found in 10-20% of humans which may result in a subtle increase in agonist induced activity (Beyer *et al.*, 2004). N150D is equivalent to N152D, a rare (<1%) polymorphism in the human MOR which shows roughly 30% of the WT cell surface receptor density (Befort *et al.*, 2001). N188K, equivalent to N190K in human MOR, causes an 80% decrease in cell surface expression, likely due to increased instability as opposed to constitutive activity (Fortin *et al.*, 2010). All variants and the WT rMOR were cloned in pLVX_mCherry_2A, packaged in lentiviral particles, and used to infect HEK293 cells at 0.1 MOI, so that a majority of cells were infected with and expressed only one copy of the receptor. We tested several doxycycline induction times and determined that a 56 hour induction with a 12 hour rest resulted in a high level of surface expression with low variance (Figure S6.8). At 68 hours, differences in surface expression were clearly discernible for N150D and N188K, whereas N38D caused at most a mild decrease in expression, which is in agreement with previous findings (Figure 6.2A). We next mixed together cells expressing either the WT or the N188K variant of rMOR, sorted the mixture into 4 bins of AF488 fluorescence, and extracted genomic DNA from the cells in each bin (Figure 6.2C).

The rMOR constructs were amplified from around 3500 cells worth of DNA (~20 ng), yielding bands of the expected size. The bands were gel purified and Sanger sequenced. Peak Picker software was used to compare peak heights at N188 (base 564) for all four bins. The sequencing-based reconstruction of the distributions of N188K and WT rMOR surface expression correctly shows the N188K population with lower fluorescence than WT, with the majority of the population falling in bins 3 and 4 (Figure 6.2C)

Having verified that mixed MOR populations are separable by flow cytometry and sequencing, we next constructed a library of myc-tagged human MOR single mutants by overlap extension PCR (OEP) (Heckman and Pease, 2007) using pairs of mutagenic primers designed to substitute each codon with the random codon NNS. Initial sequencing of the products from a few positions demonstrated a strong CG bias that was not completely alleviated by a lower annealing temperature and was traced back to the mutagenic primers (Table S6.1). As CG bias with machine mixed oligos shows batch to batch variability, a plate of 48 primer pairs was ordered and bias was determined by spike-in to a MiSeq run, showing very little bias in any of the primer pairs. The full length hMOR library can't be sequenced in one read on most next generation sequencing platforms, so the library was barcoded (15 random nucleotides) and barcodes were linked to hMOR variants by tag-directed read grouping or subassembly (Hiatt *et al.*, 2010). The subassembled library covers 91% of the single mutations in hMOR with a small amount of non-uniformity (Figure 6.3). We infected HEK293 cells with the lentiviral library, induced hMOR expression, labeled the receptors, and sorted cells into bins of AF488 fluorescence as above (Figure 6.4A). We were concerned about the high recombination rate of lentiviral vectors during packaging, so we amplified the hMOR construct from each bin's genomic DNA and shotgun sequenced the amplicon. Compared to the reads from the subassembly, the reads from the sorted library had a much higher number of mutations, as would be expected if new variants were generated by recombination between NNS single mutants (Figure 6.4B).

Since a high recombination rate breaks the link between barcode and variant in addition to generating new variants, we abandoned the lentiviral system in favor of recombinase-based integration. We tested the efficiency of 2 different recombinases: FLP, a tyrosine recombinase from *Saccharomyces cerevisiae* which catalyzes the reversible recombination between FRT sites (Dymecki, 1996), and Bxb1 integrase, a serine integrase from the phage Bxb1 which catalyzes an irreversible recombination between AttP and AttB sites, leading to the generation of AttL and AttR sites, which are no longer targets of the integrase(Xu *et al.*, 2013). Both recombinases can be used for the genomic integration of a circular plasmid. We first subcloned the barcoded library into a vector containing the FRT site immediately followed by a hygromycin resistance gene (hyg^R) lacking its start codon. Integration of the plasmid into the FLP-In cell line (ThermoFisher) puts hyg^R in frame with a start codon in the genomic landing pad and

expression is driven from a CMV promoter. Successful integration can be selected for by treatment with hygromycin. Both WT hMOR and the N188K variant of rMOR were successfully integrated by this strategy and showed the expected surface expression (Figure 6.5A). However, integration efficiency with FLP was never more than 0.02% (~2000 cells from a T225 flask), which, given the uniformity of the hMOR library, was not high enough to obtain sufficient coverage of single mutants (Figure 6.3C). In order to develop another assay with which to test functional selectivity in the MOR, we also introduced a GFP reporter of AP1 expression into the FLP-In cell line, since AP1 is activated by the JNK->cJun pathway and nuclear localization of phospho-JNK was previously shown to occur in response to morphine treatment but not fentanyl treatment (unpublished data, Jamie Kuhar (Chavkin lab), University of Washington). FLP-in AP-1 reporter cells were generated by infection with a lentiviral vector and reporter activity was verified by treatment with TPA (Figure 6.5B). Neither 10uM morphine nor 10uM fentanyl activated the AP-1 reporter after treatment for 3 hours.

To test the Bxb recombinase, we deleted TRE3G from the lentiviral library and cloned the Bxb AttB site upstream of mCherry. We used a HEK293 cell line with a genomic landing pad consisting of the Bxb AttP site and a copy of EGFP with its expression driven by the TRE3G promoter (Figure 6.6A). Integration of the target vector into this cell line by co-transfection with a Bxb1 expression vector puts the mCherry/hMOR library fusion downstream of a TRE3G promoter and upstream of EGFP, disrupting EGFP expression and causing the expression of the mCherry/hMOR fusion. Successful integrants can be selected by flow cytometry as a population of EGFP-/mCherry+ cells after induction with doxycycline (Figure 6.6B). We integrated the WT hMOR into the Bxb1 AttP containing HEK293 cells, and verified that we could detect surface expression. We then treated the cells expressing WT hMOR with 10uM fentanyl, 10uM DAMGO, or vehicle and verified that internalization occurred in response to both DAMGO and fentanyl (Figure 6.6C). In a separate experiment in FLP-In cells expressing WT hMOR, 1 hour of 10uM morphine treatment did not cause internalization, verifying the feasibility of measuring functional selectivity in this system by differential internalization (Figure 6.6C).

We then attempted to deliver the barcoded hMOR library to HEK293 cells by integration with Bxb1. Integration efficiencies with Bxb1 ranged from 1- 5%, which is 50-250 fold higher than FLP. By transfection without the Bxb1 expression vector, we found some background expression of mCherry off of the unintegrated plasmid on the first day of sorting, but this background expression became lower over time and was reduced to near background levels after 2-3 weeks (Figure 6.6). This means that to obtain a population of pure integrants, a second flow based selection for EGFP-/mCherry+ cells is required after 2 weeks. We also found that prolonged passaging of these cells without doxycycline lead

to the silencing of the landing pad locus, leading to the loss of roughly 96 % of integrants over 2 weeks (Figure 6.7A). Culturing in the presence of doxycycline mitigated this loss (Figure 6.7B).

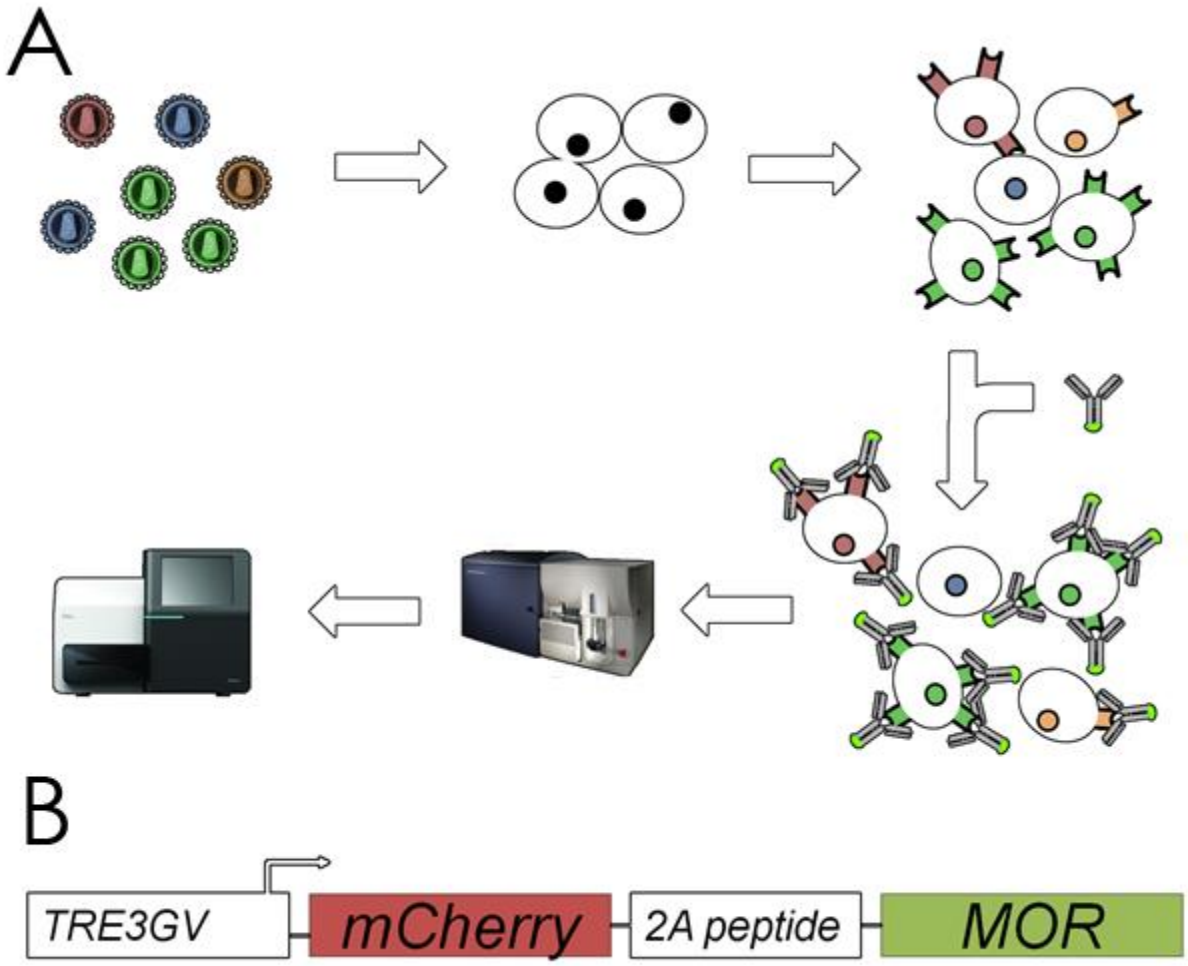


Figure 6.1 An overview of the GPCR expression assay.

(A) The mutant library of myc-tagged MORs is packaged into lentiviral particles which are used to infect TetOn HEK 293 cells at a low MOI so that most cells express only one receptor. An antibody to the myc-epitope is used in conjunction with an Alexafluor 488 conjugated secondary antibody to visualize cell surface expression of the MOR variants. The cells are sorted into fluorescence bins with the FACSAria II, and DNA is extracted and sequenced from each bin on a high throughput sequencing platform (pictured: Illumina MiSeq). (B) Tetracycline regulated expression of the MOR fused to mCherry via the self-cleaving 2A peptide.

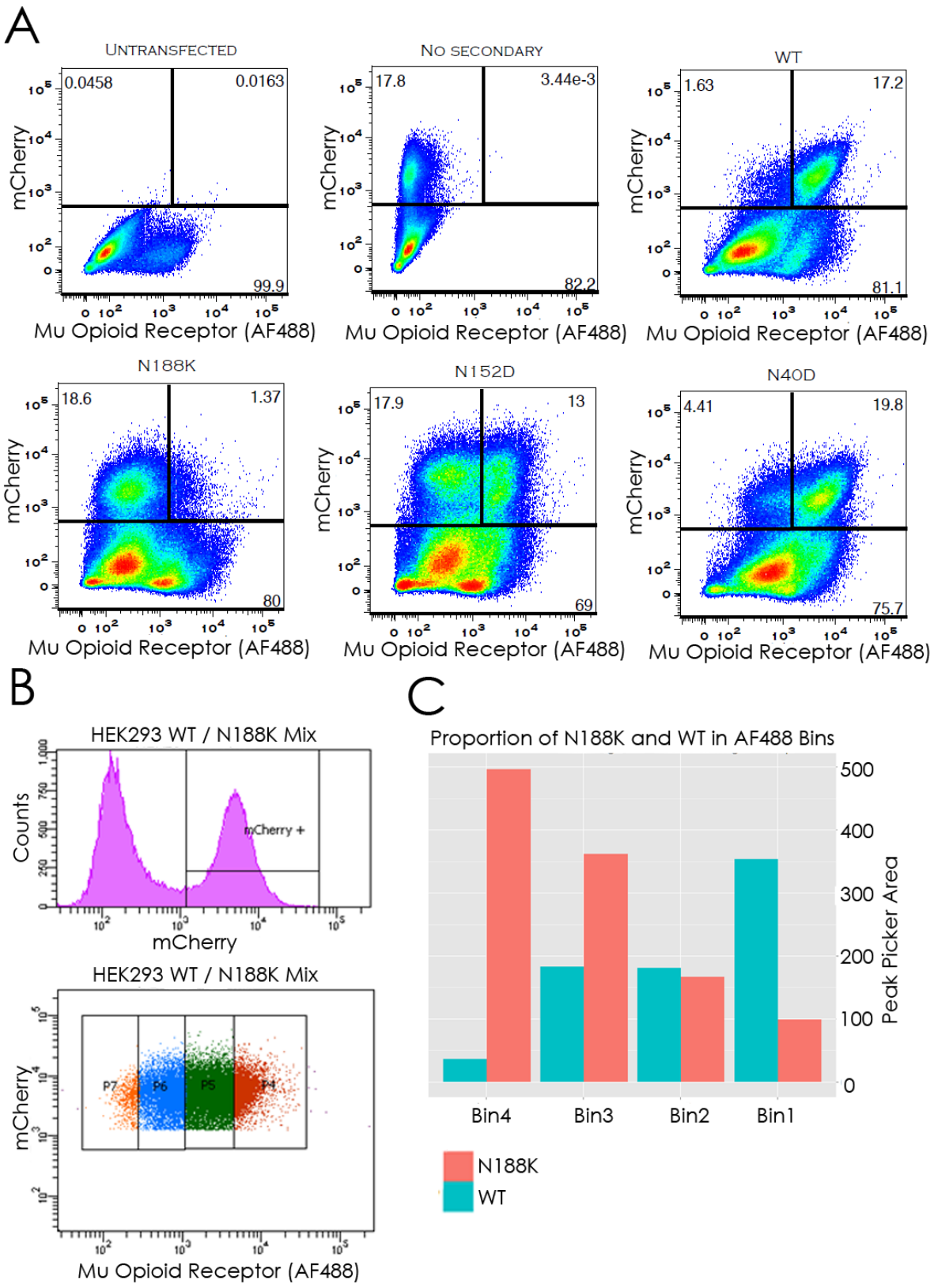

A
UNTRANSFECTED
0.0458
0.0163
99.9
NO SECONDARY
17.8
3.44e-3
82.2
WT
1.63
17.2
81.1
N188K
18.6
1.37
80
N152D
17.9
13
69
N40D
4.41
19.8
75.7
mCherry
Mu Opioid Receptor (AF488)
B
HEK293 WT / N188K Mix
Counts
mCherry +
mCherry
HEK293 WT / N188K Mix
P7
P6
P5
P4
mCherry
Mu Opioid Receptor (AF488)
C
Proportion of N188K and WT in AF488 Bins
Peak Picker Area
Bin4
Bin3
Bin2
Bin1
N188K
WT

Figure 6.2. Quantification of MOR surface expression by flow cytometry followed by DNA sequencing

(A) Flow cytometry of the WT MOR and variants with different levels of reduced surface expression. (B) Sort of a mixture of HEK293 cells expressing either WT or the N188K variant of rMOR. (C) Sanger sequencing of the DNA from each bin obtained in the sort shown in Fig 10. The peak areas at base pair 564 (aa position 188) were used to determine the relative abundances of the WT MOR and the N188K mutant in each bin. As expected from previous flow cytometry analysis, N188K shows higher abundances in the lower fluorescence bins (bins 3 and 4), and the WT MOR shows higher abundances in the higher fluorescence bins.

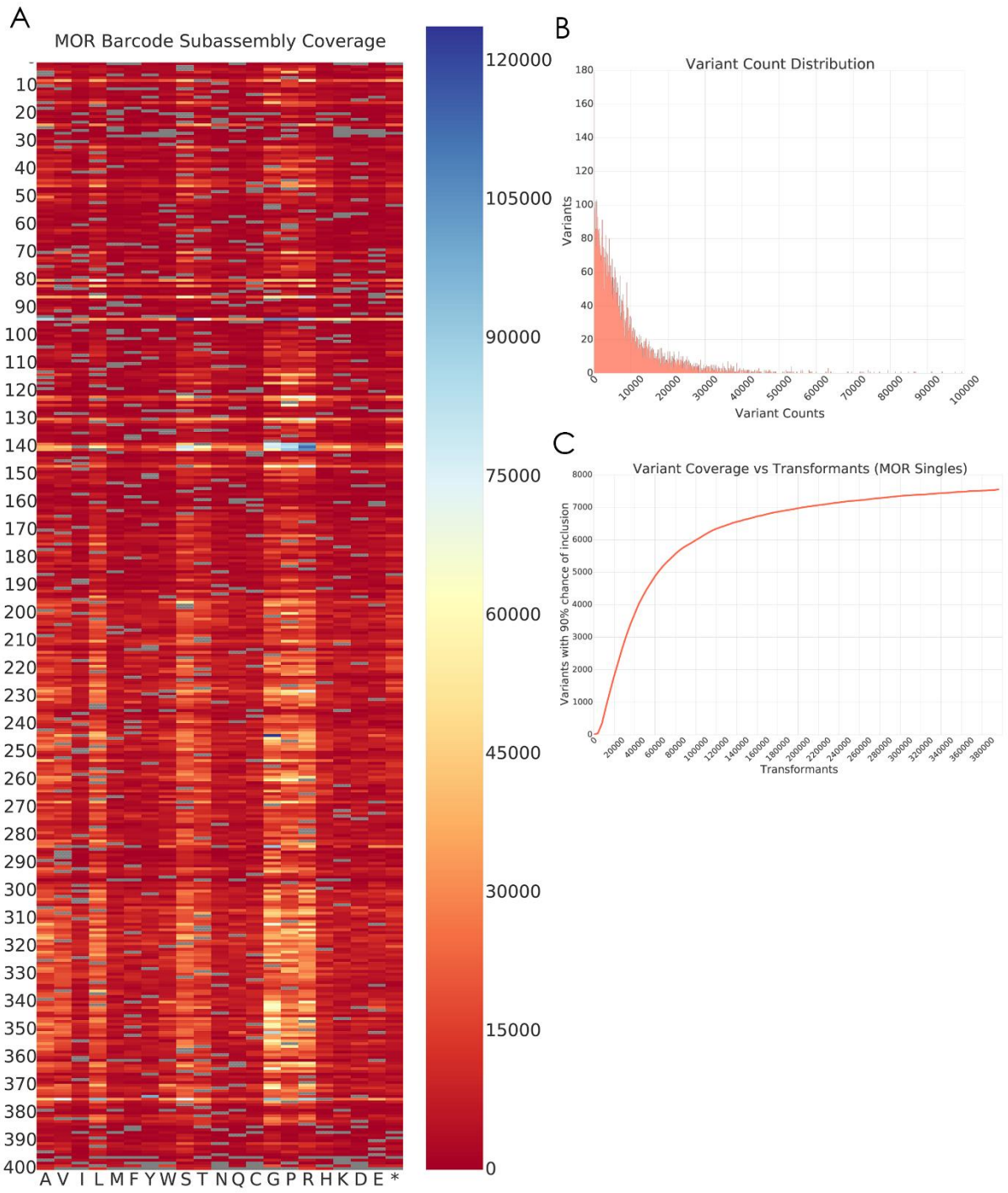


Figure 6.3 Coverage and uniformity of the subassembled hMOR library

(A) 91% of single mutation hMOR variants are found in the subassembled library. Variants are colored based on the number of times the barcode was sequenced during subassembly. (B) Uniformity of the subassembled library. A few variants are present at much higher frequencies in the library. (C) The relationship between the number of HEK293 transformants and the number of unique variants that are likely to be represented among those transformants.

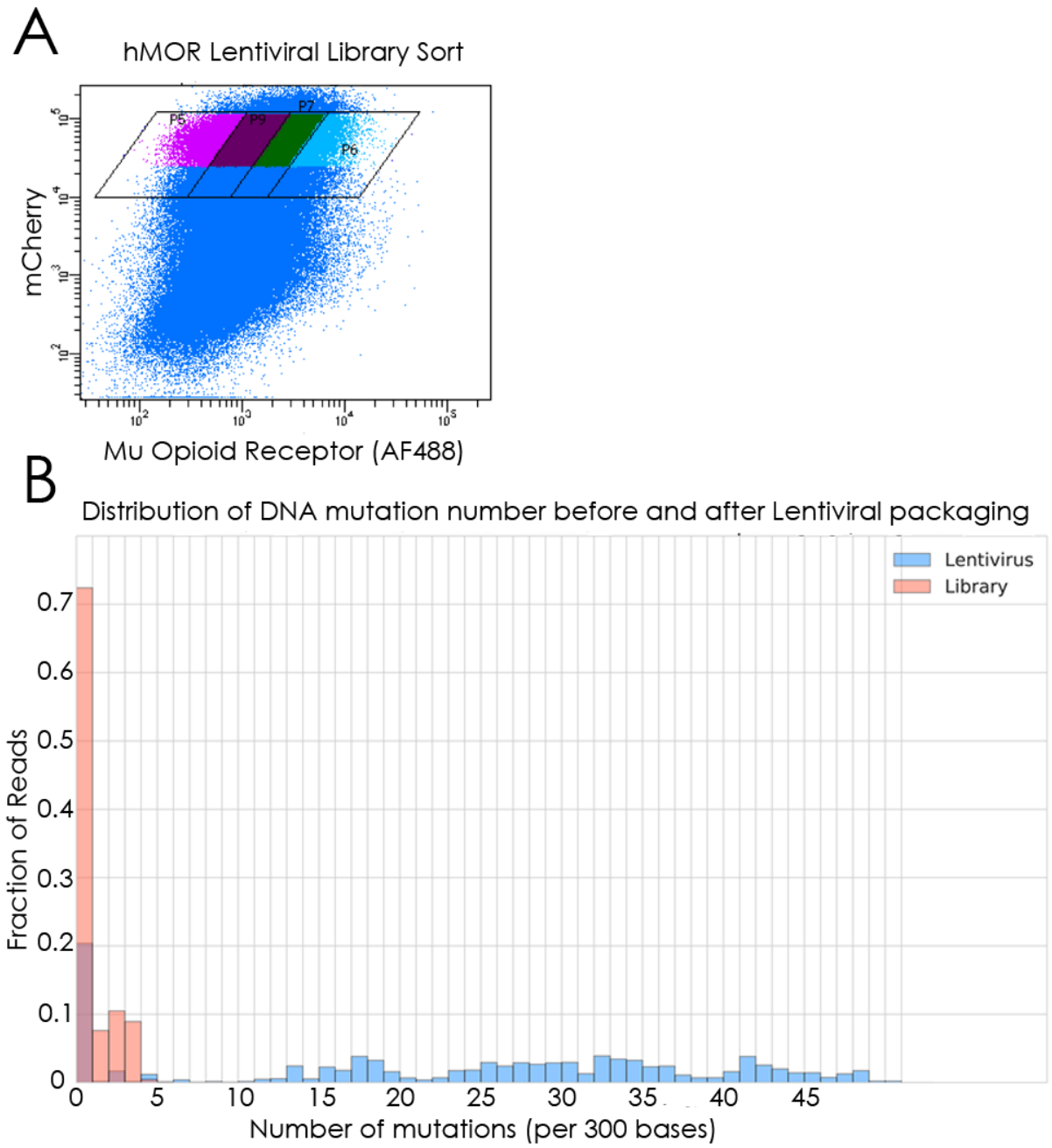


Figure 6.4 High levels of recombination in the packaged lentiviral hMOR library

(A) The lentiviral hMOR library was sorted into 4 bins of AF488 fluorescence. (B) Higher numbers of mutations in hMOR after packaging into lentivirus. Forward and reverse reads were merged and only those with 100% agreement were aligned to the MOR.

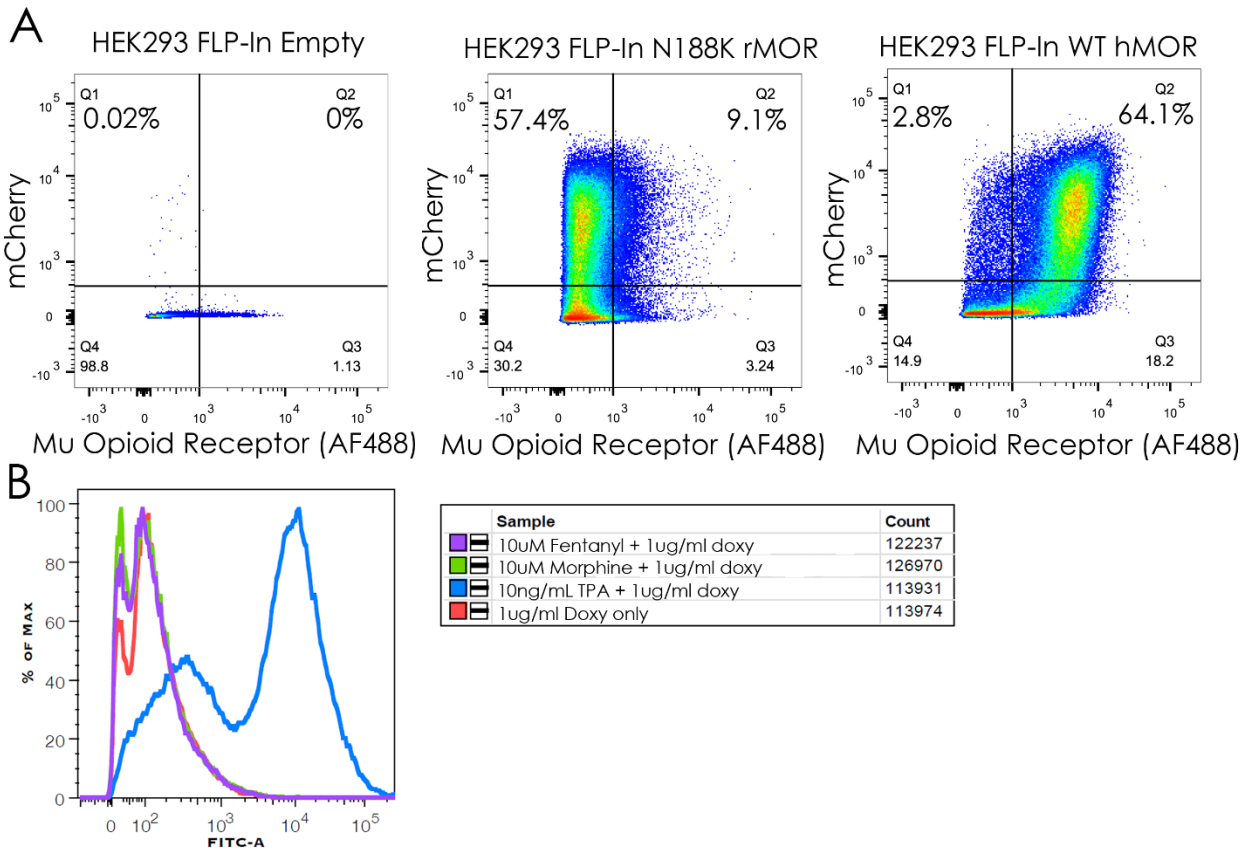


Figure 6.5 Assay development with FLP recombinase

(A) Expression of FLP integrated N188K rMOR and WT hMOR can be distinguished. (B) FLP-In cells containing a GFP AP1 reporter and expressing WT MOR are activated by TPA but not by either morphine or fentanyl.

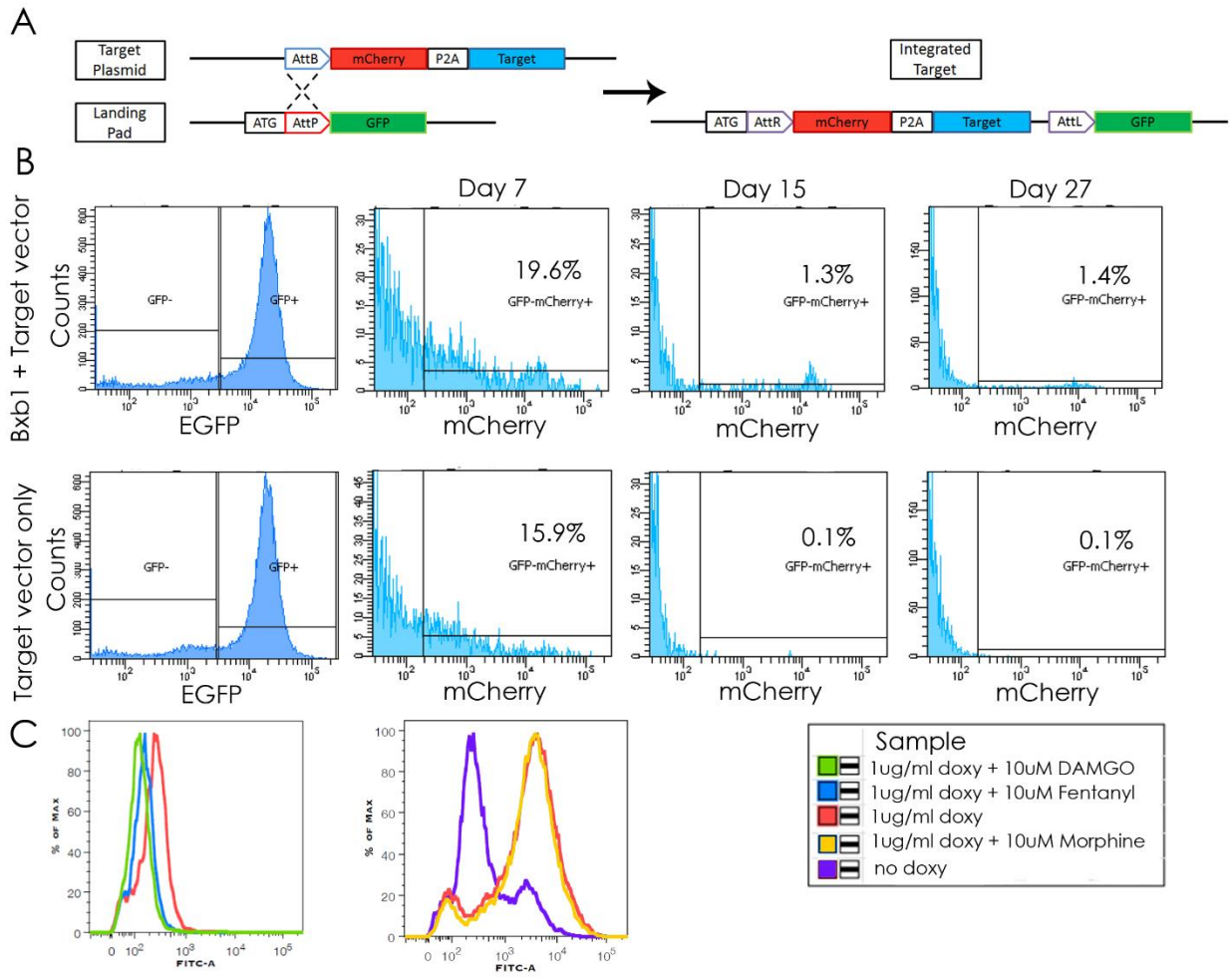


Figure 6.6 Integration system using Bxb recombinase

(A) The genomic Bxb1 AttP landing pad expresses EGFP until expression is disrupted by integration of the target construct. (B) Background expression of the mCherry translation reporter off of the target plasmid starts high, but decreases over time. (C) 1 hour treatments with either 10uM DAMGO or 10uM fentanyl (in Bxb integrated WT MOR cells) cause receptor internalization, but 10uM morphine (in FLP-In WT MOR cells) does not.

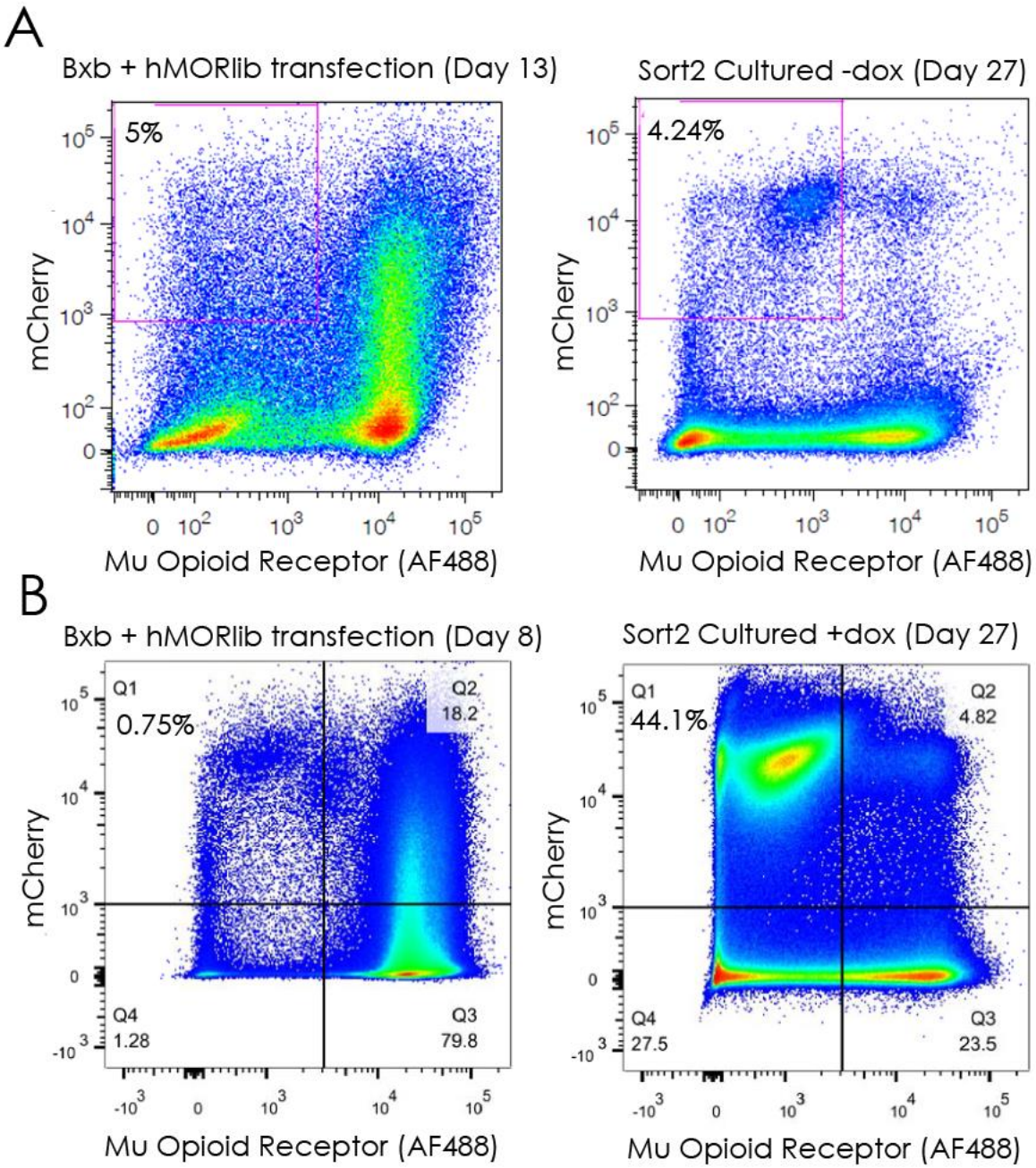


Figure 6.7 Silencing of the Bxb landing pad in the absence of doxycycline

(A) After transfection with Bxb and the hMOR library vector, the GFP-/mCherry+ population was sorted. Resorting this population after growth for 2 weeks without doxycycline yielded only 4.2% GFP-/mCherry+ cells, likely due in part to silencing of the integration locus. (B) Resorting after growth in doxycycline yields a much higher fraction of GFP-/mCherry+ cells (44.1%)

6.4 DISCUSSION

Deep mutational scanning is the most promising tool for the interpretation of rare and unique variation in the human population. However, there are many cellular phenotypes that are context dependent and hence impossible to assay accurately in model organisms. Here, we've identified Bxb1 as a relatively high efficiency tool for the faithful integration of libraries of genetic variants and so have partially overcome a major barrier to DMS in mammalian systems. Despite this, library delivery remains the biggest bottleneck in mammalian DMS and further increases in efficiency are probably necessary in order to scale the method for genome-wide application.

Our assay was developed with the measurement of functional selectivity in mind. One of the simplest ways of doing this is to measure changes, due to genetic variation, in the known internalization bias of fentanyl treatment over morphine treatment. Cell surface expression levels, measured here, are a necessary baseline when measuring changes in internalization. We have shown that internalization can be detected in HEK293 cells expressing hMOR from an integrated Bxb1 landing pad (Figure 6.6C), but the difference was fairly small, likely due to silencing of the hMOR expression locus due to prolonged culturing in the absence of doxycycline (Figure 6.6B). We also hoped to detect changes in biased agonism through the use of an AP1 reporter. Unfortunately, both morphine and fentanyl failed to stimulate AP-1 activity despite the known activation of JNK2 by morphine treatment (Melief *et al.*, 2010; Mittal *et al.*, 2012). One explanation for this result is that activated JNK2 failed to increase nuclear levels of the active phosphorylated cJun, which would otherwise lead to AP1 activity. This is consistent with Mittal *et al.*, who only saw increased levels of nuclear phospho-cJun in the absence of β-arrestin II, a result that they hypothesized is due to release and activation of JNK3 (rather than JNK2), which is normally binds to β-arrestin II. Another possibility is that the 3 hour time point is too early to detect AP1 activation.

Bxb1 catalyzes recombination between AttB and AttP sites with high enough efficiency to generate 20,000-40,000 integrants in a single transformation. This integrated library can now be used to measure the effects of genetic variation in the MOR on surface expression and internalization in response to morphine and fentanyl. Though lentivirus might still be viable for short, complex libraries, Bxb1 is the better option for performing mammalian DMS with longer, single mutant saturation libraries.

6.5 METHODS

6.5.1 *Plasmids*

pLVX_mCherry_2A is a lentiviral vector in which the tetracycline activated promoter TRE3G drives expression of an mCherry fusion with the self-cleaving peptide F2A. It was created by switching the order of mCherry and the F2A peptide in pLVX-TRE3G_2A_mCherry, a lentiviral vector created by Martin Wohlfahrt at the Fred Hutchinson Vector Core that itself was derived from pLVX-TRE3G (Clontech catalog#631191, 631193). pcDNA3_mycMOR, containing the myc-tagged rat Mu Opioid Receptor (rMOR) cDNA was a gift from the lab of Charlie Chavkin. In order to create both pLVX_rMOR_2A_mCherry and pLVX_mCherry_2A_rMOR (gene fusion order switched), flanking NotI and EcoRI sites were added to the rMOR from pcDNA3_mycMOR by PCR either with primers MM_NotI_f(AAAAAAGCGGCCGCACCATGGAGCAGAAACTCATCTCTGAAGAGGATCTGGGATCT) and MM_EcoRI_r2(AAAAAAGAATTCGTAGGGCAATGGAGCAGTTTCTGCCTCCAGAT), or primers 2A_MM_NotI_f(AAAAAAGCGGCCGCCATGGAGCAGAAACTCATCTCTGAAGAGGATCTGGGATCT) and 2A_MM_EcoRI_r(AAAAAAGAATTCTTAGGGCAATGGAGCAGTTTCTGCCTCCAGAT). The resulting PCR products were cloned in frame either immediately upstream of the F2A/mCherry fusion of pLVX-TRE3G_2A_mCherry, or downstream of the mCherry/F2A fusion of pLVX_mCherry_2A. rMOR variants N40D, N152D, and N188K were generated from both pLVX_rMOR_2A_mCherry and pLVX_mCherry_2A_rMOR by Overlap Extension PCR (OEP) (Heckman and Pease, 2007) using flanking primers MM_NotI_f/MM_EcoRI_r2 or 2A_MM_NotI_f/2A_MM_EcoRI_r, respectively, and mutagenic primer pairs MM_N40D_f(ccacgttgatggcgaccagtccgatccat)/MM_N40D_r(atggatcggactggtcgccatcaacgtgg), MM_N152D_f(cgtgatctcaatagattactacgacatgttcaccagcatattcacc)/MM_N152D_r(ggtgaatatgctggtgaacatgtcgtagtaatctattgagatcacg), and MM_N188K_f(cgaaatgccaaaatcgtcaaagtctgcaactggatcctc)/MM_N188K_r(gaggatccagttgcagactttgacgattttggcatttcg).

In order to generate pLVX_mCherry_2A_hMOR, we first introduced 2 synonymous mutations at positions 1101 and 1104 of the human Mu Opioid Receptor OPRM1 cDNA (ThermoFisher Scientific IOH61930) to remove an EcoRI site (positions 1100-1105 changed from GAATTC to GGATCC). We then added an N-terminal myc tag and flanking NotI and EcoRI restriction sites using primers 2A_hMM_NotI_f(AAAAAAGCGGCCGCCATGGAGCAGAAACTCATCTCTGAAGAGGATCTGGG

ATCTgacagcagcgctgcccccacgaacgcca) and
2A_hMM_EcoRI_r(AAAAAGAATTCttagggcaacggagcagtttctgcttccagattttcta), and cloned the product in frame immediately downstream of the mCherry/F2A fusion of pLVX_mCherry_2A.

To create pDY_FRT_hMOR, an insert containing a ClaI site was generated by a fill-in reaction between oligos
FRT_add_claI_f(CTAGTTGAGCGATCCTCCGTTAACAATCGATTATTGTTACCAATGACGTAGGG) and
FRT_add_claI_r(CGCGCCCTACGTCATTGGTAACAATAATCGATTGTTAACGGAGGATCGCTCAA) and subsequently cloned into pFRT_LacZeo (ThermoFisher Scientific V601520) by digestion with SpeI and AscI to create pDY_FRT_ClaI. The TRE3G->mCherry->2A->hMOR fragment from pLVX_mCherry_2A_hMOR was then subcloned into pDY_FRT_ClaI using ClaI and AscI. The pOG44 FLP-Recombinase expression plasmid was also from ThermoFisher (Cat V600520).

To generate pLVX_Bxb_mCherry_2A_hMOR, a construct was created to remove the TRE3G promoter from pLVX_mCherry_2A_hMOR, as well as to add the Bxb1 recognition site AttB and to replace the F2A self-cleaving peptide with P2A. A 5'->XhoI->Bxb1 AttB->Kozak->mCherry->P2A->3' fragment was generated by amplifying mCherry from pLVX_mCherry_2A using primers
pLVX_Bxb1_mCherry_f(CTCGAGCCGGCTTGTCGACGACGGCGGTCTCCGTCGTCAGGATCATCCGCCACCATGGTGAGCAAGGGCGAGGAGGATAACATG) and
pLVX_Bxb1_mCherry_r(CAGCAGGCTGAAGTTAGTAGCTCCGCTTCCCTTGTACAGCTCGTCCATGCCGCCGGTGGA) and a 5'->P2A->NotI->3' fragment was generated by amplifying P2A(GGAAGCGGAGCTACTAACTTCAGCCTGCTGAAGCAGGCTGGAGACGTGGAGGAGAACCCTGGACCT) with primers
pLVX_P2A_f(TCCACCGGCGGCATGGACGAGCTGTACAAGGGAAGCGGAGCTACTAACTTCAGCCTGCTG) and
pLVX_P2A_r(AGAGATGAGTTTCTGCTCCATGGCGGCCGCAGGTCCAGGGTTCTCCTCCACGTCTCCAGC) and these 2 fragments (which have 30bp overlap) were mixed and amplified with the flanking primers pLVX_Bxb1_mCherry_f and pLVX_P2A_r to generate the full 5'->XhoI->Bxb1 AttB->Kozak->mCherry->P2A->NotI->3' construct. This construct was cloned into pGEM and a correct clone was identified by Sanger sequencing with the M13F and M13R universal primers. 4ug of the correct construct (in pGEM) was digested with XhoI and NotI and subcloned into the library plasmid pLVX_mCherry_2A_MORlib_bc between the XhoI and NotI sites, replacing the TRE3G promoter and the F2A peptide. The Bxb expression plasmid pCAG-NLS-HA-Bxb1 was a gift from the Fowler lab (Addgene 51271).

6.5.2 *Construction of hMOR libraries*

For all 399 internal codons in hMOR, pairs of mutagenic primers that substitute the WT codon for the degenerate codon NNS were designed in silico using a Matlab script from Firnberg *et al.*, 2014. Primer pairs were ordered from IDT in 96 well format with machine-mixed bases. The hMOR variant library was constructed by Overlap Extension PCR (Heckman and Pease, 2007) using separate reactions for each codon. 5' library fragments were generated by PCR from 3ng of the pLVX_mCherry_2A_hMOR template using 1ul of the reverse mutagenic primer (10uM) and 1ul of the constant primer 2A_MM_front_f (10uM). 3' library fragments were generated using the forward mutagenic primer and primer 2A_MM_back_r. PCR reactions were assembled in 96 well plates on a Biomek NX^p liquid handler (Beckman Coulter) using 0.2U Phusion polymerase and amplified for 21 cycles (98C 10s, 55C 30s, 72C 30s). 5' and 3' fragments were mixed (1ul each) separately for each codon and amplified using 0.75ul 2A_MM_front_f (10uM), 0.75ul 2A_MM_back_r (10uM), and 6.25ul KAPA HiFi HotStart ReadyMix (2x) for 25 cycles (98C 20s, 60C 15s, 72C 1m10s). The presence of full length variants was verified by gel electrophoresis, and 35 failed or low yield positions were regenerated by hand.

Full length variants were quantified using a Nanodrop 1000 (Thermo Scientific) and mixed in equal proportion. Mixed full length variants were subcloned into pLVX_mCherry_2A using NotI and XhoI. Low codon bias was verified by colony PCR (96 colonies) and Sanger sequencing. 15bp random barcodes were added upstream of the TRE3G promoter by mixing 1ul 100uM pLVX_F_barcode_gibson(cgggtttattacagggacagcagagatccagtttatcgatNNNNNNNNNNNNNNNCTTCCC TGCCCACCACCAGCTCCGTTGC) and 1ul 100uM pLVX_R_barcode_gibson(cttcatacgttctctatcactgatagggagtaaactcgagGGGACGGGTGGACTCGGGTTAG GCGGAgGCAACGGAGCTGGTGGTG) in a fill-in reaction using Kapa polymerase over 5 cycles (98C 30 s, 1C/s down to 52C, 72C 4 min). The resulting barcode-containing product was inserted within the ClaI site of the library plasmid by Gibson assembly, and a transformation into DH10B electromax cells (NEB) yielding approximately 150,000 barcoded variants was selected for subassembly (pLVX_mCherry_2A_MORlib_bc).

The library was sub-cloned into pDY_FRT_ClaI using ClaI and AscI yielding pDY_FRT_MORlib_bc. pLVX_Bxb_mCherry_2A_MORlib_bc was created in the same way as pLVX_Bxb_mCherry_2A_hMOR (see above).

6.5.3 *Subassembly to match hMOR variants and barcodes*

The addition of short barcodes to each plasmid in the MOR library allows each MOR variant to be identified by a short 15bp sequence during high throughput sequencing, circumventing limitations in next generation sequencing read lengths. In order to link the identities of barcodes and MOR variants that are present on the same plasmid, the variant associated with each barcode must be subassembled (Hiatt *et al.*, 2010). To accomplish this, pLVX_mCherry_2A_MORlib_bc was digested with either NotI (3ug) or EcoRI (3ug) and purified with the DNA Clean & Concentrator-5 kit (Zymo Research D4003). 1.7 pmol of the NotI digestion and 1.9 pmol of the EcoRI digestion were separately digested with 1.8 U and 1.7U Bal-31 (NEB Cat. M0213S) in 50ul reactions (25ul 2X Bal-31 buffer) and 7.1 ul were taken at 0, 2.5, 5, 10, 15, 20, and 25 minutes and stopped in the DNA binding buffer from the DNA Clean & Concentrator-5 kit. Time points 0, 2.5, 5, and 10 were mixed as long fragments (separately for NotI and EcoRI digestions) and time points 15, 20, and 25 were mixed as short fragments. EcoRI long and short fragments were digested with ClaI, purified, and end polished with the End-it DNA End Repair kit (Epicentre, Cat ER0720). NotI long and short fragments were digested with XhoI, purified, and end polished. 10ng of DNA from each condition (NotI->XhoI, long and short; EcoRI->ClaI, long and short) were ligated overnight at 16C in 250 ul reactions (5000 U T4 DNA ligase). NotI->XhoI fragments were amplified with Phusion polymerase on a MiniOpticon (BioRad Cat CFB-3120) for 18 cycles (reactions were monitored and pulled at mid-log phase) with primers MOR_5'_subass_F(AATGATACGGCGACCACCGAGATCTACACacagggacagcagagatccagtttatcgat) and pLVX_5'_subass_R2 (CAAGCAGAAGACGGCATACGAGATtgatccttcgaagattcctgtccttttctttggagccagaga) to add Illumina P5 and P7 adapters. EcoRI->ClaI fragments were similarly amplified with primers pLVX_3'_subass_F_try6(AATGATACGGCGACCACCGAGATCTACACctcttcatacgttctctatcactgataggg agtaaa) and pLVX_3'_subass_R_try6(CAAGCAGAAGACGGCATACGAGATcaccagcatattcaccctctgcacc). NotI->XhoI and EcoRI->ClaI PCR products were quantified with a Qubit dsDNA HS kit (ThermoFisher cat Q32851) and sequenced on both the MiSeq (Illumina Cat SY-410-1003) and NextSeq (Illumina Cat SY-415-1001) using sequencing primers pLVX_3'_subass_F_seq_try2(tcgatacattgcagtctgccaccctgtcaaggc) and tRNA_ind_hiseq_2(GCAACGGAGCTGGTGGTGGGCAGGGAAG) for EcoRI->ClaI or pLVX_5'_subass_seq_R (TGATCCTTCGAAGATTCCTGTCCTTTTCTTTGGAGCCAGAGA) and tRNA_ind_hiseq_2.

For each barcode, MOR variant reads were aligned to the WT hMOR using BWA v0.6.2 (bwasw command, all options default) and piled up using a custom python script. The script parses the SAM files

from BWA, removes barcodes with any low quality bases (quality score <25), uses the alignments to assign the number of times a specific, high quality nucleotide is seen at each position in the MOR, and makes base calls at each position if more than 80% of reads agree at that position. The output is a dictionary linking each fully subassembled barcode to its variant's identity.

6.5.4 *HEK293 cell culture and integration of MOR constructs*

Three HEK293 strains were used. Cells expressing the tetracycline transactivator Tet-On 3G were purchased from Clonetech (Cat 631182). HEK293 cells containing the FLP recombinase target site (FRT) were purchased from ThermoFisher Scientific (Cat R75007) and the Tet-On 3G transactivator was integrated by infection with lentivirus (Clonetech Cat 0055VCT). HEK293 Cells containing the Bxb landing pad and the Tet-On 3G transactivator were a gift from Kenneth Matreyek in Doug Fowler's lab. All cells were grown on TC- treated cell culture flasks and maintained at 37C / 5% CO2 in DMEM (4.5g/L D-Glucose, L-Glutamine, 110mg/L Sodium Pyruvate) with 10% Tet-free FBS (Clonetech Cat 631106). Cells were passaged at 80% confluence using 0.25% trypsin. The Bxb landing pad cell line should be maintained in 0.1-1 ug/ml doxycycline to maintain expression from the landing pad locus.

Packaging and titering of pLVX_mCherry_2A_MORlib_bc, pLVX_mCherry_2A_hMOR, pLVX_mCherry_2A_rMOR, pLVX_rMOR_2A_mCherry, and the N40D, N152D, and N188K variants were performed at the Fred Hutchinson Vector Core according to the manufacturer's instructions (Clonetech Lenti-X, Cat. 631349). Tet-On 3G expressing HEK293 cells were infected at 0.1 MOI in the presence of 8ug/ml protamine sulfate and media was replaced after 12-16 hours.

Flp-In HEK293 cells were transfected at ~50-80% confluence in T25 flasks using lipofectamine 3000 (Invitrogen Cat L300008). 2ug target plasmid DNA + 1ug pOG44 were mixed with 6ul P3000 in Optimem (125ul total). 6ul lipofectamine3000 was separately mixed with 119ul Optimem. After 5 minutes at room temperature, the DNA/P3000 solution was mixed with the lipofectamine3000 solution. After 30 minutes, the mixture was added to the culture flask and the media was changed after 6-8 hours. 24 hours after transfection, the cells were moved to 30C for 24 hours, then moved back to 37C for 24 hours, then split 1:4 into DMEM containing 100ug/ml hygromycin. Colonies were visible after 10 days and were ready to split after 14-16 days. For T225 flasks, all components of the transfection were scaled up 10 fold (3 fold for T75).

The integrated Bxb landing pad in the Bxb cell line contains the Bxb AttP site immediately downstream of the TRE3G promoter and immediately upstream of EGFP. Before integration, EGFP expression is Tet-regulated. After integration of the target plasmid, the gene of interest is inserted

downstream of the TRE3G promoter, disrupting EGFP expression. Bxb landing pad HEK293 cells were transfected as above except that doxycycline was removed from the media on day 1 and 3ug/9ug/30ug (T25, T75, T225) of the Bxb expression plasmid (pCAG-NLS-HA-Bxb1) was transfected alone on day 1. On day 3, the cells are transfected again with 3ug/9ug/30ug of the Bxb AttB-containing target plasmid. On day 4, the cells are split 1:4 into DMEM containing 1ug/ml doxycycline. On day 7-8, the transfected cells are sorted for integration (GFP-/mCherry+) and passaged for two weeks in DMEM+100 I.U/ml Pen/Strep+1ug/ml doxycycline. After two weeks, background expression of mCherry off of the unintegrated target plasmid is close to background and nearly all GFP-/mCherry+ cells are stable integrants.

Doxycycline dose was titrated by treating transfected HEK293 cells with 0.01, 0.05, 0.1, 0.5, 1, and 10 ug/ml for 24-72 hours, and all subsequent inductions were done at 1ug/ml doxycycline for >48 hours.

6.5.5 *Cell surface MOR labeling and flow cytometry*

All cell sorting was done on a BD FACSAria II (BD Biosciences 643188) and some flow cytometry was done on a BD LSR II (BD Biosciences) in the UW Pathology Flow Cytometry Core Facility. For cell surface labeling of the myc-tagged MOR, receptor expression was induced for >48 hours with 1ug/ml doxycycline. The cells were then trypsinized and resuspended in DMEM, washed twice with 5ml ice cold PBS+2%BSA, and then labeled with 1: 2000 Myc-Tag mouse mAb (cell signaling technology 9B11) in PBS+2%BSA for 45 minutes (50ul/10^6 cells). The cells were then washed 3x in 1ml ice cold PBS+2%BSA, and then labeled with 1:100 AlexaFluor 488 goat anti mouse (Jackson Immunoresearch 115-545-164) in PBS+2%BSA for 45 minutes (50ul/10^6 cells). Finally, the cells were washed 3x in 1ml ice cold PBS+2%BSA and resuspended in PBS+0.1uM DAPI. AlexaFluor 488 fluorescence was detected on the FITC channel and gain was set so that the WT MOR population was centered at ~10,000. mCherry fluorescence was assessed on the PE Texas Red channel and gain was set so that the induced integrated population was centered at ~10,000. DAPI fluorescence was detected on the BV450 channel and was used to gate out dead or dieing cells. For WT and mutant (N40D, N152D, N188K) rMOR controls, FITC gates were set on separate WT and N188K samples so that the WT population was split roughly in half by the higher two gates and the N188K population was split roughly in half by the lower two gates. The mixed WT/N188K population was then split into the 4 fluorescence bins and sorted into ice cold PBS.

For sorting the lentiviral hMOR library, cells were labeled as above, and gates were set as above using WT hMOR and N188K rMOR controls. For sorting the Bxb integrated hMOR library, gates were set so that the all 4 fluorescence bins contained roughly equal numbers of cells.

6.5.6 *Library preparation for sequencing*

Genomic DNA was extracted immediately after sorting using the DNeasy blood and tissue kit (Qiagen Cat 69504). The Bxb library barcode was amplified from 1-4ug of genomic DNA with Kapa polymerase and SYBR green using constant forward primer Bxb_F_p5 (AATGATACGGCGACCACCGAGATCTACACGgattagtgaacggatctcgacggtatcgcc), which binds to the genomic Bxb landing pad, and one of 4 variably indexed reverse primers that bind within the integrated target plasmid:
A_Bxb_R_UMI_p7(CAAGCAGAAGACGGCATACGAGATtccgtacgcTTNNNNNNNNNNCGTCGCC GTCCAGCTCGACCAG),
B_Bxb_R_UMI_p7(CAAGCAGAAGACGGCATACGAGATcgtacatcaaTTNNNNNNNNNNCGTCGCC GTCCAGCTCGACCAG),
C_Bxb_R_UMI_p7(CAAGCAGAAGACGGCATACGAGATaatgccaattTTNNNNNNNNNNCGTCGCC GTCCAGCTCGACCAG), and
D_Bxb_R_UMI_p7(CAAGCAGAAGACGGCATACGAGATctgagagacaTTNNNNNNNNNN CGTCGCCGTCCAGCTCGACCAG). Fluorescence was monitored after each cycle (98C 10s, 60C 30S, 72C 1 min) on a MiniOpticon and products were pulled at mid log phase (24-30 cycles). Samples were quantified using the Kapa Library Quantification Kit (Kapa Biosystems Cat KK4824) and sequenced on a NextSeq (Illumina) using a 75 cycle High output kit and primers Bxb_bcseq_F2(acatacaaactaaagaactacaaaaacaaattacaaaaattcaaaattttcgggtttattacagggacagcagagatccagtttatc gat), Bxb_bcseq_ind (cccatcctggtcgagctggacggcgacg), and tRNA_ind_hiseq_2 (reverse primer, GCAACGGAGCTGGTGGTGGGCAGGGAAG).

6.5.7 *Scoring cell surface expression of library variants*

Barcode sequences for each of the 4 bins were split, tallied (using a minimum q30 phred score cutoff), and matched to variants in the subassembly using a custom python script. Reads were converted to frequencies in each bin by dividing by the total number of reads in the corresponding bin. Read frequencies were converted to cell number estimates by multiplying by the number of cells collected in the corresponding bin during flow cytometry. An approximate cellular fluorescence score for each barcode (or each variant after combining degenerate barcodes) was obtained by multiplying the fraction

of cells for the barcode (or variant) in each bin by the median GFP/ median RFP for all cells in the corresponding bin and adding the results.

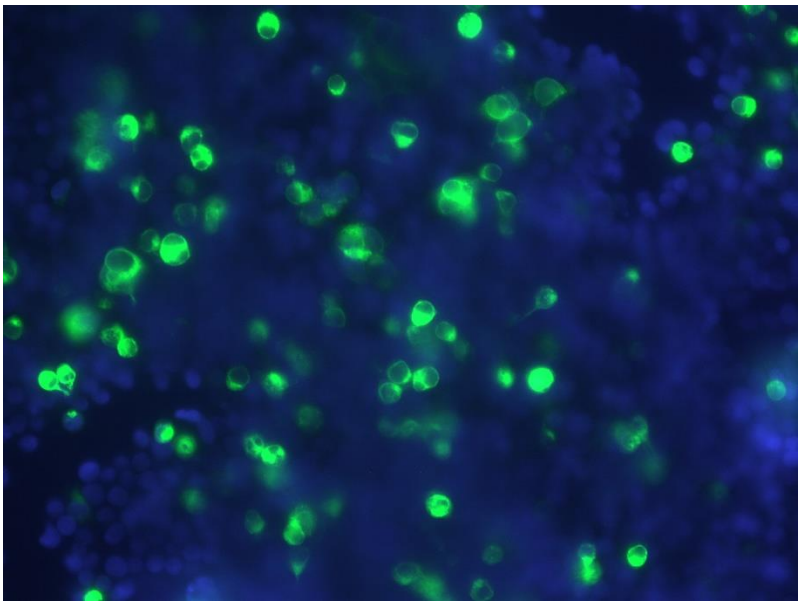


Figure S6.8 Verification of rMOR surface expression in HEK293 cells

Surface expression of myc-tagged rMOR from a transient transfection of HEK293 cells with the pcDNA3-mycMOR plasmid. Green, AF488 secondary antibody, Blue, DAPI

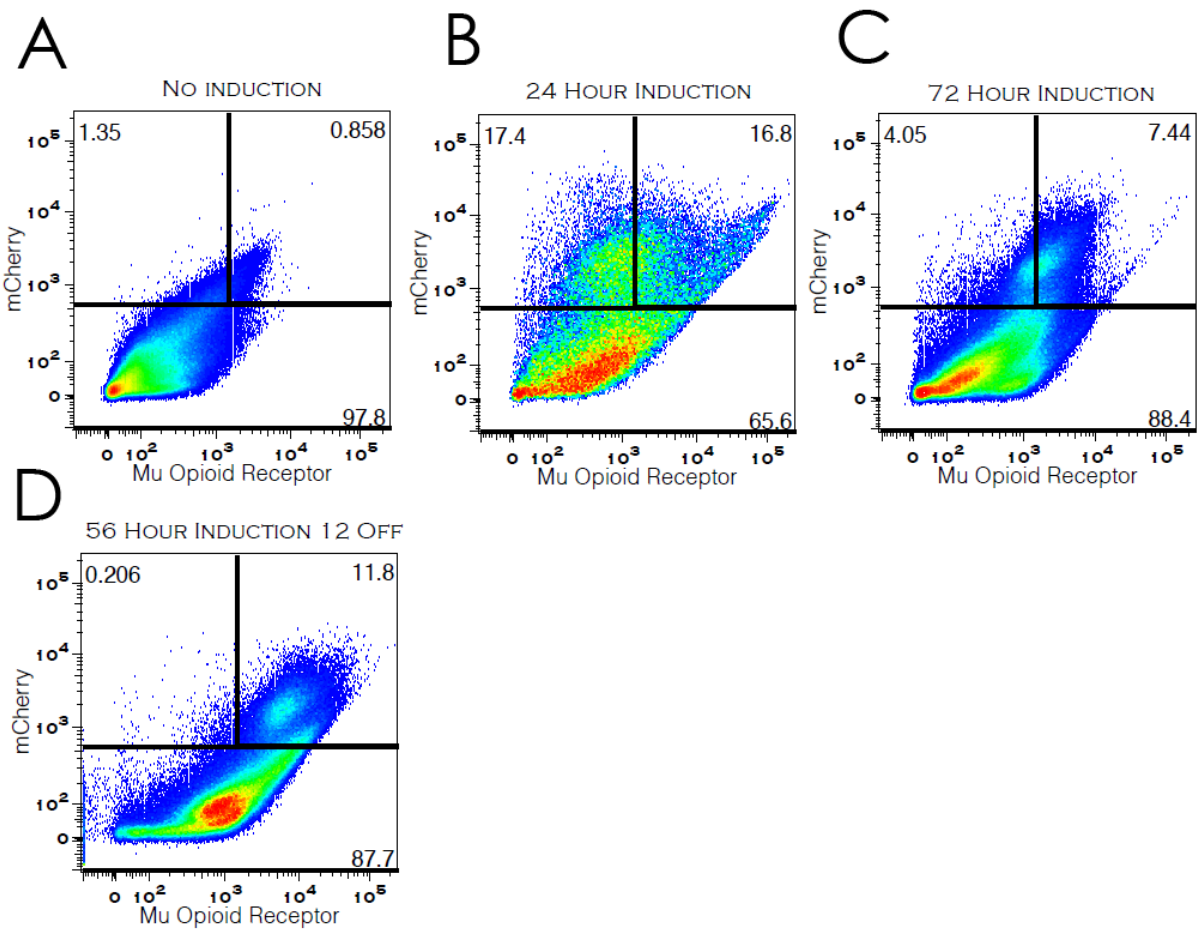


Figure S6.9 The effect of induction time on surface expression.

mCherry fluorescence, which marks translation, vs AlexaFluor488 fluorescence, which marks labeling of cell surface MOR after 0 (A), 24 (B), 72 (C), and 56 hours (with 12 hours off) (D) of 1ug/ml doxycycline induction of the TRE3G promoter

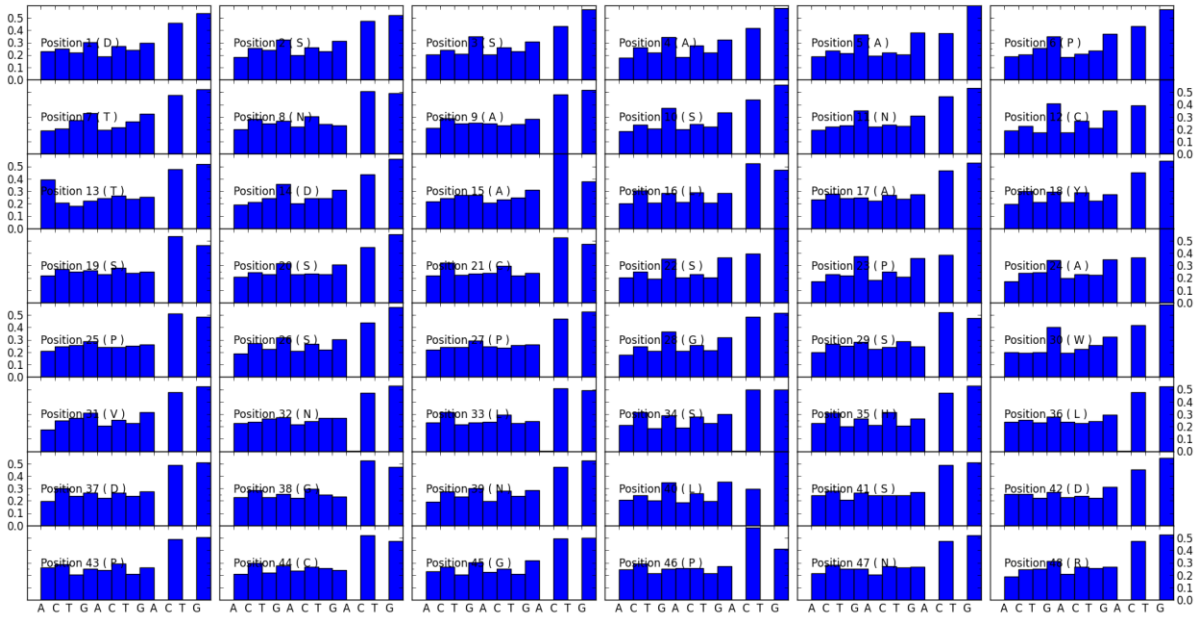


Figure S6.10 Bias in machine mixed mutagenic primers measured by high throughput sequencing

IDT machine mixed primers can show severe CG bias (see Table S6.1), but this bias shows batch to batch variability. IDT does not measure bias, but in this case offered to replace any plates that we found to be biased by sequencing. These graphs show the actual percentage, measured by high throughput sequencing, of each nucleotide at all 3 mutagenic positions of primers designed to substitute each of the first 48 codons of hMOR with NNS.

Table S6.1- Bias in mutagenic IDT oligos

	70C Annealing	60C Annealing	IDT Oligo
First Position			
A	8.7% (2/23)	20% (6/30)	20.5% (8/39)
T	4.4% (1/23)	10% (6/30)	17.9% (7/39)
C	56.5% (13/23)	40% (12/30)	15.4% (6/39)
G	30.4% (7/23)	30% (9/30)	46.2% (18/39)
Second Position			
A	8.7% (2/23)	6.7% (2/30)	15.4% (6/39)
T	8.7% (2/23)	20% (6/30)	17.9% (7/39)
C	65.2% (15/23)	43.3% (13/30)	7.7% (3/39)
G	17.4% (4/23)	30% (9/30)	59% (23/39)
Third Position			
A	21.7% (5/23)	10% (3/30)	2.6% (1/39)
T	4.4% (1/23)	20% (6/30)	2.6% (1/39)
C	56.5% (13/23)	36.6% (11/30)	38.4% (15/39)
G	17.4% (4/23)	33.3% (10/30)	56.4% (22/39)

After performing the two-step PCR library creation strategy on one codon, there was a distinct CG bias which was not completely corrected by lowering the annealing temperature. The mutagenic primers were then cloned into pGEM and sequenced. The G bias in the mutagenic oligos is likely responsible for the CG bias in the cloned inserts. This can be avoided by asking for hand mixed bases from IDT.

Chapter 7. DISCUSSION

The ability to functionally characterize large numbers of pooled variants provides an unprecedented view of the genotype-phenotype map, with many potential applications in the study of gene function, development, pathophysiology, and evolution. Through the application of deep mutational scanning, we have demonstrated a surprising robustness to mutation across several different genes, predicted sites of intermolecular interactions, generated the first large-scale, empirical estimates of the functional impact of first order epistasis, found specific examples of structural flexibility in proteins and tRNA, and generated improved predictions of a function tightly linked to the pathology of breast cancer.

7.1 DEEP MUTATION SCANNING AND ORGANISMAL PHENOTYPES

Deep mutational scanning and genome wide association studies are the high throughput extensions of reverse and forward genetics. GWASs, like forward genetic screens, start with the identification of an interesting phenotype and then attempt to identify its genetic determinants. Despite this fundamental similarity, the differences between GWAS and forward genetic screens make them somewhat complementary. Forward genetic screens enrich for large effect variants by experimental mutagenesis, but are much less likely to identify incompletely penetrant variants. GWAS, on the other hand, can identify variants with a larger range of phenotypic effect sizes, but variants with smaller effect sizes must be present at correspondingly higher frequencies in the studied population. Both methods are more likely to identify variants with larger phenotypic effects, though GWAS will miss even large effect variants if they are too rare. This makes both GWASs and forward genetic screens ill-suited for identifying the genetic determinants of complex phenotypes, which involve large numbers of genetic loci, each usually with a small phenotypic effect. DMS, as a high throughput application of reverse genetics, retains all the benefits of reverse genetics over forward genetics. By systematically analyzing all mutations in a single gene, DMS can identify phenotypically relevant mutations without the constraints on effect size and frequency in natural populations. These characteristics mean DMS and some related methods for studying promoters, enhancers, and other regulatory sequences are well suited for the interpretation of both small effect variants and the ever increasing number of rare and private variants. However, since DMS studies generally map genotypic variation onto low level phenotypes, or endophenotypes, such as cell growth and biochemical activities, these studies may in many cases be insufficient for predicting organismal phenotypes, especially in complex organisms such as humans. There are clearly exceptions in which endophenotypes are highly predictive of the organismal phenotype of interest, and these make for high value DMS targets. This is the case for BRCA1, in which homology

directed DNA repair (HDR) activity predicts increased breast cancer risk. However, mutations which abrogate HDR function in BRCA1 are specific, but not sensitive, as there are many other genes involved in breast cancer risk.

In the absence of a clear connection between lower level phenotypes and organismal phenotype, DMS data could potentially be used in conjunction with GWAS or other genotype/phenotype association studies to increase predictive power (Young and Fields, 2015). By abstracting genotypic variation to gene- or pathway-level endophenotypic variation, associations could instead be found between a smaller number of endophenotypes and the organismal phenotype of interest, increasing the number of observations and the statistical power to detect each association. A similar replacement strategy could be used in models of phenotypic variation that account for the marginal effect of all SNPs regardless of individual significance (Chatterjee *et al.*, 2013; Quang *et al.*, 2015; Stephan *et al.*, 2015). There are many factors that might limit the effectiveness of this strategy, some of which are shared with GWAS and not necessarily solved by the incorporation of DMS data, such as effects due to intermolecular epistasis and gene by environment interactions. Other factors include the accuracy and reproducibility of the DMS data and the choice in assayed phenotype, though this later factor may be mitigated somewhat by a correlation between different phenotypes for the same gene, as we observed for BRCA1. Regardless, it is clear that for many traits, the common variants accessible to identification by GWAS are insufficient to explain heritability, and methods which incorporate the effects of rare genetic variants may help remedy this (Harrison, 2015).

7.2 PROBING GENE BY ENVIRONMENT INTERACTIONS

DMS generates genotype-phenotype maps under experimental conditions, but organisms and their constitutive genes have evolved to function across a diverse set of conditions. It is likely that a subset of apparently neutral variants under one set of conditions might have fitness effects under another set of conditions. Conversely, large fitness effects observed in our experiments might be neutral in different genetic and environmental backgrounds. By performing each assay under different conditions, which in many cases requires little additional effort, this weakness can become a strength, providing information that can be used in determining the reason for the fitness effect. For example, we used this strategy to identify targets of the Rapid tRNA Decay (RTD) pathway by performing our flow based assay in a yeast strain lacking *met22*, an integral component of the RTD pathway. Variants throughout the tRNA were rescued by the *met22* deletion, providing at least a partial explanation for their fitness effects under WT conditions and demonstrating a greatly expanded role for RTD in tRNA quality control. Likewise, destabilizing mutations may in some cases be identified by fitness effects that vary with

temperature (data not shown). This systematic dissection of gene by environment interactions is unique to DMS methods. These effects are much harder to examine using conventional GWAS, since there is no opportunity to control the occurrence of individual genotypes across environmental conditions. Though it isn't feasible to test variants under the full range of possible physiological conditions, changes to temperature, culture conditions, and related genes can still be highly informative. We found that even in a single condition, clustering residues of PAB1 based on the fitness effects of all amino acid substitutions successfully grouped RNA-binding residues, core residues, and surface residues. Incorporating fitness effects across conditions might be expected to further improve the functional clustering of residues, with potential utility for engineering new gene functions and dissecting the function of pleiotropic genes.

7.3 TECHNICAL CHALLENGES

Despite the demonstrated and potential utility of DMS, there are still some barriers to its widespread adoption and there is substantial room for improvement in methods for library creation and library delivery. There are currently several methods for creating libraries of variants, each with trade-offs in cost, labor, length, and uniformity. The most common is the doped oligo approach, used in the studies of SUP4oc and PAB1, in which errors are introduced at a fixed rate randomly during DNA synthesis. The main drawbacks of this method are the size limitation (<200bp) and the cost, which is currently slightly higher than most other methods. In addition, the random introduction of mutations causes some reduction in library uniformity, which is the "flatness" of the variant frequency distribution. As errors are introduced one at a time, substitutions to codons with only one nucleotide difference from wild type end up at higher frequencies than substitutions to codons with two nucleotide differences. For double mutants especially, this limits the effective size of the library, since much larger assays and much more sequencing is required to functionally score very low frequency variants. The size limitation could potentially be overcome at a substantially increased cost by building and then mixing 2 libraries, each offset by about 30bp. By not mutagenizing the first and last 30bp of each fragment, the 2 complementary libraries could each be created from multiple mutagenized fragments using Gibson assembly, with mutations in the overlap regions present at only half frequency after mixing.

Probably the simplest method for library creation is error prone PCR, which uses an error prone polymerase during PCR to generate libraries of mutants. The method is as simple and fast as PCR, the mutation rate can be varied to an extent by varying levels of input DNA and $MgCl_2$, and relatively long libraries can be generated with this method, but bias in the types of substitutions made by the error prone polymerase causes even more non-uniformity than the doped oligo approach. The bias can be mitigated somewhat by using multiple error prone polymerases with different mutational spectra (Vanhercke *et al.*,

2005). Error prone polymerases also tend to create insertions and deletions, which are usually not desirable since frameshifts are likely to completely abolish function.

Another approach is overlap extension PCR (OEP), which I used to create the MOR library. This method uses separate mutagenic primer pairs for each codon position, allowing the incorporation of NNS at each position. The 5' and 3' fragments of the gene are amplified separately with the mutagenic primers and then overlapped and extended to regenerate the whole mutagenized gene. The main advantage of this method is that it allows much longer libraries to be created at a lower cost. The main disadvantages are that it requires the setup of 3 reactions for each mutagenized position, which is labor intensive for longer genes, and that the creation of double and triple mutants requires recursive library construction, which is even more labor intensive. On the other hand, the uniformity of OEP can be the best of all current methods, though it is dependent on the uniformity of the mutagenic primers, which tends to show batch to batch variability. As DNA concentration can be measured before mixing, position by position uniformity is very good for this method. The PALS method used in creating the BRCA1 libraries also enables the creation of long libraries, in this case using large numbers of primers synthesized on an array. It also has better price and labor scaling than OEP for libraries >1000bp. The main disadvantage is that uniformity can be poor, which becomes a bigger problem in larger libraries. No method is clearly superior, though doped oligos seem to provide a good trade-off in labor, cost, and uniformity for smaller targets.

DMS has in most cases been used for libraries <1000bp long, and there are several barriers to the application of DMS to larger genes. The first is that generating large, uniform libraries at low cost is currently not possible with any existing method. Large libraries are also more likely to be subject to bottlenecks during selection, resulting in variant drop out due to insufficient sampling, and this problem is exacerbated by non-uniformity. Finally, next generation sequencing platforms are currently limited in either throughput or read length. Illumina platforms can generate very high read numbers, making them the preferred platform for most DMS assays, but read length is limited to at most 600bp. This means that longer variants must be tagged with short barcodes that can be sequenced on Illumina machines. Barcode and variant identities are linked by the repeated sequencing of the barcode and different regions of the variant, usually accessed by digestion with a processive exonuclease for varying lengths of time. This process of linking barcode to variant is called tag-directed read grouping or subassembly (Hiatt *et al.*, 2010). Since DNA fragments longer than 1000bp can't form the necessary clusters on Illumina flow cells, some method, which is usually library specific, must be developed to reduce the barcode/variant fragment size for larger targets. The cost and difficulty of subassembly increases with the length, complexity, and non-uniformity of the library. The Pacific Biosciences platform yields longer reads, but at a significantly higher cost/read. Improvements in library uniformity and the development of new

sequencing methods that allow for longer read lengths will both be necessary to facilitate the application of DMS to larger full-length genes. In the mean time, these genes can be functionally characterized by separately analyzing smaller libraries that tile across the gene.

Most applications of DMS have so far been performed in phage and single-cell model organisms. Although these systems can be used successfully to study heterologous proteins, like BRCA1, many endophenotypes that are relevant to disease in humans are organism- or cell- type specific. Ideally, for these phenotypes, DMS should be carried out in a mammalian cell system. Functional selectivity in the MOR is one such phenotype, as it is dependent on the presence of intracellular signaling proteins, including β arrestin, that are not present in yeast. In attempting to port the method to HEK293 cells, I encountered several difficulties, most notably in library delivery. DMS requires that only one variant is expressed in each cell so that selection can be applied separately to each. Since plasmids are not replicated and transmitted by mammalian cells, and since all mammalian cell transfection methods deliver large numbers of plasmids to each cell, variants must be genomically integrated to guarantee the presence of only one variant in each cell. There are several methods for integrating foreign DNA into mammalian cells, including lentiviral vectors and sequence specific recombinases. Lentiviral vectors are problematic for barcoded libraries due to high levels of recombination during packaging. However, for smaller libraries that can be fully sequenced in one read, and for libraries in which increased complexity due to recombination can be tolerated, lentiviral vectors offer good integration efficiency, and the multiplicity of infection can be tuned to guarantee that most cells contain one variant. Of the recombinases, Bxb seems to be the most efficient, though efficiency is still usually <5%. This will likely be the bottleneck for larger libraries in mammalian cell systems, since even large 225cm^2 cell culture plates can only accommodate ~20 million cells. Transfections in these large plates require over 30ug of DNA and will result in at most 1 million integrants at 5% efficiency. Another potential bottleneck in mammalian cells is genome size. Since the human genome is about 250 times the size of the yeast genome, the library variants must be amplified from a much larger input of genomic DNA, which can be limiting for larger libraries. Though DMS in mammalian cell systems is currently possible for smaller libraries, improvements in high throughput sequencing read lengths and integration efficiencies would greatly facilitate the adoption of this tool.

7.4 OPPORTUNITIES

DMS is still in its infancy, and many potential applications of this tool have yet to be explored. By measuring functional effects across phenotypes and in different backgrounds, we can experimentally assess the hypothesized modularity or functional partitioning of residues in different types of

proteins(Halabi *et al.*, 2009). DMS also allows gene by gene and gene by environment interactions to be dissected at high resolution, but no one has yet performed a simultaneous selection on variant libraries of interacting genes. We have found that only a minority of single mutants display high amounts of functional non-additivity when combined within a single gene, but whether the cumulative effects of higher order epistasis, intermolecular epistasis, and environmental interactions preclude accurate prediction of phenotype remains to be seen. As more sequence-function datasets are generated for interacting and functionally related genes, we may begin to address these uncertainties for select low-level phenotypes. We found a small number of mutations that interact extensively with other mutations. If these positions can be identified by sequence context or functional context, then accurate phenotypic predictions may still be possible even in a setting of widespread, high magnitude epistasis. By applying DMS to phenotypes with strong connections to pathological processes, we might begin to determine the actual applicability and promise of personalized medicine.

Chapter 8. PUBLICATIONS

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

Adam, S. A., Nakagawa, T., Swanson, M. S., Woodruff, T. K., and Dreyfuss, G. (1986). mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. Mol. Cell. Biol. *6*, 2932–2943.

Adkar, B. V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., Swarnkar, M. K., Gokhale, R. S., and Varadarajan, R. (2012). Protein model discrimination using mutational sensitivity derived from deep sequencing. Structure *20*, 371–381.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Agris, P. F., Vendeix, F. A. P., and Graham, W. D. (2007). tRNA's Wobble Decoding of the Genome: 40 Years of Modification. J. Mol. Biol. *366*, 1–13.

Alexandrov, A., Chernyakov, I., Gu, W., Hiley, S. L., Hughes, T. R., Grayhack, E. J., and Phizicky, E. M. (2006). Rapid tRNA Decay Can Result from Lack of Nonessential Modifications. Mol. Cell *21*, 87–96.

Amrani, N., Ghosh, S., Mangus, D. A., and Jacobson, A. (2008). Translation factors promote the formation of two states of the closed-loop mRNP. Nature *453*, 1276-U85.

Araya, C. L., and Fowler, D. M. (2011). Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. *29*, 435–442.

Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., and Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc. Natl. Acad. Sci. *109*, 16858–16863.

Arttamangkul, S., Quillinan, N., Low, M. J., Zastrow, M. von, Pintar, J., and Williams, J. T. (2008). Differential Activation and Trafficking of μ-Opioid Receptors in Brain Slices. Mol. Pharmacol. *74*, 972–979.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. *38*, W529–W533.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. *79*, 137–158.

Baer, B. W., and Kornberg, R. D. (1983). The protein responsible for the repeating structure of cytoplasmic poly(A)-ribonucleoprotein. J. Cell Biol. *96*, 717–721.

Ballesteros, J. A., and Weinstein, H. (1995). [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. In: Methods in Neurosciences, ed. S. C. Sealfon, Academic Press, 366–428.

Basavappa, R., and Sigler, P. B. (1991). The 3 A crystal structure of yeast initiator tRNA: functional implications in initiator/elongator discrimination. EMBO J. *10*, 3105.

Befort, K., Filliol, D., Décaillot, F. M., Gavériaux-Ruff, C., Hoehe, M. R., and Kieffer, B. L. (2001). A Single Nucleotide Polymorphic Mutation in the Human μ-Opioid Receptor Severely Impairs Receptor Signaling. J. Biol. Chem. *276*, 3130–3137.

Berg, K. A., Maayani, S., Goldfarb, J., Scaramellini, C., Leff, P., and Clarke, W. P. (1998). Effector Pathway-Dependent Relative Efficacy at Serotonin Type 2A and 2C Receptors: Evidence for Agonist-Directed Trafficking of Receptor Stimulus. Mol. Pharmacol. *54*, 94–104.

Besenbacher, S. *et al.* (2015). Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat. Commun. *6*.

Beyer, A., Koch, T., Schröder, H., Schulz, S., and Höllt, V. (2004). Effect of the A118G polymorphism on binding affinity, potency and agonist-mediated endocytosis, desensitization, and resensitization of the human mu-opioid receptor. J. Neurochem. *89*, 553–560.

Bohn, L. M., Gainetdinov, R. R., Lin, F.-T., Lefkowitz, R. J., and Caron, M. G. (2000). μ-Opioid receptor desensitization by β-arrestin-2 determines morphine tolerance but not dependence. Nature *408*, 720–723.

Bohn, L. M., Lefkowitz, R. J., Gainetdinov, R. R., Peppel, K., Caron, M. G., and Lin, F.-T. (1999). Enhanced Morphine Analgesia in Mice Lacking β-Arrestin 2. Science *286*, 2495–2498.

Bonetti, B., Fu, L., Moon, J., and Bedwell, D. M. (1995). The Efficiency of Translation Termination is Determined by a Synergistic Interplay Between Upstream and Downstream Sequences inSaccharomyces cerevisiae. J. Mol. Biol. *251*, 334–345.

Bordner, A. J., and Abagyan, R. (2005). Statistical analysis and prediction of protein–protein interfaces. Proteins Struct. Funct. Bioinforma. *60*, 353–366.

Bottema, C., Ketterling, R., Ii, S., Yoon, H., Phillips, J., and Sommer, S. (1991). Missense Mutations and Evolutionary Conservation of Amino-Acids - Evidence That Many of the Amino-Acids in Factor-Ix Function as Spacer Elements. Am. J. Hum. Genet. *49*, 820–838.

Bouwman, P. *et al.* (2013). A High-Throughput Functional Complementation Assay for Classification of BRCA1 Missense Variants. Cancer Discov. *3*, 1142–1155.

Brzovic, P. S., Keeffe, J. R., Nishikawa, H., Miyamoto, K., Fox, D., Fukuda, M., Ohta, T., and Klevit, R. (2003). Binding and recognition in the assembly of an active BRCA1/BARD1 ubiquitin-ligase complex. Proc. Natl. Acad. Sci. *100*, 5646–5651.

Brzovic, P. S., Rajagopal, P., Hoyt, D. W., King, M.-C., and Klevit, R. E. (2001). Structure of a BRCA1–BARD1 heterodimeric RING–RING complex. Nat. Struct. Mol. Biol. *8*, 833–837.

Burd, C. G., Matunis, E. L., and Dreyfuss, G. (1991). The multiple RNA-binding domains of the mRNA poly(A)-binding protein have different RNA-binding activities. Mol. Cell. Biol. *11*, 3419–3424.

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., and Huang, E. S. (2004). Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. *13*, 190–202.

Cagney, G., Uetz, P., and Fields, S. (2000). [1] High-throughput screening for protein-protein interactions using two-hybrid assay. In: Methods in Enzymology, ed. S. D. E. and J. N. A. Jeremy Thorner, Academic Press, 3–14.

Callahan, B., Neher, R. A., Bachtrog, D., Andolfatto, P., and Shraiman, B. I. (2011). Correlated Evolution of Nearby Residues in Drosophilid Proteins. PLOS Genet *7*, e1001315.

Capra, J. A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875–1882.

Cawley, G. C., and Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res *11*, 2079–2107.

Chakrabarti, S., and Lanczycki, C. J. (2007). Analysis and prediction of functionally important sites in proteins. Protein Sci. *16*, 4–13.

Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S. J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405.

Chen, X.-T. *et al.* (2013). Structure–Activity Relationships and Discovery of a G Protein Biased μ Opioid Receptor Ligand, [(3-Methoxythiophen-2-yl)methyl]({2-[(9R)-9-(pyridin-2-yl)-6-oxaspiro-[4.5]decan-9-yl]ethyl})amine (TRV130), for the Treatment of Acute Severe Pain. J. Med. Chem. *56*, 8019–8031.

Cheng, G., Qian, B., Samudrala, R., and Baker, D. (2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res. *33*, 5861–5867.

Cherezov, V. *et al.* (2007). High-Resolution Crystal Structure of an Engineered Human β2-Adrenergic G Protein–Coupled Receptor. Science *318*, 1258–1265.

Chernyakov, I., Whipple, J. M., Kotelawala, L., Grayhack, E. J., and Phizicky, E. M. (2008). Degradation of several hypomodified mature tRNA species in Saccharomyces cerevisiae is mediated by Met22 and the 5′–3′ exonucleases Rat1 and Xrn1. Genes Dev. *22*, 1369–1380.

Chiba, N., and Parvin, J. D. (2001). Redistribution of BRCA1 among Four Different Protein Complexes following Replication Blockage. J. Biol. Chem. *276*, 38549–38554.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J. *5*, 823–826.

Chou, P., and Fasman, G. (1978). Empirical Predictions of Protein Conformation. Annu. Rev. Biochem. *47*, 251–276.

Christensen, D. E., Brzovic, P. S., and Klevit, R. E. (2007). E2–BRCA1 RING interactions dictate synthesis of mono- or specific polyubiquitin chain linkages. Nat. Struct. Mol. Biol. *14*, 941–948.

Clarkson, B. K., Gilbert, W. V., and Doudna, J. A. (2010). Functional Overlap between eIF4G Isoforms in Saccharomyces cerevisiae. PLOS ONE *5*, e9114.

Clery, A., Blatter, M., and Allain, F. H.-T. (2008). RNA recognition motifs: boring? Not quite. Curr. Opin. Struct. Biol. *18*, 290–298.

Cochella, L., and Green, R. (2005). An Active Role for tRNA in Decoding Beyond Codon:Anticodon Pairing. Science *308*, 1178–1180.

Consortium, E. A. *et al.* (2015). Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv, 30338.

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. Genome Res. *14*, 1188–1190.

Cunningham, B. C., and Wells, J. A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science *244*, 1081–1085.

Daniel Gietz, R., and Woods, R. A. (2002). Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. In: Methods in Enzymology, ed. C. G. and G. R. Fink, Academic Press, 87–96.

dbSNP Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD). National Center for Biotechnology Information, National Library of Medicine. {dbSNP Build ID: 144}.

Dean, K. M., and Grayhack, E. J. (2012). RNA-ID, a highly sensitive and robust method to identify cis-regulatory sequences using superfolder GFP and a fluorescence-based assay. RNA *18*, 2335–2344.

Deardorff, J. A., and Sachs, A. B. (1997). Differential effects of aromatic and charged residue substitutions in the RNA binding domains of the yeast poly(A) binding protein. J. Mol. Biol. *269*, 67–81.

Deo, R. C., Bonanno, J. B., Sonenberg, N., and Burley, S. K. (1999). Recognition of polyadenylate RNA by the poly(A)-binding protein. Cell *98*, 835–845.

Dereeper, A. *et al.* (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. *36*, W465–W469.

Deschenes-Furry, J., Perrone-Bizzozero, N., and Jasmin, B. J. (2006). The RNA-binding protein HuD: a regulator of neuronal differentiation, maintenance and plasticity. Bioessays *28*, 822–833.

Dewe, J. M., Whipple, J. M., Chernyakov, I., Jaramillo, L. N., and Phizicky, E. M. (2012). The yeast rapid tRNA decay pathway competes with elongation factor 1A for substrate tRNAs and acts on tRNAs lacking one or more of several modifications. RNA *18*, 1886–1896.

DeWire, S. M. *et al.* (2013). A G Protein-Biased Ligand at the μ-Opioid Receptor Is Potently Analgesic with Reduced Gastrointestinal and Respiratory Dysfunction Compared with Morphine. J. Pharmacol. Exp. Ther. *344*, 708–717.

Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat. Rev. Mol. Cell Biol. *3*, 195–205.

Dreyfuss, G., Swanson, M., and Pinolroma, S. (1988). Heterogeneous Nuclear Ribonucleoprotein-Particles and the Pathway of Messenger-Rna Formation. Trends Biochem. Sci. *13*, 86–91.

Drost, R. *et al.* (2011). BRCA1 RING Function Is Essential for Tumor Suppression but Dispensable for Therapy Resistance. Cancer Cell *20*, 797–809.

Dunham, M. J., and Fowler, D. M. (2013). Contemporary, yeast-based approaches to understanding human genetic variation. Curr. Opin. Genet. Dev. *23*.

Dymecki, S. M. (1996). Flp recombinase promotes site-specific DNA recombination in embryonic stem cells and transgenic mice. Proc. Natl. Acad. Sci. *93*, 6191–6196.

Dyson, H. J., Wright, P. E., and Scheraga, H. A. (2006). The role of hydrophobic interactions in initiation and propagation of protein folding. Proc. Natl. Acad. Sci. *103*, 13057–13061.

Easton, D. F. *et al.* (2007). A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer–Predisposition Genes. Am. J. Hum. Genet. *81*, 873–883.

Eggington, J. m. *et al.* (2014). A comprehensive laboratory-based program for classification of variants of uncertain significance in hereditary cancer genes. Clin. Genet. *86*, 229–237.

Erkmann, J. A., and Kutay, U. (2004). Nuclear export of mRNA: from the site of transcription to the cytoplasm. Exp. Cell Res. *296*, 12–20.

Fechter, P., Rudinger-Thirion, J., Théobald-Dietrich, A., and Giegé, R. (2000). Identity of tRNA for Yeast Tyrosyl-tRNA Synthetase: Tyrosylation Is More Sensitive to Identity Nucleotides Than to Structural Features. Biochemistry (Mosc.) *39*, 1725–1733.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. Int. J. Cancer *136*, E359–E386.

Finch A, Beiner M, Lubinski J, and et al (2006). SAlpingo-oophorectomy and the risk of ovarian, fallopian tube, and peritoneal cancers in women with a brca1 or brca2 mutation. JAMA *296*, 185–192.

Firnberg, E., Labonte, J. W., Gray, J. J., and Ostermeier, M. (2014). A comprehensive, high-resolution map of a gene's fitness landscape. Mol. Biol. Evol. *31*, 1581–1592.

Forsyth, C. M. *et al.* (2013). Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. In: MAbs, Taylor & Francis, 523–532.

Fortin, J.-P., Ci, L., Schroeder, J., Goldstein, C., Montefusco, M. C., Peter, I., Reis, S. E., Huggins, G. S., Beinborn, M., and Kopin, A. S. (2010). The μ-Opioid Receptor Variant N190K Is Unresponsive to Peptide Agonists yet Can be Rescued by Small-Molecule Drugs. Mol. Pharmacol. *78*, 837–845.

Foulkes, W. D., Knoppers, B. M., and Turnbull, C. (2016). Population genetic testing for cancer susceptibility: founder mutations to genomes. Nat. Rev. Clin. Oncol. *13*, 41–54.

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., and Fields, S. (2010a). High-resolution mapping of protein sequence-function relationships. Nat. Methods *7*, 741-U108.

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., and Fields, S. (2010b). High-resolution mapping of protein sequence-function relationships. Nat. Methods *7*, 741–746.

Fowler, D. M., Araya, C. L., Gerard, W., and Fields, S. (2011). Enrich: software for analysis of protein function by enrichment and depletion of variants. Bioinformatics *27*, 3430–3431.

Fowler, D. M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Gari, E., Piedrafita, L., Aldea, M., and Herrero, E. (1997). A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in Saccharomyces cerevisiae. Yeast *13*, 837–848.

Gautheret, D., Damberger, S. H., and Gutell, R. R. (1995). Identification of Base-triples in RNA using Comparative Sequence Analysis. J. Mol. Biol. *248*, 27–43.

Geslain, R., Martin, F., Camasses, A., and Eriani, G. (2003). A yeast knockout strain to discriminate between active and inactive tRNA molecules. Nucleic Acids Res. *31*, 4729–4737.

Gesty-Palmer, D. *et al.* (2006). Distinct β-Arrestin- and G Protein-dependent Pathways for Parathyroid Hormone Receptor-stimulated ERK1/2 Activation. J. Biol. Chem. *281*, 10856–10864.

Giaever, G. *et al.* (2002). Functional profiling of the Saccharomyces cerevisiae genome. Nature *418*, 387–391.

Giegé, R., Jühling, F., Pütz, J., Stadler, P., Sauter, C., and Florentz, C. (2012). Structure of transfer RNAs: similarity and variability. Wiley Interdiscip. Rev. RNA *3*, 37–61.

Giegé, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. Nucleic Acids Res. *26*, 5017–5035.

Goodenbour, J. M., and Pan, T. (2006). Diversity of tRNA genes in eukaryotes. Nucleic Acids Res. *34*, 6137–6146.

Goyer, C., Altmann, M., Lee, H. S., Blanc, A., Deshmukh, M., Woolford, J. L., Trachsel, H., and Sonenberg, N. (1993). TIF4631 and TIF4632: two yeast genes encoding the high-molecular-weight subunits of the cap-binding protein complex (eukaryotic initiation factor 4F) contain an RNA recognition motif-like sequence and carry out an essential function. Mol. Cell. Biol. *13*, 4860–4874.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science *185*, 862–864.

Gribaldo, S., and Philippe, H. (2002). Ancient Phylogenetic Relationships. Theor. Popul. Biol. *61*, 391–408.

Groer, C. E., Tidgewell, K., Moyer, R. A., Harding, W. W., Rothman, R. B., Prisinzano, T. E., and Bohn, L. M. (2007). An Opioid Agonist that Does Not Induce μ-Opioid Receptor—Arrestin Interactions or Receptor Internalization. Mol. Pharmacol. *71*, 549–557.

Gudipati, R. K., Xu, Z., Lebreton, A., Séraphin, B., Steinmetz, L. M., Jacquier, A., and Libri, D. (2012). Extensive Degradation of RNA Precursors by the Exosome in Wild-Type Cells. Mol. Cell *48*, 409–421.

Gueldener, U., Heinisch, J., Koehler, G. J., Voss, D., and Hegemann, J. H. (2002). A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. Nucleic Acids Res. *30*, e23–e23.

Guidugli, L. *et al.* (2014). Functional Assays for Analysis of Variants of Uncertain Significance in BRCA2. Hum. Mutat. *35*, 151–164.

Güldener, U., Heck, S., Fiedler, T., Beinhauer, J., and Hegemann, J. H. (1996). A New Efficient Gene Disruption Cassette for Repeated Use in Budding Yeast. Nucleic Acids Res. *24*, 2519–2524.

Gurwitz, D., Haring, R., Heldman, E., Fraser, C. M., Manor, D., and Fisher, A. (1994). Discrete activation of transduction pathways associated with acetylcholine m1 receptor by several muscarinic ligands. Eur. J. Pharmacol. *267*, 21–31.

Guy, M. P., Young, D. L., Payea, M. J., Zhang, X., Kon, Y., Dean, K. M., Grayhack, E. J., Mathews, D. H., Fields, S., and Phizicky, E. M. (2014). Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. Genes Dev. *28*, 1721–1732.

Haberstock-Debic, H., Kim, K.-A., Yu, Y. J., and Zastrow, M. von (2005). Morphine Promotes Rapid, Arrestin-Dependent Endocytosis of μ-Opioid Receptors in Striatal Neurons. J. Neurosci. *25*, 7847–7857.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell *138*, 774–786.

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., and King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science *250*, 1684–1689.

Hall, M. J., Reid, J. E., Burbidge, L. A., Pruss, D., Deffenbaugh, A. M., Frye, C., Wenstrup, R. J., Ward, B. E., Scholl, T. A., and Noll, W. W. (2009). BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. Cancer *115*, 2222–2233.

Han, M., Smith, S. O., and Sakmar, T. P. (1998). Constitutive Activation of Opsin by Mutation of Methionine 257 on Transmembrane Helix 6. Biochemistry (Mosc.) *37*, 8253–8261.

Hansen, D. F., Vallurupalli, P., and Kay, L. E. (2008). Using relaxation dispersion NMR spectroscopy to determine structures of excited, invisible protein states. J. Biomol. NMR *41*, 113–120.

Harrison, J. S., and Burton, R. S. (2006). Tracing Hybrid Incompatibilities to Single Amino Acid Substitutions. Mol. Biol. Evol. *23*, 559–564.

Harrison, P. J. (2015). Recent genetic findings in schizophrenia and their therapeutic relevance. J. Psychopharmacol. Oxf. Engl. *29*, 85–96.

Hashizume, R., Fukuda, M., Maeda, I., Nishikawa, H., Oyake, D., Yabuki, Y., Ogata, H., and Ohta, T. (2001). The RING Heterodimer BRCA1-BARD1 Is a Ubiquitin Ligase Inactivated by a Breast Cancer-derived Mutation. J. Biol. Chem. *276*, 14537–14540.

Heckman, K. L., and Pease, L. R. (2007). Gene splicing and mutagenesis by PCR-driven overlap extension. Nat. Protoc. *2*, 924–932.

Heemskerk-Gerritsen, B. a. M., Menke-Pluijmers, M. B. E., Jager, A., Tilanus-Linthorst, M. M. A., Koppert, L. B., Obdeijn, I. M. A., Deurzen, C. H. M. van, Collée, J. M., Seynaeve, C., and Hooning, M. J. (2013). Substantial breast cancer risk reduction and potential survival benefit after bilateral mastectomy when compared with surveillance in healthy BRCA1 and BRCA2 mutation carriers: a prospective analysis. Ann. Oncol. *24*, 2029–2035.

Hiatt, J. B., Patwardhan, R. P., Turner, E. H., Lee, C., and Shendure, J. (2010). Parallel, tag-directed assembly of locally derived short sequence reads. Nat. Methods *7*, 119–122.

Hietpas, R. T., Jensen, J. D., and Bolon, D. N. (2011). Experimental illumination of a fitness landscape. Proc. Natl. Acad. Sci. *108*, 7896–7901.

Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. Nat. Genet. *43*, 487–489.

Hoon, M. J. L. de, Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453–1454.

Hopper, A. K. (2013). Transfer RNA Post-Transcriptional Processing, Turnover, and Subcellular Dynamics in the Yeast Saccharomyces cerevisiae. Genetics *194*, 43–67.

Horovitz, A. (1996). Double-mutant cycles: a powerful tool for analyzing protein structure and function. Fold. Des. *1*, R121–R126.

Hou, Y. M., and Schimmel, P. (1992). Novel transfer RNAs that are active in Escherichia coli. Biochemistry (Mosc.) *31*, 4157–4160.

Imataka, H. (1998). A newly identified N-terminal amino acid sequence of human eIF4G binds poly(A)-binding protein and functions in poly(A)-dependent translation. EMBO J. *17*, 7480–7489.

Jackman, J. E., Montange, R. K., Malik, H. S., and Phizicky, E. M. (2003). Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. RNA *9*, 574–585.

Jacobson, A., and Favreau, M. (1983). Possible Involvement of poly(A) in protein synthesks. Nucleic Acids Res. *11*, 6353–6368.

James, P., Halladay, J., and Craig, E. A. (1996). Genomic Libraries and a Host Strain Designed for Highly Efficient Two-Hybrid Selection in Yeast. Genetics *144*, 1425–1436.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. Nat. Rev. Genet. *14*, 249–261.

Kadaba, S., Krueger, A., Trice, T., Krecic, A. M., Hinnebusch, A. G., and Anderson, J. (2004). Nuclear surveillance and degradation of hypomodified initiator tRNAMet in S. cerevisiae. Genes Dev. *18*, 1227–1240.

Kahsai, A. W., Xiao, K., Rajagopal, S., Ahn, S., Shukla, A. K., Sun, J., Oas, T. G., and Lefkowitz, R. J. (2011). Multiple ligand-specific conformations of the β2-adrenergic receptor. Nat. Chem. Biol. *7*, 692–700.

Kaiser, M. W., and Brow, D. A. (1995). Lethal mutations in a yeast U6 RNA gene B block promoter element identify essential contacts with transcription factor-IIIC. J. Biol. Chem. *270*, 11398–11405.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. *9*, 286–298.

Kessler, S. H., and Sachs, A. B. (1998). RNA recognition motif 2 of yeast Pab1p is required for its functional interaction with eukaryotic translation initiation factor 4G. Mol. Cell. Biol. *18*, 51–57.

Kim, I., Miller, C. R., Young, D. L., and Fields, S. (2013). High-throughput analysis of in vivo protein stability. Mol. Cell. Proteomics *12*, 3370–3378.

Kim, J. H., Lee, S.-R., Li, L.-H., Park, H.-J., Park, J.-H., Lee, K. Y., Kim, M.-K., Shin, B. A., and Choi, S.-Y. (2011). High Cleavage Efficiency of a 2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice. PLOS ONE *6*, e18556.

Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H. J., Seeman, N. C., and Rich, A. (1974). Three-Dimensional Tertiary Structure of Yeast Phenylalanine Transfer RNA. Science *185*, 435–440.

King, M.-C., Marks, J. H., and Mandell, J. B. (2003). Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. Science *302*, 643–646.

King M, Levy-Lahad E, and Lahad A (2014). Population-based screening for brca1 and brca2: 2014 lasker award. JAMA *312*, 1091–1092.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S., and Shendure, J. (2015). Massively parallel single-amino-acid mutagenesis. Nat. Methods *12*, 203–206.

Kohalmi, L., and Kunz, B. A. (1992). In vitro mutagenesis of the yeast SUP4-o gene to identify all substitutions that can be detected in vivo with the SUP4-o system. Environ. Mol. Mutagen. *19*, 282–287.

Kohout, T. A., Nicholas, S. L., Perry, S. J., Reinhart, G., Junger, S., and Struthers, R. S. (2004). Differential Desensitization, Receptor Phosphorylation, β-Arrestin Recruitment, and ERK1/2 Activation by the Two Endogenous Ligands for the CC Chemokine Receptor 7. J. Biol. Chem. *279*, 23214–23222.

Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. (2002). Dobzhansky–Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. *99*, 14878–14883.

Koski, R. A., Clarkson, S. G., Kurjan, J., Hall, B. D., and Smith, M. (1980). Mutations of the yeast SUP4 tRNATyr Locus: Transcription of the mutant genes in vitro. Cell *22*, 415–425.

Kramer, E. B., and Hopper, A. K. (2013). Retrograde transfer RNA nuclear import provides a new level of tRNA quality control in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. *110*, 21042–21047.

Kramer, E. B., Vallabhaneni, H., Mayer, L. M., and Farabaugh, P. J. (2010). A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. RNA *16*, 1797–1808.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*.

Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdel, S., and Wahle, E. (2009). Poly(A) Tail Length Is Controlled by the Nuclear Poly(A)-binding Protein Regulating the Interaction between Poly(A) Polymerase and the Cleavage and Polyadenylation Specificity Factor. J. Biol. Chem. *284*, 22803–22814.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.

Kurjan, J., and Hall, B. D. (1982). Mutations at the Saccharomyces cerevisiae SUP4 tRNATyr Locus: Isolation, Genetic Fine-Structure Mapping, and Correlation with Physical Structure. Mol. Cell. Biol. *2*, 1501–1513.

Kurjan, J., Hall, B. D., Gillam, S., and Smith, M. (1980). Mutations at the yeast SUP4 tRNATyr locus: DNA sequence changes in mutants lacking suppressor activity. Cell *20*, 701–709.

Kuroyanagi, H., Watanabe, Y., and Hagiwara, M. (2013). CELF Family RNA–Binding Protein UNC-75 Regulates Two Sets of Mutually Exclusive Exons of the unc-32 Gene in Neuron-Specific Manners in Caenorhabditis elegans. PLOS Genet *9*, e1003337.

Kurrasch-Orbaugh, D. M., Watts, V. J., Barker, E. L., and Nichols, D. E. (2003). Serotonin 5-Hydroxytryptamine2A Receptor-Coupled Phospholipase C and Phospholipase A2 Signaling Pathways Have Different Receptor Reserves. J. Pharmacol. Exp. Ther. *304*, 229–237.

Kwan, S. S., and Brow, D. A. (2005). The N- and C-terminal RNA recognition motifs of splicing factor Prp24 have distinct functions in U6 RNA binding. RNA *11*, 808–820.

Lagerström, M. C., and Schiöth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. Nat. Rev. Drug Discov. *7*, 339–357.

Lamb, K., Tidgewell, K., Simpson, D. S., Bohn, L. M., and Prisinzano, T. E. (2012). Antinociceptive effects of herkinorin, a MOP receptor agonist derived from salvinorin A in the formalin test in rats: New

concepts in mu opioid receptor pharmacology: From a symposium on new concepts in mu-opioid pharmacology. Drug Alcohol Depend. *121*, 181–188.

Ledoux, S., Olejniczak, M., and Uhlenbeck, O. C. (2009). A sequence element that tunes Escherichia coli tRNAAlaGGC to ensure accurate decoding. Nat. Struct. Mol. Biol. *16*, 359–364.

Lefkowitz, R. J., Roth, J., Pricer, W., and Pastan, I. (1970). ACTH Receptors in the Adrenal: Specific Binding of ACTH-125I and Its Relation to Adenyl Cyclase. Proc. Natl. Acad. Sci. U. S. A. *65*, 745–752.

Lefkowitz, R. J., and Shenoy, S. K. (2005). Transduction of Receptor Signals by ß-Arrestins. Science *308*, 512–517.

Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-Watson–Crick base pairs and their associated isostericity matrices. Nucleic Acids Res. *30*, 3497–3531.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. J. Mol. Biol. *257*, 342–358.

Lim, W. A., and Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of λrepresser. Nature *339*, 31–36.

Lindor, N. M., Guidugli, L., Wang, X., Vallée, M. P., Monteiro, A. N. A., Tavtigian, S., Goldgar, D. E., and Couch, F. J. (2012). A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS). Hum. Mutat. *33*, 8–21.

Ling, J., Reynolds, N., and Ibba, M. (2009). Aminoacyl-tRNA Synthesis and Translational Quality Control. Annu. Rev. Microbiol. *63*, 61–78.

Lovell, S. C., and Robertson, D. L. (2010). An Integrated View of Molecular Coevolution in Protein–Protein Interactions. Mol. Biol. Evol. *27*, 2567–2575.

Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. *8*, 479–490.

Luttrell, L. M. *et al.* (1999). β-Arrestin-Dependent Formation of β2 Adrenergic Receptor-Src Protein Kinase Complexes. Science *283*, 655–661.

Lynch, H. T., and Krush, A. J. (1971). Carcinoma of the breast and ovary in three families. Surg. Gynecol. Obstet. *133*, 644–648.

Lynch, H. T., Snyder, C., and Casey, M. J. (2013). Hereditary ovarian and breast cancer: what have we learned? Ann. Oncol. *24*, viii83-viii95.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database *2011*, bar009.

Mangus, D. A., Evans, M. C., and Jacobson, A. (2003). Poly (A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. Genome Biol *4*, 223.

Marck, C., and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA *8*, 1189–1232.

Marck, C., Kachouri-Lafond, R., Lafontaine, I., Westhof, E., Dujon, B., and Grosjean, H. (2006). The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. Nucleic Acids Res. *34*, 1816–1835.

Marini, N. J., Thomas, P. D., and Rine, J. (2010). The Use of Orthologous Sequences to Predict the Impact of Amino Acid Substitutions on Protein Function. PLOS Genet *6*, e1000968.

Maris, C., Dominguez, C., and Allain, F. H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. *272*, 2118–2131.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc. Natl. Acad. Sci. U. S. A. *101*, 7287–7292.

McLaughlin Jr, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., and Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly (A)-binding protein. RNA *19*, 1537–1551.

Melamed, D., Young, D. L., Miller, C. R., and Fields, S. (2015). Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites. PLoS Genet. *11*, e1004918–e1004918.

Melief, E. J., Miyatake, M., Bruchas, M. R., and Chavkin, C. (2010). Ligand-directed c-Jun N-terminal kinase activation disrupts opioid receptor signaling. Proc. Natl. Acad. Sci. *107*, 11608–11613.

Melnikov, A., Rogov, P., Wang, L., Gnirke, A., and Mikkelsen, T. S. (2014). Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res., gku511.

Miki, Y. *et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66–71.

Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. Proc. Natl. Acad. Sci. U. S. A. *102*, 10930–10935.

Mirny, L. A., and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1. J. Mol. Biol. *291*, 177–196.

Mittal, N., Tan, M., Egbuta, O., Desai, N., Crawford, C., Xie, C.-W., Evans, C., and Walwyn, W. (2012). Evidence that Behavioral Phenotypes of Morphine in β-arr2|[minus]|/|[minus]| Mice Are Due to the Unmasking of JNK Signaling. Neuropsychopharmacology *37*, 1953–1962.

Mody, A., Weiner, J., and Ramanathan, S. (2009). Modularity of MAP kinases allows deformation of their signalling pathways. Nat. Cell Biol. *11*, 484–491.

Morris, J. R., Pangon, L., Boutell, C., Katagiri, T., Keep, N. H., and Solomon, E. (2006). Genetic analysis of BRCA1 ubiquitin ligase activity and its relationship to breast cancer susceptibility. Hum. Mol. Genet. *15*, 599–606.

Moutsianas, L. *et al.* (2015). The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease.

Moynahan, M. E., Chiu, J. W., Koller, B. H., and Jasin, M. (1999). Brca1 Controls Homology-Directed DNA Repair. Mol. Cell *4*, 511–518.

Mumberg, D., Müller, R., and Funk, M. (1995). Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. Gene *156*, 119–122.

Murphy, F. V., Ramakrishnan, V., Malkiewicz, A., and Agris, P. F. (2004). The role of modifications in codon discrimination by tRNALysUUU. Nat. Struct. Mol. Biol. *11*, 1186–1191.

Musier-Forsyth, K., Usman, N., Scaringe, S., Doudna, J., Green, R., and Schimmel, P. (1991). Specificity for aminoacylation of an RNA helix: an unpaired, exocyclic amino group in the minor groove. Science *253*, 784–786.

Muto, Y., and Yokoyama, S. (2012). Structural insight into RNA recognition motifs: versatile molecular Lego building blocks for biological systems. Wiley Interdiscip. Rev. RNA *3*, 229–246.

Ng, P. C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Nimwegen, E. van, Crutchfield, J. P., and Huynen, M. (1999). Neutral evolution of mutational robustness. Proc. Natl. Acad. Sci. *96*, 9716–9720.

Normanly, J., Ogden, R. C., Horvath, S. J., and Abelson, J. (1986). Changing the identity of a transfer RNA. Nature *321*, 213–219.

Orioli, A., Pascali, C., Pagano, A., Teichmann, M., and Dieci, G. (2012). RNA polymerase III transcription control elements: Themes and variations. Gene *493*, 185–194.

Orr, H. A. (1995). The Population Genetics of Speciation: The Evolution of Hybrid Incompatibilities. Genetics *139*, 1805–1813.

Otero, L. J. (1999). The yeast poly(A)-binding protein Pab1p stimulates invitro poly(A)-dependent and cap-dependent translation by distinct mechanisms. EMBO J. *18*, 3153–3163.

Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. Elife *3*, e02030.

Park, E.-H., Walker, S. E., Lee, J. M., Rothenburg, S., Lorsch, J. R., and Hinnebusch, A. G. (2011). Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1 center dot PABP mRNPs in vivo. Embo J. *30*, 302–316.

Patwardhan, R. P. *et al.* (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat. Biotechnol. *30*, 265–270.

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat. Biotechnol. *27*, 1173–1175.

Pavlidis, P., and Noble, W. S. (2003). Matrix2png: a utility for visualizing matrix data. Bioinformatics *19*, 295–296.

Pearson, D., Willis, I., Hottinger, H., Bell, J., Kumar, A., Leupold, U., and Söll, D. (1985). Mutations preventing expression of sup3 tRNASer nonsense suppressors of Schizosaccharomyces pombe. Mol. Cell. Biol. *5*, 808–815.

Pruitt, K. D., Tatusova, T., Brown, G. R., and Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. *40*, D130–D135.

Pupo, A. S., Duarte, D. A., Lima, V., Teixeira, L. B., Parreiras-e-Silva, L. T., and Costa-Neto, C. M. Recent updates on GPCR biased agonism. Pharmacol. Res.

Pütz, J., Florentz, C., Benseler, F., and Giegé, R. (1994). A single methyl group prevents the mischarging of a tRNA. Nat. Struct. Mol. Biol. *1*, 580–582.

Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics *31*, 761–763.

Raehal, K. M., Walker, J. K. L., and Bohn, L. M. (2005). Morphine Side Effects in β-Arrestin 2 Knockout Mice. J. Pharmacol. Exp. Ther. *314*, 1195–1201.

Randau, L., Stanley, B. J., Kohlway, A., Mechta, S., Xiong, Y., and Söll, D. (2009). A Cytidine Deaminase Edits C to U in Transfer RNAs in Archaea. Science *324*, 657–659.

Ransburgh, D. J. R., Chiba, N., Ishioka, C., Toland, A. E., and Parvin, J. D. (2010). Identification of Breast Tumor Mutations in BRCA1 That Abolish Its Function in Homologous DNA Recombination. Cancer Res. *70*, 988–995.

Raote, I., Bhattacharyya, S., and Panicker, M. M. (2013). Functional Selectivity in Serotonin Receptor 2A (5-HT2A) Endocytosis, Recycling, and Phosphorylation. Mol. Pharmacol. *83*, 42–50.

Rasmussen, S. G. F. *et al.* (2011a). Crystal structure of the β2 adrenergic receptor-Gs protein complex. Nature *477*, 549–555.

Rasmussen, S. G. F. *et al.* (2011b). Structure of a nanobody-stabilized active state of the β2 adrenoceptor. Nature *469*, 175–180.

Reid, L. J., Shakya, R., Modi, A. P., Lokshin, M., Cheng, J.-T., Jasin, M., Baer, R., and Ludwig, T. (2008). E3 ligase activity of BRCA1 is not essential for mammalian cell viability or homology-directed repair of double-strand DNA breaks. Proc. Natl. Acad. Sci. *105*, 20876–20881.

Reuter, J. S., and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics *11*, 129.

Richardson, R., Denis, C. L., Zhang, C., Nielsen, M. E. O., Chiang, Y.-C., Kierkegaard, M., Wang, X., Lee, D. J., Andersen, J. S., and Yao, G. (2012). Mass spectrometric identification of proteins that interact through specific domains of the poly(A) binding protein. Mol. Genet. Genomics *287*, 711–730.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics *12*, 77.

Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D., and Bolon, D. N. (2013). Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J. Mol. Biol. *425*, 1363–1377.

Rosenbaum, D. M. *et al.* (2007). GPCR Engineering Yields High-Resolution Structural Insights into β2-Adrenergic Receptor Function. Science *318*, 1266–1273.

Ruiz-Pesini, E., Lott, M. T., Procaccio, V., Poole, J. C., Brandon, M. C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., and Wallace, D. C. (2007). An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res. *35*, D823–D828.

Saadatmand, S. *et al.* (2014). Relevance and efficacy of breast cancer screening in BRCA1 and BRCA2 mutation carriers above 60 years: A national cohort study. Int. J. Cancer *135*, 2940–2949.

Sachs, A. B., Bond, M. W., and Kornberg, R. D. (1986). A single gene from yeast for both nuclear and cytoplasmic polyadenylate-binding proteins: Domain structure and expression. Cell *45*, 827–835.

Sachs, A. B., Davis, R. W., and Kornberg, R. D. (1987). A single domain of yeast poly(A)-binding protein is necessary and sufficient for RNA binding and cell viability. Mol. Cell. Biol. *7*, 3268–3276.

Safaee, N., Kozlov, G., Noronha, A. M., Xie, J., Wilds, C. J., and Gehring, K. (2012). Interdomain Allostery Promotes Assembly of the Poly(A) mRNA Complex with PABP and eIF4G. Mol. Cell *48*, 375–386.

Saldanha, A. J. (2004). Java Treeview—extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Schlinkmann, K. M., Honegger, A., Türeci, E., Robison, K. E., Lipovšek, D., and Plückthun, A. (2012). Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. Proc. Natl. Acad. Sci. *109*, 9810–9815.

Schmeing, T. M., Voorhees, R. M., Kelley, A. C., Gao, Y.-G., Murphy, F. V., Weir, J. R., and Ramakrishnan, V. (2009). The Crystal Structure of the Ribosome Bound to EF-Tu and Aminoacyl-tRNA. Science *326*, 688–694.

Schrader, J. M., Chapman, S. J., and Uhlenbeck, O. C. (2009). Understanding the Sequence Specificity of tRNA Binding to Elongation Factor Tu using tRNA Mutagenesis. J. Mol. Biol. *386*, 1255–1264.

Schultz, D. W., and Yarus, M. (1994). tRNA Structure and Ribosomal Function. J. Mol. Biol. *235*, 1381–1394.

Shakya, R. *et al.* (2011). BRCA1 Tumor Suppression Depends on BRCT Phosphoprotein Binding, But Not Its E3 Ligase Activity. Science *334*, 525–528.

Shalem, O., Sharon, E., Lubliner, S., Regev, I., Lotan-Pompan, M., Yakhini, Z., and Segal, E. (2015). Systematic Dissection of the Sequence Determinants of Gene 3'End Mediated Expression Control.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Shepotinovskaya, I., and Uhlenbeck, O. C. (2013). tRNA residues evolved to promote translational accuracy. RNA *19*, 510–516.

Sievers, F. *et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

Simon, R., and Roychowdhury, S. (2013). Implementing personalized cancer genomics in clinical trials. Nat. Rev. Drug Discov. *12*, 358–369.

Sowa, M. E., He, W., Slep, K. C., Kercher, M. A., Lichtarge, O., and Wensel, T. G. (2001). Prediction and confirmation of a site critical for effector regulation of RGS domain activity. Nat. Struct. Mol. Biol. *8*, 234–237.

Soylemez, O., and Kondrashov, F. A. (2012). Estimating the Rate of Irreversibility in Protein Evolution. Genome Biol. Evol. *4*, 1213–1222.

Spongier, D., Waeber, C., Pantaloni, C., Holsboer, F., Bockaert, J., Seeburgt, P. H., and Journot, L. (1993). Differential signal transduction by five splice variants of the PACAP receptor. Nature *365*, 170–175.

Starita, L. M., Pruneda, J. N., Lo, R. S., Fowler, D. M., Kim, H. J., Hiatt, J. B., Shendure, J., Brzovic, P. S., Fields, S., and Klevit, R. E. (2013). Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proc. Natl. Acad. Sci. *110*, E1263–E1272.

Starita, L. M., Young, D. L., Islam, M., Kitzman, J. O., Gullingsrud, J., Hause, R. J., Fowler, D. M., Parvin, J. D., Shendure, J., and Fields, S. (2015). Massively parallel functional analysis of BRCA1 RING domain variants. Genetics, genetics–115.

Stephan, J., Stegle, O., and Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. Nat. Commun. *6*.

Sterner, T., Jansen, M., and Hou, Y. M. (1995). Structural and functional accommodation of nucleotide variations at a conserved tRNA tertiary base pair. RNA *1*, 841–851.

Stevens, R. C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., Rosen, H., and Wüthrich, K. (2013). The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. Nat. Rev. Drug Discov. *12*, 25–34.

Stiffler, M. A., Hekstra, D. R., and Ranganathan, R. (2015). Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. Cell *160*, 882–892.

Stone, E. A., and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. *15*, 978–986.

Süel, G. M., Lockless, S. W., Wall, M. A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Mol. Biol. *10*, 59–69.

Suzuki, T., Nagao, A., and Suzuki, T. (2011). Human Mitochondrial tRNAs: Biogenesis, Function, Structural Aspects, and Diseases. Annu. Rev. Genet. *45*, 299–329.

Swanson, M. S., Nakagawa, T. Y., LeVan, K., and Dreyfuss, G. (1987). Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. Mol. Cell. Biol. *7*, 1731–1739.

Tarun, S. Z., and Sachs, A. B. (1996). Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. EMBO J. *15*, 7168–7177.

Tarun, S. Z., Wells, S. E., Deardorff, J. A., and Sachs, A. B. (1997). Translation initiation factor eIF4G mediates in vitro poly(A) tail-dependent translation. Proc. Natl. Acad. Sci. *94*, 9046–9051.

Tavtigian, S. V., Deffenbaugh, A. M., Yin, L., Judkins, T., Scholl, T., Samollow, P. B., Silva, D. de, Zharkikh, A., and Thomas, A. (2006). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J. Med. Genet. *43*, 295–305.

Thyagarajan, B., and Bloom, J. D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife *3*, e03300.

Towler, W. I., Zhang, J., Ransburgh, D. J. R., Toland, A. E., Ishioka, C., Chiba, N., and Parvin, J. D. (2013). Analysis of BRCA1 Variants in Double-Strand Break Repair by Homologous Recombination and Single-Strand Annealing. Hum. Mutat. *34*, 439–445.

Traxlmayr, M. W., Hasenhindl, C., Hackl, M., Stadlmayr, G., Rybka, J. D., Borth, N., Grillari, J., Rüker, F., and Obinger, C. (2012). Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. J. Mol. Biol. *423*, 397–412.

Unal, H., and Karnik, S. S. (2012). Domain coupling in GPCRs: the engine for induced conformational changes. Trends Pharmacol. Sci. *33*, 79–88.

Valle, M., Zavialov, A., Li, W., Stagg, S. M., Sengupta, J., Nielsen, R. C., Nissen, P., Harvey, S. C., Ehrenberg, M., and Frank, J. (2003). Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. Nat. Struct. Mol. Biol. *10*, 899–906.

Vanhercke, T., Ampe, C., Tirry, L., and Denolf, P. (2005). Reducing mutational bias in random protein libraries. Anal. Biochem. *339*, 9–14.

Vega, A. *et al.* (2001). The R71G BRCA1 is a founder Spanish mutation and leads to aberrant splicing of the transcript. Hum. Mutat. *17*, 520–521.

Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F., and Babu, M. M. (2013). Molecular signatures of G-protein-coupled receptors. Nature *494*, 185–194.

Viscusi, E. R., Webster, L., Kuss, M., Daniels, S., Bolognese, J. A., Zuckerman, S., Soergel, D. G., Subach, R. A., Cook, E., and Skobieranda, F. (2016). A randomized, phase 2 study investigating TRV130, a biased ligand of the μ-opioid receptor, for the intravenous treatment of acute pain. Pain *157*, 264–272.

Wacker, D. *et al.* (2013). Structural Features for Functional Selectivity at Serotonin Receptors. Science *340*, 615–619.

Wagner, A. (2012). The role of robustness in phenotypic adaptation and innovation. Proc. R. Soc. Lond. B Biol. Sci. *279*, 1249–1258.

Walhout, A. J. M., Boulton, S. J., and Vidal, M. (2000). Yeast Two-Hybrid Systems and Protein Interaction Mapping Projects for Yeast and Worm. Yeast Chichester Engl. *17*, 88–94.

Warne, T., Moukhametzianov, R., Baker, J. G., Nehmé, R., Edwards, P. C., Leslie, A. G. W., Schertler, G. F. X., and Tate, C. G. (2011). The structural basis for agonist and partial agonist action on a β1-adrenergic receptor. Nature *469*, 241–244.

Wei, H., Ahn, S., Shenoy, S. K., Karnik, S. S., Hunyady, L., Luttrell, L. M., and Lefkowitz, R. J. (2003). Independent β-arrestin 2 and G protein-mediated pathways for angiotensin II activation of extracellular signal-regulated kinases 1 and 2. Proc. Natl. Acad. Sci. *100*, 10782–10787.

Wells, S. E., Hillner, P. E., Vale, R. D., and Sachs, A. B. (1998). Circularization of mRNA by Eukaryotic Translation Initiation Factors. Mol. Cell *2*, 135–140.

Westhof, E., Dumas, P., and Moras, D. (1985). Crystallographic refinement of yeast aspartic acid transfer RNA. J. Mol. Biol. *184*, 119–145.

Whipple, J. M., Lane, E. A., Chernyakov, I., D'Silva, S., and Phizicky, E. M. (2011). The yeast rapid tRNA decay pathway primarily monitors the structural integrity of the acceptor and T-stems of mature tRNA. Genes Dev. *25*, 1173–1184.

Whistler, J. L., and Zastrow, M. von (1998). Morphine-activated opioid receptors elude desensitization by β-arrestin. Proc. Natl. Acad. Sci. *95*, 9914–9919.

White, J. F. *et al.* (2012). Structure of the agonist-bound neurotensin receptor. Nature *490*, 508–513.

Wilden, U., Hall, S. W., and Kühn, H. (1986). Phosphodiesterase activation by photoexcited rhodopsin is quenched when rhodopsin is phosphorylated and binds the intrinsic 48-kDa protein of rod outer segments. Proc. Natl. Acad. Sci. *83*, 1174–1178.

Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D., and Wishart, D. S. (2003). VADAR: a web server for quantitative evaluation of protein structure quality. Nucleic Acids Res. *31*, 3316–3319.

Williamson, R. M. (1995). Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J. Theor. Biol. *174*, 179–188.

Wu, W., Sato, K., Koike, A., Nishikawa, H., Koizumi, H., Venkitaraman, A. R., and Ohta, T. (2010). HERC2 Is an E3 Ligase That Targets BRCA1 for Degradation. Cancer Res. *70*, 6384–6392.

Xie, G., Gross, A. K., and Oprian, D. D. (2003). An Opsin Mutant with Increased Thermal Stability. Biochemistry (Mosc.) *42*, 1995–2001.

Xu, Z., Thomas, L., Davies, B., Chalmers, R., Smith, M., and Brown, W. (2013). Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. BMC Biotechnol. *13*, 87.

Yan, W., and Francklyn, C. (1994). Cytosine 73 is a discriminator nucleotide in vivo for histidyl-tRNA in Escherichia coli. J. Biol. Chem. *269*, 10022–10027.

Yao, G., Chiang, Y.-C., Zhang, C., Lee, D. J., Laue, T. M., and Denis, C. L. (2007). PAB1 Self-Association Precludes Its Binding to Poly(A), Thereby Accelerating CCR4 Deadenylation In Vivo. Mol. Cell. Biol. *27*, 6243–6253.

Yarham, J. W., Elson, J. L., Blakely, E. L., McFarland, R., and Taylor, R. W. (2010). Mitochondrial tRNA mutations and disease. Wiley Interdiscip. Rev. - RNA *1*, 304–324.

Young, D. L., and Fields, S. (2015). The role of functional data in interpreting the effects of genetic variation. Mol. Biol. Cell *26*, 3904–3908.

Yurgelun, M. B., Hiller, E., and Garber, J. E. (2015). Population-Wide Screening for Germline BRCA1 and BRCA2 Mutations: Too Much of a Good Thing? J. Clin. Oncol. *33*, 3092–3095.

Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011a). NUPACK: Analysis and design of nucleic acid systems. J. Comput. Chem. *32*, 170–173.

Zadeh, J. N., Wolfe, B. R., and Pierce, N. A. (2011b). Nucleic acid sequence design via efficient ensemble defect optimization. J. Comput. Chem. *32*, 439–452.

Zhang, N., Osborn, M., Gitsham, P., Yen, K., Miller, J. R., and Oliver, S. G. (2003). Using yeast to place human genes in functional categories. Gene *303*, 121–129.

Zhou, J., Lancaster, L., Donohue, J. P., and Noller, H. F. (2013). Crystal Structures of EF-G–Ribosome Complexes Trapped in Intermediate States of Translocation. Science *340*, 1236086.