Using deep mutagenesis to understand genetic and physical interactions

Júlia Domingo Espinós

TESI DOCTORAL UPF 2019

Dr. Ben Lehner

GENETIC SYSTEMS

SYSTEMS BIOLOGY DEPARTMENT

CENTRE FOR GENOMIC REGULATION (CRG)

Al Marc,

Acknowledgements

Ben, because in March of 2015 you gave me the best birthday present when you told me that I could start the PhD in your lab. For hiring me despite not bringing any fellowship. For your incredible guidance. For bringing together the amazing people of the Lehner lab. You are an outstanding scientist and mentor, and I feel extremely lucky and thankful for having been under your supervision during these last four years.

Pablo, the one and only Pablo. Por estar siempre dispuesto a ayudar. Por las mil y una english corrections. Porque contigo es difícil no reírse. Por nuestras charlas en *spanglish* científico. No podría haber tenido más suerte en encontrar un compañero de PhD como tu.

Guillaume, for your patience and thoughtful mentoring. Until it was my moment to teach someone, I hadn't realized how much I learned from you. For always finding the right words to cheer me up.

Jörn, because working hand-by-hand with you is too easy. Because, quite simply, I like to spend time and chat with you during a coffee, or even better, a beer.

Cici, for being the most joyful and positive person I have ever known. For your enthusiasm and determination.

Benni, for your happiness and passion for science. For always being there when I needed advice. For being a role model.

Solip, for being the strongest woman I have ever known. For your strange sense of humor. I miss you.

Cristina, por procurar ayudarme en todo lo que he necesitado durante el doctorado. Por tus risas y ganas de procrastinar tan necesarias.

Mirko, por esas charlas trascendentales sobre la vida a las 8 de la noche en el lab. Because the lab is not the same without you. For introducing Alessa to the lab—she is the epitome of kindness.

Andre, for always coming back to us. The lab needs you, and so do I.

Marcos, for the coffee and chocolate-time in the terrace. Because you are one of the most brilliant minds I have ever known.

Aaron, for your wisdom. Per compartir amb mi la teva passió per la cultura catalana.

Jeni, Adam, Kadri and Fran, the first to leave when I arrived. For making of the adaptation to the lab a smooth transition.

Sarah, Thomas and Mishan, the last ones to join. For being always willing to listen and help.

Laure, for reminding me how much I enjoy teaching. For your enthusiasm to learn and good taste in music.

To my thesis committee—Manu, Lucas and Guillaume—for your compromise and feedback given in our annual meetings

To the graduate committee I had the pleasure to belong as a PhD representative. Imma, Anna and Fátima, for caring so much for PhD students. Hima, Hana, Bogu, Fer and Claudia, with whom I shared my responsibilities, for making this an engaging activity.

Laura. Perquè amb tu m'enduc una amiga de per vida. Per sempre estar disposada a escoltar i proveir-me dels consells més savis.

Antonio, Vicky, Yami, Demi, Javi y Laura. Por ser mi laboratorio adoptivo. Por todas las cenas, borracheras y risas.

Cate, Fer, Paul, Claudia, Moritz, Dand, Fran and Hannah. For all the beers in Bambú and the endless board games nights. For supporting each other in the toughest moments of the PhD.

Ari, una troballa d'últim moment durant el meu doctorat, però quina troballa!

The CRG community, for providing such a wonderful environment to work and socialise. For all the retreats, beer sessions, volleyball matches and football games.

Lab 325. Lucas, for convincing Ben to hire me. Lorena, por tu mentoraje y amistad. Carlos, por nuestras ideas locas de experimentos 'perturbadores'. Leticia, per la teva inquietud de voler saber y explicar. Alsu, because you are crazy funny. Miki, per la teva perseverancia i entusiasme.

A totes aquelles de l'equip de fútbol 'Game of Troncas'. Per ser garantia d'esvair els maldecaps momentanis del doctorat. Perquè, el que inicialment semblava ser un senzill equip de fútbol, ha acabat convertint-se en una fantàstica amistat.

A tots aquells dels grups de 'Pixus', 'Planazos' i 'Juntos y revueltos'. A tota la bona gent de Poblenou. Per totes les Festes Majors, Kaspas, Suecades, calçotades y Bagís que hem passat junts. I per totes aquelles que vindran.

Als meus grans amics. Maria, Laura, Sergi, Frede, Sandra, Mariona, Rafa, Berta, Aleix, Pablo i Ruda. Per simplement ser els meus amics, estar allà quan més us he necessitat. Perquè el moment de reunir-nos plegats és el moment més esperat de la setmana.

Famílies de Manuel i Montero, per rebre'm amb els braços oberts. Especialment al Jordi i a l'Antonia, per fer-me sentir com una filla més de la família.

Famílies Domingo i Espinós. Per totes les reunions familiars. Per els cap de setmana a la Cerdanya. Per les nits de Nochebuena. Per tot l'encoratjament que rebo de tots vosaltres quan parlo de la meva ciència.

Avi i abuelita. Por querernos y cuidarnos con devoción, a Marc y a mi. Por tantos tuppers que nos han salvado en momentos de apuro. Por vuestra inmensa generosidad.

Joana i Roger. Perquè fer vista enrere i adonar-me com ens hem fet grans m'omple de joia. Per haver portat a la familia dues persones tant meravelloses com el Marc i la Bea. Perquè em fa feliç veure-us feliços.

Pare i mare. Pel tot el suport incondicional que sempre he rebut de part vostre. Per com us estimeu als Marc i a tots els meus amics. Perquè sé que sempre podré comptar amb vosaltres. Perquè sou clarament el meu model a seguir. Us estimo.

Marc. Perquè em costa expressar amb paraules tot que sento per tu. Perquè tens el do únic de fer esvaïr totes les meves preocupacions amb una de les teves abraçades. Perquè ets generós i bondadós. Perquè si pogués retrocedir en el temps, sense pensar-m'ho ni un segon, tornaria a reviure totes les experiències que he viscut al teu costat, les bones i les dolentes. Per el futur que ens espera junts. Perquè t'estimo incondicionalment, fins al cel i les estrelles. Anar i tornar.

Abstract

The first aim of this thesis was to tackle a core question in biology—to understand how large numbers of mutations combine together to influence phenotypes. In order to do so, we built a combinatorially-complete library of naturally occurring variants in a yeast tRNA. For the first time in any gene, we could quantify the extent of which both the effects of individual mutations and the interactions between pairs of mutations change across a large number of closely-related genotypes. We found that all mutations switch from beneficial to detrimental effects and all interactions switch from positive to negative in different backgrounds. Secondly, with the use of systematic mutagenesis, protein complementation assays and deep sequencing, we developed a new experimental methodology to map the interaction interfaces of physically interacting proteins at amino acid resolution. The approach works by quantifying the effects of mutations on both protein binding and stability, resulting in a high resolution map of an interaction interface.

Resum

El primer objectiu d'aquesta tesi va ser abordar una qüestió fonamental en biologia: entendre com la combinació d'un gran nombre de mutacions poden produir canvis en el fenotip. Per tal d'investigar aquest problema, vam construir una col·lecció de totes les variants genètiques observades en l'evolució d'un ARNt del llevat. Per primera vegada en un anàlisi d'un gen complet, vam quantificar com els efectes de les mutacions individuals, així com les interaccions entre parelles de mutacions, canvien el fenotip. Els resultats mostren que en diferent contextos genètics, totes les mutacions poden ser beneficioses o deletèries. De la mateixa manera, les interactions entre parelles de mutacions exageren o atenuen els efectes de les mutacions individuals depenent del context genètic on ocorren. En segon lloc, utilitzant tècniques de mutagènesi sistemàtica, mètodes de complementació de proteïnes i seqüenciació d'ADN, hem desenvolupat una nova metodologia experimental per identificar a resolució d'aminoàcid la interfície de contacte entre dues proteïnes. La metodologia es basa en quantificar els efectes de mutacions que alteren la unió de les dues proteïnes, bé degut a que alteren l'estabilitat d'una d'elles, o bé perquè canvien l'afinitat de l'una per l'altre.

Abbreviations

A	Adenine
aa	Amino acid
AD	Transactivation domain
Arg	Arginine
aaRS	Aminoacyl tRNA synthetase
AP–MS	Affinity-Purification Mass Spectrometry
C	Cytosine
C. elegans	*Caenorhabditis elegans*
CL-MS	Cross-Linking Mass Spectrometry
Cryo-EM	Cryogenic Electron Microscopy
DBD	DNA binding domain
DHFR	Dihydrofolate reductase
DMS	Deep Mutational Scanning
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
FACS	Fluorescence-Activated Cell Sorting
G	Guanine
GFP	Green fluorescence protein
D. melanogaster	Drosophila melanogaster
H. sapiens	*Homo sapiens*
I	Inosine
Indel	Insertion or deletion
L	Liter
mL	Milliliter
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
MTX	Methotrexate
NMR	Nuclear Magnetic Resonance crystallography
PCA	Protein-fragment Complementation Assay
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank

PPI	Protein-protein interaction
RNA	Ribonucleic acid
RTD	Rapid tRNA decay
S. cerevisiae	*Saccharomyces cerevisiae*
T	Thymine
tRNA	Transfer RNA
UAS	Upstream Activating Sequence
uL	Microliter
Y2H	Yeast Two-Hybrid

Table of contents

List of figures

1. Introduction

1. INTRODUCTION

1.1. DNA, mutations and directed mutagenesis

Phenotypic variability among living organisms occurs due to two main reasons. One is our life experiences: what we eat, our health habits, where we grow up or what pathogens are we exposed to. Another reason is the genetic information encoded in their genome. This was observed by scientists even before the discovery of the molecule carrying this information, the DNA.

By the mid late nineteenth century, Gregor Mendel performed extensive experiments on how physical traits of sweet pea plants were transmitted across generations, and described the 'unit of heredity' as a particle that did not change and was passed to the offspring. However, it was not until the beginning of the twentieth century that DNA was discovered to be the repository of genetic information (Sutton 1903; Avery et al. 1944). Afterwards, the discovery of the structure of the DNA (Crick & Watson 1953; Maddox 2003) marked a milestone in the history of science and gave rise to modern molecular biology.

DNA is a molecule made of two chains of nucleotides (A, C, G and T) which coil around each other to form a double helix. Genomes are composed of long stretches of DNA carrying the genetic instructions used in the growth, development, functioning and reproduction of an organism. These instructions are encrypted in the form of genes and their products (e.g. proteins or non-coding RNAs), which interact with each other in a dynamic and coordinated manner allowing cells to function. Changes in the genetic material, or *mutations*, can occur naturally during DNA replication but also sporadically in non-replicating DNA.

Mutations can alter gene products and/or their regulation and thus are the ultimate source of inherited phenotypic variation. For the last century, science has strived to develop methods to generate mutations in a controlled and directed manner. Such ability opens the possibility to understand the consequences of mutations and the function of the genes in which they occur.

Perhaps surprisingly, the capacity to induce mutations was devised prior to the discovery of the structure of the DNA molecule itself. In the early 1920s, Hermann Müller found that high temperatures had the ability to mutate genes. A few years later, in 1927, he demonstrated the causal link to mutation after exposing fruit flies to relatively high doses of X-rays (Muller 1927; Muller 1928). Others (Auerbach & Robson 1946) proved the mutational effects of chemical components such as mustard gas.

Since the characterization of *mutagens*—physical or chemical agents that create changes in the genetic material—experimental geneticist induced random mutations in purified DNA or organisms to investigate their consequences through the observation of mutant phenotypes. However, this approach is inefficient due to the randomness and low probability of mutations of interest, in addition to the confounding effects from irrelevant secondary mutations.

An improved and more targeted forward genetics mutagenesis approach was transposon mutagenesis (Ruvkun & Ausubel 1981). This method results in one unique insertion mutation per genome, allowing single hit mutations, the incorporation of selectable markers during strain construction, and the recovery of the gene of interest after mutagenesis by transposon-tagging (Seifert et al. 1986).

However, the major advance in site-directed mutagenesis happened with the development of the Nobel prize-winning technique that uses mutant DNA oligomers that anneal to the complementary template DNA and act as primers for polymerases to perform elongation (Hutchison et al. 1978). This advance, together with the development of the Polymerase Chain Reaction (PCR) in the early 1980s (Bartlett & Stirling 2003; Shampo & Kyle 2002), allowed scientist to mutagenize DNA with high precision, and relatively little effort.

In the late 1980s, the rapid advances in the fields of mutagenesis, such as the development of targeted random mutagenesis by error prone PCR (Cadwell & Joyce 1992), experimental molecular evolution (Eigen 1985), and the conceptualization of fitness landscapes as a way to visualize and explain how a population can evolve through sequence space (Wright 1932; Kauffman 1993), came together to develop experimental strategies that mimic the evolutionary process —combining genetic change and selection—in order to modify or improve the function of a particular gene.

This process, termed *directed evolution*, involved the generation of large libraries of proteins that differ by few point mutations that were selected for a particular function in an iterative manner. Directed evolution has been used for adapting enzymes into unusual environments (Chen & Arnold 1993), catalyzing new reactions (Moore & Arnold 1996), or evolving antibodies, with the aim of producing new pharmaceuticals (McCafferty et al. 1990; Winter et al. 1994). However, in this process of optimization for a particular function only the few 'winning' versions of the original protein are identified.

With the emergence of next-generation sequencing technologies and the substantial reduction of its cost, it became possible to not only identify the few selected mutants, but explore the entire initial population of variants that undergo selection. It was not until 2010-2011 when the work by Fowler et al. (2010), Ernst et al. (2010) and Hietpas et al. (2011) collectively pioneered a technology called *Deep Mutational Scanning* (DMS), a technique that allows the systematic interrogation of the effects of thousands of mutations in a single experiment. DMS combines the construction of large mutant libraries that are subjected to selection for a specific function and finally deep sequenced before and after selection to obtain functional scores for each of the mutant variants. These seminal papers have since inspired a growing number of similar efforts by other groups, and to date, DMS has been used to interrogate the effects of mutations and their combinations in proteins (Araya et al. 2012; Fujino et al. 2012; Whitehead et al. 2012; Starita et al. 2013; Melamed et al. 2013; Olson et al. 2014; Firnberg et al. 2014; Kitzman et al. 2015; Bank et al. 2015; Aakre et al. 2015; Palmer et al. 2015; Starita et al. 2015; Doud & Bloom 2016; Mavor et al. 2016; Majithia et al.

2016; Sarkisyan et al. 2016; Starr et al. 2017; Diss & Lehner 2018; Staller et al. 2018; Jones et al. 2019; Bolognesi et al. 2019; Li et al. 2019), non-coding RNAs (Guy et al. 2014; Li et al. 2016; Puchta et al. 2016; Payea et al. 2018), regulatory regions (Patwardhan et al. 2012; Dvir et al. 2013; Rich et al. 2016; Cuperus et al. 2017; Maricque et al. 2018), or introns and exons that alter splicing (Ke et al. 2011; Julien et al. 2016; Bhagavatula et al. 2017; Braun et al. 2018; Baeza-Centurion et al. 2019).

1.1.1. Deep mutational scanning: mutagenesis, selection and next-generation sequencing

Although DMS experiments have been used to interrogate the effects of hundreds or thousands of mutations in a variety of gene products, they all share a common experimental structure that can be broken down into three main steps: (I) mutant library construction, (II) a selection experiment and (III) deep sequencing to quantify changes in the frequency of different variants (**Figure 1**). During the last decade, depending on the experimental system used or the purpose of the screen, several strategies for each of these steps have been described.

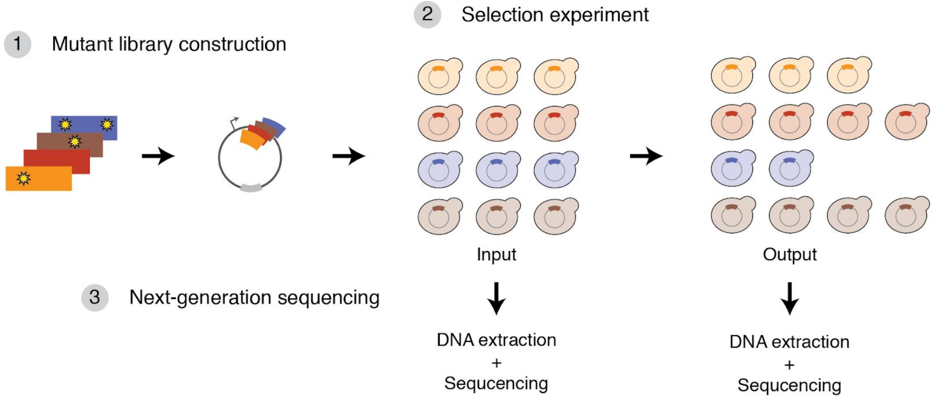


Figure 1: Schematic representation of a DMS experiment.

To obtain a library of mutants, a fair number of saturation mutagenesis methods have been applied in DMS studies—some more technically challenging than others. The simplest method is error-prone PCR amplification using low fidelity *Taq* polymerases (Cadwell & Joyce 1994; Mohan et al. 2011). Although a cheap and easy procedure, especially suited for long stretches of DNA, the preference for transitions over transversions in this method leads to uneven representations of mutations in the mutant library.

'Doped' oligonucleotide synthesis, in which the targeted region is synthesized with a tunable error rate, can overcome this last limitation. Another alternative, albeit more expensive, is the synthesis of collections of oligonucleotides containing all the versions of the gene of interest (named 'oligonucleotide pools'). However, in both doped oligonucleotides and oligonucleotide pools, frameshifting deletion errors limit the length of sequence that can be directly synthesized. Thus, oligonucleotide synthesis has been used to mutagenize small fragments of proteins (Starita et al. 2013), short RNA molecules (Li et al. 2016; Puchta et al. 2016; Julien et al. 2016; Hayden et al. 2015) or small combinatorially complete libraries (Poelwijk et al. 2017; Baeza-Centurion et al. 2019). Both error-prone PCR and doped oligos result in the generation of point mutations that occur at different frequencies, where the wild-type sequence is more represented than single point mutations, that occur more frequently than double mutations. As such, in the case of proteins, not all possible amino acid replacements can be easily created.

Methods that overcome this limitation are scaled-up versions of site-directed mutagenesis approaches. These include PCR-based approaches (Jain & Varadarajan 2014; Papworth et al. 1994; Kitzman et al. 2015) that use oligonucleotides carrying degeneracy codons. The most popular degeneracies are NNK or NNS, where K denotes either G or T, whereas S denotes either G or C. These two options only enable 32 out of all 64 possible codons, but each covers all 20 possible amino acids while avoiding two of the three possible stop codons (TGA and TAA). However, PCR-based library construction has limited scalability since each PCR reaction carrying one degenerate oligonucleotide has to be performed independently and needs to be afterwards assembled.

Methods such as Kunkel mutagenesis (Kunkel 1985) or its derived, PFunkel (Firnberg & Ostermeier 2012), overcome this limitation obtaining mutagenized plasmids in a single reaction tube. These methods use an *E. coli* strain that has been modified to produce high levels of uridine and lacks the ability to excise these bases from DNA. A phage vector carrying the desired template sequence is transfected into the cells resulting in its replication with a high uracil incorporation rate. The 'uracilated' template can be PCR amplified with primers containing the mutations of interest and subsequently amplified in regular *E. coli* strains that will degrade the uracilated template, thus enriching the mutant copies.

Other developed methods with a similar principle, such as nicking mutagenesis (Wrenbeck et al. 2016), avoid the preparation of uracil containing ssDNA, which can be highly variable (Sambrook et al. 1989). Very recently it has been shown that unamplified oligonucleotide pools can be used as codon-degenerate primers in plasmid-based one-pot mutagenesis techniques to prepare site-saturation mutagenesis libraries from plasmid DNA with near-complete coverage of the desired mutations with few off-target mutations (Medina-Cucurella et al. 2019).

In addition to all the mutagenesis strategies discussed here, complete variant libraries are also recently becoming commercially available. While this method is the most convenient in terms of library coverage, mutational efficiency and control over the number of mutations introduced, it is by far the most expensive option. However it is possible that with increased interest in gene synthesis applications, these options may become more affordable in the future. Finally, the recent development of template-directed mutagenesis techniques using CRISPR/Cas9 (Findlay et al. 2018; Sharon et al. 2018) are enabling the investigation of the effects of mutations in several chromosomal contexts in a single assay.

Mutant library construction precedes the selection assay, the most central and crucial part in a DMS experiment (**Figure 2**). Selection strategies can vary extensively depending on the type of phenotype screened, and can be classified into four broad categories: (I) *in vitro* display assays, (II) cell viability or growth, (III) cell sorting methods often based on the expression of fluorescently labeled reporters, and (IV) RNA-seq methodologies.

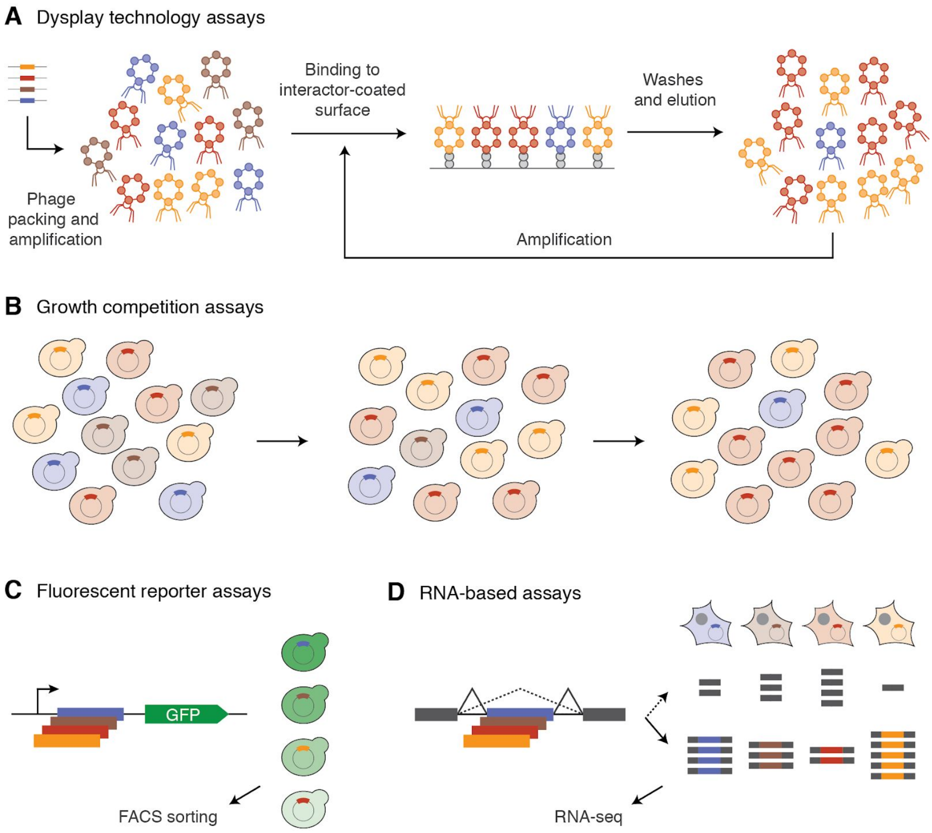

A Dysplay technology assays
Binding to interactor-coated surface
Washes and elution
Phage packing and amplification
Amplification
B Growth competition assays
C Fluorescent reporter assays
GFP
FACS sorting
D RNA-based assays
RNA-seq

Figure 2: Selection assays used in DMS. **A**. Display technology assays. In phage display, the bacteriophage displays a library of protein variants that are fused to its capsid proteins. The input phage library is subjected to several rounds of selection involving binding to an immobilized interactor, washing to remove unbound phage and final elution of the bound phage, which can be afterwards amplified for the next selection round. **B**. Growth competition assays. Variants in the libraries are enriched or depleted in the population because they provide a growth advantage or disadvantage to the cell. **C**. Fluorescent reporter assays. Cells are sorted depending on the fluorescence emitted by a reporter whose abundance is proportional to the activity of the mutated gene. **D**. RNA-seq based assays. mRNA abundance is used as a phenotype to measure how mutations in exons alter alternative splicing.

The first selection category includes surface display technologies (**Figure 2A**), such as phage display (Parmley & Smith 1988; Scott & Smith 1990) or yeast display (Boder & Wittrup 1997), which couple the genetic information of a given variant to the physical protein itself that is fused to host membrane component, and thus anchored into the host surface. The different protein variants are selected according to their affinity to bind an immobilized interactor. Variants that are unable to bind the interactor-coated surface are washed away and thus depleted from the initial population. This can be done in multiple rounds, as the genetic information can be replicated via viral propagation in bacteria or transformed into yeast (for phage and yeast display respectively) after selection. One of the first DMS studies by Fowler and colleagues (2010) employed phage display to analyse the binding of the WW domain of YAP65 to its cognate peptide target. Similarly, *in vitro* mRNA display and ribosome display (Roberts 1999) couple the genetic information to the protein variant by linking it to its mRNA progenitor (Olson et al. 2014; Fujino et al. 2012).

Another selection strategy is the use of growth competition assays (**Figure 2B**). In these experiments, the variants in the libraries are enriched or depleted in the population because they provide a growth advantage or disadvantage to the host, respectively. This occurs because either mutations alter the functionality of an essential gene (e.g. Hsp90 in yeast (Hietpas et al. 2011; Bank et al. 2016; Mishra et al. 2016)) or a conditionally essential gene (e.g. beta-lactamase in *E. coli* in the presence of Ampicillin (Firnberg et al. 2014; Stiffler et al. 2019)), or because growth is coupled to a specific activity of the gene. For instance, yeast growth can be coupled to the strength of binding between two proteins when systems such as protein-fragment complementation assays are used (see **section 1.3.2**). One example is the DMS that quantified the effects of >120,000 pairs of point mutations on the formation of the AP-1 transcription factor complex between the FOS and JUN proto-oncogenes (Diss & Lehner 2018).

Another selection mechanism is the use of fluorescence-activated cell sorting (FACS). Here, fluorescence reporters whose abundance are proportional to the activity of the studied gene allow the sorting of cells accordingly (**Figure 2C**). This last approach has been used to understand the role of synonymous mutations (Bhagavatula et al. 2017)—mutation in regulatory regions that alter the expression of a fluorescence reporter (e.g. mutating a promoter (Kwasnieski et al. 2012) or 5' untranslated regulatory region (Dvir et al. 2013; Cuperus et al. 2017)) or mutation in proteins that regulate transcription (e.g. a transcriptional repressor, (Li et al. 2019)). Fluorescence based assays have also been used to measure the steady-state abundance of protein variants (Matreyek et al. 2018) or how mutations alter the fluorescence of the reporter itself (Sarkisyan et al. 2016).

Finally, some DMS studies use mRNA abundance as a phenotype (**Figure 2D**). This has been proven useful to systematically understand the effects of mutations on alternative splicing, both in introns or exons (Julien et al. 2016; Braun et al. 2018; Baeza-Centurion et al. 2019; Ke et al. 2011), as well as the effect of non-coding variation on enhancer activity (Patwardhan et al. 2012).

It is important to note that, although I have classified the different selection strategies into four categories, several different selection assays have been developed since the emergence of the first DMS studies. Some of the assays can be generalizable to many proteins, as is the case of protein abundance assays (Matreyek et al. 2018). Others are designed to assay very particular gene functions, sometimes of the same gene. For instance, almost all single amino acid variants of the Ring domain of BRCA1 have been systematically analysed for their effect on E3 ubiquitin ligase activity (Starita et al. 2015), BARD1 ring domain binding (Starita et al. 2015) or homology-directed DNA repair (HDR) function (Findlay et al. 2018).

The last experimental step in a DMS experiment is deep sequencing. Next-generation sequencing can be considered the key technological advance that made DMS possible. It allows (I) the identification of all the variants generated in the experiment as well as (II) high-throughput quantification of the frequency of each variant in the library before and after selection that can be later transformed into a functional score (e.g. fitness or protein abundance). The sequencing strategies vary depending on the complexity of the mutant library and the length of the mutagenized sequence.

Some studies have used a fairly simple approach by performing deep shotgun sequencing of the libraries (Whitehead et al. 2012; Ernst et al. 2010). However, a major problem with this approach is that, without knowing which reads originate from which DNA molecule, each read can only be considered by itself; hence, this approach has been used to map single mutational effects. Reads from amplified regions that do not contain mutations yield no information and are wasted, and as the mutagenized region of the gene increases in length, the percentage of usable sequencing data decreases and. Consequently, more reads are necessary to ensure proper coverage per variant.

Mutagenizing a region of a gene that does not exceed the length of sequencing reads solves this issue and allows the study of more than one mutation per gene. Although the later has been the most popular approach in DMS studies (Olson et al. 2014; Li et al. 2019; Li et al. 2016; Hayden et al. 2015; Julien et al. 2016; Melamed et al. 2013; Araya et al. 2012), it limits the size of the mutant library to few hundred base pairs.

This can be overcome by using two different strategies. The first, 'gene tiling', involves dividing the gene into multiple 'tiles', each of which is effectively treated as a distinct gene. Each tile is independently mutagenized, subjected to selection, and sequenced; later the data is merged and normalized to generate the sequence-function map of the full gene (Kowalsky, Klesmith, et al. 2015; Bolognesi et al. 2019). Although the sequencing approach of gene tiling is straight forward, several selection assays are needed, and the combination of mutations between the different tiles is challenging.

The second approach involves associating molecular barcodes with each variant in the DMS library (Kitzman et al. 2015). While this simplifies the readout of the experiment (as only the barcodes need to be sequenced and counted), it adds the requirement to identify which barcode belongs to which genotype. In most cases this is addressed using 'subassembly' (Hiatt et al. 2010)—a high-throughput amplicon sequencing approach based on attaching random tags to amplicons. The DNA is amplified, sheared and ligated to adapters, so that paired-end sequencing can be used to identify the random tag together with each read. This allows reads to be sorted according to which original tagged molecule they belong to, which enables assemblies for each molecule to be computed. The resulting high-quality assembled reads are long enough to cover both the long mutagenized gene and barcode. Although barcoded mutant libraries are technically more challenging to build and sequence, they provide less noisy estimates per variant. If the library of barcodes is complex enough (i.e. each barcode differs at several nucleotides from any other barcode), sequencing errors can be estimated and corrected for. Usually several barcodes are associated to the same gene variant, which allows the identification of outlier barcode-variant associations (e.g. cells gaining background mutation that alters the screen phenotype during selection) that can be afterwards discarded.

A fourth common step of a DMS experiment is the computational analysis that comes after sequencing. Given the variety of mutagenesis, selection and sequencing approaches, most studies use custom scripts to process the sequencing data and calculate functional scores per variant. In the last decade, few software packages, some more sophisticated than others, have been developed to analyse and visualize DMS data (Fowler et al. 2011; Hietpas et al. 2011; Bloom 2015; Rubin et al. 2017). One example is Enrich2 (Rubin et al. 2017), that uses a statistical model to generate error estimates for each variant enrichment score, which not only captures the error resulting from the consistency between replicates (biological and technical variation), but also takes into account the sampling error owing to the low number of read counts for each variant in each replicate sample. Thus, the score of a particular variant can be more accurately calculated if the average of the different replicate estimates are previously normalized by their sampling error (i.e. if one of the replicates has less total read counts, estimates from that replicate will be less confidently estimated than the others).

The emergence of diverse DMS studies that use alternative mutant library construction, selection and sequencing strategies requires the development of a generic framework to be able to compare datasets from different laboratories. One step towards that goal is providing platforms that provide guidance for the experimental design (Matuszewski et al. 2016), help to process the raw sequencing data to obtain better estimates and errors (Rubin et al. 2017) as well as facilitate the identification of possible 'bottlenecks' that can bias downstream analyses (Faure et al. 2019).

1.2. Understanding genetic interactions using deep mutagenesis

Mutations can have different effects when occurring in different individuals. This observation draws from the fact that mutation outcome is dependent on the genetic background it occurs—a phenomenon known as *epistasis* (i.e. a given mutation may somehow be interacting with other mutations of the individual). DMS has allowed the systematic study of the prevalence, causes and consequences of epistasis.

This section of the thesis addresses the concept of epistasis, how can be experimentally studied and the lessons learned from it. It will also describe basic concepts of the molecule which is at the center of this work: tRNA.

1.2.1. Epistasis: The genetic context dependency of mutations

This section of the introduction takes the form of a review written by myself and Pablo Baeza Centurión, another PhD student of Ben Lehner's laboratory. I wrote the sections of the review related to specific epistasis, while Pablo wrote the sections discussing nonspecific epistasis. The remaining part of the text has substantial contributions from both of us. The review was published in August 2019 in *Annual Reviews*.

Domingo J, Baeza-Centurion P, Lehner B. The Causes and Consequences of Genetic Interactions (Epistasis). Annual review of genomics and human genetics. 2019;20(1):433–60. DOI: 10.1146/annurev-genom-083118-014857

1.2.2. Background: tRNAs

Transfer RNAs (tRNAs) are adaptor molecules composed of RNA present in all living organisms, that serve as the physical link between the mRNA and the amino acid sequence of proteins. tRNAs do this by carrying amino acids to the ribosome, the cellular machinery responsible for synthesizing proteins, directed by a 3-nucleotide sequence (codon) of the messenger RNA (mRNA). As such, tRNAs are essential components of translation—the biological synthesis of new proteins in accordance with the genetic code. Each tRNA is covalently attached to one amino acid that corresponds to the *anticodon* sequence (the three complementary bases to the mRNA codon) by the aminoacyl tRNA synthetase (aaRS) (Ling et al. 2009). During protein synthesis, elongation factors deliver the aminoacylated tRNAs to the ribosome. If the tRNA's anticodon matches the mRNA codon, the ribosome catalyses the peptide bond reaction between the newly delivered amino-acid to the one of the tRNA that was already bound, thus elongating the polypeptide chain.

Because the genetic code contains multiple codons that encode for the same amino acid, there are several tRNA molecules bearing different anticodons that carry the same amino acid. Based on their aminoacylation identity, all tRNAs are subdivided into 20 accepting groups. Each group comprises several tRNAs (*isoacceptors*) that translate synonymous codons, which usually vary by the third position. From prokaryotes to eukaryotes, tRNA genes tend to be present from one to multiple copies in the genomes, the number of gene copies for each tRNA family (tRNAs with the same anticodon) varying widely from species to species (Marck & Grosjean 2002). It has also been shown that the concentration of tRNA isoacceptors is determined by the number of gene copies of that family (Tuller et al. 2010).

Since the relative amounts of each tRNA isoacceptor relate to protein translation efficiency, tRNA gene content might explain codon usage bias (the unequal frequency of synonymous codons in the genome). This is the case for several unicellular organisms, including yeast, where tRNA copy number correlates with codon usage (Percudani et al. 1997; Ikemura 1981;

Kanaya et al. 2001). However, in higher eukaryotes this correlation is weaker (Kanaya et al. 2001; dos Reis et al. 2004), although it is more pronounced when modifications in the anticodon positions that expand the decoding capacity are taken into account (e.g. A-to-I, where I is able to wobble with A, C and U) (Novoa et al. 2012).

Finally, the tRNA pool is quite robust to genetic perturbations. A study in yeast showed that only ~20% of 204/275 yeast tRNA gene deletions had no appreciable phenotype in rich conditions (Bloom-Ackermann et al. 2014). This robustness to tRNA gene deletion was often facilitated through compensatory effects of other tRNAs within and between tRNA families. For instance, compensations between isoacceptors which operate via wobble interactions. However, under more severe or stressful conditions the percentage of deleterious phenotypes increases (Bloom-Ackermann et al. 2014). This can relate to the changes in tRNA abundance during stress conditions, where the expression level of some of the compensatory tRNAs change (Torrent et al. 2018).

The life cycle of a tRNA is quite complex, undergoing several post-transcriptional processes from its transcription until its binding with the ribosome. Cytoplasmic tRNAs are transcribed in the nucleus by DNA-dependent RNA polymerase III (Pol III), promoted by highly conserved sequence elements, A and B blocks, located within the transcribed region (Galli et al. 1981). tRNAs are transcribed as precursor molecules (pre-tRNAs) that undergo an elaborate set of post-transcriptional alterations to generate a mature tRNA. They are transcribed with ~12 extra leader nucleotides in both the 5' and 3’ ends. The first post-transcriptional step involves the removal of the 5’ leader followed by the removal of the 3’ extension (Hopper & Phizicky 2003). After the excision of both leader nucleotides, in eukaryotic tRNAs a CCA sequence is added to the 3’ terminus, a process that is required for tRNA aminoacylation (most prokaryotic tRNAs are already encoded with the CCA sequence).

Some tRNAs contain introns located one base 3’ of the anticodon, which need to be spliced out to function properly. The amount of tRNAs that contain introns vary extensively in the tree of life, humans having ~5% of their tRNAs harbouring introns (similar to *C. elegans*, *D. melanogaster* or mouse), while in yeast this percentage can be higher than 20% (61 out of 275) (Chan & Lowe 2009; Chan & Lowe 2016). tRNA introns are differently spliced in humans compared to yeast. While in vertebrates splicing occurs in two sequential steps in the nucleus (intron removal and 3'-5' ligation), in yeast tRNA intron splicing is performed in three steps (intron removal, 5'-3' exon ligation and residual 2' phosphate removal) in the cytoplasm (Sarkar & Hopper 1998). Interestingly, some introns are required for the modification of tRNA nucleosides, the most extensive and crucial processes in tRNA post-transcriptional modifications. There are 93 known different tRNA modifications across all kingdoms of life (El Yacoubi et al. 2012; Limbach et al. 1994), with ~25 of them occurring in yeast (Phizicky & Hopper 2010). tRNA modification serve diverse functions, including tRNA discrimination (e.g. distinction of the tRNA-Met for initiation or elongation steps (Aström & Byström 1994)), translation fidelity via codon-anticodon interaction (e.g. A-to-I editing of the wobble position that extends the codon-anticodon interaction capabilities (Gerber & Keller 1999; Gerber & Keller 2001)), maintenance of reading frame (Waas et al. 2007) and tRNA stability (Alexandrov et al. 2006).

tRNAs need to be folded into a particular structure to be aminoacylated. This tRNA secondary structure consists of four hydrogen bonded stems and associated loops (acceptor stem, D-arm, anticodon stem, variable loop and T-arm from 5' to 3') (**Figure 3B**). Nucleotides from the D- and T-loops come together and interact via coaxial stacking interactions to give rise to the canonical L-shape of tRNAs, where one branch contains the amino acid acceptor stem over the T-arm, and the other perpendicular branch is formed by the stack of the anticodon and the D-arm (**Figure 3A**).

tRNAs are aminoacylated by aaRS, which attaches the amino acid to the 3' end of the tRNA. There is one aaRS for each amino acid, aaRSs being highly selective for their cognate tRNAs. aaRS charge the amino acid to their cognate tRNAs by recognizing the anticodon as well as specific modifications (Agris et al. 2007), which sometimes lie outside the anticodon region. Amino acids encoded by six different codons (e.g. arginine, leucine and serine) have a particularly high variability in their anticodon sequences.

Thus, the set of identity elements for these particular tRNAs to be recognized by their cognate aaRS extend beyond the anticodon and acceptor stem (e.g. identity elements in D-loop or large-variable arm) to maintain strict specificity and accurately decode mRNA (Hendrickson 2001; Achsel & Gross 1993; Ling et al. 2009).

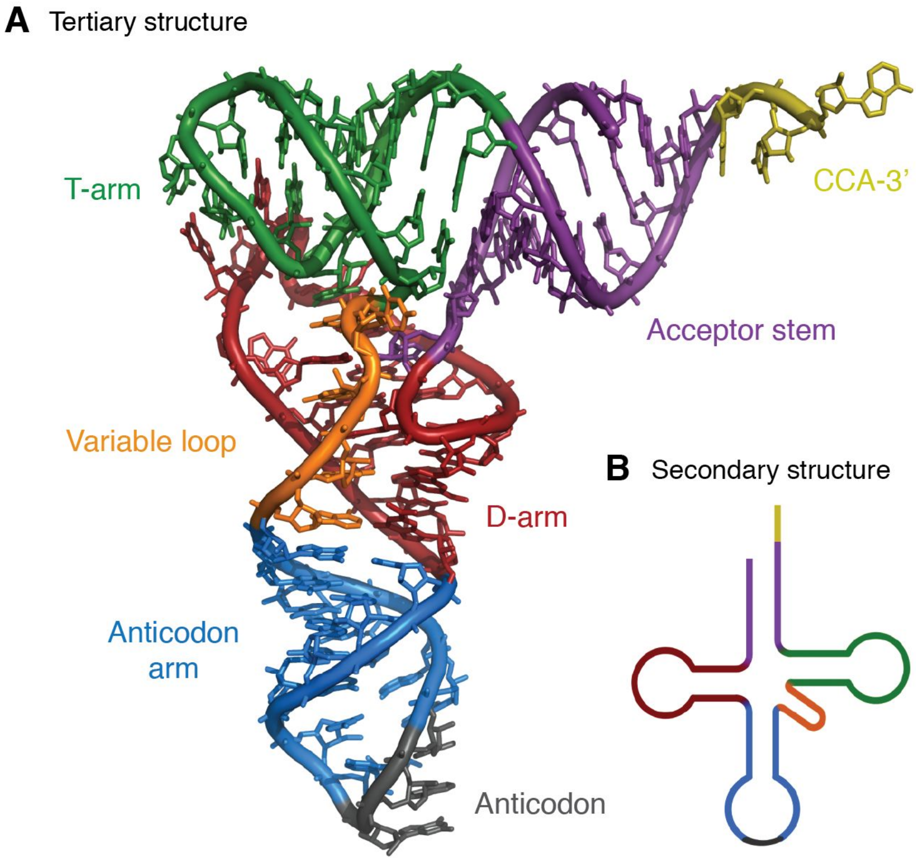


Figure 3: Structure of a tRNA molecule. **A**. Tertiary and **B**. secondary structure of the yeast tRNA-Phe (PDB entry 1ehz). CCA tail in yellow, Acceptor stem in purple, Variable loop in orange, D-arm in red, Anticodon arm in blue with the anticodon in grey, T-arm in green. Adapted from the Wikipedia entry on transfer RNA.

Mature tRNAs are very stable (estimated half life of ~9 hours in yeast) (Phizicky & Hopper 2010) and cells possess multiple pathways to degrade tRNAs that are inappropriately processed, modified or folded. 3'-5' exonucleolytic degradation by the nuclear exome serves as a quality control pathway to monitor for both appropriate tRNA nuclear modifications as well as 3' end maturation, while 5'-3' exonucleolytic degradation by the rapid tRNA degradation pathway targets mature tRNAs that lack one or more body modifications or are destabilized (Wolin et al. 2012; Megel et al. 2015). So, by the time the tRNA is incorporated in the ribosome to elongate the polypeptide chain, it has interacted with an extensive number of proteins. In yeast, >100 proteins have been described to interact and modify tRNAs (Hopper 2013).

Given the peculiar fold of tRNAs, the amount of post-transcriptional modifications they undergo and their interaction with other macromolecules, their primary sequence is highly conserved between species, suggesting strong functional constraints. This suggests that tRNAs would be largely intolerant to mutation. Indeed, some tRNA mutations have been shown to be deleterious in yeast (Kurjan et al. 1980) and numerous mutations in mitochondrial and cytoplasmic tRNAs have been associated to human diseases, including developmental disorders (Yarham et al. 2010; Lant et al. 2019). Sine tRNAs are short genes, typically shorter than 90 bp, DMS provides an unprecedented way to assess the consequences of mutations in a systematic manner.

The first DMS on a tRNA molecule was performed by Phizicky, Fields and colleagues, where they constructed a library of single and double nucleotide substitutions of the suppressor tRNA SUP4$_{oc}$ (tRNA-Tyr-G34U) and selected it for tRNA activity (Guy et al. 2014). The library of SUP4$_{oc}$ variants was introduced in a strain containing a GFP reporter with an ochre mutation, which allowed the sorting of fluorescence-activated cells based on their level of nonsense suppression (i.e. only functional SUP4$_{oc}$ tRNA variants that can rewrite the ochre stop codon with a tyrosine allow GFP expression). Unexpectedly, this showed that the tRNA tolerated ~37% of all single point mutations along the gene body, where positions that did not tolerate mutations where tertiary pairs involved in the three-dimensional fold of the tRNA. The tRNA also showed substantial epistasis between pairs of mutations, with a large excess of negative epistasis over positive epistasis (~7% and 1.5% of all tested ~25,000 double mutants, respectively). This excess was mainly due to complete non-functional double mutants where both singles were highly functional, suggesting that the tRNA can tolerate single mutations at multiple locations with little loss of function, but is extremely sensitive to a second mutation. The majority of positive epistasis cases were explained by the restoration of base pairing in stems, and a minority suggested alternative functional structural conformations. Still, not all positive interactions could be explained by structural features. In the same study, Guy et al. also evaluated which mutant variants were susceptive to rapid tRNA decay (RTD)—the 5'-3' tRNA exonucleolytic degradation machinery—by transforming the library of $SUP4_{oc}$ tRNA variants into a reporter strain in which RTD was inactivated (Chernyakov et al. 2008). Even though RTD was thought to degrade tRNAs that have exposed 5'ends, the results showed that mutations that sensitise $SUP4_{oc}$ to RTD were found to be located throughout the entire gene body, including the anticodon stem. This suggested that RTD monitors the integrity of the entire tRNA molecule, probably through tRNA stability, making these degradation process a major factor in determining the sequence limits to tRNA function (Guy et al. 2014). Finally the same $SUP4_{oc}$ library was subjected to selection under high temperature conditions revealing that the effect of most mutations was enhanced and that temperature sensitivity was associated with RTD susceptibility, consistent with the previous findings that RTD acts upon destabilized tRNAs (Payea et al. 2018).

Another DMS study on a different tRNA gene showed similar findings (Li et al. 2016). The target gene was the single-copy arginine yeast tRNA (tRNA-Arg(CUU)), the deletion of which impairs growth in a high temperature environment (37ºC). Similarly to $SUP4_{oc}$, the arginine tRNA was quite resistant to mutations, where only 9 of the 69 sites mutated, including the three anticodon positions, did not tolerate any substitution. Almost half of all mutation pairs exhibited statistically significant epistasis, which had a strong negative bias, except when the mutations occurred at Watson-Crick paired sites. Fitness of tRNA variants was broadly correlated with the predicted fraction of correctly folded tRNA molecules (considering alternative secondary structures with favorable predicted folding energies). Nevertheless, in concordance with the SUP_{oc} library, neither folding stability, secondary nor tertiary structure conformation can explain all the epistatic effects found, pointing towards additional mechanisms that alter tRNA function, for instance tRNA modifications.

1.3. Identifying protein-protein interaction interfaces by deep mutagenesis

The systematic study of mutations has not only provided an understanding of the consequences of mutations, but has also been extensively used for technological applications such as genome editing (Khan 2019; Rodríguez-Rodríguez et al. 2019), protein engineering (Brannigan & Wilkinson 2002) and determination of macromolecular structures (Chiasson & Fowler 2019; Schmiedel & Lehner 2019; Rollins et al. 2019).

This section of the introduction, reviews the concepts and methodologies necessary to understand a new application of deep mutational scans developed by us: the rapid identification of protein-protein interaction interfaces at amino acid resolution.

1.3.1. Interactomes and the three-dimensional structures of protein complexes

Genes and their products do not act in isolation, but rather interact with each other in nearly all biological processes of living organisms (Barabási & Oltvai 2004). These dynamic and intricate networks, also named *interactomes*, are composed by 'nodes'—the different molecules—and their mutual physical interactions—'edges'. Ongoing protein–protein interaction mapping efforts have identified a substantial amount of the 'edges' in the interactome of humans (Rolland et al. 2014), and similar advances have been made for model organisms (Tarassov et al. 2008; Li et al. 2004; Giot et al. 2003). Mutations that perturb the interactome are often the cause of disease (Vidal et al. 2011; David et al. 2012). These alter interactions either by (I) disrupting entire gene products (node removal) or (II) by altering some of their interactions (gain or deletion of one or more edges). The proportion of mutations that can be classified into these two categories is still under debate (Sahni et al. 2013). Truncating mutations, including out-of-frame indels and nonsense mutations, will likely lead to a node removal perturbation. However, non-synonymous single nucleotide polymorphisms, the most common mutations in Mendelian human heritable diseases (Stenson et al. 2017; Wang & Moult 2001), can either disrupt protein interactions ('edgetic' mutation) or destabilize the protein (node removal). Studies that have functionally profiled several thousand missense mutations across a varied spectrum of Mendelian disorders using interaction assays have identified both node removal and edgetic perturbations in roughly equal proportions (Zhong et al. 2009; Sahni et al. 2015).

Since most proteins affected in disease exert their function through interaction with other proteins (Kar et al. 2009), the development of drugs that target a specific interaction would provide an advantage compared to those that completely ablade protein activity (Duran-Frigola et al. 2013). However, targeting protein-protein interactions with small molecules is currently challenging for several reasons. First, most interaction surfaces are usually large and flat, involving many polar and hydrophobic interactions which makes difficult the binding of a small molecule (Jin et al. 2014). Second, the lack of knowledge of the residues that constitute a macromolecule interface, as well as poor understanding of the binding energies of these, hinders the discovery of new drugs. Thus, despite their importance in disease and pharmacology, most three-dimensional structural contacts between proteins remain unknown (Mosca, Céol, et al. 2013). To date, 51% of the ~25,000 non-redundant protein structures available in the Protein Data Bank (PDB) obtained by NMR or X-Ray crystallography, are protein complexes (Marsh & Teichmann 2015). However, the majority of these protein complexes (77%) are homomeric complexes, in contrast with the cells scenario, where most protein complexes are heteromeric. This is evidenced by the composition of crystal structures of complexes purified from native tissues, which are enriched for heteromeric protein complexes compared to those obtained using recombinantly produced proteins (Perica et al. 2012; Marsh & Teichmann 2014).

Other experimental techniques, such as cryogenic electron microscopy (cryo-EM), have helped to extend the current repertoire of three-dimensional protein complex structures, particularly with very large structures, complexes and heteromers with multiple distinct subunits. Complementary to the experimental determination, the structure of complexes can be modeled by homology in the same way as for individual proteins. This is certainly possible since it has been shown that most *interologues* (i.e. homologous interacting pairs) do indeed interact in the same way (Aloy et al. 2003). This means that, as for monomers, the high resolution three-dimensional structure of a given protein-protein interaction can be used to model all the interactions that involve homologous proteins and for which the binding has been experimentally confirmed (Mosca, Céol, et al. 2013). These models can then be complemented with low-resolution structural information, whenever it is available, to build the most complete possible model.

However, interaction templates are only available for a limited number of interactions and thus, to get a more complete picture of the interactome, it is necessary to apply methodologies that are template-independent. One of them is computational docking, which aims to predict the structure of a complex formed by two interacting proteins starting from the structures of the individual components (Schneider & Zacharias 2011; Ritchie 2008; Mosca et al. 2009). Traditional approaches to protein-protein docking ('template-free docking'), sample the binding modes of two proteins with no *a priori* knowledge of the structure of the complex. Template-free methods can yield good models of protein complexes (Lensink & Wodak 2010), but their ability to sample the conformational space is limited, and the multiple possible correct solutions generate many false positives. Further developed docking approaches, such as 'template-based docking', utilise local structural templates can help *ab-initio* docking protocols to provide more reliable three-dimensional models (Sinha et al. 2010; Kundrotas & Vakser 2013; Günther et al. 2007; Szilagyi & Zhang 2014). Although it has been suggested that it might be already structural templates to model nearly all complexes for which we have structural information of the interacting components (Kundrotas et al. 2012), template-based docking strategies still struggle to obtain good quality models when the sequence similarity between homologs is low (Negroni et al. 2014).

To date, there are experimental o modeled structures available for only 35% of the known protein interactions in *E. coli* (Mosca, Céol, et al. 2013; Kundrotas et al. 2010). In humans the numbers are even lower, with only 10% of protein interactions having a suggested three-dimensional complex structure. Thus, alternative techniques are necessary to define at amino acid resolution the interface of contact between interacting proteins.

1.3.2. High-throughput mapping of protein-protein interactions

Before knowing which residues constitute the interaction interface of a protein, it is necessary to know which proteins physically interact with each other. Therefore, building comprehensive 'reference' interactome networks is the first step for identifying protein-protein interaction interfaces. It was not until the early 2000s that it became technically feasible to map protein-protein interactions systematically, where hundreds or thousands of proteins are tested for thousands of physical interactions. The first interactome network generated was in *Saccharomyces cerevisiae* in 2000 (Uetz et al. 2000), soon to be followed by the *Drosophila melanogaster* (Giot et al. 2003), *Caenorhabditis elegans* (Li et al. 2004), and *Homo sapiens* (Rual et al. 2005).

There exist two major approaches to systematically map protein-protein interactions: (I) the mapping of complexes by affinity purification followed by mass spectrometry (AP-MS) (Walzthoeni et al. 2013) and (II) the identification of binary direct interactions by yeast two-hybrid (Y2H) (Fields & Song 1989; Brückner et al. 2009) or protein-fragment complementation (PCA) methodologies (Michnick 2003; Pelletier & Michnick 1997). These two strategies are fundamentally different in the kind of interactome data they produce. While AP-MS uses direct affinity between a bait protein and other proteins present in the biological sample to 'pull-down' direct or indirect interacting partners that can be later identified using mass spectrometry (Walzthoeni et al. 2013; Guruharsha et al. 2011; Krogan et al. 2006; Havugimana et al. 2012), Y2H and PCA interrogate direct interactions between two proteins.

Y2H was originally designed to detect protein-protein interactions using the yeast GAL4 transcriptional activator (Fields & Song 1989). GAL4 contains an N-terminal DNA-binding domain and a C-terminal transactivation domain. Both domains are independently stable and functional: the DNA-binding domain binding the GAL1 upstream activating sequence (UAS), and the transactivation domain, which activates transcription if brought into the vicinity of transcription start site. When two interacting proteins are fused to the DNA-binding and transactivation domains, respectively, their association creates a chimeric transcription factor, turning on gene expression of a reporter downstream of the GAL1 UAS (**Figure 4**). The reporter can consist of a chromogenic enzyme such as b-galactosidase, for selection on colored colonies, or prototrophic markers, which select for cell survival.

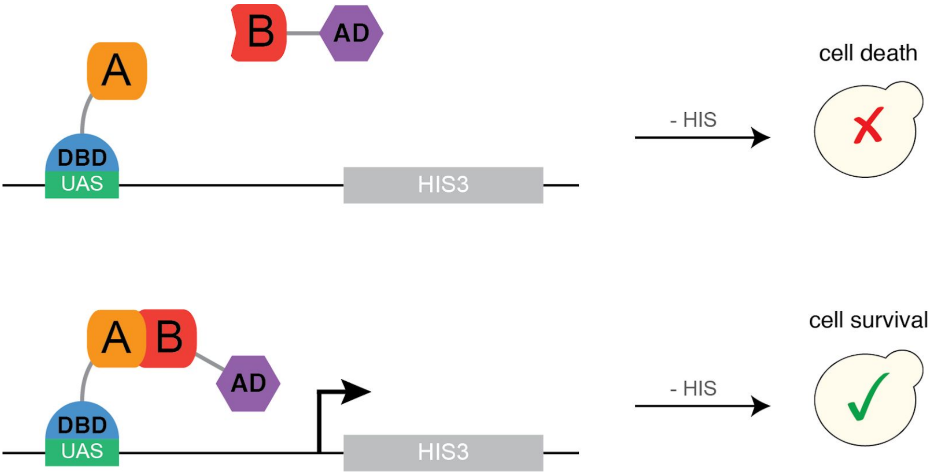


Figure 4: The classical yeast two-hybrid (Y2H) system. The protein of interest A is fused to the DNA binding domain (DBD) and the potential interactor protein B is fused to the transactivation domain (AD). The DBD-A fusion protein binds the upstream activator sequence (UAS) of the promoter. If A and B interact, the AD-B fusion protein is recruited to the promoter reconstituting a functional transcription factor. This allows the expression of the prototrophic marker HIS3 and thus, cell survival in a media depleted of histidine.

The invention of the Y2H technique not only triggered thousands of studies on protein-protein interactions but also spearheaded the development of variations of the original method with different purposes (Vidal 1999). For instance, it has been applied to detect other kinds of macromolecular interactions such as DNA-protein (one-hybrid) (Wilson et al. 1991; Inouye et al. 1994), RNA-protein (RNA-based three-hybrid) (SenGupta et al. 1996) and small molecule-protein interactions (ligand-based three-hybrid) (Licitra & Liu 1996).

Despite its great popularity, the biggest disadvantage of the classical Y2H system is the obligatory nuclear localization of the proteins and, hence, their site of interaction. PCA (protein-fragment complementation assay) was developed to overcome this limitation (Pelletier & Michnick 1997; Johnsson & Varshavsky 1994). In PCA the two proteins of interest ('bait' and 'prey') are fused to two non-active fragments of a reporter (**Figure 5**). If 'bait' and 'prey' interact, they bring together the two fragments of the reporter protein in close proximity, which allows the formation of a functional reporter protein whose activity can be measured or coupled to cell growth. Several reporters have been described, including ubiquitin (Johnsson & Varshavsky 1994; Dünkler et al. 2012), cytosine deaminase (FCY1) (Ear & Michnick 2009), beta-lactamase (Park et al. 2007) and GFP (Barnard & Timson 2010; Cabantous et al. 2013). One of the most effective reporters is the modified murine dihydrofolate reductase (DHFR) (Pelletier et al. 1998; Tarassov et al. 2008), which confers resistance to the chemical methotrexate (MTX). Therefore, the interaction between two proteins of interest can be detected and measured as cellular growth on media supplemented with MTX. DHFR-PCA is highly quantitative because the growth rate is correlated to the abundance of the complementation complex (Freschi et al. 2013; Levy et al. 2014).

DHFR-PCA has also been used to measure local concentration of proteins based on the strength of their nonspecific interactions with a neutral reporter protein (Levy et al. 2014). To achieve this, one DHFR fragment is fused to a protein of interest and the other fragment is very highly expressed alone so that the PCA signal reports on the concentration of the first protein fusion because of random protein encounters. This approach was used to quantify the concentration of all yeast proteins, with extremely good agreement with orthogonal methods (Levy et al. 2014).

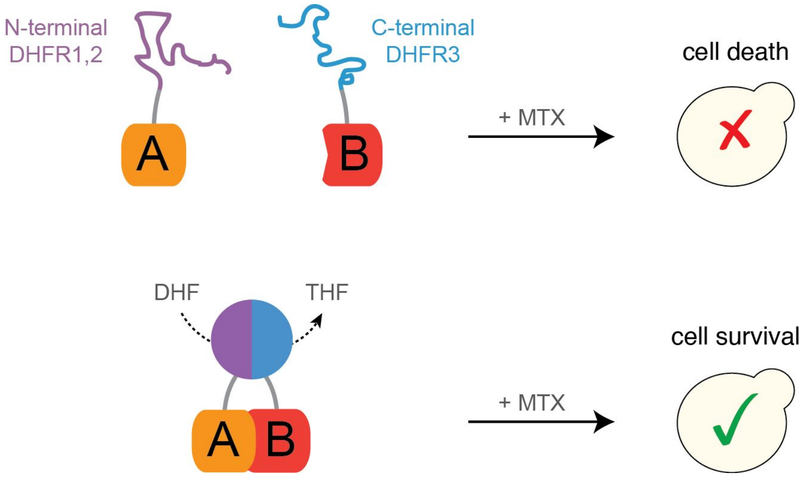


Figure 5: Protein-fragment Complementation Assay (PCA). The putative interacting proteins A and B are fused to the N- and C-terminal fragments of the murine methotrexate (MTX) resistant dihydrofolate reductase enzyme (DHFR). If A and B are in spatial proximity in the cell, the two complementary fragments fold together, resulting in the reconstituted enzyme. MTX inhibits the essential yeast DHFR, which converts dihydrofolate (DHF) to tetrahydrofolate (THF) for the *de novo* synthesis of pyrimidines. Thus, only when protein A and B interact, cells can grow in the presence of MTX in the media.

Since the implementation of these high-throughput technologies, the interactome for humans (Rual et al. 2005; Stelzl et al. 2005; Rolland et al. 2014; Ewing et al. 2007; Havugimana et al. 2012; Malovannaya et al. 2011) and other organisms (Arabidopsis Interactome Mapping Consortium 2011; Tarassov et al. 2008; Das et al. 2013; Yu et al. 2008; Rajagopala et al. 2014; Li et al. 2004; Simonis et al. 2009; Giot et al. 2003; Guruharsha et al. 2011) has expanded in both quantity and quality, reaching >100,000 identified interactions within the human proteome. However, the structural interfaces that mediate most protein interactions still remain unknown.

One step towards defining which regions of proteins are responsible for the binding with others is fragment-based interacting technologies. Ever since Y2H technologies were first implemented, fragmenting an interacting protein into domain-sized pieces has been used to determine the minimal region required for an interaction (Staub et al. 1996; Albers et al. 2005; Boxem et al. 2008). Staub et al. found that the WW domain of Nedd4 binds to the proline-rich motifs of epithelial sodium channels, an interaction impaired in Liddle's syndrome, a hereditary form of hypertension (Staub et al. 1996). The fragment-based approach in Y2H has been used in a systematic manner, for example, to map the minimal interacting regions of 200 embryogenesis proteins in *C. elegans* (Boxem et al. 2008). This strategy is also more sensitive, detecting a higher number of interactions than when using full length proteins. Instead of fusing full-length proteins that may not properly fold in yeast or remain in a conformation that does not allow binding, the use of multiple fragments for each protein in a 'fragment library' increases the probability that at least one fusion product will be capable of interacting in the assay.

Other methodologies that combine very different technologies, such as cross-linking mass spectrometry (CL-MS) (Yu & Huang 2018), lead to similar low-resolution complex structures. However, none of these methods allow the identification at amino acid resolution of protein interfaces, and the use of systematic mutagenesis was proposed as an alternative strategy towards this goal.

1.3.3. Using mutations to identify protein-protein interaction interfaces

Current mutagenesis methods to map the contact interface between proteins take advantage of the methodologies that identify interacting protein pairs. The original Y2H concept was turned upside down to develop the reverse yeast two-hybrid system in order to identify mutations that dissociate macromolecular interactions (Vidal, Brachmann, et al. 1996). In such a system, the interaction between two proteins of interest is deleterious for growth, and only a mutation that disrupts the protein-protein interaction confers a growth advantage. The combination of this technique with systematic site-directed mutagenesis (sometimes referred to as 'reverse edgetics' approach) allowed the identification of important structural elements of proteins that determine complex formation (Vidal, Braun, et al.

1996; Dreze et al. 2009; Woodsmith et al. 2017). For instance, the application of this technique to the conserved transcription factor E2F and its interaction partner DP1, identified a putative helix conserved among E2F family members that was relevant for their interaction (Vidal, Braun, et al. 1996). More recently, systematic and high-throughput versions of this methodology have been used to systematically identify >1,000 interaction-disrupting amino acid mutations across eight subunits of the BBSome, the major human cilia protein complex associated with the pleiotropic genetic disorder Bardet-Biedl syndrome (Woodsmith et al. 2017).

Other methods that are based on the same principles are used to map the epitope between antibodies and antigens. These methods involve generating libraries of antibodies (or antigen neutralizers) by parallel programmed mutagenesis that can be later screened for antigen binding affinity by cell surface display followed by fluorescence-activated cell sorting (FACS). Initially, this method was used to optimize inhibitors that better neutralized antigens (Whitehead et al. 2012). Follow up studies applied this mutagenesis method to identify the interface of the antigen where antibodies were bound (Kowalsky, Faber, et al. 2015; Van Blarcom et al. 2015; Doolan & Colby 2015; Wang et al. 2017). This technique assumes that aberrantly folded proteins do not make it to the surface because of the yeast secretion quality control system. However, this is not always the case (Whitehead et al. 2012), possibly due to the small size of the displayed proteins, and additional structural restrictions are required to be taken into account to discriminate the surface epitope (Van Blarcom et al. 2015). Both reverse edgetics and cell surface display systems, cannot always distinguish between an edgetic mutation that alters the protein binding from another that alters the folding or stability of the entire protein. Thus, they cannot be used alone to identify protein interaction interfaces at high resolution.

A few studies have been able to distinguish whether mutations are affecting a protein interaction directly (by altering the affinity of the interaction) or indirectly (by altering the concentration of a protein). So far two strategies have been developed for this purpose. The first involves measuring both molecular phenotypes, binding and stability. This approach was used to individually screen 204 disease-related mutations in 51 proteins that had known interactors (Wei et al. 2014). To test if the mutations affected stability, each allelic variant was tagged with GFP and fluorescence was quantified in a plate reader. The ability to bind a known interactor was measured by Y2H. This procedure estimated that half of the disease mutations disrupted the protein interaction with its partner, and 42% of the instances it was due to altering the affinity of the interaction and not the stability of the protein. Although this is consistent with previous computational and experimental analyses (Zhong et al. 2009; Sahni et al. 2015), the sparse number of mutations tested for each protein does not allow the entire protein-protein interaction interface to be identified.

The second method involves computationally inferring the energies of folding and binding of all single amino acid mutations from the fitness effects of double mutants that have been screened in a binding assay. This was recently reported by Otwinowsky (2018), who fitted a three-state thermodynamic model (bound and folded, unbound and folded and unfolded states) to the deep mutational scanning data of the 56-residue protein G B1 domain, where all single and nearly all 500,000 double mutants were measured for their binding to immunoglobulin G (Olson et al. 2014). The energies of folding for 812 mutations were later confirmed by an independent work that generated *in vitro* stability data for nearly every single mutant of the same domain (Nisthal et al. 2019). Although an impressive achievement, this approach has only been applied to this one complex dataset. Generating extensive double mutant libraries for most proteins is extremely challenging, costly and time consuming.

Finally, one promising computational approach is to use evolutionary analysis of the covariation between amino acid to identify close residue contacts across protein interactions, which was first used more than 20 years ago (Göbel et al. 1994; Pazos & Valencia 2001), and subsequently adapted to identify protein interactions (Pazos & Valencia 2002). The correlated evolution of pairs of residues in interacting proteins has also proven useful for inferring the contacting residues in protein-protein interactions (Ovchinnikov et al. 2014; Hopf et al. 2014; Weigt et al. 2009; Cong et al. 2019). Still, identifying coevolving amino acid changes between pairs of proteins is not trivial. First, large multiple sequence alignments (MSA) are needed for both of the interacting proteins. Second, the interaction between the two proteins must be conserved, so only organisms that contain both orthologs can contribute to the MSA. This issue was initially solved by using pairs of bacterial genes in conserved and proximal chromosomal locations (Ovchinnikov et al. 2014; Hopf et al. 2014). Only very recently have such methods been applied to the entire proteome of *E. coli* and *M. tuberculosis* (Cong et al. 2019). This latest work identified 804 and 911 protein-protein interactions (with computationally docked three dimensional structures) for each corresponding species, where hundreds of these were previously uncharacterized. Using coevolution to unravel interaction networks and protein-protein interaction interfaces in eukaryotic proteomes is still a challenge. It will likely require more genome sequence data on less complex eukaryotes spanning wider evolutionary distances as well as an improved methodologies to distinguish orthologous from paralogous genes.

2. Results

2.1. Pairwise and higher-order genetic interactions during the evolution of a tRNA

The first section of the results of this thesis takes the form from an article that was published in *Nature* in May 2018. I designed and performed all the experiments and computational analyses of this manuscript. Guillaume Diss contributed to the design of some experiments and computational analyses.

Domingo J, Diss G, Lehner B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature (London). 2018;558(7708):117–21. DOI: 10.1038/s41586-018-0170-7

2.2. Rapid high-resolution mapping of protein interaction interfaces by deep mutagenesis

The second section of the results of this thesis takes the form from a preprint that is about to be submitted. I share the first authorship of this manuscript with Jörn M. Schmiedel. I designed and performed all the experiments. Jörn processed the raw sequencing data. We both performed the computational analyses and figures. Guillaume Diss participated in the design of the plasmid constructs and sequencing strategy.

Domingo, J.*, Schmiedel, J.M.*, Diss, G. & Lehner, B., 2019. Rapid high-resolution mapping of protein interaction interfaces by deep mutagenesis. *In prep.*

Rapid high-resolution mapping of protein interaction interfaces by deep mutagenesis

Júlia Domingo[1]*, Jörn M. Schmiedel[1]*, Guillaume Diss[1,2], Ben Lehner[1,3,4†]

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2] Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland.
[3] Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
*equal contribution to the work
†e-mail: ben.lehner@crg.eu

Abstract

Protein interactions mediate most cellular processes and are frequently disrupted in human disease. However, the interaction interfaces of most proteins remain unidentified, making the interpretation of disease variants—whether they disrupt interactions ('edges') or entire proteins ('nodes') in networks—a challenge. Here we present a fast and simple experimental method—*DoubleDeepPCA*—to map the interaction interfaces of proteins at high resolution using deep mutagenesis. The approach works by quantifying the effects of mutations on both protein binding and stability, resulting in a high resolution map of an interaction interface. Our approach offers a new opportunity to identify and map protein interactions interfaces in a systematic manner.

Main text

Physical interactions between proteins are fundamental to nearly all biological processes. Protein-protein interactions have been systematically mapped in multiple species[1–10]. However, the structural interfaces that mediate most protein interactions remain unknown[11]. Since protein interactions are critical regulatory events in physiology and disease, and they represent an important target space for pharmacological intervention[12], identifying their binding interfaces is important for drug discovery and protein engineering.

Although there have been impressive advances in determining the three-dimensional structure of protein complexes[13], only 35% of the known protein interactions in *E. coli* have structural models (~1450/4200 protein interactions)[11]. In humans the numbers are even lower, with only 10% of protein interactions having a suggested structural model[11]. Alternative strategies such as inferring interactions from the evolutionary covariation of residue pairs between interaction partners can predict contacting residues[14–17]. However, this approach has only been applied to the *E. coli* and *M. tuberculosis* proteomes (identifying 804 and 911 protein-protein interactions of the 5.4 and 3.9 million tested respectively) because identifying covarying residues between two proteins is not trivial—it requires large multiple sequence alignments for both proteins, together with properly paired orthologous proteins that maintain the physical interaction, which becomes a challenge for eukaryotic genomes[14,15,17].

Deep mutational scanning (DMS)—a technique that combines DNA synthesis, selection and deep sequencing to measure the effects of thousands of mutations on gene function in a single experiment[18]—offers a systematic approach to identify mutations that disrupt protein interactions. A recently developed approach combined yeast two-hybrid selection with massively parallel programmed mutagenesis to systematically identify mutations that disrupt protein interactions[19]. Nonetheless, such an approach cannot distinguish whether an interaction is affected because of disruption of the specific interaction interface (the 'edge') or because of disruption of the folding or stability of the entire protein (the 'node')[20,21], and thus cannot be used alone to identify protein interaction interfaces at high resolution. Techniques that combine deep mutational scanning with yeast surface display followed by fluorescence-activated cell sorting (FACS) for epitope mapping[22–25] suffer from the same limitation, and typically require additional information, such structural feature restrictions, to refine the antibody-antigen contact interface[25].

To distinguish whether mutations affect a protein interaction directly (by altering the affinity of the interaction) or indirectly (by altering the concentration of a protein) requires both molecular phenotypes to be quantified. This approach was used to test whether 204 disease-causing mutations affected the abundance and/or interactions of 51 proteins by plate-based GFP fluorescence assay to quantify abundance and a yeast two-hybrid assay to quantify protein interactions[26]. Similar to previous computational and experimental analyses[21,27], this estimated that ~49% of disease mutations disrupt protein interactions, with 42% of those altering the interaction with the partner without affecting the stability of the protein. However the sparse number of mutations tested for each protein did not allow the identification of the interaction interface.

An alternative approach to directly measuring the effects of mutations on both binding and stability is to infer them from the interactions between mutations in double mutants[28]. This approach was suggested by Otwinoski[28] who fitted a three-state thermodynamic model to data for the binding of nearly all possible 500,000 double mutants of the protein G B1 domain to IgG[29], allowing the underlying changes in the energies of folding for 812 mutations to be successfully estimated[30]. Although an impressive achievement, generating such an extensive double mutant dataset for most proteins is extremely challenging and time consuming.

Here we present *DoubleDeepPCA*, a high-throughput mutagenesis method that quantifies the effects of mutations on both the stability and the binding of a protein to an interaction partner to rapidly identify the residues involved in the interaction interface. *DoubleDeepPCA* uses two different protein-fragment complementation assays (PCA)[31,32] (**Figure 1a**). The first assay, *deepPCA*[33], quantifies how mutations alter the interactions between two proteins. In this assay, a protein A and its binding partner B (a protein or peptide ligand), are fused to the C- and N-terminal fragments of the murine methotrexate-resistant dihydrofolate reductase (DHFR) enzyme, respectively, and are expressed in yeast. If the two proteins interact, the active DHFR enzyme is reconstituted, which allows growth in the presence of methotrexate. However, if mutations in protein A destabilise the protein or disrupt the protein interaction, the two DHFR fragments can not complement each other and yeast will not grow when methotrexate is present (**Figure 1a**). In the second assay, *stabilityPCA*, the C-terminal DHFR fragment is fused to protein A and the other N-terminal fragment is very highly expressed alone so that the PCA signal reports on the concentration of the first protein fusion because of random protein encounters[34]. Consequently, only mutations that alter the stability of protein A will impair yeast growth (**Figure 1a**).

To quantify the effects of mutations in a protein of interested on both protein stability and interaction partner binding in a systematic manner, first a library of mutants is created that is subsequently cloned into the two different assay plasmids to create a fusion protein with the DHFR3 fragment (**Figure 1b**). After cloning the library of mutants, the pool of plasmids is transformed into yeast and cells are first allowed to grow in non-selective conditions in the absence of methotrexate for ~4 generations (input) before they are switched to selective conditions under the presence of methotrexate for ~5 generations (output) (**Figure 1c**). Deep sequencing of DNA extracts from input and output cell populations is then used to obtain fitness estimates from selection-induced frequency changes for all assayed protein A variants (see Methods). Finally, by comparing the fitness effects of mutations between the two assays we can distinguish which mutations disrupt the protein-ligand interaction by destabilizing one member of the protein complex or directly affecting the binding by altering the residues involved in the interaction interface.

We applied *DoubleDeepPCA* to the C-terminal SH3 domain of the human growth factor receptor-bound protein 2 (GRB2), which binds the proline-rich linear peptide of the GRB2 associated-binding protein 2 (GAB2) and for which the structure of the complex has been previously determined by X-Ray crystallography[35]. To test if this domain could be screened using the *DoubleDeepPCA* methodology, we cloned the wild-type GRB2 SH3 into the *stabilityPCA* and *deepPCA* plasmids (the latter containing the linear peptide of GAB2) and assessed if the growth phenotype in yeast in the presence of methotrexate was rescued. Both constructs allow yeast to grow, indicating that GRB2 SH3 alone is stable and binds the linear peptide of GAB2 (**Supplementary Figure 1a**). Next, we cloned seven GRB2 single mutants in residues on the GRB2-GAB2 binding interface into both assay plasmids. Whereas 3/7 mutations have a deleterious growth phenotype when assessed for binding to GAB2 (*deepPCA*), only one mutation shows a deleterious growth phenotype when assessing GRB2 stability (*stabilityPCA*) (**Supplementary Figure 1a**). These results show that our setup allows the effect of mutations to be assayed across a wide range of growth rates and the mode by which mutations alter the interaction between two proteins to be identified.

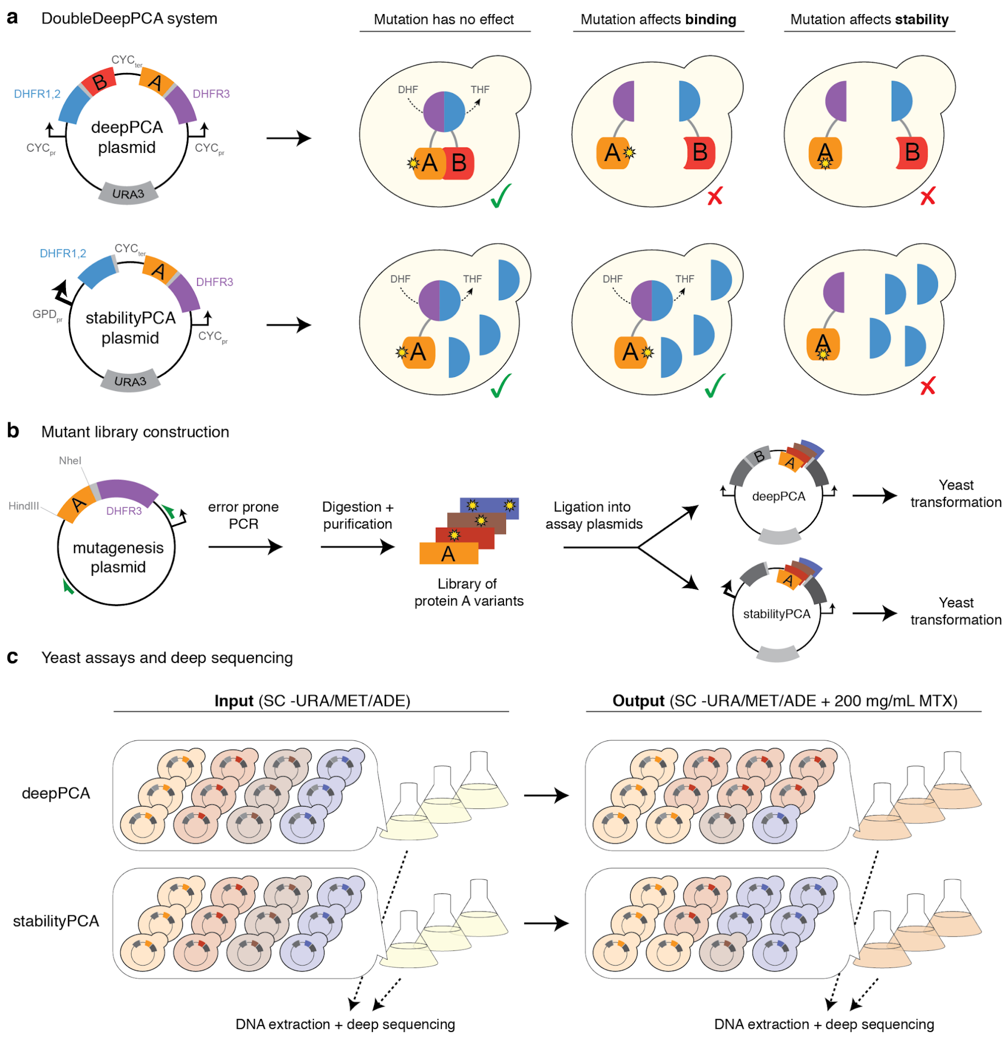


Figure 1: *DoubleDeepPCA* basis and workflow. (**a**) In the *deepPCA* assay two proteins that interact are fused to two halves of the DHFR enzyme, which is necessary for yeast growth in the presence of methotrexate. Mutations that affect the stability of one of the proteins or alter the binding affinity between them, will impair growth. In the *stabilityPCA* assay one of the enzymes halves is overexpressed so that the reconstituted complex is only dependent on the concentration of protein A fused to the other half. Only mutations that affect stability will impair yeast growth. (**b**) Construction of the mutant libraries of the protein of interest by error prone PCR and digestion-ligation assembly. (**c**) Highthoughput yeast assay. After yeast transformation, the libraries of protein variants are subjected to selection in the presence of methotrexate per triplicate for the two independent assays. The different

variant plasmids are extracted and deep sequenced before and after selection to obtain fitness estimates for each mutant in the two conditions.

To analyse thousands of mutations in a high-throughput manner, we generated a library of GRB2 mutants by error prone PCR (**Figure 1b**). This yielded an average of 1.64 nucleotide mutations in the 171bp long SH3 GRB2 coding sequence (**Supplementary Figure 2a**), with a bias towards A>T, A>G and G>A nucleotide substitutions (**Supplementary Figure 2b**). We subjected this protein domain to both selection assays in triplicate (**Figure 1c**). After filtering for low quality sequencing data and low count variants (see Methods) we obtained a total of 527 and 580 single amino acid mutations (~50% of the total possible 56*19=1,064 single amino acid substitutions) in the *stabilityPCA* and *deepPCA* assay respectively, with an average fitness correlation between replicates of 0.9 and 0.95 (**Supplementary Figure 3**, top). On average, each of the 56 residues of the GRB2 SH3 domain contained 11 mutations per position (range 8 to 14). Out of the mutations created by error prone PCR, eleven intersect with the 14 individually constructed variants, and effects estimated from deep sequencing and individual growth measurements assays are in excellent agreement (Spearman correlation ρ = 0.94, p= ,**Supplementary Figure 1b**). We also obtained fitness estimates for 17,464 double amino acid mutations shared between both assays (~3% of the total possible (56*19*55*19)/2=555,940 double amino acid substitutions), with an average correlation between replicates of 0.87 (**Supplementary Figure 3**, bottom).

We found that in both the *stabilityPCA* and *deepPCA* assays the distribution of single amino acid substitution effects is bimodal, with the majority of mutations having very little or no effect on growth, whereas 25% of mutations at least halve the growth rate per generation (fitness < 0.5) (**Figure 2a**, left). Mutations that have the biggest impact on fitness are enriched for core residues in both assays. However, while detrimental mutations in the stability assay are depleted in positions that are close to the ligand (< 5Å, minimal distance of any two heavy atoms in residues of GRB2 and GAB2), in the *deepPCA* assay detrimental mutations are enriched for both core and ligand binding positions (**Figure 2a**, right and **Figure 2e**). When directly comparing the effects of single mutants between the two assays, we found that positions close to the ligand contain mutations that have little to no effect in the stability assay but are often detrimental in the binding assay (**Figure 2b**). A similar pattern is observed with double amino acid mutants, where mutations in position pairs that are on average closer to the ligand are more detrimental in the binding than in the stability assay (**Supplementary Figure 4**). This suggests that, by combining the fitness scores from both assays, it should be feasible to identify which residues of GRB2 are in direct contact with GAB2.

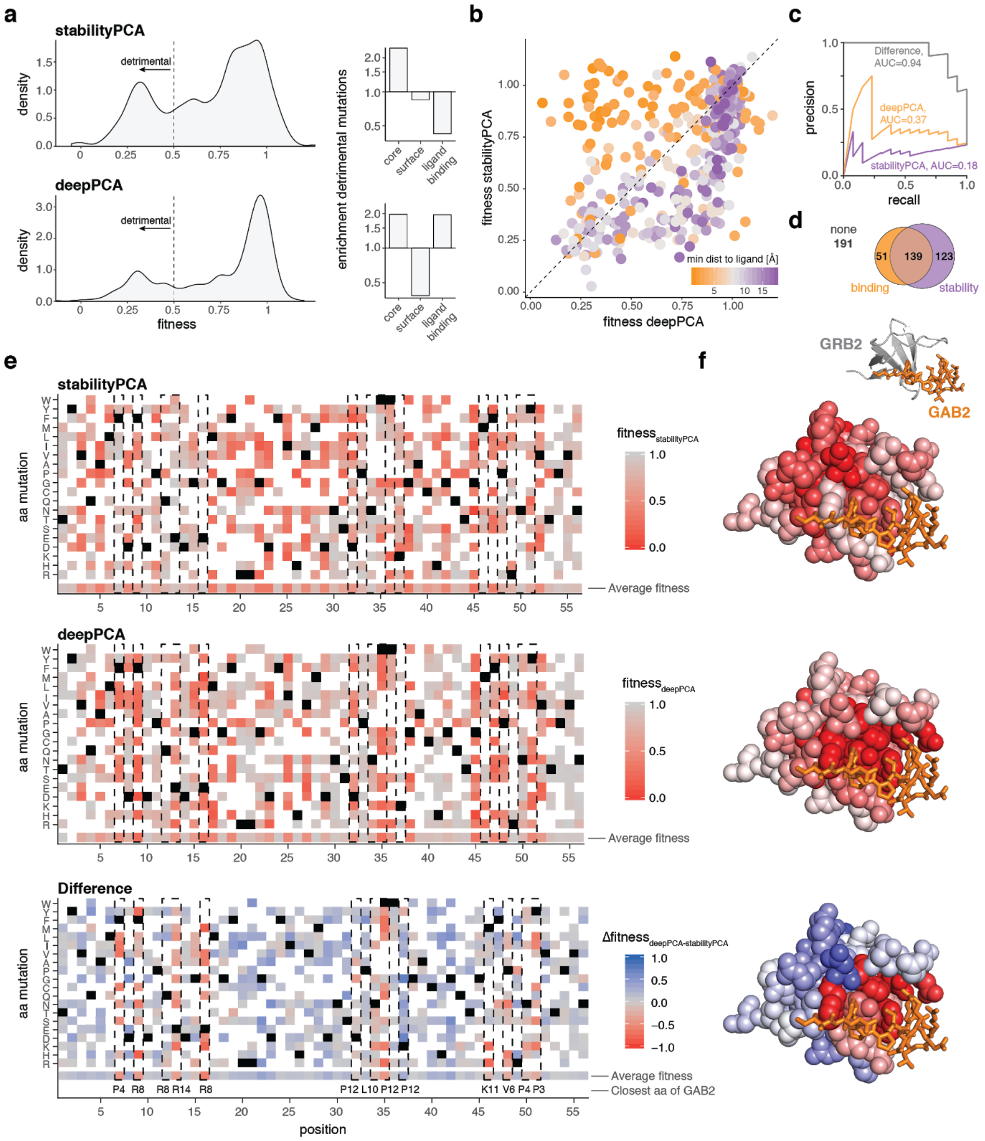


Figure 2: Identification of the GRB2-GAB2 interaction interface at amino acid resolution using single amino acid mutations. (**a**) Distribution of single mutant fitness values in both assays (left). Enrichment of detrimental mutations in the protein core (Relative accessible surface area, RSA <= 10), ligand-binding surface (<5Å from the most proximal GAB2 residue) or the rest of the surface (right). The enrichment corresponds to the fraction of detrimental mutations in a particular category over the fraction of detrimental mutations in all GRB2 positions. (**b**) Comparison of single mutant fitness values

in the stability assay compared to the binding assay. Mutations are coloured by the distance to the closest GAB2 residue (Å, using heavy atoms only). (**c**) Precision-recall curve for identifying GAB2 contacting GRB2 residues (<5Å) from the most deleterious mutations per position. AUC = Area Under the Curve. (**d**) Number of single amino acid mutations that affect only stability, only binding, both or none (see Methods). (**e**) Heat map of the fitness effect of GRB2 mutations in the stability (top) and binding (middle) assays, as well as the difference of the two assays (bottom). The lowest row in the heat map corresponds to the average fitness scores per position. Columns highlighted are at <5Å distance of the closest labeled GAB2 residue found at the bottom heatmap (**f**) Crystal structure of GRB2 (grey) bound to GAB2 (orange, PDB accession 2vwf). Residues of GRB2 are coloured by the average fitness per position.

Indeed, a simple subtraction of fitness scores (Δfitness = fitness$_{deepPCA}$ - fitness$_{stabilityPCA}$, with negative values indicating that a mutation is more detrimental in the binding assay than in the stability assay) provides enough information to identify the interaction interface (**Figure 2e,** bottom heatmap). Nine out of 56 residues have significantly negative Δfitness distributions (false discovery rate adjusted p-value $p < 0.05$, Student's t-test), all of which are in contact with GAB2 (<5Å, 11 residues in total, **Supplementary Figure 5a**). Similar, when ranking residues either by their most deleterious mutation effects or by their average fitness effects, top ranked residues are enriched for those contacting GAB2 (area under the precision recall curve AUC$_{prc}$=0.94 or AUC$_{prc}$=0.89, respectively, **Figure 2c** and **Supplementary Figure 5b**). Fitness scores from either the *stabilityPCA* or the *deepPCA* assays alone provive little discriminatory power (AUC$_{prc}$=0.18 and 0.37, respectively, when using most deleterious effects per position or AUC$_{prcAUC}$=0.15 and 0.51, respectively when using average fitness effects). A distance threshold might not be the most accurate way of defining a contact interface, which might lead to false negatives. This seems to be the case of Lys37, which on average has a negative Δfitness score (**Figure 2e**). The side chain of Lys37 is pointing away at a distance of >4Å from GAB2 Pro12, which is being stabilized by an aromatic interaction with GRB2 Trp35 (**Supplementary Figure 7a**). Colouring each residue of the GRB2 SH3 domain in complex with the linear peptide of GAB2 using the average fitness or Δfitness score clearly highlights how the combination of the two assays can identify at amino acid resolution the interface of contact between GRB2 and its ligand (**Figure 2f, Supplementary Figure 6**).

Our methodology also provides information about the interactions that mediate binding between a protein and its ligand. For instance, positions Phe7 and Tyr52 of GRB2 SH3 only tolerate amino acid substitutions that allow aromatic (Tyr, Phe and His) or amino/aromatic interactions (His and Gln), whereas all other mutations are deleterious (negative for Δfitness, **Figure 2e**). This can be explained because those positions are interacting with Pro3 and Pro4 of GAB2 through aromatic interactions (**Supplementary Figure 7b**). Also it provides mechanistic insights of what causes the disruption of the protein-ligand interaction. An example is the residue Met46, which can tolerate all amino acid substitutions but to positively charged residues (**Figure 2e**). The closest residue of GAB1 to GRB2 Met46 is a Lys11, which would lead to a repulsive electrostatic interaction when a positively charged amino acid occupies position 46 of the SH3 domain.

Finally, our method also allows us to evaluate the biochemical pleiotropy of mutations, that is, how often a mutation affects more than one biochemical parameter. In particular, we can use the ∆fitness score to address the question of how often mutations that affect a specific function of a protein (binding to an interaction partner) also alter its stability or folding[36]. Of 504 mutations quantified in both assays, 123 (24%) affect stability only, 139 (28%) affect stability and binding, and only 51 (10%) affect binding to GAB2 only ($p < 0.05$, Student's t-test, **Figure 2d**). The level of pleiotropy varies across aa positions (**Supplementary Figure 5c**), with mutations at some positions (e.g. positions 17, 25 ,40 and 45) typically affecting both binding affinity and stability whereas mutations at other positions (e.g. positions 16 and 50) only affect binding affinity but not stability.

In conclusion, we have developed a method that quickly and robustly identifies the contact interface between a protein and its interaction partner at amino acid resolution *in vivo* by decoupling the effect of mutations on binding and stability. The method also provides insights into the type of interactions occurring between a protein and its ligand. Although we could obtain accurate fitness estimates for ~18,000 mutations, including single and double amino acid substitutions, single amino acid mutations (50% of possible single mutants, 0.6% of synthesized mutations) were sufficient to identify the contacting residues of the GRB2 SH3 domain with GAB2. This indicates that less complex libraries of single amino acid mutations would provide enough information for interface mapping, potentially further reducing cost and effort of the method. *DoubleDeepPCA* offers a new opportunity to quickly identify and map in a systematic manner the interaction interfaces of proteins with known binding partners.

Methods

1. Media and buffers used

1.1. LB

10 g/L Bacto-tryptone, 5 g/L Yeast extract, 10 g/L NaCl. Autoclaved 20 min at 120ºC.

1.2. YPDA

20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract, 40 mg/L adenine sulfate. Autoclaved 20 min at 120ºC.

1.3. Plate mixture

40% PEG3350, 100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0. Filter sterilized (0.2 mm Nylon membrane, ThermoScientific).

1.4. Recovery medium

YPD (20 g/L glucose, 20 g/L Peptone, 10 g/L Yeast extract) + 0.5 M sorbitol. Filter sterilized.

1.5. SC -URA

6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.77 g/L complete supplement mixture drop-out without uracil. Filter sterilized.

1.6. SC -URA/MET/ADE

6.7 g/L Yeast Nitrogen base without amino acid, 20 g/L glucose, 0.74 g/L complete supplement mixture drop-out without uracil, adenine and methionine. Filter sterilized.

1.7. Competition medium

SC –ura/ade/met + 200 mg/mL methotrexate (BioShop Canada Inc., Canada), 2% DMSO.

1.8. DNA extraction buffer

2% Triton-X, 1% SDS, 100mM NaCl, 10mM Tris-HCl pH8, 1mM EDTA pH8

2. Generic plasmids construction

Three generic plasmids were constructed to be able to assay any protein of interest by *deepPCA* or *stabilityPCA*: the *deepPCA* plasmid (pGJJ001), the *stabilityPCA* plasmid (pGJJ045) and the mutagenesis plasmid (pGJJ003).

The first two plasmids were derived from plasmid pGD110[33], which carries two halves of the murine Methotrexate-resistant DHFR (DHFR1,2 and DHFR3) with C-terminus $(GGGGS)_4$ linker fusions under the expression of CYC promoters and a shared CYC terminator (URA3 cassette plasmid for yeast auxotrophic selection during the selection assays). pGJJ001 had the same structure as pGD110 but with a barcode cloning site upstream of the CYC promoter driving expression of DHFR3 in case a barcode-variant association sequencing strategy was necessary. To construct the plasmid, pGD110 was amplified in 3 different fragments using primer pairs oGJJ001-oGJJ002, oGJJ003-oGJJ083 and oGJJ82-oGJJ016 (**Supplementary Table 1**) , which were then assembled by Gibson reaction (prepared in house) at 50ºC for one hour. The *stabilityPCA* plasmid pGJJ045 was constructed by Gibson assembly by substituting the CYC promoter driving expression of the half DHFR1,2 for the GPD promoter using primer pairs oGJJ47-oGD087 and oGJJ46-oGD089 (**Supplementary Table 1**).

The generic mutagenesis plasmid (pGJJ003) was created to harbour a landing site with HindIII and AvrII restriction sites so that the CYC promoter and DHFR3 fused to any protein of interest could be cloned into it to perform error prone PCR or other alternative mutagenesis strategies. It was derived from pUC19, to avoid future plasmid selection in yeast if not properly purified (not containing any yeast auxotrophic cassette). The plasmid was also reduced in size (the lacZα fragment was deleted) to increase the efficiency of *E. coli* transformation and three synonymous mutations were introduced in the ampicillin resistance cassette (*bla* gene) to remove specific restriction sites. pGJJ003 was built using three fragment Gibson assembly. Two fragments were amplified (primer pairs oGJJ008-9 and oGJJ010-11) from pUC19 that introduced the barcode landing site. The third fragment with the *bla* sequence with the synonymous mutations was synthesized as a dsDNA gene block (gbGJJ001, GeneScript, **Supplementary Table 2**).

3. GRB2 plasmids construction

To construct the GRB2 *stabilityPCA* plasmid (pGJJ034) the 56 amino acid long SH3 domain of GRB2 (from amino acid 159 to 224 of the human protein GRB2) was fused to the C-terminus of the DHFR3 fragment of the *stabilityPCA* plasmid (pGJJ045). To do so the SH3 domain was amplified by PCR reaction using primer pair oGJJ012-13 to introduce the flanking HindIII and NheI restriction sites and then cloned into the digested pGJJ045 plasmid using T4 Ligase (NEB).

To construct the GRB2 *deepPCA* plasmid, first the sequence of GAB2 containing the linear peptide (32 amino acids long, amino acid 498 to 530 of the human GAB2 protein) was fused to the fragment DHFR1,2. GAB2 was amplified using primer pair oGJJ014-15, which introduced flanking BamHI and SpeI restriction sites. Both the PCR product and the *deepPCA* plasmid (pGJJ001) were digested and purified. The assembly of the new *deepPCA* plasmid with GAB2 (pGJJ006) was obtained by ligation using T4 Ligase. After validation by Sanger sequencing, pGJJ006 was digested with HindIII and NheI restriction enzymes and cloned with the GRB2 SH3 domain to obtain the final wild-type GRB2 *deepPCA* plasmid pGJJ034 (with both GRB2 SH3 and GAB2 linear peptide fused to both fragments of DHFR).

4. Selection assay controls

Seven single amino acid mutations of the GRB2 SH3 domain were tested for yeast growth in the presence of MTX on both stability and binding assays. To construct these plasmids, 7 gene blocks containing the mutated versions of GRB2 were synthesized with HindIII and NheI flanking restriction sites (gbGJJ002-8, Twist, **Supplementary Table 2**). The gene blocks were digested, purified and assembled into the previously digested and purified pGJJ006 and pGJJ045 (*deepPCA* plasmid with GAB2 and *stabilityPCA* plasmid, respectively).

A previously assayed interaction between the leucine zippers FOS and JUN[33], was used as positive control for the binding assay. The empty *stabilityPCA* plasmid (unfused DHFR3) served as positive control for the stability assay.

pGJJ025, a *deepPCA* plasmid that contains GRB2 fused to DHFR3 but no ligand fused to DHFR1,2, served as negative control for the binding assay. This plasmid was obtained as mentioned above, by digestion-ligation assembly. For the stability assay negative control, the Deg1 region (degron) from the yeast Matα2 protein[37], a degradation signal recognized by the proteolytic machinery, was fused to the DHFR3 fragment. The degron was amplified from yeast genomic DNA by PCR reaction using oligos oGJJ028-29, which added the flanking restriction enzymes to obtain the final plasmid pGJJ054 by digestion-ligation assembly as reported previously.

5. Growth rate measurements of individual constructs

The GRB2 wild-type and mutant constructs, as well as the positive and negative controls, were individually transformed into yeast following a small scale high efficiency yeast transformation (See section 7.1, same protocol but scaling down the volume 0.0003X). After transformation different colonies for each construct were picked and grown independently in 500 uL of SC -URA/MET/ADE overnight at 30ºC using a 96 deep well plate. The following morning the optical density (OD) of each well was measured using a Tecan Infinite M Plex plate reader (Tecan, Switzerland). Cultures in each well were diluted to an OD_{600nm} of 0.1. For the selection experiment, 5 uL of the diluted cells were added into 95 ul of

1.053X competition media (SC -URA/MET/ADE + 200 mg/mL methotrexate) to obtain 100 uL of starting culture at OD_{600nm} = 0.005 of uL of per well. The culture was grown for 60h at 30ºC in a Tecan plate reader where OD_{600nm} measurements were taken every 15 minutes. The growth rate in each well was obtained by calculating the slope of the exponential phase of the growth curve (slope of a linear fit of the log10(OD_{600nm}) against time).

6. GRB2 mutagenesis library construction

6.1. Cloning GRB2 into the mutagenesis plasmid

The CYC promoter and DHFR3 C-terminally fused to GRB2 was cloned into the mutagenesis plasmid (pGJJ003) by digestion-ligation protocol. The CYC-DHFR3-GRB2 insert was obtained by digesting the plasmid pGJJ025 (*deepPCA* plasmid with GRB2 tagged to DHFR3) with HindIII-HF and AvrII (New England Biolabs) and purifying the correct size band using the QIAquick Gel Extraction Kit (QIAGEN). The mutagenesis plasmid pGJJ003 was digested with HindIII-HF and AvrII and purified with the MinElute PCR Purification Kit (QIAGEN). The GRB2 mutagenesis plasmid (pGJJ043) was assembled by ligation reaction (T4 Ligase, New England Biolabs) following the manufacturer's protocol. After transformation into NEB10-beta High Efficiency competent cells, the plasmid sequence was verified by Sanger Sequencing (GATC, Eurofins Genomics).

6.2. Error prone PCR

The error prone PCR reaction was done using the GeneMorph II Mutagenesis Kit (Agilent Technologies) following the manufacturer's protocol. Primers oGJJ048-152 were used to amplify 1,011 bp of the GRB2 mutagenesis plasmid (pGJJ043). A single PCR reaction of 50 uL was run using 0.91 ng of template plasmid pGJJ043 (reaching the lowest recommended amount of plasmid by the manufacturer, lower plasmid amount increases the mutation frequency), with an annealing temperature of 56.4ºC (previously determined by gradient PCR), 1 minute and 10 seconds of extension time for 25 cycles. The PCR product was later run on an agarose gel for band confirmation and purified using the MinElute PCR Purification Kit (QIAGEN).

6.3. GRB2 mutagenesis library cloning into the assay plasmids

The product of the error prone PCR reaction was used to build the GRB2 libraries for both assays. The entire error prone PCR product was digested with HindIII-HF and NheI-HF restriction enzymes. The correct band with the GRB2 SH3 domain was purified using the MinElute Gel Extraction Kit (QIAGEN) and the sample was quantified using a Qubit fluorometer. Both, the *stabilityPCA* plasmid (pGJJ045) and the *deepPCA* plasmid containing the ligand GAB2 fused to DHFR1,2 (pGJJ006) were also digested with the same enzymes and purified using the QIAquick Gel Extraction Kit (QIAGEN).

The assembly of GRB2 variants in both assay plasmids was done overnight by temperature-cycle ligation[38] using T4 ligase (New England Biolabs) according to the manufacturer's protocol, 67 fmol of backbone and 200 fmol of insert in a 33.3 uL reaction. The ligation was desalted by dialysis using membrane filters for 1h and later concentrated 3.3X using a SpeedVac concentrator (Thermo Scientific).

Four microlitres of the concentrated assembled libraries were transformed into 100 ul of NEB 10β High-efficiency Electrocompetent *E. coli* cells according to the manufacturer's protocol. Cells were allowed to recover in SOC medium (NEB 10β Stable Outgrowth Medium) for 30 minutes and later transferred to 200 mL of LB medium with ampicillin 4X overnight. The total number of transformants was estimated to be 1.33×10^7 and 1.36×10^7 for the *deepPCA* and *stabilityPCA* libraries respectively. 100 mL of saturated *E. coli* culture was harvested next morning to extract the plasmid library using the QIAfilter Plasmid Midi Kit (QIAGEN).

7. Methotrexate selection assays

7.1. Large-scale yeast transformation

The high-efficiency yeast transformation protocol was derived from[33]. For each of the two assays (*deepPCA* and *stabilityPCA*) three independent pre-cultures of BY4742 were grown in 80 mL standard YPDA at 30ºC overnight. The next morning, the cultures were diluted into 700 mL of pre-wormed YPDA at an OD_{600nm} = 0.3. The three cultures were incubated at 30ºC for 4 hours. After growth, the cells were harvested and centrifuged for 5 minutes at 3,000g, washed with sterile water and later with SORB medium (100mM LiOAc, 10mM Tris pH 8.0, 1mM EDTA, 1M sorbitol). The cells were resuspended in 34.4 mL of SORB and incubated at room temperature for 30 minutes. After incubation, 700 µL of 10mg/mL boiled salmon sperm DNA (Agilent Genomics) was added to each tube of cells, as well as 14 µg of plasmid library. After gentle mixing, cells were split in four tubes of ~8.8 mL of cells and 35 mL of Plate Mixture (100mM LiOAc, 10mM Tris-HCl pH 8, 1mM EDTA/NaOH, pH 8, 40% PEG3350) were added to each tube to be incubated at room temperature for 30 more minutes. 3.5 mL of DMSO was added to each tube and the cells were then heat shocked at 42ºC for 20 minutes (inverting tubes from time to time to ensure homogenous heat transfer). After heat shock, cells were centrifuged and re-suspended in ~50 mL of recovery media and allowed to recover for 1 hour at 30ºC. Next cells were again centrifuged, washed with SC-URA medium and re-suspended in 800 mL SC -URA. After homogenization by stirring, 10 uL were plated on SC -URA Petri dishes and incubated for ~48 hours at 30ºC to measure the transformation efficiency. The three independent liquid cultures were grown at 30ºC for ~48 hours until saturation. In both assays, between 1.2x10^7 to 1.8x10^7 transformants were obtained across replicates, which ensured that all possible single and most double mutant genotypes were represented on average more than 10 times in the yeast population.

7.2. Selection assays

For each of the assays (*deepPCA* or *stabilityPCA*), each of the growth competitions was performed right after yeast transformation, following the same protocol but on different days. After the first cycle of post-transformation plasmid selection, a second plasmid selection cycle (input) was performed by inoculating 800 mL of SC -URA/MET/ADE at a starting OD_{600nm} = 0.1 with the saturated culture. Cells were grown for 4 generations at 30ºC under constant agitation at 200 rpm. This allowed the pool of mutants to be amplified and enter the exponential growth phase. The competition cycle (output) was then started by inoculating cells from the input cycle into 800 mL of competition media (SC -URA/MET/ADE + 200 mg/mL Methotrexate) so that the starting OD_{600nm} was 0.05. For that, the adequate volume of cells were collected, centrifuged at 3,000 rpm for 5 minutes and resuspended in the pre-warmed output media. Meanwhile, two times ~380 mL of each input replicate cultures were harvested by centrifugation for 5 min at 5,000g at 4ºC using a JLA 10.500 rotor. Yeast cells were washed with water, pelleted and stored at -20ºC for later DNA extraction. After ~5-5.2 generations of competition cycle, two times ~400 mL of each output replicate cultures were harvested by centrifugation for 5 min at 5,000g at 4ºC, washed twice with water and pelleted to be stored at -20ºC.

8. DNA extraction and plasmid quantification

Cell pellets (six tubes, three input and three output replicates) were re-suspended in 4 mL of DNA extraction buffer, frozen by dry ice-ethanol bath and incubated at 62ºC water bath twice. Subsequently, 4 mL of Phenol/Chloro/Isoamyl 25:24:1 (equilibrated in 10mM Tris-HCl, 1mM EDTA, pH8) was added, together with 4 g of acid-washed glass beads (Sigma Aldrich) and the samples were vortexed for 10 minutes. Samples were centrifuged at RT for 30 minutes at 4,000 rpm and the aqueous phase was transferred into new tubes. The same step was repeated twice. 0.4 mL of NaOAc 3M and 8.8 mL of pre-chilled absolute ethanol were added to the aqueous phase. The samples were gently mixed and incubated at -20ºC for 30 minutes. After that, they were centrifuged for 30 min at full speed at 4ºC to precipitate the DNA. The ethanol was removed and the DNA pellet was allowed to dry overnight at RT. DNA pellets were resuspended in 2.4 mL TE 1X and treated with 20 uL of RNaseA (10mg/mL, Thermo Scientific) for 30 minutes at 37ºC. To desalt and concentrate the DNA solutions, QIAEX II Gel Extraction Kit was used (200 µL of QIAEX II beads). The samples were washed twice with PE buffer and eluted 500 µL of 10 mM Tris-HCl buffer, pH 8.5. Finally, plasmid concentrations in the total DNA extract were quantified by qPCR using the primer pair oGJJ152-oGJJ153, which binds in the *ori* region of the plasmids.

9. Sequencing library preparation

The sequencing libraries were constructed in two consecutive PCR reactions. The first PCR (PCR1) was designed to amplify the mutated protein of interest and to increase the nucleotide complexity of the first sequenced bases by introducing frame-shift bases between the adapters and the sequencing region of interest. The second PCR (PCR2) was necessary to add the remainder of the Illumina adapter sequences.

For each independent input/output replicate of any of the two yeast assays (12 DNA samples in total, 6 input/output replicates per assay) a total of 16 50 uL PCR1 reactions were performed (two entire 96 well plates) using Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs) according to the manufacturer's protocol. Each reaction contained 6.25x10^{7} template plasmid molecules from the DNA extractions (1x10^{9} total molecules of plasmid per DNA sample) and 25 pmol of pooled frame-shift primers oGJJ52/54-58 and oGJJ84-89 (forward and reverse primers were independently pooled according to the nucleotide diversity of each oligo, **Supplementary Table 1**). The PCR reaction was set to 60ºC annealing temperature, 10 seconds of extension time and run for 15 cycles. Excess primers were removed by adding 2 mL of ExoSAP-IT (Affymetrix) and incubating for 20 min at 37 ℃ followed by an inactivation for 15 min at 80 ℃. The 16 PCRs of each sample were then pooled and purified using the MinElute PCR Purification Kit (QIAGEN) according to the manufacturer's protocol, using 2 columns per sample. DNA was eluted twice in 33.33 uL of EB buffer and pooled for each sample.

For each of the 12 independent samples, 8 50uL PCR2 reactions were run using Hot Start High-Fidelity DNA Polymerase using 2.5 uL of the previous purified PCR1. In this second PCR the remaining parts of the Illumina adapters were added to the library amplicon. The forward primer (5' P5 Illumina adapter) was the same for all samples, while the reverse primer (3' P7 Illumina adapter) differed by the barcode index (**Supplementary Table 1**), to subsequently pool all the samples together and demultiplex them after deep sequencing. 10 cycles of PCR2 were run at 62ºC of annealing temperature and 15 seconds of extension time. All 8 reactions from the same samples were pooled together and run on a 2% agarose gel to be quantified. After quantification, the 12 subsamples with different Illumina indexes were pooled together in an equimolar ratio, run on a gel, purified using the QIAEX II Gel Extraction Kit and subjected to 125 bp paired-end sequencing on an Illumina HiSeq 2500v5 sequencer at the CRG Genomics Core Facility.

10. From sequencing read counts to fitness estimates

Demultiplexed FASTQ paired-end sequencing read files from one HiSeq 2500v5 lane were processed using a custom pipeline to obtain variant counts per input/output replicate (https://github.com/lehner-lab/DiMSum,[39]). In short, constant regions of reads were trimmed using cutadapt[40], paired-end reads merged using Usearch[41], merged reads were discarded if the minimum merged Phred score for base calls was below 25, and unique variants per input/output replicate were counted using FASTX-toolkit (http://hannonlab.cshl.ed/fastx_toolkit/). Only variants with no more than two amino acid mutations (and two more nucleotide mutations than amino acid mutations) were retained, resulting in a total of 124 million reads across the 12 replicates of both assays (range 5.4 to 13.3 million reads per replicate). Read counts for synonymous variants were summed and a threshold of an average of 20 read counts across input replicates in each assay was imposed to filter low quality variants (variants with fewer input reads cannot span the whole fitness range), resulting in count data for 527 and 580 single mutants and 19985 and 24952 double mutants in the stability and binding assay data, respectively. Fitness of each mutant variant f_i per replicate selection was calculated as the growth rate relative to the GRB2 wild-type variant as $f_i = \frac{log_2\left(c_i^{out}/c_i^{in}\right) + d}{log_2\left(c_{wt}^{out}/c_{wt}^{in}\right) + d}$, with c^{out} as variant frequency in output sample (read count normalized by total read count, either for variant *i* or wild-type *wt*) and c^{in} as variant frequency in input sample and *d* as the number of doublings of the yeast population during the competition experiment (5.1 for stability assay, 5.25 for binding assay), as inferred from OD measurements. An error estimate for each fitness value was calculated as $\sigma_i = \sqrt{log_2(e)} \times \sqrt{1/r_i^{out} + 1/r_i^{in}} \,/\, \left(log_2\left(c_{wt}^{out}/c_{wt}^{in}\right) \;+\; d\right)$. Finally, fitness of variants across replicate selections was calculated as weighted averages of replicate selection fitness values, with weights according to error estimates plus a constant 5% replicate error term.

The Δfitness scores were obtained by subtracting the fitness scores of the stability assay from fitness scores of the binding assay.

11. PDB metrics calculations

A crystal structure of GRB2 bound to GAB2 peptide (PDB entry 2vwf[35]) was used to calculate minimal distances of each GRB2 residue to any GAB2 residue (only heavy atoms considered). Relative solvent accessibility surface area (RSA) for GRB2 when not bound to GAB2 was calculated using freeSASA[42]. Residues were classified as ‘core’ when RSA is smaller 0.1, or ‘surface’ otherwise (0.1 < RSA < 1). Moreover, 13/46 surface residues were classified as ‘ligand binding’, as their minimal distance to a GAB2 residue is smaller 5Å.

12. Statistical testing

Whether positions affect stability and/or binding was assessed using one-sided Student’s t-test with Benjamini-Hochberg false discovery rate adjustment (R functions t.test and p.adjust). A position was classified as affecting binding if its mutations were both biased towards lower fitness in the deepPCA assay (fitness < 1) as well as in Δfitness scores (*deepPCA* - *stabilityPCA* fitness, Δfitness < 0). A position was classified as affecting stability if its mutations were biased towards lower fitness in the *stabilityPCA* assay (fitness < 1).

Test for individual mutations were calculated similarly, with the exception that Student’s t-test statistic was calculated as $\frac{f_i - \mu}{\sigma_i}$ (using R function pt), with two degrees of freedom and $\mu = 1$ for *deepPCA* and *stabilityPCA* assays and $\mu = 0$ for Δfitness scores and condition $f_i < \mu$.

Area under the precision recall curve was calculated using R function pr.curve (library PRROC) either using ascending ordering of positions by their minimal fitness value (as shown in FIgure 2c) or their average fitness values (Supplementary Figure 5b).

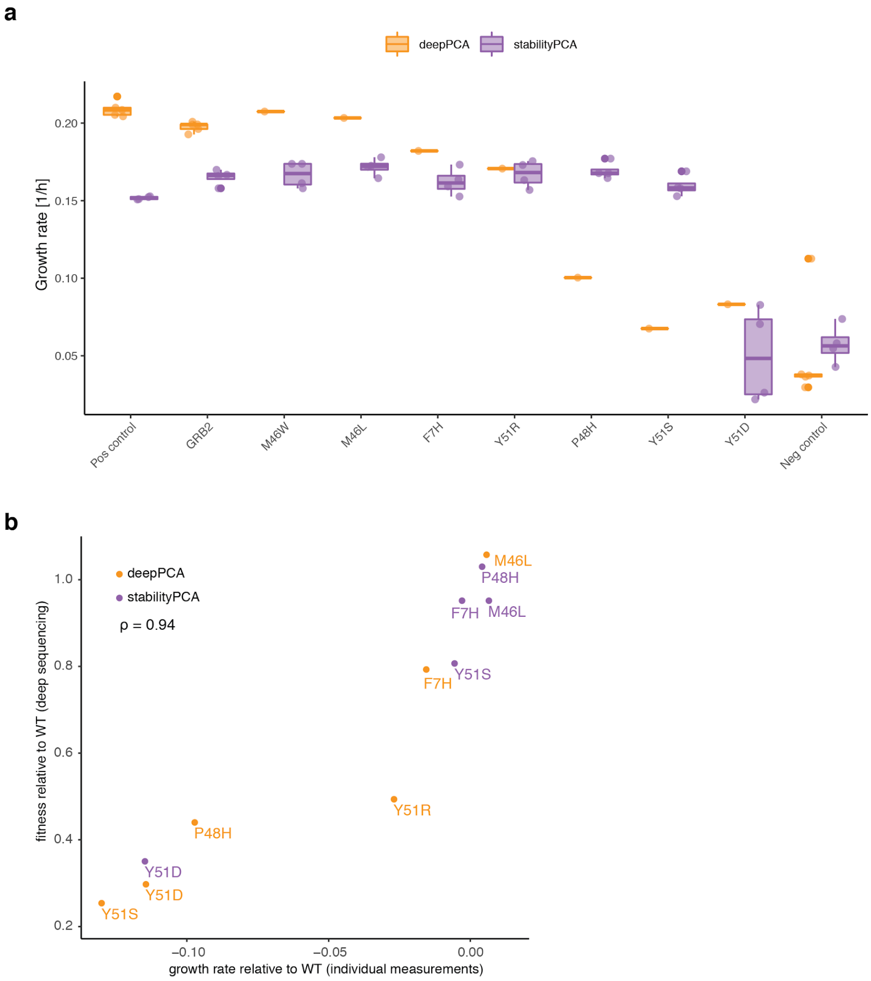
Acknowledgments

This work was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2017-89488-P and SEV-2012-0208), the Bettencourt Schueller Foundation, Agencia de Gestio d’Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322.), and the CERCA Program/Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership and the Centro de Excelencia Severo Ochoa. This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 752809 (J.M.S.).

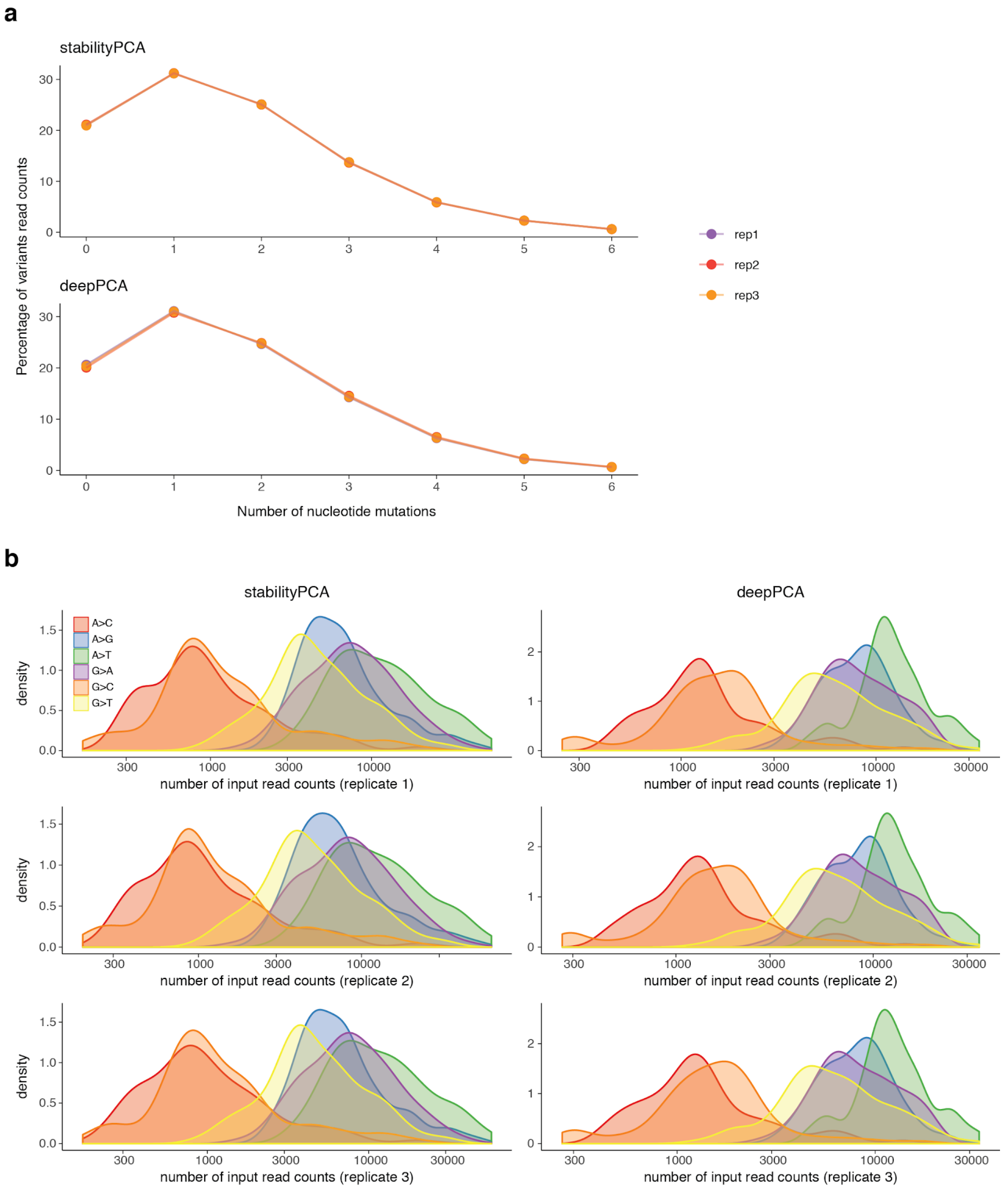
Author contributions

All authors conceived the project. J.D. and G.D. designed the plasmid constructs and sequencing strategy. J.D. performed the experiments. J.M.S. processed the sequencing data. J.M.S and J.D. did the computational analyses and figures. J.D., J.M.S. and B.L. wrote the manuscript.

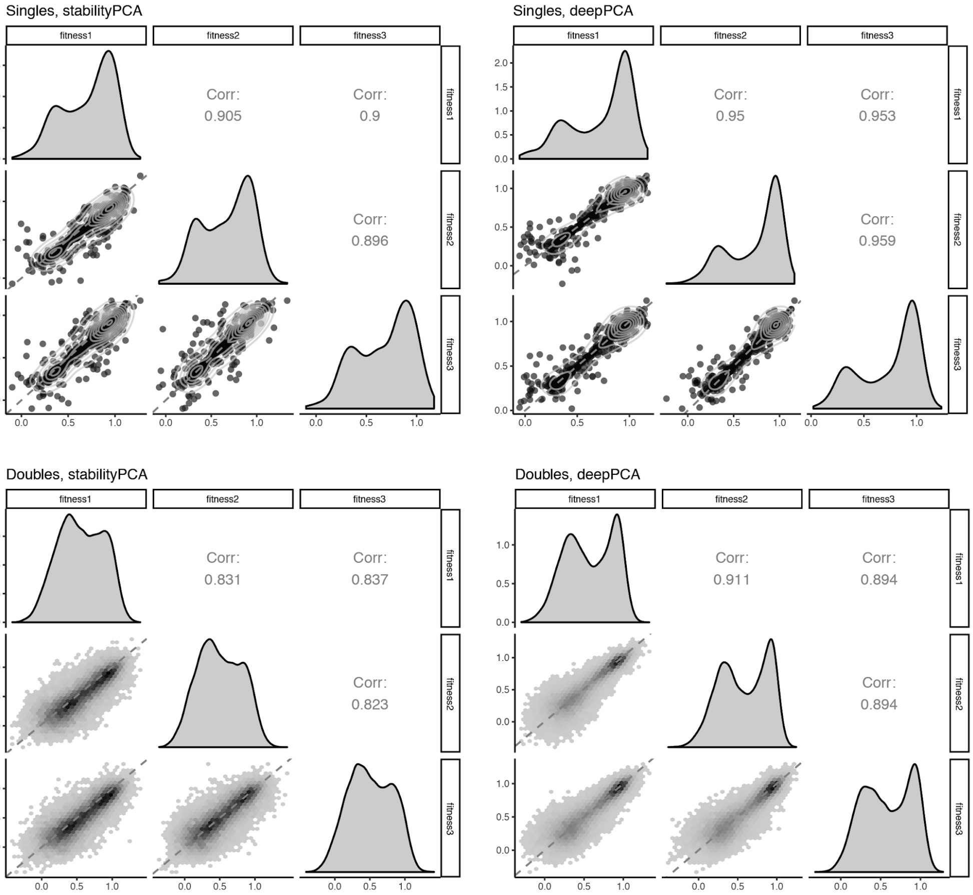
Supplementary Figures



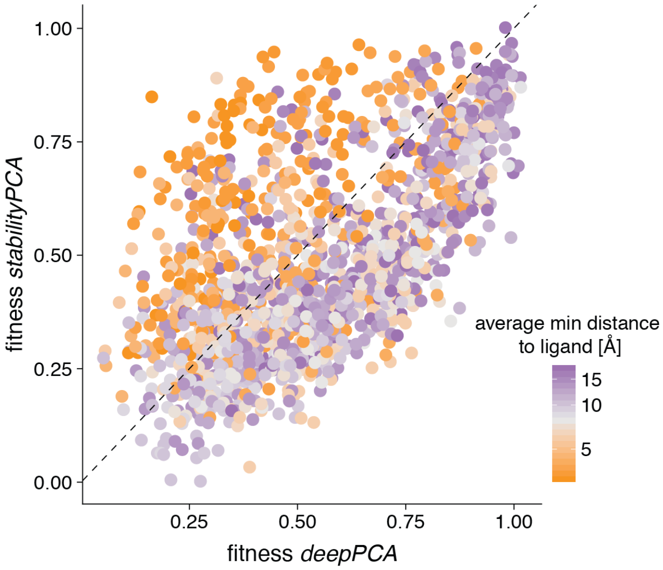
Supplementary Figure 1: Individual validation of GRB2 mutants for both *deepPCA* and *stabilityPCA* assays. (**a**) Individual growth rate measurements of independently constructed single mutants of GRB2. Growth rates are calculated as the slope of a linear fit of the $\log_{10}(OD_{600nm})$ against time during the exponential phase. For the binding assay, the positive control corresponds to the leucine zippers FOS and JUN fused to the two DHFR fragments, and the negative control is the wild-type GRB2 SH3 domain fused to DHFR3 in the *deepPCA* plasmid in the absence of its ligand GAB2 on the other DHFR fragment. For the stability assay, the positive control corresponds to the *stabilityPCA* plasmid itself, and in the negative control DHFR3 is fused to the degron sequence of the Matα2 protein. (**b**) Comparison of individually measured growth rates to fitness calculated from deep sequencing (Spearman correlation coefficient $\rho = 0.94$).



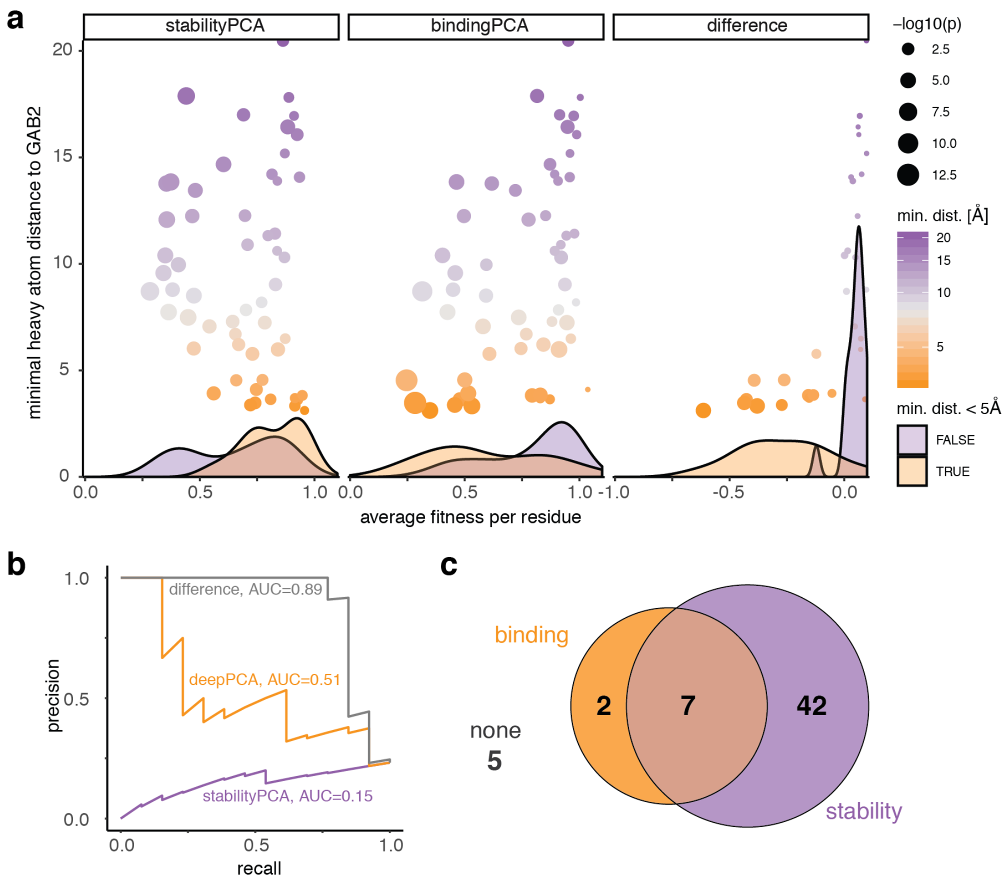
Supplementary Figure 2: Error prone PCR mutational frequency and bias. (**a**) Distribution of number of mutations per GRB2 coding sequence in the input libraries (before methotrexate selection). (**b**) Distribution of number of sequencing read counts grouped by mutation types.



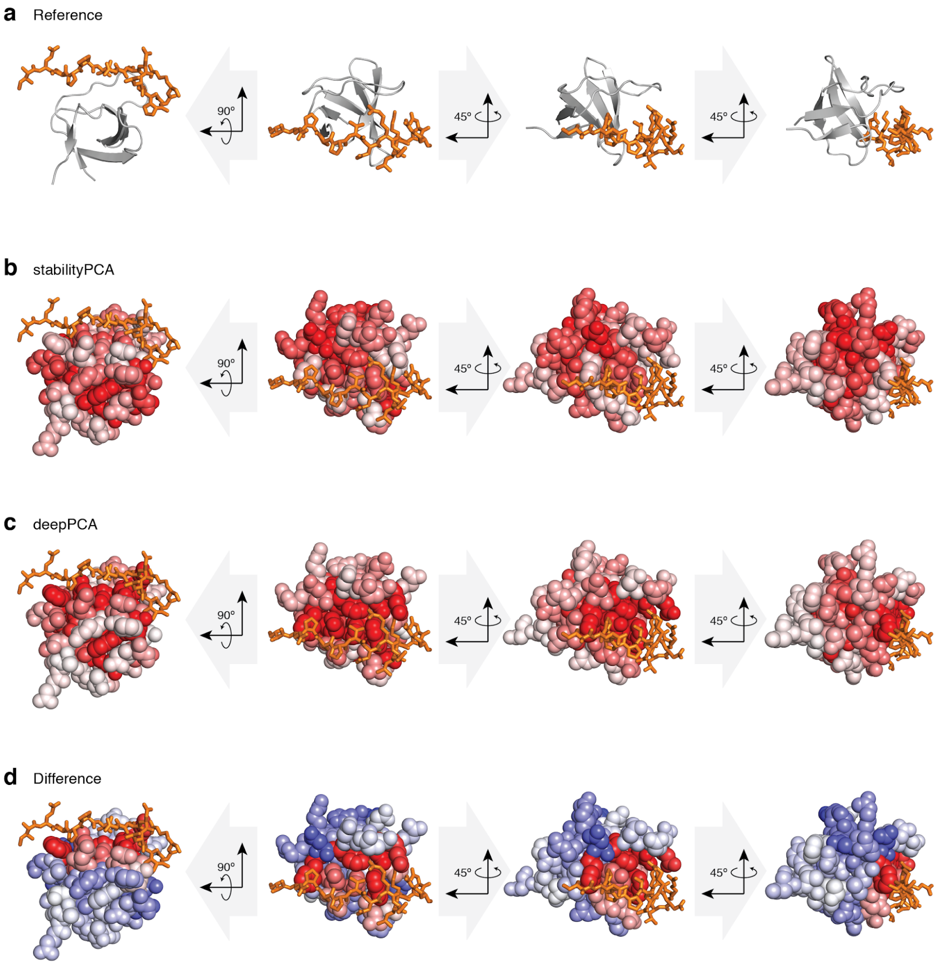
Supplementary Figure 3: Correlation of fitness estimates between biological replicates in both assays, for single and double amino acid mutations. Pearson correlation coefficient is indicated.



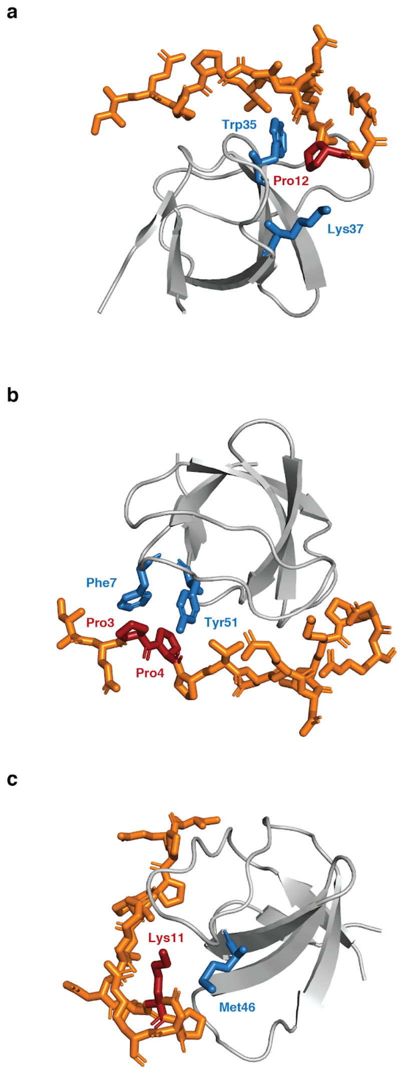
Supplementary Figure 4: Comparison of double mutant fitness in the binding and stability assays. Each genotype is coloured by the average distance of the two GRB2 amino acid positions to their most proximal GAB2 residue.



Supplementary Figure 5: (**a**) Relationship between average fitness per residue and distance from GAB2. Size of dots corresponds to false discovery rate adjusted p-value that mutations at the position affect the fitness (fitness < 1). Lower densities give fitness distributions for GRB2 residues in contact with GAB2 (<5Å, yellow) or not (blue). (**b**) Precision-recall curve when using the average fitness per position to classify residues that are less than 5Å from the closest residue of GAB2. (**c**) Classification of GRB2 positions that on average alter only stability, only binding (see Methods).



Supplementary Figure 6: Crystal structure of GRB2 SH3 bound to the linear peptide of GAB2 (**a**) Reference structure of the complex (PDB entry 2vwf [35]). GRB2 is coloured by fitness scores in the (**b**) stability assay, (**c**) binding assay and (**d**) Δfitness scores.



Supplementary Figure 7: *DoubleDeepPCA* identifies the type of interactions occurring between the GRB2 and GAB2 residues. GRB2 in grey (with highlighted residues in blue) and the ligand peptide of GAB2 in orange (with highlighted residues in red) (**a**) Lys37 is not identified as a contacting residue although found at <5Å from GAB2 Pro12. Pro12 is likely stabilized by GRB2 Trp35 instead, so amino acid substitutions at position 37 do not have any deleterious effect on binding. (**b**) Position 7 and 51 of GRB2 only allow aromatic residues that stabilize the interaction with GAB2 through aromatic interactions with Pro3 and Pro4. (**c**) GRB2 position 46 is near the positively charged Lys11 of GAB2, which can lead to a repulsive electrostatic interaction when Met46 is substituted for positively charged residues.

References

1. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

2. Rajagopala, S. V. *et al.* The binary protein-protein interaction landscape of Escherichia coli. *Nat. Biotechnol.* **32**, 285–290 (2014).

3. Simonis, N. *et al.* Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods* **6**, 47–54 (2009).

4. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).

5. Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).

6. Das, J. *et al.* Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Sci. Signal.* **6**, ra38 (2013).

7. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **440**, 637–643 (2006).

8. Havugimana, P. C. *et al.* A Census of Human Soluble Protein Complexes. *Cell* **150**, 1068–1081 (2012).

9. Guruharsha, K. G. *et al.* A Protein Complex Network of Drosophila melanogaster. *Cell* **147**, 690–703 (2011).

10. Arabidopsis Interactome Mapping Consortium. Evidence for network evolution in an Arabidopsis interactome map. *Science* **333**, 601–607 (2011).

11. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).

12. Jin, L., Wang, W. & Fang, G. Targeting Protein-Protein Interaction by Small Molecules. *Annual Review of Pharmacology and Toxicology* **54**, 435–456 (2014).

13. Marsh, J. A. & Teichmann, S. A. Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.* **84**, 551–575 (2015).

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

15. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, (2014).

16. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 67–72 (2009).

17. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by

proteome coevolution. *Science* **365**, 185–189 (2019).

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

19. Woodsmith, J. *et al.* Protein interaction perturbation profiling at amino-acid resolution. *Nat. Methods* **14**, 1213–1221 (2017).

20. Sahni, N. *et al.* Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* **23**, 649–657 (2013).

21. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).

22. Wang, X. *et al.* Fine Epitope Mapping of Two Antibodies Neutralizing the Bordetella Adenylate Cyclase Toxin. *Biochemistry* **56**, 1324–1336 (2017).

23. Doolan, K. M. & Colby, D. W. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *J. Mol. Biol.* **427**, 328–340 (2015).

24. Kowalsky, C. A. *et al.* Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *J. Biol. Chem.* **290**, 26457–26470 (2015).

25. Van Blarcom, T. *et al.* Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing. *J. Mol. Biol.* **427**, 1513–1534 (2015).

26. Wei, X. *et al.* A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).

27. Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).

28. Otwinowski, J. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).

29. Olson, C. A., Anders Olson, C., Wu, N. C. & Sun, R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology* **24**, 2643–2651 (2014).

30. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. doi:10.1101/484949

31. Michnick, S. W. Protein fragment complementation strategies for biochemical network mapping. *Curr. Opin. Biotechnol.* **14**, 610–617 (2003).

32. Pelletier, J. N., Campbell-Valois, F. X. & Michnick, S. W. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 12141–12146 (1998).

33. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* **7**, (2018).

34. Levy, E. D., Kowarzyk, J. & Michnick, S. W. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell Rep.* **7**, 1333–1340 (2014).

35. Harkiolaki, M. *et al.* Distinct binding modes of two epitopes in Gab2 that interact with the SH3C domain of Grb2. *Structure* **17**, 809–822 (2009).

36. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).

37. Johnson, P. R., Swanson, R., Rakhilina, L. & Hochstrasser, M. Degradation signal masking by heterodimerization of MATalpha2 and MATa1 blocks their mutual destruction by the ubiquitin-proteasome pathway. *Cell* **94**, 217–227 (1998).

38. Lund, A. H., Duch, M. & Pedersen, F. S. Increased cloning efficiency by temperature-cycle ligation. *Nucleic Acids Res.* **24**, 800–801 (1996).

39. Bolognesi, B., Faure, A. J., Seuma, M. & Schmiedel, J. M. The mutational landscape of a prion-like domain. *bioRxiv* (2019).

40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

41. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

42. Mitternacht, S. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Res.* **5**, 189 (2016).

3. Conclusions

3. CONCLUSIONS

The conclusions drawn from the results **section 2.1** are the following:

- The effects of all single point mutations that occurred during the evolution of a yeast tRNA can change from beneficial to deleterious depending on the genetic context in which they occur.
- Similarly to single mutations, genetic interactions can switch from positive to negative in different backgrounds due to the presence of higher-order epistasis.
- Accurate genetic prediction can be achieved with sparse models that include single mutation effects, pairwise and higher-order interactions that have been averaged across different genetic backgrounds.
- The abundance of sign epistasis in the tRNA landscape limits the number of accessible paths between genotypes.

The conclusions drawn from the results **section 2.2** are the following:

- We have developed a new methodology—*DoubleDeepPCA*—which combines systematic mutagenesis, protein-fragment complementation assays and deep sequencing for the rapid identification of the contact interface between proteins.
- A library containing ~50% of all possible single amino acid mutations of the SH3 domain of GRB2 was sufficient to map with high accuracy the interaction interface with the linear peptide of GAB2.
- Most mutations in GRB2 that disrupt the binding with GAB2 do so by destabilising the protein rather than affecting binding affinity alone.

4. DISCUSSION

4. DISCUSSION

The following section discusses the relevance of the results presented herein for the understanding of genetic and physical interactions using deep mutagenesis. It also addresses the potential of the methodologies used and the findings found for future matters of investigation. As shown in the results **section 2**, my research has focused on two topics: (I) exploring the relevance of higher-order genetic interactions during the evolution of a tRNA molecule and (II) the development of a methodology to map protein interaction interfaces at single amino acid resolution.

4.1. Extensive genetic interactions in a tRNA molecule

In order to understand how higher-order combinations of mutations combine together to influence phenotypes, we used a yeast tRNA molecule as a model system. Building a combinatorially-complete library of few substitutions that naturally occur during the evolution of this tRNA we could, for the first time in any gene, quantify the extent to which both the effects of individual mutations and the interactions between pairs of mutations change across closely-related genotypes. We found that all mutations can switch from beneficial to detrimental effects and all pairs switched from interacting positively to interacting negatively in different backgrounds.

4.1.1. The implications of pairwise and higher-order epistasis

Since the outcome of mutations depends on their coexistence with other mutations (epistasis) it is difficult to know beforehand what will happen when a mutation occurs. This suggests that predicting changes in phenotype from genotype changes might be a challenging task if genetic interactions are not taken into account. This is indeed what we saw when trying to predict the fitness effects of combinations of mutations in the yeast tRNA-Arg(CCU). First, we showed that the effects of mutations are substantially more informative for genetic prediction when they are measured and average across several different genetic backgrounds. Thus, evaluating the effects of mutations in a single reference individual would be of limited application for genetic prediction in another. Second, to a certain extent, pairwise and higher-order interactions are required for accurate predictions, making the mapping of phenotype from genotype a more challenging task than previously thought. However, in other molecular systems it has been shown that the use of mechanistic or non-mechanistic models that capture the nonspecific epistatic component (i.e. adapting the null model for how two mutations combine rather than using the additive combination of single mutant effects), reduces the number of interaction terms needed to make accurate genetic predictions (Diss & Lehner 2018; Kemble et al. 2018; Baeza-Centurion et al. 2019; Li et al. 2019). From these observations, two questions arise. First, can we identify the null model that captures the nonspecific component of epistasis in this tRNA? Second, even if we account for the global shape of genotype-phenotype map, are pairwise and higher-order genetic interactions indispensable for genetic prediction?

The most straightforward hint of global nonspecific epistasis in a DMS dataset is the clear systematic bias when epistatic terms are plotted against the fitness effects of single mutants, independently of their identity (e.g. in a protein-protein interaction, positive interactions tend to occur between mutations that weaken the interaction or when strength-increasing and strength-decreasing mutations are combined (Diss & Lehner 2018)). In the case of a combinatorial dataset, the evidence of global epistasis can be noticed when the effects of mutations follow a particular trend depending on the phenotype of the background they occur in (e.g. changes in the percentage of inclusion of an alternatively-spliced exon which depends on the starting inclusion level of the background genotype (Baeza-Centurion et al. 2019)). Surprisingly, we cannot find a clear systematic trend in the tRNA combinatorial dataset (Results **section 2.1**, **Extended Data Figure 2c** and **5a**). A first explanation could be that global epistasis does not exist in this tRNA. However, this is very unlikely given that the biophysical effects of mutations tend not to relate linearly with a measured phenotype. One of the clearest examples is thermodynamic stability, for which the changes in the free energy of folding of molecules and the fraction of correctly folded molecules covary in a nonlinear fashion. tRNAs are highly structured molecules and their three-dimensional conformation is essential for their role in protein synthesis.

Another explanation could be that we are just observing a small part of the genotype to phenotype map which coincides to be linearly related with the underlying additive traits. This could occur if all substitutions observed in the extant species would only have mildly deleterious effects on the yeast fitness, compared to other mutations in the tRNA. It would be interesting to repeat the experiment, but using random mutations rather than the ones observed in extant species.

A further explanation could be that the mapping from the underlying biophysical trait to fitness is far more complex than a single non-linearity. As explained in the introduction, tRNAs are molecules that are processed in many ways and interact with a huge number of proteins within the cell to properly function. The effect of mutations in each of these different crucial steps in the life cycle of a tRNA (which I imagine as different 'layers' of underlying additive traits) might not relate linearly to one another, and even less to the observed phenotype (fitness), leaving a 'noisy kind-of-linear' relationship between epistasis and background fitness effect. Thus, the convolution of the mapping between all these layers can be very difficult to disentangle by just observing a final fitness phenotype. Measuring the effects of mutations in much more concrete phenotypes (e.g. tRNA abundance, 5'-3' degradation propensity, changes in one post-transcriptional modification, etc.) would likely reveal some of these global epistatic trends (see **section 4.1.2** below). Finally, accounting for the global shape of the genotype–phenotype map not always provides a solution to the problems associated with genetic prediction. Even when global trends are taken into account, many significant pairwise and higher-order interactions may remain (specific interactions), hindering the prediction of phenotypes from genotypes (Sailer & Harms 2017a; Sailer & Harms 2017b; Poelwijk et al. 2019; Sailer & Harms 2017c).

We also found that in this tRNA landscape, sign and reciprocal epistasis is abundant, limiting the number of accessible paths between genotypes. This highlights the contingency of new mutations on previous mutations acquired. However, one must be cautious when extrapolating these implications to the evolutionary process itself. The rugged tRNA landscape we have characterized is one particular 'snapshot' of the multiple conformations that this can have. For instance, this experiment was done in one specific environment, opposite to the situation in nature, where the environmental conditions can fluctuate and change. The extent to which genetic interactions (especially sign epistasis) change across environmental conditions is still a question that needs to be addressed in a systematic manner.

Finally, one of the most key questions is the extent to which the conclusions we made from this tRNA will also apply to other molecules, including proteins and other RNA molecules. Do we have enough data to answer this question? So far, very few combinatorially-complete datasets of a large number of mutations have been reported (Palmer et al. 2015; Starr et al. 2017; Baeza-Centurion et al. 2019; Pokusaeva et al. 2019; Poelwijk et al. 2019; Wu et al. 2016). In some proteins, similar patterns can be observed, where in the presence of higher-order interactions, epistasis between two sites vary across different genetic backgrounds (Wu et al. 2016; Pokusaeva et al. 2019), even after factoring out global epistasis (Poelwijk et al. 2019). Still, with these few observations (which have been reached with different analytical methods), it is difficult to draw any conclusions. Thus, we need to generate more combinatorially-complete libraries in a variety of proteins and RNAs, together with the use of models that correctly estimate errors and biases in such kind of data (Rubin et al. 2017; Faure et al. 2019).

4.1.2. Unraveling the underlying mechanism of pairwise and higher-order interactions.

Even though our tRNA data can identify some very clear mechanistic insights into why mutations interact in a tRNA molecule (e.g. most robust positive interactions are mutations that restore Watson–Crick base pairs), we are missing the mechanistic basis for most of them. Identifying the molecular mechanism of genetic interactions from our tRNA library (without any clear global tendency of the effects of mutations) can be difficult for several reasons. One is the sparsity of mutations. We mutated 10 positions with a total of 14 possible substitutions that are sparsely located in different regions of the tRNA (in the acceptor and anticodon stems and in the variable loop). A possible solution to this issue would be to create a library of single, double and triple mutants in a particular region of the tRNA (only the anticodon arm for instance), covering more positions and nucleotide substitutions and allowing the measurement of higher-order interactions (third-order).

The mechanistic understanding of interactions is also hindered by 'biochemical pleiotropy', that is, whenever mutations can affect more than one biochemical parameter. As reviewed in the introduction section (see **section 1.2.2**), tRNAs have a complex life cycle, in which they interact with several dozens of proteins from their transcription until their contact with the ribosome. A mutation in the body of the tRNA could increase the transcription rate by creating a higher binding affinity motif for TFIIIC in the internal promoter B-box, but at the same time could incapacitate that position to be modified, increasing its propensity to be targeted by the 3'-5' exonucleolytic degradation machinery. Since the only phenotypic read out in our assay is fitness, and the effects of mutations can propagate through all these different biochemical 'layers', distinguishing the biological mechanism behind the effect of a mutation is difficult. To reduce this complexity, one could try to quantify the effect of mutations tackling each one of these different phenotypes at a time. For instance, we could do tRNA-seq (Zheng et al. 2015; Orioli 2017) on the yeast tRNA-Arg(CCU) library to measure the expression of each variant. Since most yeast genes encoding for tRNA modification enzymes are unessential in yeast (Hopper 2013), the library of tRNA variants could be assayed in different deletion strains. Similarly to Guy et al. (2014), where they measured the sensitivity of mutations to degradation by inactivating the rapid tRNA decay pathway, assaying the library of tRNA variants in strains which lack one of the tRNA modification enzymes would reveal which mutations and interactions change in the absence of a specific tRNA modification. Very recently, it has been shown that nanopore technology can be used to accurately detect the m6A RNA modifications in native yeast RNA sequences (Liu et al. 2019). It would be interesting to see if this can be extrapolated to tRNAs, which would allow to uncover how mutations alter tRNA modifications along the entire gene body. However, it can be a particularly difficult challenge, given the high error rate of this sequencing technology and the low genetic diversity in our tRNA library. Finally, it would be interesting to monitor how changes in the tRNA affect its interaction with other proteins in a more direct way. This could be achieved using yeast three-hybrid assays (Jaeger et al. 2004), but so far, only one functional three-hybrid system isolated from a couple of tRNAs and aminoacyl-tRNA synthetase have been reported (Zheng et al. 2004).

4.2. Development of a new methodology to map protein interaction interfaces

As shown in the results **section 2.2**, we have developed a new experimental methodology—*DoubleDeepPCA*—that takes advantage of systematic mutagenesis, protein complementation assays and deep sequencing to map the interaction interface of proteins at amino acid resolution. By quantifying the effects of mutations on both protein binding and stability, we could determine with high accuracy the surface of contact between the SH3 domain of GRB2 and the linear peptide of GAB2 without using any prior structural information.

4.2.1. Advantages of *DoubleDeepPCA* for the identification of protein interaction interfaces

This methodology provides some advantages that make it suitable for determining protein-protein interaction interfaces in a systematic manner. First, contrary to experimental approaches such as X-ray or NMR, *DoubleDeepPCA* is an *in vivo* method. The effects of mutations on binding and stability are measured inside the cell, so the mapped interaction interface between two proteins might reflect the one occurring in physiological conditions. Currently available structural information on protein complexes is heavily biased towards rigid proteins, because molecule flexibility tends to increase the difficulty of crystallization, which results in lower-resolution crystal structures (Marsh & Teichmann 2015). *DoubleDeepPCA* overcomes this problem by not requiring crystallization. Second, the phenotypes screened in our assay—growth coupled to protein stability and binding—are independent of the function of the protein of interest. Since the only requirement for the proper function in our method is a stable folding of the protein of interest in the yeast cell, the assay is highly generalizable, allowing the screening of protein interactions from any species other than yeast (e.g. human GRB2 SH3 domain). Third, since *DoubleDeepPCA* is based on protein-fragment complementation assay, target proteins do not need to be confined in the nucleus to trigger an interaction signal (such in the case of yeast two-hybrid techniques) but can be located in other cell compartments, such as the cytosol. Fourth, this methodology is highly quantitative and it opens up the possibility of identifying mutations that might strengthen the interaction between the two proteins. Fifth, a protein of interest with multiple interactors would require one binding assay for each of the different binding partners, but only one screen for stability. Sixth, since *DoubleDeepPCA* does not require prior information on homologous sequences, it can be useful to identify the contacting residues between proteins that have a recent evolutionary history, which are not accessible candidates for coevolutionary analysis (they have too short MSA with few homologous sequences) (Ovchinnikov et al. 2014; Cong et al. 2019). *DoubleDeepPCA* would also prove useful in cases where interacting proteins share less than 30% sequence identity with their homologs, for which attempts of homology modeling or *ab-initio* docking methodologies have shown limited power (Mosca, Pons, et al. 2013; Negroni et al. 2014).

One of the main advantages of our technique is its scalability. Only a few mutations per position are required to identify the amino acids involved in the protein-protein interaction interface (e.g. ~50% of all total possible 1,064 single amino acid substitutions in the case of GRB2 SH3), which simplifies and reduces the difficulty and cost of the entire experiment. The complexity of a single amino acid mutant library is far lower than a double mutant amino acid library (e.g. for GRB2 SH3 $56 \cdot 19 = 1{,}064$ singles compared to $(56 \cdot 19 \cdot 55 \cdot 19)/2 = 555{,}940$ doubles). This makes the construction of the mutant library quick and straightforward (e.g. using one-pot single-day saturation mutagenesis (Wrenbeck et al. 2016)). It also reduces the entire cost of the experiment because it requires less sequencing, currently (without considering the researcher's salary) the most expensive part of a DMS. For instance, in the error prone PCR mutant library of GRB2, which we targeted to cover for most single and some double amino acid mutations, less than 50% of the sequencing reads belonged to the wild-type and the single amino acid mutations, with an average coverage of >6,000 paired-end reads per single amino acid mutant. Additionally, since the reduction in library complexity translates into a reduction of the variant pool population size (which in turn translates into lower cell culture volumes), assays can be conducted in parallel for several proteins of interest with the respective binding partners. For instance, if we would like to perform the assay on several protein domains of the same size as the SH3 of GRB2, the complexity of the target libraries would be $56 \cdot 19/0.6 = 1{,}773$—the number of variants needed to get all single amino acid mutations assuming that nicking mutagenesis approaches (Wrenbeck et al. 2016) usually leave ~60% of variants in the pool with exactly one codon mutated. Such complexity would require a total of $\sim 2 \cdot 10^4$ to $2 \cdot 10^5$ transformed yeast cells (to ensure that each variant is incorporated in 10 to 100 different yeast cells). This efficiency is relatively easy to achieve using the standard yeast lab strain, so multiple transformations for different proteins and replicates could be performed at once. Finally, the competition could be performed in small cell culture volumes of 5 mL. Since in 1 mL of minimal media at an $OD_{600nm} = 1$ contains $\sim 1.8 \cdot 10^7$ yeast cells, each variant would be covered >2,000 times at the beginning of the selection experiment if the starting OD_{600nm} is 0.05 ($1.8 \cdot 10^7 \cdot 0.05 \cdot 5/1{,}773 = 2{,}538$). Thus, the binding and stability assay for 4 different proteins domains could be performed in triplicate in a single 24 deep well plate.

Additionally, *DoubleDeepPCA* provides further information for the understanding of physical interactions of proteins. On one hand, by analysing the effects of different amino acid substitutions in the positions of the interacting interface, it is possible to discriminate the type of protein-ligand interactions and identify the mechanisms of disruption between proteins. On the other hand, it allows discriminating which mutations disrupt an interaction due to the destabilization of one of the proteins (i.e. node removal) or due to the alteration of the affinity between two proteins (i.e. edgetic perturbation). This can help to understand the role of disease mutations and grant relevant knowledge for the design and engineering of proteins. Finally, as shown by previous work (Otwinowski 2018; Li et al. 2019), fitting thermodynamic models to the data generated by *DoubleDeepPCA* offers the opportunity to infer how mutations change the energies of folding and binding of proteins *in vivo*. This can reveal binding and stability threshold robustness, where some mutations alter the free energy of binding or folding by a certain amount, without impairing bound and folded states of the proteins (i.e. a set of 'hidden' mutations that do change the free energy but by a small amount that does not affect the assay).

4.2.2. Limitations of the *DoubleDeepPCA* methodology

Although this methodology provides some advantages that make it useful as an orthogonal approach to determine protein interaction interfaces from currently existing methodologies, it has some limitations. The first limitation, in comparison to computational approaches, is the time and resources required to obtain the data. Initially, it is necessary to clone the wild-type interacting protein in the constructs (and the respective positive and controls if available) to ensure that the proteins are suited for the assay (i.e. the protein of interest correctly binds the partner and is stable providing a good growth signal in both *deepPCA* and *stabilityPCA* assays), which usually takes about two weeks (including cloning, Sanger sequencing confirmation of each construct and the methotrexate assay in 96 well plates for at least 50 hours). The following step involves generating the mutant library (which usually takes two or three days) and cloning it into the assays plasmids (two more days). After that, the pool of variants are transformed into yeast for the selection assay. Since single mutant libraries are less complex and the cell culture volumes required are small, both binding and stability assays can be performed in parallel. The steps from yeast transformation to DNA extraction take another week of work. Finally, we need to add two or three weeks for the sequencing library preparation, the next-generation sequencing and the basic processing of the raw data to obtain fitness and Δfitness scores. In total, from the protein candidate selection to the first preliminary results, it would take 45-60 days to identify the residues involved in the interaction between two proteins. This time can be reduced since most of the steps in the protocol can be parallelised for libraries of different proteins.

It still needs to be shown whether it is possible to identify a small binding interface of a large protein. So far, we have applied the methodology to a relatively short protein domain of 56 amino acids. As the number of mutations introduced can be expected to scale linearly with protein length, the library of mutants required would increase in complexity. However, efficient mutagenesis protocols have been used to obtain single amino acid mutant libraries of long proteins (e.g. the TEM-1 beta-lactamase (Firnberg et al. 2014) or the beta-2 adrenergic receptor β_2AR (Jones et al. 2019) of 2,583 and 7,828 amino acids respectively). Variant pools for longer proteins would need to be kept at larger population sizes at all times to avoid bottlenecking the complexity of the library, requiring cell cultures bigger than a few milliliters. Nevertheless, the complexity of a single amino acid mutant library for a long protein (e.g. 19 · 2,000 = 38,000 possible single mutants for a 2,000 amino acids long protein) would not be as intricate as that of a library of double amino acid mutations of a short protein domain (e.g. 56 · 19 + (19 · 56 · 19 · 55)/2 = 557,004 single and double amino acid mutations for a 56 amino acid long protein domain (Olson et al. 2014)). Longer proteins would definitely require a reformulation of the sequencing strategy, since paired end reads would not be able to cover the entire mutated coding region. Since associating unique barcodes to the protein variants has been shown to solve this issue (Kitzman et al. 2015) (see **section 1.1.1**), in both the stability and binding assay plasmids we introduced a barcode cloning site for the rapid incorporation of random barcodes.

Another limitation of this technique is that is not applicable to all protein pairs. *DoubleDeepPCA* requires stable growth of the wild-type proteins in both the *deepPCA* and *stabilityPCA* assays, and sometimes this is not the case for some proteins (see **Figure X**). Some proteins allow growth in the stability assay, but do not provide a proper PCA signal for the binding one (such as the Ras binding domain of BRAF or RAF1 tested for binding with HRAS). This could be because the interaction partner is not well-expressed in yeast or because one or both proteins require some post-translational modifications. It can also occur in the opposite situation, where the binding assay works fine, but the stability assay does not provide a good readout. This is the case for the BRCA1 ring domain, which is marginally stable in yeast, consistent with data from other methods in mammalian cells (Matreyek et al. 2018). Increasing the expression of the protein fused to the DHFR fragment might provide a better signal.

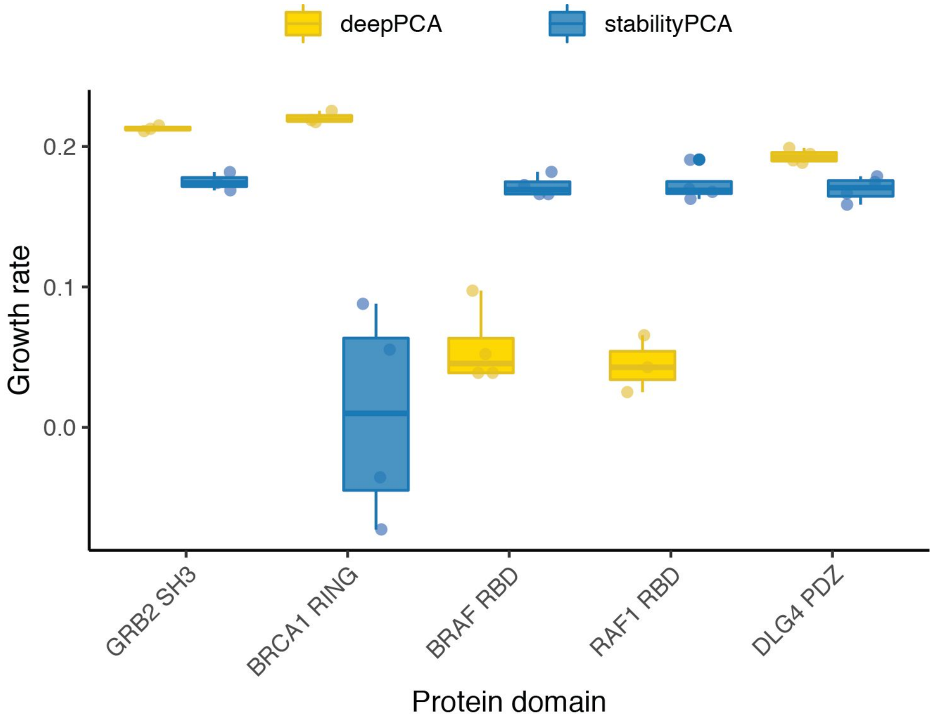


Figure X: *DoubleDeepPCA* is not applicable to all protein pairs. Individual growth rate measurements of five different protein domains with its respective interactor (labeled between brackets). Growth rates are calculated as the slope of a linear fit of the $\log_{10}(OD_{600nm})$ against time during the exponential phase.

However, there might be other reasons that can difficult the applicability of the assay to other proteins. For instance, some proteins might tolerate a C-terminal fusion of the DHFR fragment instead of an N-terminal fusion, the one currently used in our method. Also, our current methodology cannot be applied to transmembrane proteins, since it has been shown previously that in the *stabilityPCA* assay the overexpressed DHFR1,2 fragment is mainly found in the cytosol so the abundances of proteins localized in other subcellular compartments, such as membranes, are underestimated (Levy et al. 2014). This issue can be solved by fusing a transmembrane helix to the DHFR1,2, which would target it to a certain membrane (e.g. plasma membrane), increasing the PCA signal since the protein of interest is located at the same location (Levy et al. 2014). Thus, in order to extend the repertoire of possible proteins to be targeted by our assay, with particular interest for transmembrane proteins, we could build vectors that contain different transmembrane helices and localize the DHFR fragments to membrane compartments. Obtaining the *stabilityPCA* plasmid that contains the optimal combination of these three features (C-tag or N-tag, correct promoter expression level and presence or absence of different localization transmembrane helices) would require testing each protein in each of these different vectors. This could represent a substantial amount of work if we were to combine, for instance, the two different tags, with five promoters with varying expression level and five different transmembrane helices (to localize DHFR in the mitochondria, endoplasmic reticulum, Golgi apparatus, lysosome or plasma membrane), leaving a total of 2 ▪ 5 ▪ 5 = 50 different vector combinations. To speed up this process we could introduce a barcode in each of these different vectors that would uniquely identify each of them, and simultaneously clone our proteins of interest into these pool of vectors. By doing a small scale methotrexate assay and shallow sequencing with paired-end reads the barcode and part of the coding region of the tested protein, we could identify the best vector-protein combination for a bunch of proteins of interest in a single experiment.

4.2.3. Additional applications of *DoubleDeepPCA*

Finally, our methodology could be used for alternative purposes. One of them is the identification of allosteric sites in proteins. Allostery is the phenomenon whereby a perturbation by an effector molecule on one site of a protein leads to a functional change at another remote site. Proteins exist as an ensemble of different conformations. Allostery arises when binding at one site alters the free energy of one of these and changes how structures are distributed across the ensemble (Sailer & Harms 2017c; Gunasekaran et al. 2004; Motlagh et al. 2014). Thus, if we apply our assay on a protein that changes its structural conformation upon binding, the distal residues that energetically contribute the new conformation of the ensemble (i.e. contribute to the binding energy with the other protein) would appear as hit residues as the ones located directly at the interface. Even if the structure of the bound proteins was unknown (and only the monomeric structures were available), the allosteric site distant to the binding interface could be identified by clustering all the hit positions identified in the assay by their pairwise three-dimensional distances. Residues in the allosteric site would cluster apart from the ones on the binding interface. The identification of allosteric sites in proteins would provide a huge advantage to the development of new therapeutics (Nussinov & Tsai 2013). Drugs that target allosteric sites are highly specific since they do not bind to active sites of proteins, which tend to be highly conserved across protein families. They also allow the modulation of protein activity, rather than completely killing it, and they act only when the target protein is functioning. As a proof of concept, we would apply *DoubleDeepPCA* to a PDZ domain, which can be assayed for binding and stability (Figure X), and has been shown to have an allosteric site located on the opposite side of the ligand-binding pocket (Peterson et al. 2004; Jr et al. 2012).

Another possibility that this methodology offers is the identification of resistance mutations to protein interaction inhibitors. Protein-protein interactions are attractive drug targets since alterations on the edges of the interactome are the cause of many disorders (Zhong et al. 2009; Sahni et al. 2015). Drugs that target protein interactions present some advantages over more traditional protein inhibitors, one of them being a more specific form of regulation that can avoid side effects due to node removal (Duran-Frigola et al. 2013). An example is the antiapoptotic Bcl-2 protein, which is usually overexpressed in solid human tumors and contributes to cancer progression by binding the BH3 domains of Bax and Bak proteins and blocking their proapoptotic function. Contrary to Bcl2-inhibitors, which have been shown to be nonspecific (i.e. altering other cellular targets) and generate adverse toxicity effects (Zinzalla & Thurston 2009), drugs that target Bcl2 interactions serve as anti-cancer treatment without the off-target effects (Gandhi et al. 2011). Identifying mutations that impair the protein interaction inhibitory effect of a certain drug would help to identify mutations that might present resistance to the drug treatment. *DoubleDeepPCA* could be used for that purpose by adding protein-protein interaction inhibitor in the selection assay. Since the wild-type proteins would be unbound in presence of the inhibitor, the current positive assay of *DoubleDeepPCA* would lead to low-confidence estimates of the wild-type reference, thus a negative selection strategy where binding is deleterious for growth would be required. The yeast cytosine deaminase (yCD) can be used as a reporter in PCA for both positive and negative selection (Ear & Michnick 2009). The deletion of the FCY1 gene encoding the enzyme, which converts cytosine into uracil, renders the strain defective for the pyrimidine salvage pathway unable to *the novo* synthesize uracil, thus cannot grow in the absence of uracil. In the positive selection assay, the interaction of two proteins brings the complementary fragments of yCD into proximity, allowing them to fold, reconstitute its catalytic activity and restore cell growth. The negative selection is achieved when the same strain is treated with 5-fluorocytosine (5-FC), a nontoxic compound that is converted to toxic 5-fluorouridine triphosphate (5-FUTP) in a pathway that depends on yCD activity. Thus, the binding of two proteins fused to the yCD fragments impairs cell growth in media complemented with uracil and 5-FC. Using the yCD as a PCA reporter instead of DHFR would allow to run both positive and negative selection assays using the same mutant library.

Finally, our methodology could be used to guide computational docking strategies for the three-dimensional determination of protein complexes. Mapping the residues responsible for the physical interaction in one of the two proteins would reduce the conformational space to be sampled (i.e. focusing the docking in a reduced area of the protein's surface). Information about the type of protein-protein interactions that occur between the two proteins could add some further structural restraints to the docking simulations. However, since the effects of mutations are only quantified in one of the two proteins, this would be of limited applicability in cases were the non-mutated interactor was longer than a linear peptide. In order to overcome this issue, one could repeat the selection assay by mutating the other interacting partner in an alanine-scan fashion (i.e. substituting all positions for alanine) together with a library of mutants in the first protein restricted to the previously identified interface residues.

References

Aakre, C.D. et al., 2015. Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell*, 163(3), pp.594–606.

Achsel, T. & Gross, H.J., 1993. Identity determinants of human tRNA(Ser): sequence elements necessary for serylation and maturation of a tRNA with a long extra arm. *The EMBO Journal*, 12(8), pp.3333–3338. Available at: http://dx.doi.org/10.1002/j.1460-2075.1993.tb06003.x.

Agris, P.F., Vendeix, F.A.P. & Graham, W.D., 2007. tRNA's wobble decoding of the genome: 40 years of modification. *Journal of molecular biology*, 366(1), pp.1–13.

Albers, M. et al., 2005. Automated yeast two-hybrid screening for nuclear receptor-interacting proteins. *Molecular & cellular proteomics: MCP*, 4(2), pp.205–213.

Alexandrov, A. et al., 2006. Rapid tRNA decay can result from lack of nonessential modifications. *Molecular cell*, 21(1), pp.87–96.

Aloy, P. et al., 2003. The relationship between sequence and interaction divergence in proteins. *Journal of molecular biology*, 332(5), pp.989–998.

Arabidopsis Interactome Mapping Consortium, 2011. Evidence for network evolution in an Arabidopsis interactome map. *Science*, 333(6042), pp.601–607.

Araya, C.L. et al., 2012. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42), pp.16858–16863.

Aström, S.U. & Byström, A.S., 1994. Rit1, a tRNA backbone-modifying enzyme that mediates initiator and elongator tRNA discrimination. *Cell*, 79(3), pp.535–546.

Auerbach, C. & Robson, J.M., 1946. Chemical production of mutations. *Nature*, 157, p.302.

Avery, O.T., Macleod, C.M. & 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. *The Journal* *of experimental medicine*, 79(2), pp.137–158.

Baeza-Centurion, P. et al., 2019. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell*, 176(3), pp.549–563.e23.

Bank, C. et al., 2015. A systematic survey of an intragenic epistatic landscape. *Molecular biology and evolution*, 32(1), pp.229–238.

Bank, C. et al., 2016. On the (un)predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 113(49), pp.14085–14090.

Barabási, A.-L. & Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Nature reviews. Genetics*, 5(2), pp.101–113.

Barnard, E. & Timson, D.J., 2010. Split-EGFP Screens for the Detection and Localisation of Protein–Protein Interactions in Living Yeast Cells. In A. Sharon, ed. *Molecular and Cell Biology Methods for Fungi*. Totowa, NJ: Humana Press, pp. 303–317.

Bartlett, J.M.S. & Stirling, D., 2003. A short history of the polymerase chain reaction. *Methods in molecular biology* , 226, pp.3–6.

Bhagavatula, G. et al., 2017. A Massively Parallel Fluorescence Assay to Characterize the Effects of Synonymous Mutations on TP53 Expression. *Molecular cancer research: MCR*, 15(10), pp.1301–1307.

Bloom-Ackermann, Z. et al., 2014. A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *PLoS genetics*, 10(1), p.e1004084.

Bloom, J.D., 2015. Software for the analysis and visualization of deep mutational scanning data. *BMC bioinformatics*, 16, p.168.

Boder, E.T. & Wittrup, K.D., 1997. Yeast surface display for screening combinatorial polypeptide libraries. *Nature biotechnology*, 15(6), pp.553–557.

Bolognesi, B. et al., 2019. The mutational landscape of a prion-like domain. *Nature communications*, 10(1), p.4162.

Boxem, M. et al., 2008. A protein domain-based interactome network for C. elegans early embryogenesis. *Cell*, 134(3), pp.534–545.

Brannigan, J.A. & Wilkinson, A.J., 2002. Protein engineering 20 years on. *Nature reviews. Molecular cell biology*, 3(12), pp.964–970.

Braun, S. et al., 2018. Decoding a cancer-relevant splicing decision in the RON proto-oncogene using high-throughput mutagenesis. *Nature communications*, 9(1), p.3315.

Brückner, A. et al., 2009. Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6), pp.2763–2788.

Cabantous, S. et al., 2013. A new protein-protein interaction sensor based on tripartite split-GFP association. *Scientific reports*, 3, p.2854.

Cadwell, R.C. & Joyce, G.F., 1994. Mutagenic PCR. *PCR methods and applications*, 3(6), pp.S136–40.

Cadwell, R.C. & Joyce, G.F., 1992. Randomization of genes by PCR mutagenesis. *PCR methods and applications*, 2(1), pp.28–33.

Chan, P.P. & Lowe, T.M., 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic acids research*, 44(D1), pp.D184–9.

Chan, P.P. & Lowe, T.M., 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic acids research*, 37(Database issue), pp.D93–7.

Chen, K. & Arnold, F.H., 1993. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12), pp.5618–5622.

Chernyakov, I., Whipple, J.M. & Kotelawala, L., 2008. Degradation of several hypomodified mature tRNA species in Saccharomyces cerevisiae is mediated by Met22 and the 5'–3' exonucleases Rat1 and Xrn1. *Genes*. Available at: http://genesdev.cshlp.org/content/22/10/1369.short.

Chiasson, M. & Fowler, D.M., 2019. Mutagenesis-based protein structure determination. *Nature genetics*, 51(7), pp.1072–1073.

Cong, Q. et al., 2019. Protein interaction networks revealed by proteome coevolution. *Science*, 365(6449), pp.185–189.

Crick, F. & Watson, J., 1953. A structure for deoxyribose nucleic acid. *Nature*. Available at: http://eduardbardaji.com/DOCENCIA/prodnat/nucleicacids03.doc.

Cuperus, J.T. et al., 2017. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome research*, 27(12), pp.2015–2024.

Das, J. et al., 2013. Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Science signaling*, 6(276), p.ra38.

David, A. et al., 2012. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human Mutation*, 33(2), pp.359–363. Available at: http://dx.doi.org/10.1002/humu.21656.

Diss, G. & Lehner, B., 2018. The genetic landscape of a physical interaction. *eLife*, 7. Available at: http://dx.doi.org/10.7554/eLife.32472.

Doolan, K.M. & Colby, D.W., 2015. Conformation-dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *Journal of molecular biology*, 427(2), pp.328–340.

Doud, M.B. & Bloom, J.D., 2016. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*, 8(6). Available at: http://dx.doi.org/10.3390/v8060155.

Dreze, M. et al., 2009. “Edgetic” perturbation of a C. elegans BCL2 ortholog. *Nature methods*, 6(11), pp.843–849.

Dünkler, A., Müller, J. & Johnsson, N., 2012. Detecting Protein–Protein Interactions with the Split-Ubiquitin Sensor. In B. Deplancke & N. Gheldof, eds. *Gene Regulatory Networks: Methods and Protocols*. Totowa, NJ: Humana Press, pp. 115–130.

Duran-Frigola, M., Mosca, R. & Aloy, P., 2013. Structural systems pharmacology: the role of 3D structures in next-generation drug development. *Chemistry & biology*, 20(5), pp.674–684.

Dvir, S. et al., 2013. Deciphering the rules by which 5’-UTR sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30), pp.E2792–801.

Ear, P.H. & Michnick, S.W., 2009. A general life-death selection strategy for dissecting protein functions. *Nature methods*, 6(11), pp.813–816.

Eigen, M., 1985. Macromolecular Evolution: Dynamical Ordering in Sequence Space. *Berichte der Bunsengesellschaft für physikalische Chemie*, 89(6), pp.658–667.

El Yacoubi, B., Bailly, M. & de Crécy-Lagard, V., 2012. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annual review of genetics*, 46, pp.69–95.

Ernst, A. et al., 2010. Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Molecular bioSystems*, 6(10), pp.1782–1790.

Ewing, R.M. et al., 2007. Large‐scale mapping of human protein–protein interactions by mass spectrometry. *Molecular systems biology*, 3(1). Available at: https://www.embopress.org/doi/full/10.1038/msb4100134 [Accessed September 1, 2019].

Faure, A.J. et al., 2019. DiMSum: a pipeline for processing deep mutational scanning data and diagnosing common experimental pathologies. *In prep.*

Fields, S. & Song, O., 1989. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230), pp.245–246.

Findlay, G.M. et al., 2018. Accurate classification of BRCA1 variants with saturation genome editing. *Nature*, 562(7726), pp.217–222.

Firnberg, E. et al., 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular biology and evolution*, 31(6), pp.1581–1592.

Firnberg, E. & Ostermeier, M., 2012. PFunkel: efficient, expansive, user-defined mutagenesis. *PloS one*, 7(12), p.e52031.

Fowler, D.M. et al., 2011. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* , 27(24), pp.3430–3431.

Freschi, L. et al., 2013. qPCA: a scalable assay to measure the perturbation of protein--protein interactions in living cells. *Molecular bioSystems*, 9(1), pp.36–43.

Fujino, Y. et al., 2012. Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochemical and biophysical research communications*, 428(3), pp.395–400.

Galli, G., Hofstetter, H. & Birnstiel, M.L., 1981. Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. *Nature*, 294(5842), pp.626–631.

Gandhi, L. et al., 2011. Phase I study of Navitoclax (ABT-263), a novel Bcl-2 family inhibitor, in patients with small-cell lung cancer and other solid tumors. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 29(7), pp.909–916.

Gerber, A.P. & Keller, W., 1999. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*, 286(5442), pp.1146–1149.

Gerber, A.P. & Keller, W., 2001. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends in biochemical sciences*, 26(6), pp.376–384.

Giot, L. et al., 2003. A protein interaction map of Drosophila melanogaster. *Science*, 302(5651), pp.1727–1736.

Göbel, U. et al., 1994. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4), pp.309–317.

Gunasekaran, K., Ma, B. & Nussinov, R., 2004. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3), pp.433–443.

Günther, S. et al., 2007. Docking without docking: ISEARCH—prediction of interactions using known interfaces. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.839–844.

Guruharsha, K.G. et al., 2011. A Protein Complex Network of Drosophila melanogaster. *Cell*, 147(3), pp.690–703.

Guy, M.P. et al., 2014. Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. *Genes & Development*, 28(15), pp.1721–1732. Available at: http://dx.doi.org/10.1101/gad.245936.114.

Havugimana, P.C. et al., 2012. A Census of Human Soluble Protein Complexes. *Cell*, 150(5), pp.1068–1081.

Hayden, E.J., Bendixsen, D.P. & Wagner, A., 2015. Intramolecular phenotypic capacitance in a modular RNA molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 112(40), pp.12444–12449.

Hendrickson, T.L., 2001. Recognizing the D-loop of transfer RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24), pp.13473–13475.

Hiatt, J.B. et al., 2010. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature methods*, 7(2), pp.119–122.

Hietpas, R.T., Jensen, J.D. & Bolon, D.N.A., 2011. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 108(19), pp.7896–7901.

Hopf, T.A. et al., 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3. Available at:

http://dx.doi.org/10.7554/eLife.03430.

Hopper, A.K., 2013. Transfer RNA post-transcriptional processing, turnover, and subcellular dynamics in the yeast Saccharomyces cerevisiae. *Genetics*, 194(1), pp.43–67.

Hopper, A.K. & Phizicky, E.M., 2003. tRNA transfers to the limelight. *Genes & development*, 17(2), pp.162–180.

Hutchison, C.A., 3rd et al., 1978. Mutagenesis at a specific position in a DNA sequence. *The Journal of biological chemistry*, 253(18), pp.6551–6560.

Ikemura, T., 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, 146(1), pp.1–21. Available at: http://dx.doi.org/10.1016/0022-2836(81)90363-6.

Inouye, C. et al., 1994. Isolation of a cDNA encoding a metal response element binding protein using a novel expression cloning procedure: the one hybrid system. *DNA and cell biology*, 13(7), pp.731–742.

Jaeger, S., Eriani, G. & Martin, F., 2004. Results and prospects of the yeast three-hybrid system. *FEBS letters*, 556(1-3), pp.7–12.

Jain, P.C. & Varadarajan, R., 2014. A rapid, efficient, and economical inverse polymerase chain reaction-based method for generating a site saturation mutant library. *Analytical biochemistry*, 449, pp.90–98.

Jin, L., Wang, W. & Fang, G., 2014. Targeting protein-protein interaction by small molecules. *Annual review of pharmacology and toxicology*, 54, pp.435–456.

Johnsson, N. & Varshavsky, A., 1994. Split ubiquitin as a sensor of protein interactions in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 91(22), pp.10340–10344.

Jones, E.M. et al., 2019. Structural and Functional Characterization of G Protein-Coupled Receptors with Deep Mutational Scanning. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/623108v1.abstract.

Jr, R.N.M. et al., 2012. The spatial architecture of protein function and adaptation. *Nature*, 491(7422), pp.138–142. Available at: http://dx.doi.org/10.1038/nature11500.

Julien, P. et al., 2016. The complete local genotype–phenotype landscape for the alternative splicing of a human exon. *Nature Communications*, 7(1). Available at:

http://dx.doi.org/10.1038/ncomms11558.

Kanaya, S. et al., 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *Journal of molecular evolution*, 53(4-5), pp.290–298.

Kar, G., Gursoy, A. & Keskin, O., 2009. Human cancer protein-protein interaction network: a structural perspective. *PLoS computational biology*, 5(12), p.e1000601.

Kauffman, S.A., 1993. *The Origins of Order: Self-organization and Selection in Evolution*, Oxford University Press.

Kemble, H. et al., 2018. Flux, toxicity and protein expression costs shape genetic interaction in a metabolic pathway. *bioRxiv*, p.362327. Available at: https://www.biorxiv.org/content/10.1101/362327v2 [Accessed September 20, 2019].

Ke, S. et al., 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome research*, 21(8), pp.1360–1374.

Khan, S.H., 2019. Genome-Editing Technologies: Concept, Pros, and Cons of Various Genome-Editing Techniques and Bioethical Concerns for Clinical Application. *Molecular therapy. Nucleic acids*, 16, pp.326–334.

Kitzman, J.O. et al., 2015. Massively parallel single-amino-acid mutagenesis. *Nature methods*, 12(3), pp.203–6, 4 p following 206.

Kowalsky, C.A., Klesmith, J.R., et al., 2015. High-resolution sequence-function mapping of full-length proteins. *PloS one*, 10(3), p.e0118193.

Kowalsky, C.A., Faber, M.S., et al., 2015. Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *The Journal of biological chemistry*, 290(44), pp.26457–26470.

Krogan, N.J. et al., 2006. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084), pp.637–643.

Kundrotas, P.J. et al., 2012. Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24), pp.9438–9441.

Kundrotas, P.J. & Vakser, I.A., 2013. Global and local structural similarity in protein-protein complexes: Implications for template-based docking. *Proteins: Structure, Function, and Bioinformatics*, 81(12), pp.2137–2142. Available at:

http://dx.doi.org/10.1002/prot.24392.

Kundrotas, P.J., Zhu, Z. & Vakser, I.A., 2010. GWIDD: Genome-wide protein docking database. *Nucleic acids research*, 38(Database issue), pp.D513–7.

Kunkel, T.A., 1985. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences of the United States of America*, 82(2), pp.488–492.

Kurjan, J. et al., 1980. Mutations at the yeast SUP4 tRNATyr locus: DNA sequence changes in mutants lacking suppressor activity. *Cell*, 20(3), pp.701–709.

Kwasnieski, J.C. et al., 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences of the United States of America*, 109(47), pp.19498–19503.

Lant, J.T. et al., 2019. Pathways to disease from natural variations in human cytoplasmic tRNAs. *The Journal of biological chemistry*, 294(14), pp.5294–5308.

Lensink, M.F. & Wodak, S.J., 2010. Docking and scoring protein interactions: CAPRI 2009. *Proteins*, 78(15), pp.3073–3084.

Levy, E.D., Kowarzyk, J. & Michnick, S.W., 2014. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. *Cell reports*, 7(4), pp.1333–1340.

Li, C. et al., 2016. The fitness landscape of a tRNA gene. *Science*, 352(6287), pp.837–840.

Licitra, E.J. & Liu, J.O., 1996. A three-hybrid system for detecting small ligand-protein receptor interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(23), pp.12817–12821.

Limbach, P.A., Crain, P.F. & McCloskey, J.A., 1994. Summary: the modified nucleosides of RNA. *Nucleic acids research*, 22(12), pp.2183–2196.

Ling, J., Reynolds, N. & Ibba, M., 2009. Aminoacyl-tRNA synthesis and translational quality control. *Annual review of microbiology*, 63, pp.61–78.

Li, S. et al., 2004. A map of the interactome network of the metazoan C. elegans. *Science*, 303(5657), pp.540–543.

Liu, H. et al., 2019. Accurate detection of m6A RNA modifications in native RNA sequences. *Nature communications*, 10(1), p.4079.

Li, X. et al., 2019. Changes in gene expression predictably shift and switch genetic interactions. *Nature communications*, 10(1), p.3886.

Maddox, B., 2003. The double helix and the'wronged heroine'. *Nature*. Available at: https://www.nature.com/articles/nature01399.

Majithia, A.R. et al., 2016. Prospective functional classification of all possible missense variants in PPARG. *Nature genetics*, 48(12), pp.1570–1575.

Malovannaya, A. et al., 2011. Analysis of the human endogenous coregulator complexome. *Cell*, 145(5), pp.787–799.

Marck, C. & 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(10), pp.1189–1232.

Maricque, B.B., Chaudhari, H.G. & Cohen, B.A., 2018. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nature biotechnology*. Available at: http://dx.doi.org/10.1038/nbt.4285.

Marsh, J.A. & Teichmann, S.A., 2014. Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS biology*, 12(5), p.e1001870.

Marsh, J.A. & Teichmann, S.A., 2015. Structure, dynamics, assembly, and evolution of protein complexes. *Annual review of biochemistry*, 84, pp.551–575.

Matreyek, K.A. et al., 2018. Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nature genetics*, 50(6), pp.874–882.

Matuszewski, S. et al., 2016. A Statistical Guide to the Design of Deep Mutational Scanning Experiments. *Genetics*, 204(1), pp.77–87.

Mavor, D. et al., 2016. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, 5. Available at: http://dx.doi.org/10.7554/eLife.15802.

McCafferty, J. et al., 1990. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature*, 348(6301), pp.552–554.

Medina-Cucurella, A.V. et al., 2019. User-defined single pot mutagenesis using unamplified oligo pools. *Protein engineering, design & selection: PEDS*, 32(1), pp.41–45.

Megel, C. et al., 2015. Surveillance and cleavage of eukaryotic tRNAs. *International journal of molecular sciences*, 16(1), pp.1873–1893.

Melamed, D. et al., 2013. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* , 19(11), pp.1537–1551.

Michnick, S.W., 2003. Protein fragment complementation strategies for biochemical network mapping. *Current opinion in biotechnology*, 14(6), pp.610–617.

Mishra, P. et al., 2016. Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. *Cell reports*, 15(3), pp.588–598.

Mohan, U., Kaushik, S. & Banerjee, U.C., 2011. PCR based random mutagenesis approach for a defined DNA sequence using the mutagenic potential of oxidized nucleotide products. *The open biotechnology journal*, 5(1), pp.21–27.

Moore, J.C. & Arnold, F.H., 1996. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature biotechnology*, 14(4), pp.458–467.

Mosca, R. et al., 2009. Pushing structural information into the yeast interactome by high-throughput protein docking experiments. *PLoS computational biology*, 5(8), p.e1000490.

Mosca, R., Pons, T., et al., 2013. Towards a detailed atlas of protein–protein interactions. *Current opinion in structural biology*, 23(6), pp.929–940.

Mosca, R., Céol, A. & Aloy, P., 2013. Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1), pp.47–53.

Motlagh, H.N. et al., 2014. The ensemble nature of allostery. *Nature*, 508(7496), pp.331–339.

Muller, H.J., 1927. ARTIFICIAL TRANSMUTATION OF THE GENE. *Science*, 66(1699), pp.84–87.

Muller, H.J., 1928. The problem of genic modification. In *Proceedings of the 5th International Congress, Supplementband of the Z. indukt. Abstamm.-u. Vererb-Lehre*. pp. 234–260.

Negroni, J., Mosca, R. & Aloy, P., 2014. Assessing the applicability of template-based protein docking in the twilight zone. *Structure* , 22(9), pp.1356–1362.

Nisthal, A. et al., 2019. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 116(33), pp.16367–16377.

Novoa, E.M. et al., 2012. A role for tRNA modifications in genome structure and codon usage. *Cell*, 149(1), pp.202–213.

Nussinov, R. & Tsai, C.-J., 2013. Allostery in disease and in drug discovery. *Cell*, 153(2), pp.293–305.

Olson, C.A., Wu, N.C. & Sun, R., 2014. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current biology: CB*, 24(22), pp.2643–2651.

Orioli, A., 2017. tRNA biology in the omics era: Stress signalling dynamics and cancer progression. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 39(3). Available at: http://dx.doi.org/10.1002/bies.201600158.

Otwinowski, J., 2018. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Molecular biology and evolution*, 35(10), pp.2345–2354.

Ovchinnikov, S., Kamisetty, H. & Baker, D., 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3, p.e02030.

Palmer, A.C. et al., 2015. Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nature communications*, 6, p.7385.

Papworth, C., Greener, A. & Braman, J., 1994. Highly efficient double-stranded, site-directed mutagenesis with the Chameleon^{TM} kit. *Strategies in molecular biology*, 7, pp.38–40.

Park, J. et al., 2007. Bacterial beta-Lactamase Fragment Complementation Strategy Can Be Used as a Method for Identifying Interacting Protein Pairs. *Journal of microbiology and biotechnology*, 17(10), p.1607.

Parmley, S.F. & Smith, G.P., 1988. Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. *Gene*, 73(2), pp.305–318.

Patwardhan, R.P. et al., 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3), pp.265–270.

Payea, M.J. et al., 2018. Widespread temperature sensitivity and tRNA decay due to mutations in a yeast tRNA. *RNA* , 24(3), pp.410–422.

Pazos, F. & Valencia, A., 2002. In silico two‐hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function, and Bioinformatics*. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10074?casa_token=7uUD_wtupNsAAAAA:eVH_BdGKA8j0IHBTSZ962q2fDKCDfdg5l_2bNUh2p6WB16_5T-Z6cq65azhHIqSoBSD0r-a5lBAGQzE.

Pazos, F. & Valencia, A., 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering, design & selection: PEDS*, 14(9), pp.609–614.

Pelletier, J.N., Campbell-Valois, F.X. & Michnick, S.W., 1998. Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. *Proceedings of the National Academy of Sciences of the United States of America*, 95(21), pp.12141–12146.

Pelletier, J.N. & Michnick, S.W., 1997. A protein complementation assay for detection of protein-protein interactions in vivo. *Protein engineering*, 10, p.89.

Percudani, R., Pavesi, A. & Ottonello, S., 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *Journal of molecular biology*, 268(2), pp.322–330.

Perica, T. et al., 2012. The emergence of protein complexes: quaternary structure, dynamics and allostery. Available at: http://www.biochemsoctrans.org/content/40/3/475.abstract.

Peterson, F.C. et al., 2004. Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB-PDZ transition. *Molecular cell*, 13(5), pp.665–676.

Phizicky, E.M. & Hopper, A.K., 2010. tRNA biology charges to the front. *Genes & development*, 24(17), pp.1832–1860.

Poelwijk, F.J., Socolich, M. & Ranganathan, R., 2017. Learning the pattern of epistasis linking genotype and phenotype in a protein. *bioRxiv*. Available at: http://dx.doi.org/10.1101/213835.

Poelwijk, F.J., Socolich, M. & Ranganathan, R., 2019. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, 10(1), p.4213.

Pokusaeva, V.O. et al., 2019. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS genetics*, 15(4), p.e1008079.

Puchta, O. et al., 2016. Network of epistatic interactions within a yeast snoRNA. *Science*, 352(6287), pp.840–844. Available at: http://dx.doi.org/10.1126/science.aaf0965.

Rajagopala, S.V. et al., 2014. The binary protein-protein interaction landscape of Escherichia coli. *Nature biotechnology*, 32(3), pp.285–290.

dos Reis, M., Savva, R. & Wernisch, L., 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, 32(17), pp.5036–5044.

Rich, M.S. et al., 2016. Comprehensive Analysis of the SUL1 Promoter of Saccharomyces cerevisiae. *Genetics*, 203(1), pp.191–202.

Ritchie, D.W., 2008. Recent progress and future directions in protein-protein docking. *Current protein & peptide science*, 9(1), pp.1–15.

Roberts, R.W., 1999. Totally in vitro protein selection using mRNA-protein fusions and ribosome display. *Current opinion in chemical biology*, 3(3), pp.268–273.

Rodríguez-Rodríguez, D.R. et al., 2019. Genome editing: A perspective on the application of CRISPR/Cas9 to study human diseases. *International journal of molecular medicine*, 43(4), pp.1559–1574.

Rolland, T. et al., 2014. A proteome-scale map of the human interactome network. *Cell*, 159(5), pp.1212–1226.

Rollins, N.J. et al., 2019. Inferring protein 3D structure from deep mutation scans. *Nature genetics*, 51(7), pp.1170–1176.

Rual, J.-F. et al., 2005. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062), pp.1173–1178.

Rubin, A.F. et al., 2017. A statistical framework for analyzing deep mutational scanning data. *Genome biology*, 18(1), p.150.

Ruvkun, G.B. & Ausubel, F.M., 1981. A general method for site-directed mutagenesis in prokaryotes. *Nature*, 289(5793), pp.85–88.

Sahni, N. et al., 2013. Edgotype: a fundamental link between genotype and phenotype. *Current opinion in genetics & development*, 23(6), pp.649–657.

Sahni, N. et al., 2015. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell*, 161(3), pp.647–660.

Sailer, Z.R. & Harms, M.J., 2017a. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics*, 205(3), pp.1079–1088.

Sailer, Z.R. & Harms, M.J., 2017b. High-order epistasis shapes evolutionary trajectories. *PLoS computational biology*, 13(5), p.e1005541.

Sailer, Z.R. & Harms, M.J., 2017c. Molecular ensembles make evolution unpredictable. *Proceedings of the National Academy of Sciences of the United States of America*, 114(45), pp.11938–11943.

Sambrook, J., Fritsch, E.F. & Maniatis, T., 1989. *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory.

Sarkar, S. & Hopper, A.K., 1998. tRNA nuclear export in saccharomyces cerevisiae: in situ hybridization analysis. *Molecular biology of the cell*, 9(11), pp.3041–3055.

Sarkisyan, K.S. et al., 2016. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603), pp.397–401.

Schmiedel, J.M. & Lehner, B., 2019. Determining protein structures using deep mutagenesis. *Nature Genetics*, 51(7), pp.1177–1186. Available at: http://dx.doi.org/10.1038/s41588-019-0431-x.

Schneider, S. & Zacharias, M., 2011. Flexible Protein-Protein Docking. *Selected Works in Bioinformatics*. Available at: http://dx.doi.org/10.5772/20865.

Scott, J.K. & Smith, G.P., 1990. Searching for peptide ligands with an epitope library. *Science*, 249(4967), pp.386–390.

Seifert, H.S. et al., 1986. Shuttle mutagenesis: a method of transposon mutagenesis for Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 83(3), pp.735–739.

SenGupta, D.J. et al., 1996. A three-hybrid system to detect RNA-protein interactions in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 93(16), pp.8496–8501.

Shampo, M.A. & Kyle, R.A., 2002. Kary B. Mullis—Nobel Laureate for Procedure to Replicate DNA. *Mayo Clinic proceedings. Mayo Clinic*, 77(7), p.606.

Sharon, E. et al., 2018. Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell*, 175(2), pp.544–557.e16.

Simonis, N. et al., 2009. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nature methods*, 6(1), pp.47–54.

Sinha, R., Kundrotas, P.J. & Vakser, I.A., 2010. Docking by structural similarity at protein-protein interfaces. *Proteins*, 78(15), pp.3235–3241.

Staller, M.V. et al., 2018. A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. *Cell systems*, 6(4), pp.444–455.e6.

Starita, L.M. et al., 2013. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14), pp.E1263–72.

Starita, L.M. et al., 2015. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*, 200(2), pp.413–422.

Starr, T.N., Picton, L.K. & Thornton, J.W., 2017. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672), pp.409–413.

Staub, O. et al., 1996. WW domains of Nedd4 bind to the proline‑rich PY motifs in the epithelial Na+ channel deleted in Liddle's syndrome. *The EMBO journal*, 15(10), pp.2371–2380.

Stelzl, U. et al., 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6), pp.957–968.

Stenson, P.D. et al., 2017. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human Genetics*, 136(6), pp.665–677. Available at: http://dx.doi.org/10.1007/s00439-017-1779-6.

Stiffler, M.A. et al., 2019. Protein structure from experimental evolution. *bioRxiv*, p.667790. Available at: https://www.biorxiv.org/content/10.1101/667790v1 [Accessed September 14, 2019].

Sutton, W.S., 1903. THE CHROMOSOMES IN HEREDITY. *The Biological bulletin*, 4(5), pp.231–250.

Szilagyi, A. & Zhang, Y., 2014. Template-based structure modeling of protein-protein interactions. *Current opinion in structural biology*, 24, pp.10–23.

Tarassov, K. et al., 2008. An in vivo map of the yeast protein interactome. *Science*, 320(5882), pp.1465–1470.

Torrent, M. et al., 2018. Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science signaling*, 11(546). Available at: http://dx.doi.org/10.1126/scisignal.aat6409.

Tuller, T. et al., 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2), pp.344–354.

Uetz, P. et al., 2000. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770), pp.623–627.

Van Blarcom, T. et al., 2015. Precise and efficient antibody epitope determination through library design, yeast display and next-generation sequencing. *Journal of molecular biology*, 427(6 Pt B), pp.1513–1534.

Vidal, M., Braun, P., et al., 1996. Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), pp.10321–10326.

Vidal, M., Brachmann, R.K., et al., 1996. Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19), pp.10315–10320.

Vidal, M., 1999. Yeast forward and reverse "n"-hybrid systems. *Nucleic Acids Research*, 27(4), pp.919–929. Available at: http://dx.doi.org/10.1093/nar/27.4.919.

Vidal, M., Cusick, M.E. & Barabási, A.-L., 2011. Interactome Networks and Human Disease. *Cell*, 144(6), pp.986–998. Available at: http://dx.doi.org/10.1016/j.cell.2011.02.016.

Waas, W.F. et al., 2007. Role of a tRNA base modification and its precursors in frameshifting in eukaryotes. *The Journal of biological chemistry*, 282(36), pp.26026–26034.

Walzthoeni, T. et al., 2013. Mass spectrometry supported determination of protein complex structure. *Current opinion in structural biology*, 23(2), pp.252–260.

Wang, X. et al., 2017. Fine Epitope Mapping of Two Antibodies Neutralizing the Bordetella Adenylate Cyclase Toxin. *Biochemistry*, 56(9), pp.1324–1336.

Wang, Z. & Moult, J., 2001. SNPs, protein structure, and disease. *Human mutation*, 17(4), pp.263–270.

Weigt, M. et al., 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1), pp.67–72.

Wei, X. et al., 2014. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS genetics*, 10(12), p.e1004819.

Whitehead, T.A. et al., 2012. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature biotechnology*, 30(6), pp.543–548.

Wilson, T.E. et al., 1991. Identification of the DNA binding site for NGFI-B by genetic selection in yeast. *Science*, 252(5010), pp.1296–1300.

Winter, G. et al., 1994. Making antibodies by phage display technology. *Annual review of immunology*, 12, pp.433–455.

Wolin, S.L., Sim, S. & Chen, X., 2012. Nuclear noncoding RNA surveillance: is the end in sight? *Trends in genetics: TIG*, 28(7), pp.306–313.

Woodsmith, J. et al., 2017. Protein interaction perturbation profiling at amino-acid resolution. *Nature methods*, 14(12), pp.1213–1221.

Wrenbeck, E.E. et al., 2016. Plasmid-based one-pot saturation mutagenesis. *Nature Methods*, 13(11), pp.928–930. Available at: http://dx.doi.org/10.1038/nmeth.4029.

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress on Genetics*, 1(8), pp.355–366.

Wu, N.C. et al., 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5. Available at: http://dx.doi.org/10.7554/eLife.16965.

Yarham, J.W. et al., 2010. Mitochondrial tRNA mutations and disease. *Wiley interdisciplinary reviews. RNA*, 1(2), pp.304–324.

Yu, C. & Huang, L., 2018. Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Analytical chemistry*, 90(1), pp.144–165.

Yu, H. et al., 2008. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898), pp.104–110.

Zheng, G. et al., 2015. Efficient and quantitative high-throughput tRNA sequencing. *Nature methods*, 12(9), pp.835–837.

Zheng, Y.-G. et al., 2004. Two distinct domains of the beta subunit of Aquifex aeolicus leucyl-tRNA synthetase are involved in tRNA binding as revealed by a three-hybrid selection. *Nucleic acids research*, 32(11), pp.3294–3303.

Zhong, Q. et al., 2009. Edgetic perturbation models of human inherited disorders. *Molecular systems biology*, 5, p.321.

Zinzalla, G. & Thurston, D.E., 2009. Targeting protein-protein interactions for therapeutic intervention: a challenge for the future. *Future medicinal chemistry*, 1(1), pp.65–93.