Extending the Atlas of Variant Effects in Human Disease Genes

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Graduate Department of Molecular Genetics
University of Toronto

Jochen Weile

July 7, 2017

Abstract

Although we now routinely sequence human genomes, we cannot yet confidently identify functional variants. Here a deep mutational scanning framework is developed that combines random codon-mutagenesis and multiplexed functional variation assays with computational imputation and regularization to yield exhaustive functional maps for human missense variants. The framework is applied to five proteins corresponding to seven human genes: *UBE2I* (encoding SUMO E2 conjugase), *SUMO1* (small ubiquitin-like modifier), *NCS1* (neuronal calcium sensor 1), *TPK1* (thiamin pyrophosphokinase), and *CALM1/2/3* (three genes encoding the protein calmodulin). The resulting functional impact scores correspond to known protein features, and serve to confidently identify pathogenic variation.

Acknowledgements

The author would like to thank Fritz Roth, Atina Coté, Jennifer Knapp, Song Sun, Marta Verby, Yingzhou Wu, Cassandra Wong, Fan Yang, Carles Pons, Patrick Aloy, Natascha van Lieshout, Anjali Gopal, Jesse Bloom, Guihong Tan, Joseph Mellor, Shan Yang, Robert Nussbaum, Douglas Fowler, Nidhi Sahni, Marc Vidal, David Hill, Amy Caudy, Lincoln Stein, and Igor Stagljar for their collaboration, help and advice. Furthermore, the author gratefully acknowledges funding to the Roth Lab by the National Institutes of Health, the Canadian Excellence Research Chairs Program, the Canadian Institute for Advanced Research and the Ontario Ministry of Research, Innovation and Science.

Contents

1. Introduction

Given the constantly improving cost and speed of genome sequencing, it is reasonable to expect that within the coming decades personal genomes will be known for a substantial part of the global populace. Unfortunately, our limited ability to interpret the variation within stands in stark contrast with this development. Even when only considering mutations in coding regions of the genome, the effects of most missense variants are not known. While a number of computational approaches exist to make predictions as to the effects of coding variants, they are currently not reliable enough for clinical use. By comparison, laboratory assays produce more trustworthy results, but until recently did not scale to the space of all possible mutations. The development of Deep Mutational Scanning [1–3] has now made this endeavour possible. In the following sections, each of these issues will be discussed in more detail.

1.1. The Genotype-Phenotype Problem

Linking genotype to phenotype is a very difficult problem. The part of the human genome we understand best are protein-coding genes, yet they only constitute a small fraction the whole. Impacts of mutations in other functional elements such as splice sites, promoters, or regulatory sequences are more difficult to assay, not to mention the vast stretches of intergenic space. While one might expect the latter to not bear functional significance *a priori*, a large number of loci identified as correlated with diseases in genome-wide association studies (GWAS) are found within these regions [4]. While many of these cases may simply be spurious findings due to linkage disequilibrium with variants in coding regions [5], more functions yet unknown may lie hidden within this vast space. But even for protein-coding sequences the problem is far from simple. Alleles with simple Mendelian behaviour are the exception rather than the rule. Most phenotypes are complex, i.e. they emerge through the interplay of many different genetic or environmental factors. Conversely, many genes are also pleiotropic, i.e. they are involved more than one mechanism. As a result

of this complexity, a mutation found in one person may not have the same effect as in another—a phenomenon called incomplete penetrance. Similarly, two different mutations within the same coding sequence will often not have the same effect either. Depending on how the translated protein is affected (e.g. catastrophic folding failure, alteration of a molecular interaction interface or active site, or a subtle change on an unused surface) the effects may differ in severity or in rare cases may even result in the emergence of new behaviours.

Given the much greater difficulty of interpreting non-coding regions, clinical applications have so far largely concentrated on protein-coding genes. Sequencing panels for known disease-associated genes and even whole-exome sequencing (WES) are widely commercially available. A number of different standards for classifying mutations with respect to their potential health impacts have been proposed. Most prominently, the American College of Medical Genetics and Genomics (ACMG) standard [6]. It defines categories stretching from "pathogenic" via "variant of uncertain significance" (VUS) to "benign". Even though the mutational landscape for a handful of genes, such as *BRCA1* are explored better than others due to their established relevance and potential for taking clinical action [7], the vast majority of clinical variants are currently classified as VUS. For example, in a recent study using gene panels assessing germline cancer risk loci [8], over 98% of missense variants have been called VUS. Not only can these uncertainties burden patients with unnecessary anxiety [7], they also call into question the value of sequencing in the clinic if the majority of findings are not actionable. With increasing use of WES over gene panels, this problem is only going to get worse. According to the 1000 Genomes Project data, every person carries 100-400 missense variants that are so rare that they have likely never been seen before in the clinic [9]. In the absence of previous observations they would automatically be added to the long list of VUS.

1.2. In silico approaches to variant function assessment

A number of algorithms exist that offer predictions as to the deleteriousness of mutations, the most prominent ones being PolyPhen-2 [10], SIFT [11] and PROVEAN [12]. PolyPhen-2 employs a machine learning method based on evolutionary conservation and protein structural features. It uses a set of previously reported pathogenic alleles as a positive training set and differences

between human genes and their mammalian homologues as a negative training set. By contrast, SIFT (Sorting Intolerant From Tolerant) only uses evolutionary conservation. The tool uses multiple sequence alignments to calculate position-specific score matrices for each gene which are then normalized and transformed into probability values. PROVEAN (PROtein Variation Effect ANalyzer) similarly only takes into account sequence alignments. However, rather than just computing a position-specific score, PROVEAN calculates the difference in alignment quality between using the wildtype or variant sequence against clusters of homologous sequences. The average distance is then interpreted as indicative of the deleteriousness of the variant.

While the three tools succeed in making good predictions, their reliability is unfortunately still not high enough to serve as a basis of clinical decision making. Song Sun and other members of the Roth Lab recently performed an independent comparison of these tools on a set of well established disease-causing variants as well as rare polymorphisms with no known disease association [13]. A high precision (the fraction of correct classifications out of all positive classifications) can be considered especially important when considering taking clinical action based on a prediction. When compared at a minimum precision level of 90%, PolyPhen-2 and PROVEAN only reach a sensitivity of 19% and 21%, respectively (where sensitivity is defined as the fraction of correct classifications out of all real existing disease variants). SIFT was not capable of achieving 90% precision at any score threshold. In concordance with these limitations, the ACMG currently considers only cases in which multiple methodologically orthogonal prediction algorithms agree as weak evidence in a supporting role for VUS re-classification [6].

1.3. Laboratory approaches to variant function assessment

An alternative to computational prediction for variant assessment is the use of laboratory assays. Many different types of assays exist that can yield potential insight into the effects of missense variants on protein function. However, most of them need to be performed individually for each protein and are not easily scalable. Two particularly useful assays in this respect are Yeast-2-Hybrid and functional complementation.

Yeast-2-Hybrid (Y2H) [14] is a binary protein interaction assay performed within the yeast *Saccharomyces cerevisiae* (Figure 1.1A). The qualifier 'bi-

nary' refers to the fact that it detects direct physical associations compared between two individual proteins as opposed to often-indirect associations like co-localization or co-complex-membership. It is based on the reconstitution of two fragments of the transcription factor Gal4. The Gal4 protein is comprised of two domains: A DNA-binding (DB) domain and an activating domain (AD). Both are required for it to successfully associate with its cognate promoter region and induce expression of a reporter gene downstream of the promoter. When two proteins X and Y are fused to the DB and AD domain respectively, a prospective interaction between X and Y leads to the reconstitution of the transcription factor and subsequently to reporter expression. In most cases, the reporter is an auxotrophy marker, such as *HIS3*, thus linking the ability of the two proteins to interact with each other to the ability of the yeast strain to grow on selective (e.g. histidine-deficient) media. When comparing different variants of the same protein interacting with the same partner, reporter expression has even been shown to be proportional to binding affinity [15]. This proportional relationship allows for quantitative interpretation of Y2H results under these specific circumstances. However, this cannot be generalized to compare different proteins.

Y2H does however suffer from a number of drawbacks. Due to the the transcription factor needing to physically associate with DNA, any protein to be examined needs to be able to enter the nucleus and function within. While the DB domain already contains a nuclear localization sequence (NLS), the AD ORF is often the fused with an additional NLS. However, this does not work for every protein [16]. A particular problem are membrane proteins which generally cannot enter the nucleus at all [17]. A variant of Y2H, MYTH (Membrane Yeast-Two-Hybrid) exists for these proteins [18, 19]. This system relies on the reconstitution of a split ubiquitin through the interaction of membrane proteins. A reconstitution of ubiquitin allows for recognition by deubiquitinases (DUBs), which cleave off a fused transcription factor that activates a reporter gene.

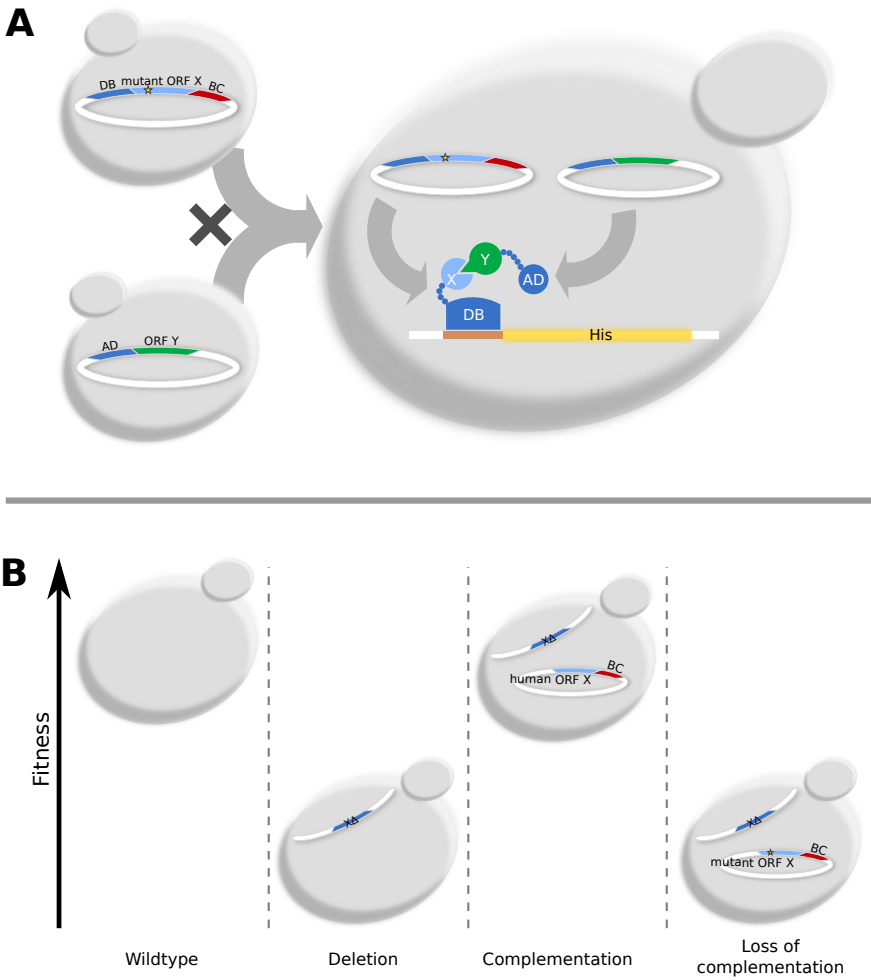

Figure 1.1.: Complementation and Yeast-2-Hybrid. A) Yeast-2-Hybrid: Strains carrying fusion of ORF X to Gal4-DB and ORF Y to Gal4-AD are mated. A successful interaction between X and Y in the diploid progeny results in reconstitution of Gal4 and thus in the expression of the HIS3 reporter, allowing for auxotrophy selection. B) Complementation: Inactivation of gene X in yeast results in a fitness defect, that is rescued by expression of X's human orthologue. A damaging variant of human X results in loss of complementation.

In the past it has often been stated that Y2H results are unreliable and suffer from low precision. One source for these claims goes back to a comparison between two early Y2H screens of the *S. cerevisiae* interactome by Ito *et al.* [20] and Uetz *et al.* [21], whose maps only overlapped by 19%. Yu *et al.* later showed that this low overlap was not due to low specificity as previously thought, but rather low sensitivity. It has been estimated that Y2H has an overall assay sensitivity of 20% [22]. That is, only one in five real existing protein interactions can be detected by Y2H. These sensitivity levels are comparable to most other binary interaction assays, such as Protein-fragment Complementation Assays (PCA) [23] or the Mammalian Protein-Protein-Interaction Trap (MaPPIT) [24].

When considering Y2H as an assay for variant function assessment it is important to consider that it does not measure all aspects of a protein's functionality, but rather only its ability to physically associate with a given interaction partner. Thus only variants that result either in major failures in protein folding or in changes to the binding binding interface could be detected. However, in a recent examination of the Y2H performance of common disease associated variants, we found that approximately two out of three disease variants in proteins with detectable interactions manifest in such a way [25].

Nonetheless, an assay that can measure the overall functionality of a protein within the cell would be preferable. Functional complementation in yeast [26, 27] offers such an option (Figure 1.1B). It based on the premise that some human genes can be used to rescue the deletion of their orthologues in yeast. That is, a fitness defect resulting from the inactivation of the yeast gene is alleviated by the artificial expression of the human gene. Therefore, any relative changes in fitness upon expressing a variant of the human gene can be interpreted as the variant's effect on the protein's overall ability to function. Song Sun and other members of the Roth Lab have recently examined the applicability of functional complementation in yeast to the assessment of disease variants [13]. They have found an astonishing predictive capacity despite yeast and humans being diverged by ~ 1 billion years. Yeast complementation outperformed *in silico* methods like PolyPhen-2 and PROVEAN in terms of disease variant prediction by a wide margin. At the 90% specificity threshold discussed in section 1.2, the complementation assay achieved a sensitivity of over 60% (as compared to 19% and 21% for the two *in silico* methods, respectively). It is consistent with these findings that the ACMG considers functional assays among the strongest sources of evidence for variant classification [6].

The only major drawback of yeast complementation is that currently only

$\sim$ 200 human genes have been found to be amenable to the assay [13]. However, in recent years CRISPR screens have revealed many genes for which growth phenotypes exist directly in human cell lines [28–30]; opening the possibility of performing functional complementation directly in these cell lines.

1.4. Deep Mutational Scanning

Complementation and Y2H promise to be useful tools in the classification of variants of uncertain significance. Yet applying them to retroactively test variants only once they have been found in the clinic would be a slow process. Instead, a proactive approach could prove to be more useful: Building an atlas of the functional effects of all possible variants before they are observed in a patient. Indeed, given the size of the human population and the frequency of *de novo* mutation [31], every missense variant that can possibly exist (and is not fundamentally incompatible with life) can be expected to occur on average in 46 individuals[1]. However, assaying all possible variants in known disease genes would require massive parallelization. Indeed such parallelization efforts have previously been described, albeit not primarily for the reclassification of VUS. The winter semester of 2010/11 saw three papers by Fowler *et al.* [1], Ernst *et al.* [2] and Hietpas *et al.* [3] that collectively pioneered a technology called Deep Mutational Scanning (DMS). DMS can be thought of as a natural extension to Alanine Scanning [32], expanding it into the space of all possible amino acid changes. These seminal papers have since inspired a growing number of similar efforts by other groups [33–61]. Tables 1.1 and 1.2 list a selection of these studies that showcase the breadth of methodologies that has since emerged. Deep Mutational Scanning, as performed in these studies, can be broken down into a number of experimental and computational components: (1) Mutagenesis; (2) Selection of functional variants; (3) Sequencing of the selected and control populations; and (4) Computational analysis. In the following sections we will review the different previous implementations of these components in detail.

[1] Back-of-envelope calculation: $\frac{7.4\text{bn humans} \times 0.6 \text{ de-novo exome SNVs}}{30\text{Mb exome} \times 3 \text{ possible SNVs}} \approx 46 \frac{\text{humans}}{\text{bp}}$

Table 1.1.: Coverage of possible variants in a selection of previous DMS studies

Year	Author	Protein	Region	Coverage
2010	Fowler *et al.* [1]	YAP65	WW domain	$\sim 100\%$
2010	Ernst *et al.* [2]	Synthetic PDZ domain	10 AAs	$\sim 100\%$
2011	Hietpas *et al.* [3]	Hsp90	9 AAs	$\sim 100\%$
2012	Fujino *et al.* [33]	Fab antibody fragment	fragment	79%
2012	Adkar *et al.* [34]	Ccdb	whole protein	$< 74\%$
2012	McLaughlin *et al.* [35]	PSD95	PDZ domain	$\sim 100\%$
2012	Schlinkmann *et al.* [36]	GPCR	whole protein	$\sim 90\%$
2012	Whitehead *et al.* [37]	Synthetic protein	51AA (whole protein)	99%
2012	Traxlmayr *et al.* [38]	IgG1	CH2/CH3 domains	$< 50\%$
2012	Wu *et al.* [39]	Neuraminidase	SNP accessible	$< 50\%$
2013	Roscoe *et al.* [40]	Ubiquitin	Whole protein	$\sim 95\%$
2013	Starita *et al.* [41]	Ub.E3 E4B	whole protein	$\sim 50\%$
2013	Procko *et al.* [42]	Synthetic protein	60 AA	$\sim 100\%$
2013	Tinberg *et al.* [43]	Synthetic protein	40AA	90%
2013	Jiang *et al.* [44]	Hsp90	Substrate binding loop	$\sim 100\%$
2013	Kim *et al.* [45]	Mat alpha	degron region	$< 50\%$
2013	Melamed *et al.* [46]	Pab1	RRM domain	$\sim 90\%$
2013	Forsyth *et al.* [47]	Antibody for EGFR	whole protein	$\sim 99\%$
2013	Wagenaar *et al.* [48]	BRAF	77 AAs	99.65%
2014	Firnberg *et al.* [49]	TEM1 β-lactamase	Whole protein	$\sim 95\%$
2014	Olson *et al.* [50]	G-protein (GB1)	IgG-binding domain	$\sim 95\%$
2014	Melnikov *et al.* [51]	APH(3')II (kinase)	Whole protein	$\sim 100\%$
2014	Bloom [52]	influenza nucleoprotein	whole protein	$> 75\%$
2014	Thyagarajan*et al.* [53]	influenza hemagglutinin	whole protein	$\sim 85\%$
2015	Stiffler *et al.* [54]	TEM1 β-lactamase	whole protein	$\sim 100\%$
2015	Doud *et al.* [55]	influenza nucleoprotein	whole protein	$\sim 100\%$
2015	Kitzman *et al.* [56]	Gal4	DB domain	$\sim 99\%$
2015	Starita *et al.* [57]	BRCA1	RING domain	$\sim 80\%$
2016	Mishra *et al.* [58]	Hsp90	ATPase domain	$\sim 99\%$
2016	Doud *et al.* [59]	Hemagglutinin	whole protein	$< 97\%$
2016	Mavor *et al.* [60]	Ubiquitin	whole protein	$\sim 99\%$
2016	Majithia *et al.* [61]	PPARγ	whole protein	$\sim 99\%$

Table 1.2.: Methods in a selection of previous DMS studies

Year	Author	Selection	System	Mutagenesis
2010	Fowler *et al.* [1]	Phage display	*in vitro*	commercial oligo pool PCR
2010	Ernst *et al.* [2]	Phage display	*in vitro*	Kunkel
2011	Hietpas *et al.* [3]	Complementation	Yeast	EMPIRIC
2012	Fujino *et al.* [33]	Ribodisplay	*in vitro*	oligo pool PCR
2012	Adkar *et al.* [34]	Toxin activity	*E. coli*	oligo pool PCR (certain codons)
2012	McLaughlin *et al.* [35]	B2H+FACS	*E. coli*	tiled oligo pool PCR (NNS)
2012	Schlinkmann *et al.* [36]	FACS	*E. coli*	oligo pool PCR (NNN)
2012	Whitehead *et al.* [37]	Yeast display	Yeast	oligo pool PCR
2012	Traxlmayr *et al.* [38]	Yeast display	Yeast	error prone PCR
2012	Wu *et al.* [39]	Oseltamivir resistance	Human/H1N1	error prone PCR
2013	Roscoe *et al.* [40]	Growth	Yeast	EMPIRIC
2013	Starita *et al.* [41]	Phage display	*in vitro*	same as Fowler *et al.* 2010
2013	Procko *et al.* [42]	Yeast display	Yeast	oligo pool PCR + error prone PCR
2013	Tinberg *et al.* [43]	Yeast display	*in vitro*	Kunkel NNK
2013	Jiang *et al.* [44]	Complementation	Yeast	EMPIRIC
2013	Kim *et al.* [45]	Degron activity	Yeast	Error-prone PCR
2013	Melamed *et al.* [46]	Complementation	Yeast	Error-prone PCR
2013	Forsyth *et al.* [47]	FACS	Human	oligo pool PCR (NNK)
2013	Wagenaar *et al.* [48]	Vemurafenib resistance	Human	EMPIRIC
2014	Firnberg *et al.* [49]	Amp resistance	E.coli	pfunkel
2014	Olson *et al.* [50]	RNA display	*in vitro*	cassette ligation (NNK/NNS)
2014	Melnikov *et al.* [51]	Kanamycin resistance	*E. coli*	commercial oligo pool PCR
2014	Bloom [52]	viral replication	Human/H1N1	oligo pool PCR (NNN)
2014	Thyagarajan*et al.* [53]	viral replication	Human/H1N1	same as Bloom 2014
2015	Stiffler *et al.* [54]	Growth	*E. coli*	same as McLaughlin *et al.*2012
2015	Doud *et al.* [55]	viral replication	Human/H1N1	same as Bloom 2014
2015	Kitzman *et al.* [56]	growth	Yeast	PALS (array+PCR)
2015	Starita *et al.* [57]	Y2H / E3 activity	Yeast	PALS (array+PCR)
2016	Mishra *et al.* [58]	growth	Yeast	EMPIRIC
2016	Doud *et al.* [59]	viral replication	Human/H1N1	same as Bloom 2014
2016	Mavor *et al.* [60]	growth	Yeast	Roscoe *et al.* 2013 library
2016	Majithia *et al.* [61]	Surface marker FACS	Human	oligo pool PCR + dUTP + indel selection

1.4.1. Mutagenesis approaches

A fair number of saturation mutagenesis methods have previously been applied in DMS studies; some more technically challenging than others. The simplest method is error-prone PCR amplification [62, 63]. While this has the advantage of being an inexpensive and facile procedure, it will only result in the generation of point mutations and as such will not generate all possible amino acid replacements. One may argue that the evaluation of VUS does not require insight into mutations outside of these variants, as they are unlikely to occur in nature. Nonetheless, exploring all possible amino acid changes offers the potential of valuable biochemical insights. Moreover, the preference for transitions over transversions in these methods leads to uneven representations of variants.

Another set of methods often employed are scaled-up versions of site-directed mutagenesis approaches [64–66], with one popular example being Kunkel mutagenesis [67]. It uses a strain of *E. coli* 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 thus uracilated template can be PCR amplified with primers containing the mutations of interest and subsequently amplified in regular *E. coli* which will degrade the uracilated template, thus enriching the mutant copies. A number of derivatives of Kunkel mutagenesis have since been developed to bring its output to a scale supporting saturated libraries, most notably Pfunkel [66]. To address the full spectrum of amino acids at a given position, oligonucleotides carrying degeneracy codons [68] are often used. Particularly popular is the use of `NNK` and `NNS` degeneracies, which have long been used in biochemistry [69,70]. Here, `S` denotes either Guanine or Cytosine and `K` denotes either Guanine or Thymine in the third position of the degenerate codon. Either of these options only enables 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`). A more recent development is the use of custom oligonucleotide arrays covering all possible (or desired) options of codon changes explicitly rather than relying on degeneracy [56]. While this option allows for the precise control of desired mutations, it is currently too expensive to be applicable for more than a handful of genes at a time.

Another saturation mutagenesis method often applied in Deep Mutational Scanning is EMPIRIC ("Extremely Methodical and Parallel Investigation of

Randomized Individual Codons") [3]. In this method, rather than using PCR amplification, oligonucleotide cassettes carrying the variants of interest are directly ligated at the appropriate positions. This is achieved by designing the underlying vector such that it omits the cassette sequence. Instead, it carries a restriction site at the equivalent position, which can be cut to create sticky ends. Pairs of oligos carrying the variants of interest can be synthesized such that they can assemble into a fitting cassette that integrates with the vector. EMPIRIC is one example of a mutagenesis method that was explicitly developed to be used in Deep Mutational Scanning. Another example is PALS ("Programmed ALlelic Series") [56], which aims to limit the number of amino acid changes per library clone to only one. Oligos carrying the variants of interest are annealed to uracilated templates and linearly amplified with strand-displacing polymerase. In a second step, the template is degraded using Uracil-DNA-Glycosylase and an antisense strand is generated in a second linear amplification step. The product is denatured and yet again hybridized with uracilated template allowing it to be extended towards the other end of the template. Finally, the template is degraded again and the now full-length mutagenized strands are amplified.

In addition to the various mutagenesis methods discussed here, it may be noted that complete variant libraries are also recently becoming commercially available via gene synthesis [71]. While this method is certainly the most convenient, 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.

1.4.2. Selection approaches

The most central component of a Deep Mutation Scan is the selection process. In section 1.3 two options were already discussed in detail: Y2H and functional complementation. There are a fair number of other options, even though many of them may not be as useful in the context of identifying disease variants. The different assays used in previous studies can be sorted into three broad categories: (i) *In vitro* display methods (such as Phage Display or Ribodisplay); (ii) Competition-based methods that couple a protein property under investigation (such as molecular interactions, toxicity, or overall functionality) to host cell fitness; and (iii) Cell sorting based on fluorescence labeled reporters.

Phage display [72] and ribodisplay [73] couple the genetic information of a given variant to the physical protein itself and select according to the protein's

ability to bind to a fixed interactor. In phage display this is achieved by the protein being displayed on the surface of a phage that contains the corresponding gene; while ribodisplay stalls a cluster of ribosomes on the variant mRNA with the corresponding protein still attached. Variants that are unable to bind to the interactor-coated surface are washed away and thus depleted. This can be done in multiple rounds, as the associated genetic information can be replicated again after selection (via viral propagation in bacteria for phage display or via PCR in ribodisplay). Fowler and colleagues employed phage display in their seminal DMS study with respect to the binding of the YAP65-WW domain to its cognate peptide target [1]. However, since display methods are only feasible for small proteins or fragments thereof, more recent studies have employed more scalable methods instead.

The most frequently applied selection mechanisms are fitness based. In these cases a particular property of the variant protein is coupled to its host cell's ability to thrive in competitive growth. Yeast-2-Hybrid and functional complementation (as introduced in section 1.3) are two examples of such methods. While Y2H couples fitness to the ability of the protein to maintain a specific protein-protein interaction, complementation does so for the proteins overall ability to perform its biological role. A popular condition-dependent extension to complementation is selection according to drug resistance [39, 48], but other fitness-based selection methods have been used in DMS as well. For example, Adkar and colleagues used the toxicity of CCDB in *E. coli* [34], while Kim and colleagues select according to degron activity by fusing the degron to an auxotrophic marker [45]. Finally, a number of DMS studies have been performed on viral genes, by selecting for virus propagation efficiency [52, 53].

Finally, another selection mechanism is the use of fluorescence-activated cell sorting (FACS) [74]. Here, surface markers whose abundance are proportional to the activity of the studied protein are targeted with fluorescently labeled antibodies, such that cells can be sorted accordingly, as has been performed by Schlinkmann *et al.* and Majithia *et al.* [36, 61].

1.4.3. Sequencing strategies

The experimental step immediately following selection in a DMS experiment is sequencing. Next-generation sequencing technology can be considered the key technological advance that made Deep Mutational Scanning possible. Many studies use a fairly simple approach by performing deep shotgun sequencing of the library [2, 3, 33]. 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, making it difficult to distinguish real mutations from sequencing error. To address this problem, different solutions have emerged. In cases where the amplicon is short enough, paired-end sequencing can be exploited to use information for variant calling. In the simplest case this is achieved by requiring both reads to agree on the base call in question, as in the case of Whitehead *et al.* [37]. A less stringent, but potentially more sensitive alternative as used by Fowler and colleagues [1] is to perform Bayesian inference on the quality scores associated with the base calls in each read pair. This way a variant may still be identified if one of the two reads reported a wildtype base call with low confidence.

Where the length of the nucleotide sequence in question exceeds the read length capabilities of short-read sequencing technologies, other strategies are required. A notable borderline case can be found in Olson *et al.* 2014 [50] where only a partial overlap between read pairs was achieved and variant calls outside of the overlap region were of lower quality. Other studies resort to more involved approaches. A popular paradigm is the association of molecular barcodes with each clone within the DMS library. While this simplifies the readout of the experiment (as only the barcodes need to be sequenced and counted), it adds the requirement of identifying which barcode belongs to which genotype. In most cases this is addressed using "subassembly" [75], a high-throughput amplicon sequencing approach based on attaching random tags to amplicons. The DNA is then 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 in terms enables assemblies for each molecule to be computed. The resulting high-quality virtual reads are long enough to cover both ORF and barcode locus.

Another barcode-based method, called EMPIRIC-BC was described by Mavor and colleagues [60], where the amplicon in question was short enough not to require subassembly. Here, a long read can cover the entire ORF, while a second, short read can identify the barcode.

An alternative approach to covering longer stretches of DNA is to subdivide them into smaller regions that can be sequenced separately from each other. For example, Doud and colleagues [59] amplify each region with primers carrying random tags. This way, if multiple reads contain the same tag, they are highly likely to originate from PCR copies of the same original molecule and can be used to make more accurate variant calls. While this approach has the

advantage of being less labour-intensive than barcoding each individual clone in the DMS library, it can only detect variants co-occurring within the same region of the sequence. Thus the library must be designed in such a way that either only a single mutation occurs within each clone or that it is large enough that effects of many co-occurring variants are averaged out.

1.4.4. Computational analysis

Most DMS studies use custom scripts to process the sequencing readout and calculate the selection advantage for each variant. Nonetheless, a few published software packages exist. The EMPIRIC mutagenesis and DMS method provides its own software package for data processing [3], though it is not generally applicable to other DMS methods. The `dms_tools` package [76] offers the same services, but is tailored more towards methods using regionally focused sequencing. Finally, `Enrich` [77] offers a generalized solution applicable to most DMS frameworks. A second version that adds a more sophisticated statistical analysis including the assessment of measurement confidence levels is currently under review [78].

1.4.5. Conclusion

When considering previous DMS studies in the context of VUS classification, a number issues become apparent. Many of these have primarily used DMS in the context of biochemistry. The assays underlying different DMS studies are quite diverse and measure different aspects of a protein's behaviour. As a consequence, they cannot be easily compared with each other. In addition, the achieved coverage of possible amino acid changes varies from map to map. Finally, many maps to do not control the quality of measurements. Therefore, the confidence levels underlying different parts of these maps are often unknown. A generalized framework that would allow for the construction of comparable, high-quality maps representing overall protein function would be of great utility.

1.5. Background: The Sumoylation Pathway

In the following chapters, we will evaluate the performance of Deep Mutational Scanning with respect to its ability to detect the effects of different variants

on overall function. However, as mentioned in section 1.1, many genes perform multiple functions and sub-functions and the proteins they encode engage in multiple interactions with other molecules. Thus, beyond the amenability of the proteins to the employed assays, an ideal testing ground would be comprised of a biological system that is both mechanistically complex and has been well studied previously in terms of structure and mechanism. This would allow for an examination of the assay's capabilities of detecting if and how a variant that damages an individual sub-function is reflected in its overall functional impact.

The Sumoylation life cycle does not only fulfill these criteria [79], but is also of great biological importance. Sumoylation is a protein modification in which a small ubiquitin-like modifier (SUMO) is covalently attached to target proteins in order to modulate their behaviour, especially in terms of localization and physical interactions [80]. Sumoylation plays an important role in a large number of cellular processes [80]. It is therefore not surprising that the core members of the pathway are essential genes [81].

Despite employing a distinct set of proteins compared to the ubiquitination machinery, the sumoylation pathway bears many close mechanistic similarities. Analogously to ubiquitin, a cascade of enzymes, E1, E2 and E3s, guide SUMO through its maturation, activation, conjugation and ligation phase [80] (Figure 1.2). After expression, SUMO is matured through cleavage of four amino acids from its C-terminus, exposing a diglycine motif. In humans, this process is performed by two peptidases, SENP1 and SENP2 (short for sentrin-specific peptidase, where sentrin is an alternative name for SUMO). Next, an E1 activation complex (UBA2-SAE1) forms a thioester bond between the SUMO C-terminal diglycine and a cysteine residue within the E1 protein under the consumption of ATP. An E2 conjugase (UBE2I) binds to the complex, so that the activated SUMO can be transfered to one of its own cysteine residues via transesterification.

The thus loaded E2 can recognize potential target proteins via an exposed motif of four amino acids. The motif is generally described as ΨKxD/E, i.e. a large hydrophobic residue, followed by a lysine, a spacer residue and an acidic residue [86]. The motif is often found in an exposed loop extending from the protein or in a disordered region [80, 87, 88]. The central lysine within the motif enters the E2's active site where it comes into contact with the SUMO diglycine. There, a peptide bond is formed between the lysine ε-amino group and the SUMO C-terminus [87]. This process can be made more efficient in the presence of E3 proteins. It is interesting to note that while only a single SUMO E2 conjugase (UBE2I) is encoded by the human genome, there are a variety

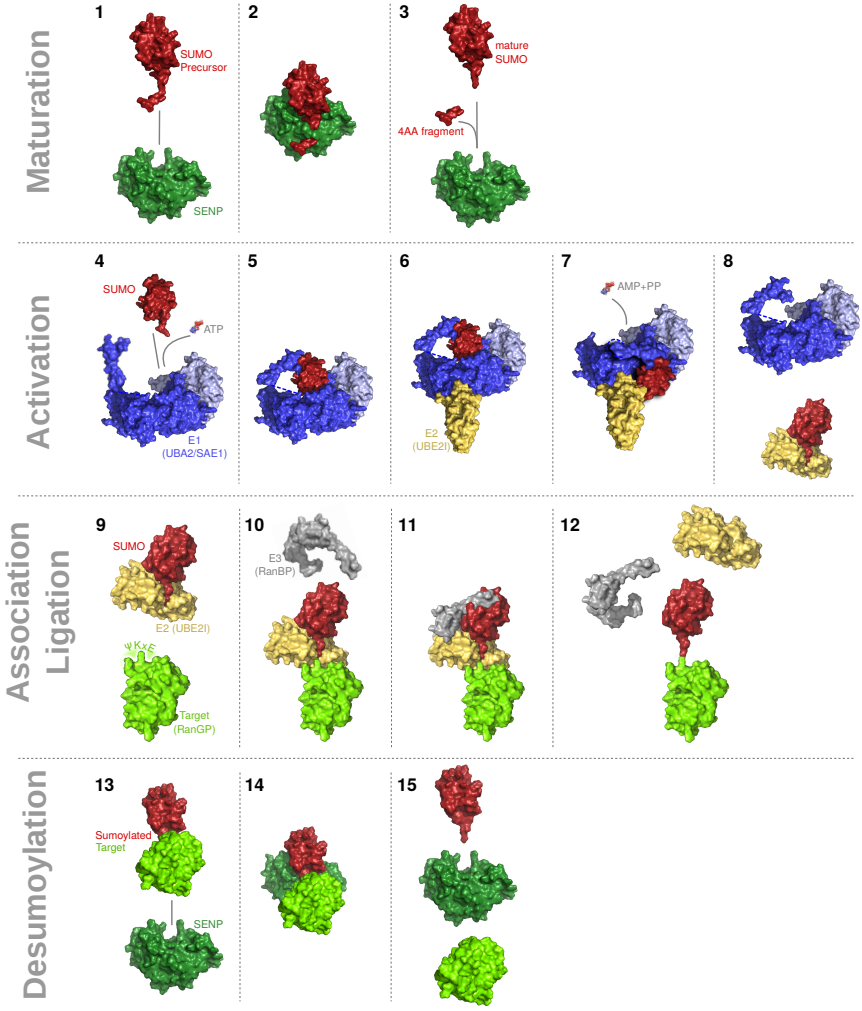

Figure 1.2.: Steps in the sumoylation cascade. SENP protease matures a SUMO precursor by cleaving off its four C-terminal residues. In the activation step, the E1 complex forms a thioester bond between SUMO and one of its cysteine residues under ATP consumption. It then transesterificates SUMO to a cysteine in the E2. The E2 recognizes potential targets via their ΨKxE motif. With the help of an E3, SUMO is then ligated to the central lysine within that motif. SENP proteases can reverse the process by hydrolysing this new peptide bond. Images were generated using data from the following PDB structures: `2G4D` [82], `3KYC` [83], `4W5V` [84], `3UIP` [85]

of different SUMO E3 ligases. Some of these work by simply stabilizing the SUMO-E2 complex, while others can outright force-feed non-canonical targets to the E2 [89].

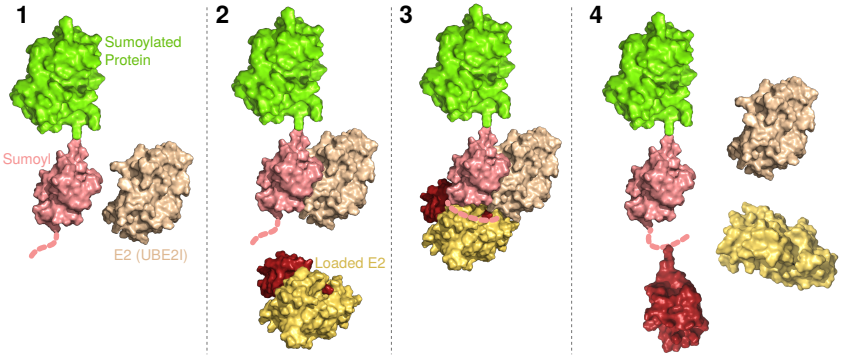

Figure 1.3.: Steps in SUMO chain formation as proposed by Alontaga and colleagues [90]. An E2 noncovalently interacts with a SUMO modification of a target protein. A second E2 carrying a covalently bound second SUMO binds the first E2-SUMO complex, allowing for the first SUMO's N-terminal tail to enter the active site, where a lysine within the tail is forms a peptide bond with the second SUMO's C-terminus. Finally, the complex dissociates, leaving behind the newly formed SUMO chain. Images were generated using data from the following PDB structures: `3UIP` [85], `4Y1L` [90]

Like ubiquitin, SUMO can also form chains (Figure 1.3). However, of the four SUMO proteins encoded by the human genome, only SUMO2 and SUMO3 are capable of doing so, as they contain a suitable lysine residue within a disordered N-terminal tail [91]. Capili and Lima previously observed that the E2 (UBE2I) and SUMO can interact in a noncovalent manner via a distinct binding interface [92]. According to a model proposed by Alontaga and colleagues [90] this interaction is a key mechanism in SUMO chain formation. The interaction recruits a second, SUMO-loaded E2 that interacts with the complex in such a manner that the lysine within the first SUMO's N-terminal tail can find its way into the active site of the second E2, where the second SUMO is

concatenated. While the role of polySUMO chains in humans are still unclear, it has been shown that yeast deficient in SUMO chain formation are unable to perform meiosis [93].

Given the complexity of the Sumoylation system, especially surrounding the E2 component, an examination of sequence-structure-function relationships becomes a multifaceted problem. Mutations could in principle affect any combination of the multiple interaction interfaces which in turn contribute in complex ways to the overall cellular phenotype. An alanine scan of the yeast SUMO E2 Ubc9 was previously performed and succeeded in identifying functionally important sites within the protein [94]. Similarly, a DMS scan of ubiquitin was previously completed [40]. While both of these projects provided great insight into the biochemistry of ubiquitin-like protein pathways, neither has produced a complete map. That is, not all possible amino acid changes were measured at high confidence levels. The Deep Mutational Scanning Framework we will discuss in chapter 2 enabled us to not only recapitulate many of the known mechansims in SUMO and its E2, but also to uncover new details about their biochemistry, as will be discussed in chapter 3.

2. A framework for comprehensive and high-fidelity Deep Mutational Scanning

The work described below represents a team effort including many members of the Roth Lab. Wet lab elements of the work were performed by Atina Coté, Jennifer Knapp, Song Sun and Marta Verby, while all computational and statistical aspects were developed and implemented by myself, except where indicated otherwise.

2.1. Introduction

Deep Mutational Scanning (DMS) [1–3], a strategy for large-scale functional testing of variants, yields functional maps describing a large fraction of substitutions for an often substantial subset of residue positions. The assays used for DMS studies are diverse, often measuring different aspects of a protein's behaviour. Functional complementation assays test a variant's impact on overall protein function by testing the variant gene's ability to rescue the phenotype caused by reduced activity of the wild type gene (or its ortholog in the case of trans-species complementation) [26, 27]. In a previous paper, Song Sun and other members of the Roth Lab have previously found cell-based functional complementation assays to accurately identify disease variants across a diverse collection of human disease genes [13].

There are many challenges to the DMS strategy. One challenge is establishment of robust interpretable assays that measure each variant's impact on the disease-relevant functions of a gene. Another is that the fraction of possible amino acid changes that are measured varies from map to map. Finally, many maps do not control for the overall quality of measurements, or estimate the quality of each measurement. The lack of a comprehensively measured map of known-quality functional impact scores limits the opportunity for confident

use of DMS maps to evaluate specific variants.

Here, a modular DMS framework will be described to generate complete, high-fidelity maps of variant function based on functional complementation. The framework employs a novel mutagenesis strategy, two alternative sequencing-based selection screens, and a machine learning strategy to impute otherwise missing parts of the map with surprising accuracy, and uses regularization to correct less confidently measured data points. The framework is evaluated with respect to its performance on the SUMO E2 conjugase *UBE2I*.

2.2. Results

When carrying out deep mutational scans of protein sequences yielding comprehensive atlases of sequence-function relationships, it is useful to describe the process in distinct stages. The framework described in the following sections can be broken down into six such stages (see Figure 2.1): 1) mutagenesis; 2) generation of a clone library; 3) selection for clones encoding a functional protein; 4) read-out of the selection results and analysis to produce an initial sequence-function map; 5) computational analysis to impute missing values; and 6) computational analysis to refine measured values based on imputation models. The framework incorporates previously-described deep mutational scanning concepts as well as new experimental components (e.g. an imputation and regularization strategy) and analytic methods. In particular, the last two stages enabling a complete and accurate DMS map have not been applied in any published DMS study.

In the following sections, I will first describe a version of the framework called DMS-BarSeq and apply it to the human SUMO conjugase *UBE2I*, exhaustively measuring the ability of protein variants to function. DMS-BarSeq provides direct variant function measurements and the ability to examine higher-order multi-mutant effects. An alternative version of the framework, DMS-TileSeq, generally captures only single-variant effects, but is less resource-intensive. After comparing DMS-TileSeq and DMS-BarSeq, the resulting maps are combined, missing data points are computationally inferred and map quality refined.

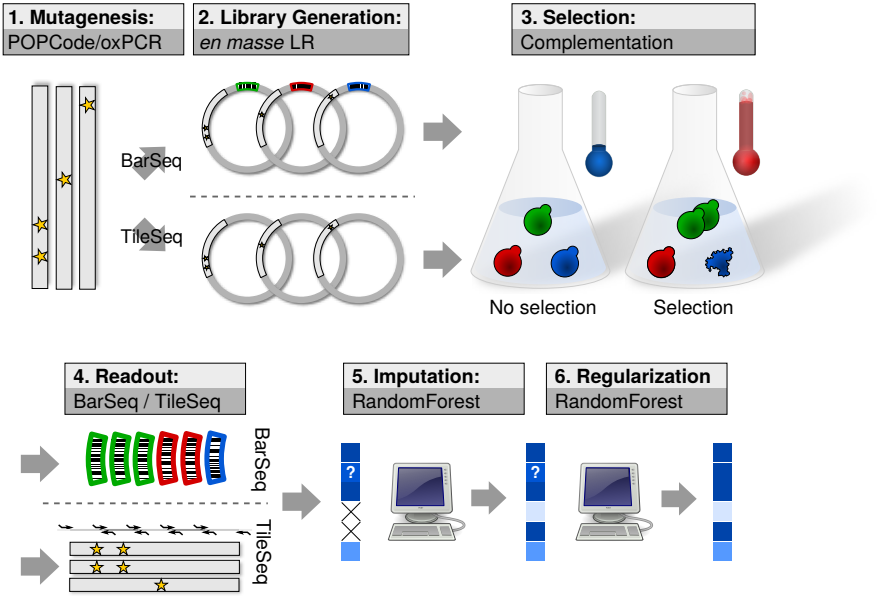

Figure 2.1.: An overview of the Deep Mutational Scanning Framework. Step 1: Using mutagenesis via POPCode and oxidized nucleotide PCR, a pool of variant ORFs is created. Step 2: A library is generated via en-masse gateway cloning. Depending on the downstream sequencing procedure either plain or barcoded expression vectors are used. Step 3: Clones compete with each other for growth under selective and control conditions. Step 4: In case of BarSeq, barcodes are sequenced and counted. In case of TileSeq, individual tiles within the ORF are amplified used in paired-end sequencing. Step 5: Machine Learning methods are used to impute the effects of missing variants. Step 6: Machine learning predictions are also used to support less confidently measured variants. (Incl. illustrations by [95, 96])

2.2.1. A barcode-based Deep Mutational Scanning strategy

As an initial test of the overall framework, we first aimed to generate a map of functional missense variation for *UBE2I*. Our goals for this map were as follows: (i) High and even coverage of the full spectrum of amino acid changes; (ii) Determination of mutant effects on overall protein functionality; (iii) High fidelity of functional effect readouts. We therefore designed the different stages of the framework accordingly.

For Stage 1 of the DMS-BarSeq framework—mutagenesis—to achieve a relatively even representation of all possible single amino acid substitutions, we wished to allow multiple mutations per clone. This would not only allow for greater mutational coverage for any given library size, but it would also offer an opportunity to discover intragenic epistatic relationships between variants. To fulfill these requirements, we developed a mutagenesis protocol (Precision Oligo-Pool based Code Alteration or POPCode) which generates random codon replacements. At the second stage—library generation—we wished to be able to track the fitness effects of each individual mutant clone rather than just average effects of mutations across the population, as this could be expected to allow for higher quality measurements. Thus, in Stage 2 of the framework, we opted to assign molecular barcodes to each clone that could be identified by sequencing. To catalogue the pairing of mutant genotypes with barcodes, we developed a novel multiplex amplicon sequencing method called KiloSeq, in collaboration with Joseph Mellor at SeqWell Inc, Boston. The selection process (Stage 3) was performed as a yeast complementation assay, to allow for determination of overall functional effects of mutations. The assay would be performed as a time series in triplicates, as this again promised to allow for higher quality of readouts Finally, Stage 4, consists of barcode sequencing and statistical analysis. All four stages will be described in further detail in the following subsections.

POPCode: A Precision Oligo Pool Codon alteration mutagenesis method

This method scales up a previously described method developed by Seyfang *et al.* [65]. To achieve complete wide coverage over the complete spectrum of possible amino acid changes in a given gene, oligonucleotides are designed such that they centre on each codon in the Open Reading Frame (ORF) and replace the target with an `NNK` degeneracy code. As explained in chapter 1 section 1.4, this has been previously used to allow all amino acid changes while reducing the

chance of generating stop codons [68].

When designing a set of suitable oligonucleotide sequences, two important criteria need to be considered: (i) The melting temperature across the complete set must be as uniform as possible as this will ensure a more even mutation rate across the ORF sequence; (ii) the degenerate codon sequence should be located as close to the centre of the oligo as permissible given the first criterium. To simplify the process of choosing an appropriate set of oligos based on these criteria, I developed a web tool that can be used to calculate the optimal solution to the given problem. The tool requires the sequence of the target ORF and flanking vector sequences, a desired average oligo length and a maximum offset parameter. The offset parameter determines how many bases can be maximally added or removed from each side of a given oligo to optimize its melting temperature.

In some cases, a moderate deviation from the average in melting temperature for some oligos cannot be avoided. To alleviate these effects, the web tool also offers a mutation rate prediction. This is based on observations from all the POPCode procedures performed as part of this work in combination with linear regression. The prediction can be used to preemptivly adjust concentrations of potentially troublesome oligos in the POPCode protocol. An additional feature in the tool, also based on the mutation rate prediction, is the automatic calculation of necessary library size to achieve a desired mutational coverage. The webtool is available online[1].

Having designed and obtained suitable oligonucleotides, the ORF sequence is PCR amplified in the presence of dUTP to generate uracil-doped template for the mutagenesis reaction. Oligonucleotide pools are then hybridized with the template. Gaps between hybridizations are filled with non-strand-displacing polymerase. Following cleanup, the uracil-doped template is incapacitated using Uracil-DNA-Glycosylase (UDG). The mutagenesis product is then amplified with primers that add `attB` sites to allow for Gateway BP cloning into entry vectors.

To accomplish mutagenesis across the entire coding region of our gene of interest, *UBE2I*, we designed a tiled collection of oligos using the web tool and applied POPCode to generate a codon-mutagenized amplicon library. In parallel, we also carried out PCR with oxidized nucleotides [63] to enable deeper representation of amino acid changes achievable from single-nucleotide changes.

[1] `http://llama.mshri.on.ca/cgi/popcodeSuite/main`

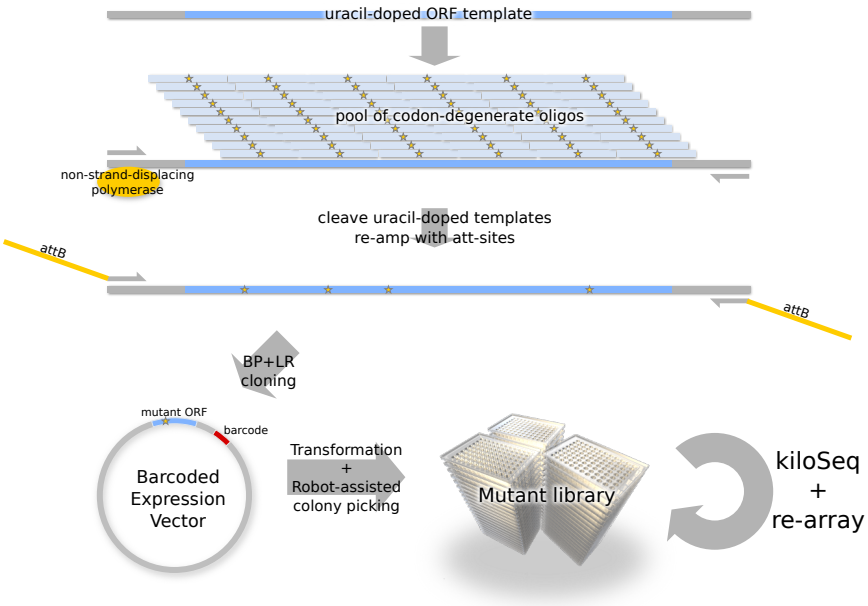

Figure 2.2.: POPCode mutagenesis and library generation. A pool of codon-denerate oligos is hybridized to a uracil-doped template, gaps between oligos are closed via non-strand-displacing polymerase, and the backbone sealed. Uracil-doped template is degraded to enrich for mutants. After mutagenesis, Gateway attB sites are added, followed by BP+LR cloning into barcoded vectors and transformation into bacteria. Finally, colonies are picked and arrayed. (Incl. illustration by [**?**])

Library generation and highly multiplexed amplicon sequencing

For Stage 2 of the framework—generation of a clone library—we employed an *en masse* recombinational cloning strategy to generate a Gateway Entry vector library of *UBE2I* variants. This library was transferred via *en masse* recombinational subcloning into a pool of randomly-barcoded plasmids enabling

expression of *UBE2I* variants in yeast. As sequencing is required to establish the full-length ORF sequence and barcode of each clone, the complementation vector is designed such that the variant ORF and the barcode locus are in close proximity to each other. Thus, only a relatively small segment of the plasmid needs to be inspected to determine the pairing of genotype and barcode.

After bacterial transformation, we proceeded to robotically pick 19,968 colonies, which were stored in 52 384-well plates. As sequencing needs to be performed to catalogue the identities of nearly 20,000 individual samples, we used a novel sequencing method called KiloSeq which combines plate-position-specific index sequences with Illumina sequencing (Figure 2.3). KiloSeq was developed in collaboration with SeqWell Inc., Boston. First, for each clone in the library, the region of interest is amplified with primers containing well-specific tags, uniquely identifying each well coordinate. This step is dependent on the use of a HydroCycler, which allows up to 4608 PCR reactions to be performed in parallel. In the next step, wells for each plate can be pooled. Nextera tagmentation using Tn5 transposase is used to break the amplicons into random fragments and simultaneously ligate them to Illumina sequencing linkers with plate-specific indices. Then the pool is re-amplified with 3'-specific primers, to enrich for fragments that contain the well tags. The resulting library is now ready for paired-end sequencing. In each pair of reads, one read will contain the well tag and the barcode locus, whereas the other will contain a fragment of the mutant ORF.

To process the results of a KiloSeq sequencing run, I developed a custom-built software pipeline, which can be divided into three phases: demultiplexing; barcode clustering; and alignment and variant calling. The first phase—demultiplexing—takes place on two levels, corresponding to library plates and the wells within those plates. Demultiplexing at plate level is performed by Illumina's `bcl2fastq` software, which resolves i5-i7 index combinations. The second phase is performed on a high performance computing cluster. Sets of read pairs are distributed across computing nodes, where they are processed by worker scripts. The well-tag within each R2 read is identified using a k-mer search algorithm, and read-pairs are sorted accordingly into bins. Each bin corresponds to one well in a given plate. At the same time, barcode sequences are extracted from the R2 reads in preparation for the next phase.

The second phase—barcode clustering—uses the extracted barcode sequences within each bin and clusters them according to their Levenstein distance [97] (i.e. the number of edit operations required to transform one into the other). This step is necessary in order to resolve possible contamination across wells

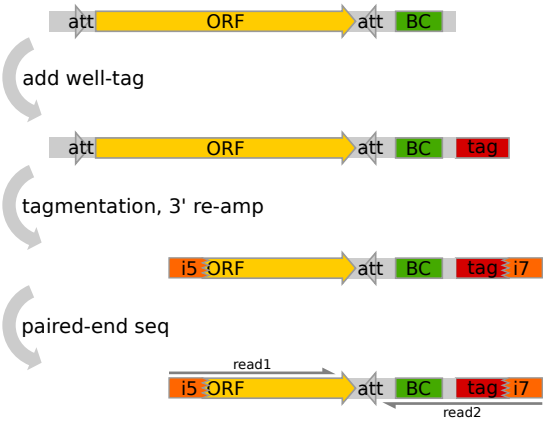

Figure 2.3.: KiloSeq schema. 1) For each library well, amplicons containing the variant ORF (gold) and Barcode locus (green) are amplified with primers adding a well-specific tag. 2) Tn5 tagmentation fragments the DNA while simultaneously adding Illumina i5/i7 linkers. 3' re-amplification enriches for fragments containing the well tags. 3) Each pair of sequencing reads now contains a fragment of ORF sequence and the associated barcode and well tag.

that occurred during library preparation. Each barcode cluster corresponds to a different clone, and the different unique sequences within each clusters correspond to different sequencing errors. The most frequently observed sequence within each cluster is interpreted as the true barcode. Finally, read pairs within each bin are again subdivided according to their respective barcode cluster.

The third phase—alignment and variant calling—is then executed for each barcode cluster within each well within each plate. The R1 reads are aligned to the template sequence and variants are called. This is complicated by the fact that the KiloSeq library preparation usually creates a certain amount of cross-contamination between wells. While single or multi-nucleotide variants are still relatively unproblematic to identify, standard tools were found to be unable to identify copy number variations (CNVs) due to these problems. I

thus developed a custom method for CNV calling, based on detecting sudden changes in read depth across the alignments. First, the individual read depth track is normalized to the average read depth across all wells the plate. Then a modified one-dimensional Sobel operator [98] is used to detect sharp edges in the signal. An example of this can be seen in Figure 2.4. Detection thresholds were optimized by comparison with Sanger sequencing.

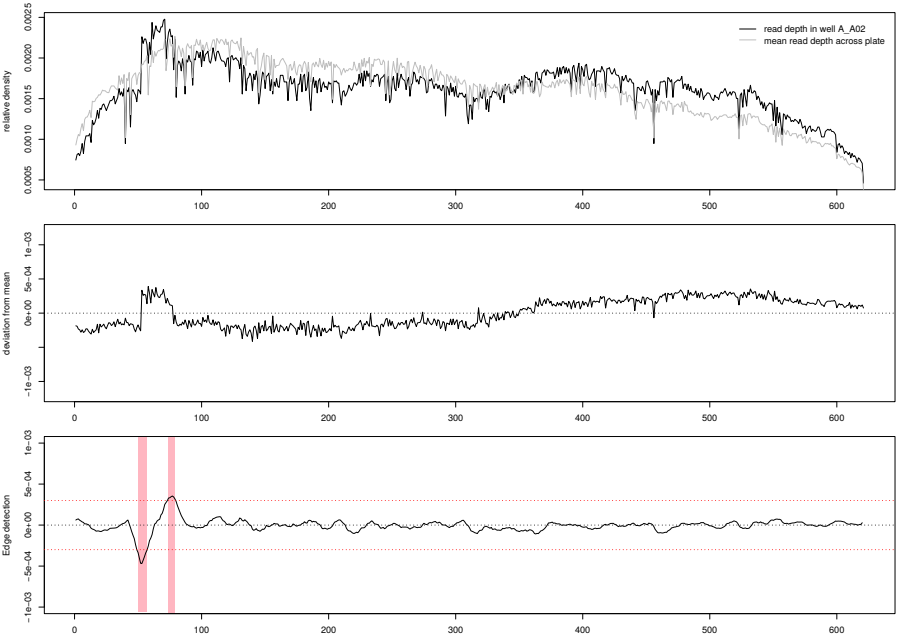

Figure 2.4.: Indel detection example. A duplication event in well `A_A02` is detected by normalizing relative read depth by the mean depth across the plate and using a Sobel operator to detect sudden changes.

After successful genotyping with kiloseq, I determined the subset of clones that (i) contained at least one missense mutation, (ii) did not contain any insertions or deletions, (iii) did not contain mutations outside of the ORF, (iii) had unique barcodes, and (iv) had sufficient read coverage during KiloSeq to

allow for confident genotyping. Over half of the clones in the library conformed to these criteria. The single largest reason for exclusion was the occurrence of indels and CNVs (Figure 2.5A).

An analysis of the mutation signatures across clones generated by POPCode revealed that two different mechanisms appear to underlie mutagenesis. When considering only mutations that change more than one base in a given codon, there is an equal chance for every possible base except in the third position, where almost no adenine or cytosine was introduced. This is consistent with the `NNK` degeneracy code used in the POPCode oligo design. By contrast, variants that change only a single base in a given codon show a strong bias for transitions over transversions. These could be introduced due to polymerase error (Figure 2.5B). This secondary source of variation is also reflected in the relative share of single nucleotide variants, which make up 56% of mutations (Figure 2.5C). As a consequence, when examining the mutation coverage across the sequence of the ORF, it is clearly visible that the share of amino acids reachable with a single nucleotide change from the respective wildtype codon is much closer to saturation than the the set of all possible amino acid changes (Figure 2.5D). Additionally, some hotspots are visible in which the mutation rate is higher, which is likely due to different hybridization efficiencies of oligos across the ORF sequence.

Using a pinning robot, we re-arrayed the subset of usable clones into a condensed final library of 40 plates. This final library comprised 6,553 *UBE2I* variants, covering different combinations of 1,848 (61% of all possible) unique amino acid changes. In preparation for the next stage, variant plasmids were pooled, together with barcoded empty vector and wild type control plasmids.

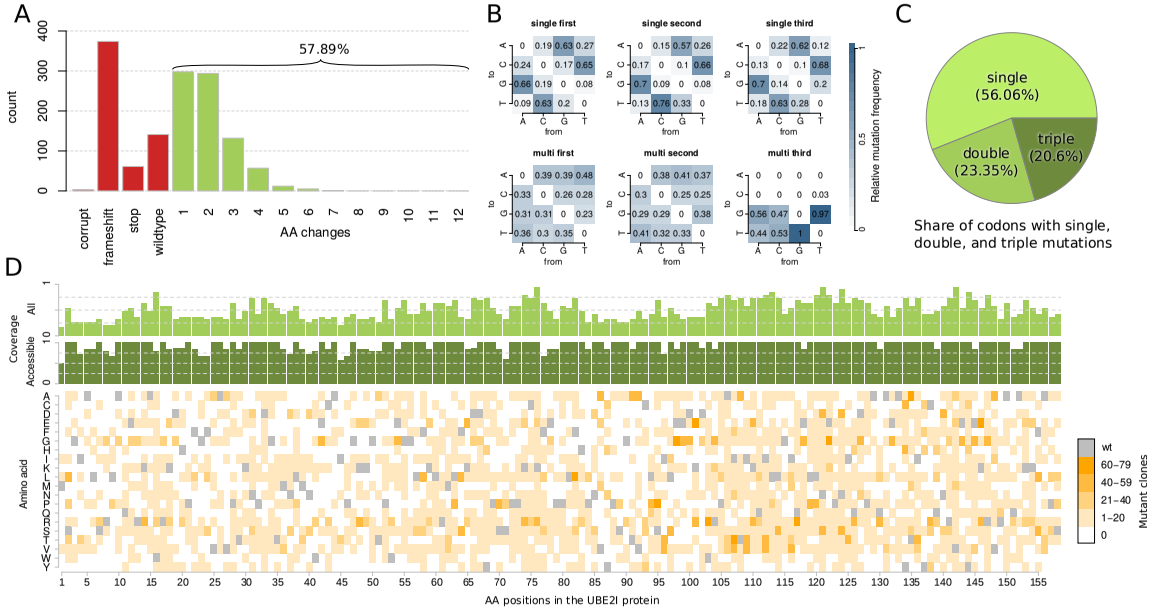

Figure 2.5.: KiloSeq-based census of the *UBE2I* POPCode library. A) Breakdown of KiloSeq results for a set of five 384 well plates of mutant clones generated by POPCode. Corrupt: Clones containing mutations outside of the ORF; Frameshift: Clones containing indels or copy number variants; Stop: Clones containing stop codons. B) Breakdown of mutations in codons. Top: Single nucleotide variants; Bottom: Multi-nucleotide variants. Columns correspond to the first, second and third position in a codon. C) Relative shares of single, double and triple nucleotide variants among all missense variants in the library. D) Coverage map of missense variants in the library. Light green track: Coverage across all possible amino acids; Dark green track: Coverage across amino acids reachable with a single nucleotide change from the wildtype codon.

Complementation screen and Barcode sequencing

For Stage 3 of the DMS-BarSeq framework—the selection of clones encoding a functional protein—we employed a previously described *S. cerevisiae* functional complementation assay [26, 27]. This assay is based a yeast strain carrying a temperature sensitive (ts) allele of the *UBE2I* orthologue *UBC9*. Expression of human *UBE2I* rescues growth at an otherwise lethal elevated temperature. As such, the fitness observed for a clone carrying a mutant allele of *UBE2I* can be interpreted as the overall ability of the variant protein to function within its biological context [13]. The plasmid library from Stage 3 was introduced into the appropriate ts strain by en-masse transformation. Pools were then grown in triplicates over a period of 48 hours at the permissive (25°C) and selective (37°C) temperatures, respectively (see Methods) and evaluated at multiple time points via high-throughput sequencing.

To facilitate the readout of the selection (Stage 4), I developed a sequence analysis pipeline. The pipeline distributes sets of read pairs across across the nodes of a high-performance computing cluster, where a k-mer search algorithm is used to identify multiplexing tags that encode the temperature and time point and replicate number associated with the sample. The same algorithm is also used to identify the barcode itself. The number of occurrences of each barcode in each sample is counted and aggregated across the cluster nodes. The frequencies at which each barcode is observed corresponds to the population size of the associated clone. This can then be used to reconstruct of individual growth curves and quantify the normalized fitness for each of the 6,553 strains (see Methods section for details). The fitness measurements are normalized to the wildtype and null controls, such that a score of 1 is equivalent to the average wildtype fitness, and 0 is equivalent to the average null control fitness.

Additional care needs to be taken to quantify the level of confidence for each fitness measurement. While comparing the three technical replicates available for each clone allows for a rough estimation of standard error, improvements can be made. Baldi and Long previously published a Bayesian method allowing for the regularization of variance estimations using prior data [99]. Two sources of prior information offer themselves: (1) The number of sequencing reads observed at time 0 of the experiment, as a low number indicates underrepresentation in the library, which is likely to result in a poor frequency estimate; and (2) the fitness estimate itself, as variance can be expected to be proportional to the mean. Indeed, when comparing both properties with the standard deviation, a clear trend is visible (Figure 2.6). After obtaining a

prior estimate via linear regression, it can be used to regularize the empirical standard deviation.

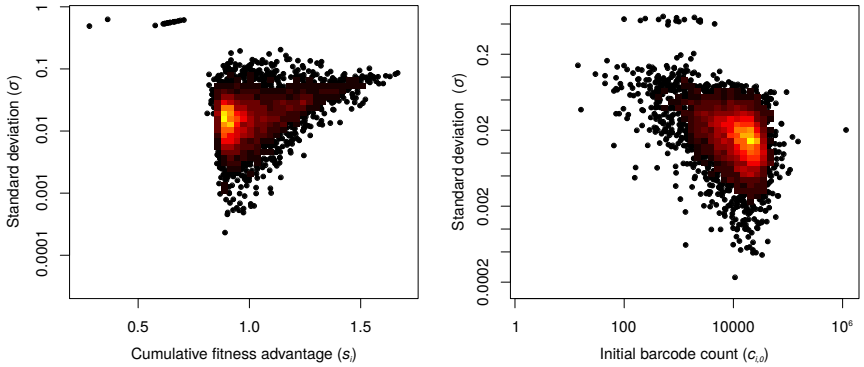

Figure 2.6.: Comparison of fitness and initial barcode count against standard deviation. Both properties can be used as prior information to improve confidence quantification.

A barcoded-based functional map of UBE2I

Before further refinement in Stages 5 and 6, I assessed the quality of complementation scores. I first examined reproducibility of scores between technical replicates (Figure 2.7A), and biological replicates (different clones carrying the same mutation; Figure 2.7B). In each case the scores were reproducible (Pearson's R of 0.97 and 0.78, respectively). We next carried out semi-quantitative manual complementation spotting assays for a subset of mutants that spanned the range of fitness scores. Complementation scores from deep mutational scanning correlated well with these small-scale tests. Indeed, agreement between the large-scale and manual scores was about the same as agreement between internal replicates of the large-scale scores (Figure 2.7B,C).

As a further sanity check, I next examined evolutionary conservation and common predictors of deleteriousness, such as PolyPhen-2 [10] and PROVEAN [12]. Although each of these measures is far from perfect in predicting the functionality of amino acid changes, they should and did each correlate with functionality

(Figure 2.7D,E,F). Finally, I confirmed that, as expected, amino acid residues on the protein surface are more tolerant to mutation than those in the protein core or within interaction interfaces (Figure 2.7G). Taken together, these observations support the biological relevance of the DMS-BarSeq approach.

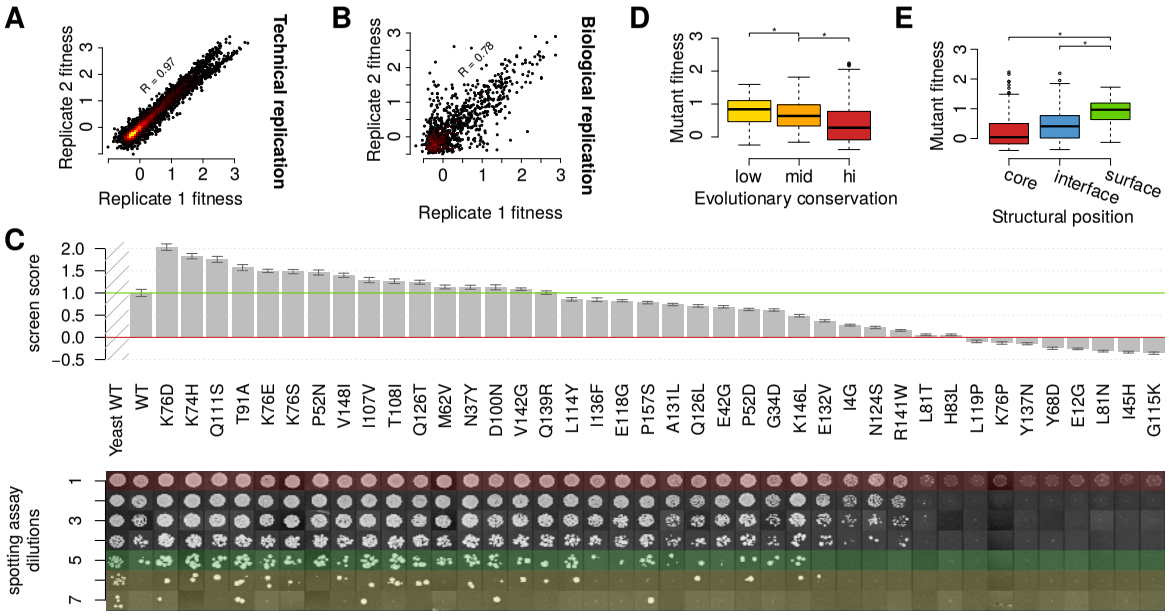

Figure 2.7.: Validation of DMS-BarSeq of UBE2I. A: Correlation between technical replicates B: Correlation between biological replicates. C: Manual complementation spotting assay compared to DMS fitness measurements. D: Comparison of fitness levels for mutations at positions with low, medium and high evolutionary conservation. E: Comparison of fitness levels for mutations at positions within the hydrophobic core, at interaction interfaces, and unused surfaces

2.2.2. An alternative strategy for DMS via tiled regional sequencing

While the DMS-BarSeq approach has many advantages (see Discussion), its performance comes at the cost of producing and maintaining an arrayed clone library, and of determining the full-length sequence of each coding region and barcode for each clone. We therefore investigated an alternative approach called DMS-TileSeq: Instead of tracking the fitness of each individual clone, we carried out *en masse* measurements of the frequency of each variant in the pool before and after selection, by deep sequencing. Sequencing was carried out for a set of short amplicon tiles that collectively encompass the complete coding region. In this way, it is possible to discern the impact of each mutation by observing the impact of selection on the abundance of clones carrying this mutation.

In terms of mutagenesis (Stage 1), DMS-TileSeq is identical to DMS-BarSeq. Given the mutagenized amplicon library, the cloning step (Stage 2) was carried out by *en masse* recombinational subcloning into complementation vectors (thus skipping the step of arraying and sequencing individual clones). This plasmid pool was next transformed *en masse* into the *ubc9-ts* strain appropriate for assessing the complementation ability of *UBE2I* variants. As with DMS-BarSeq, DMS-TileSeq employs pooled strains grown competitively (Stage 3) at the permissive and selective temperatures. However, instead of using barcode sequencing to determine the fitness associated with individual stains, we directly sequence the coding region from the clone population to determine the frequency of each variant in each pool (before and after selection). To overcome the problem of distinguishing mutations from sequencing errors, we divide the coding region into tiles such that each individual template molecule can be completely sequenced on both strands. By requiring that each variant be seen on both strands, the incidence of base-calling errors can be substantially reduced.

An important aspect of DMS-TileSeq is that it requires the library to be sufficiently complex to ensure that the effect of a mutation is determined from enough clones and averaged over enough genetic backgrounds to be reproducible. Therefore it was necessary to first validate the reliability of DMS-TileSeq in comparison to DMS-BarSeq on our established UBE2I map. Correlation between DMS-TileSeq and DMS-BarSeq was comparable to the correlation observed between biological replicates of DMS-BarSeq (Figure 2.8A), suggesting that reproducibility of DMS-TileSeq is at least comparable to that of DMS-BarSeq. DMS-TileSeq and DMS-BarSeq showed similar agreement with

complementation scores from manual assays (Figure 2.8B). Thus, DMS-TileSeq avoids the substantial cost of arraying and sequencing thousands of individual clones, while performing on par with DMS-BarSeq in terms of reliability of the functional complementation scores it produces.

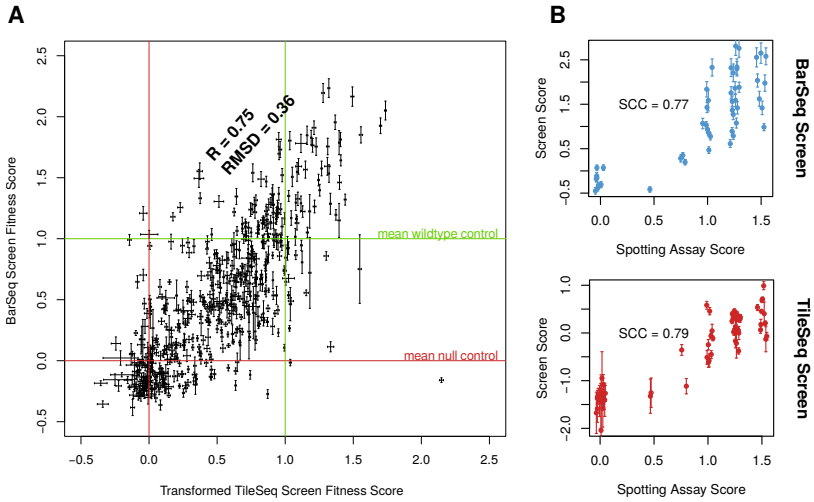

Figure 2.8.: Comparison of DMS-BarSeq to DMS-TileSeq scores. A) Scatterplot of functional scores for variants in obtained from BarSeq and TileSeq (transformed to the same scale), whisker bars show regularized standard error. B) Comparison of BarSeq (top) and TileSeq (bottom) scores to manual complementation spotting assay scores (jittered for visibility). Whisker bars show regularized standard error

2.2.3. A complete functional map of UBE2I

Having performed two independent deep mutational scans of UBE2I using functional complementation assays, we wished to integrate both results into a single comprehensive high-quality map. To accomplish this, I first combined the results of each screening approach into a joint map. This required bringing the maps onto the same scale. Using a regression-based transformation function, I

transformed the DMS-TileSeq scores to the more intuitive scale of DMS-BarSeq (where 0 corresponds to the typical score of a null mutant and 1 corresponds to the typical score of a wildtype control). I then combined scores from the two methods, giving greater weight to more confident measurements (see methods section).

Imputation and regularization of missing or less accurate data

As is the case for all previously published DMS maps, our combined map contained some entries that were poorly measured or missing (e.g., because these substitutions were underrepresented in the input clone library). To fill the gaps in the map (Stage 5 in the framework), I trained a Random Forest [100] regression model using the existing measurements in the map. The features used for the model fall into four categories: intrinsic information; conservation information; chemicophysical properties; and structural properties.

The most important intrinsic feature consists of weighted positional averages in the map. That is, for any given amino acid change, all other observed effects of variants at the same amino acid position are weighted according to their measurement confidence and are then used to form an average. A second intrinsic feature consists of the confidence-weighted average effect of all variants containing the amino acid change in question. Finally, as a third intrinsic feature I calculate the expected variant fitness predicted by a multiplicative model often applied to detect genetic interactions [101, 102]. In the absence of interaction, the fitness of a double mutant $f_{A,B}$ is expected to follow the product of the individual single mutant fitness levels $f_{A,B} \approx f_A \cdot f_B$. Thus, in cases where a double mutant (A, B) and a single mutant B is known, the fitness of A can be estimated to be $f_A \approx \frac{f_{A,B}}{f_B}$. The model is applied to all available double mutant fitness values carrying the mutation in question in combination with available complementary single mutant fitness values. As the latter two features rely on multi-mutant fitness measurements, they can only be applied where DMS-BarSeq data is available.

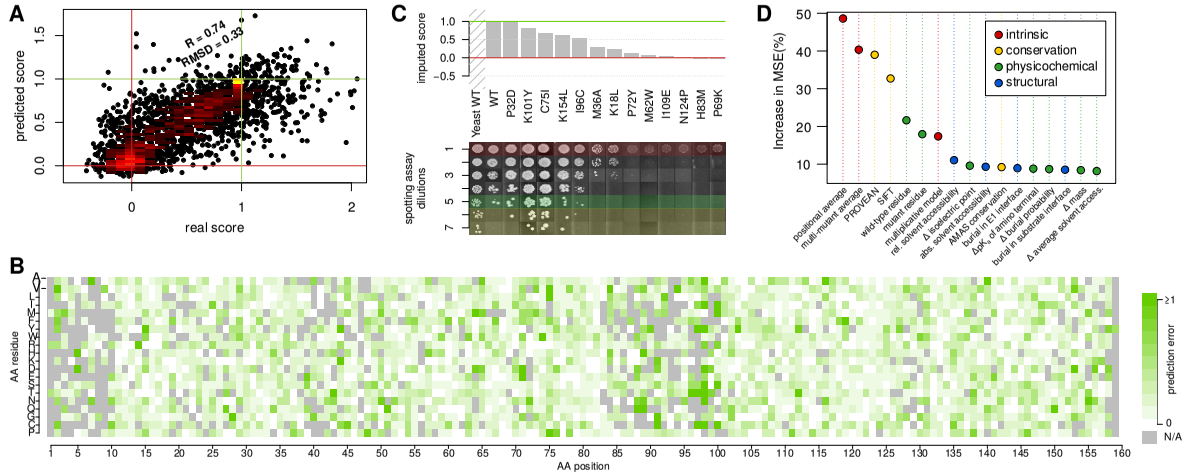

Figure 2.9.: Evaluation of machine learning imputation. A) Cross-validation correlation between measured values and machine learning predictions. B) Cross validation prediction error landscape. C) Manual complementation spotting assay compared to machine learning predictions for an independent test set of variants not present in the training data. D) Feature importance as measured by average increase in mean squared error.

The second category of features focuses on evolutionary conservation. For each amino acid change in question, this encompasses the corresponding BLOSUM62 [103], SIFT [11] and PROVEAN [12] scores, and the AMAS [104] conservation at the given position. The third category of features comprises chemicophysical properties such as mass and hydrophobicity of the original and wildtype amino acids and the difference between the two. The fourth and final category of features consists of structural properties of the affected amino acid residues, such as solvent accessibility, engagement in polar interactions and burial in interaction interfaces.

I assessed the performance of the imputation model using cross-validation. Surprisingly, I found the root-mean-squared deviation (RMSD) of imputed values to be on par with measurement error in experimentally measured data (Figure 2.9A). An examination of the prediction performance by location showed increased error in positions with lower mutation density and for variants with above-WT fitness levels (Figure 2.9B). As an additional validation step, we performed manual complementation assays for a set of UBE2I variants that were not present in the machine learning training data set and compared the results against the predictions (Figure 2.9C), again finding a surprisingly strong agreement. Notably, variants showing above wild-type level growth in the manual assay were generally predicted to be deleterious. Although above-WT complementation may indicate that a variant is adaptive in yeast, the imputation models suggested that these variants would be deleterious in humans, a hypothesis that is explored further in chapter 3.

An analysis of feature importance can be performed by comparing the increase in the mean squared prediction error upon permuting the values of a feature in question. The analysis revealed that intrinsic features were the most informative (Figure 2.9D), with the weighed position-wise average and multi-mutant average seen to be the two single most important features (49% and 40%, respectively), while the multiplicative model contributed 14%. The second most important group was conservation information, with PROVEAN and SIFT weighing in at 39% and 32%, respectively.

Finally, in stage 6 of the DMS framework, we wished to address cases in which experimental measurements were available but less confident. I implemented a regularization method, combining experimental measurements with machine-learning predicted values after dynamically weighting them according to their respective confidence levels. That means: the less confident a measurement, the stronger the regularization. Overall, most values were only adjusted minimally through regularization, with 90% of values being altered by less than 2.5% of

the score difference between null and wt controls (Figure 2.10). This reflects the fact that most values were already of high quality.

To evaluate the effect on the minority of variants that required stronger regularization, I looked for cases that were of low quality in the DMS-TileSEQ dataset, but well measured in the DMS-BarSeq experiment. This would allow me to treat the DMS-BarSeq values as a gold-standard basis of comparison when performing the regularization procedure only on the DMS-TileSeq dataset. I identified six cases that fulfilled these criteria. In all six cases regularization of DMS-TileSeq resulted in improvement, i.e. adjusted the corresponding values such that they more closely resembled the gold standard (Figure 2.10B). However, I found the changes to be still very conservative. More drastic weighting towards the machine learning prediction could have improved these cases even more.

To evaluate the complete map, we once more applied manual complementation assays to a set of variants that represented the full range of fitness scores. DMS fitness scores corresponded closely with manual assays (Figure 2.10)C, with a Spearman correlation of 0.83 between the high-throughput and low-throughput values (a slight improvement of 0.06 compared to the raw, unregularized experimental data).

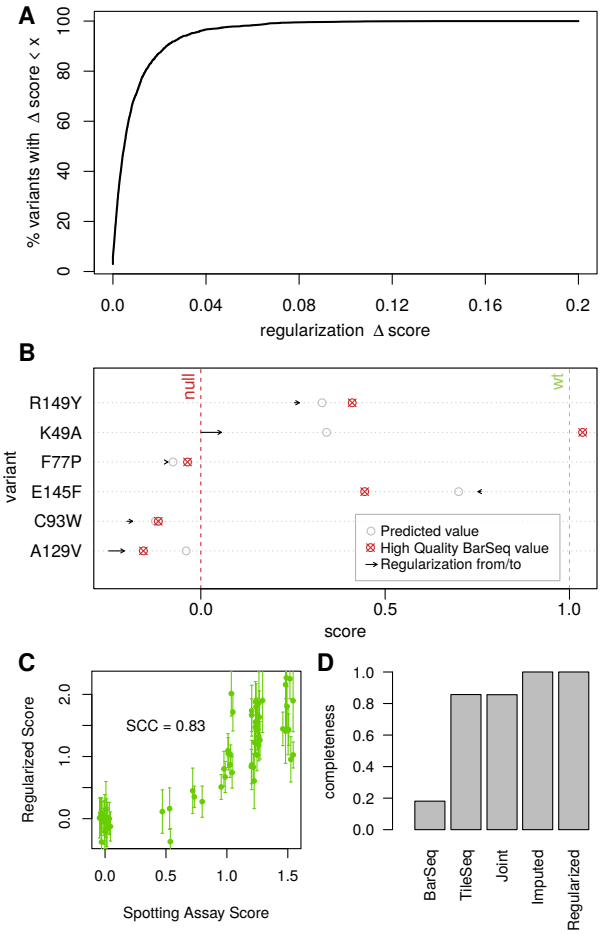

Figure 2.10.: A) Cumulative distribution of changes to the fitness scores on the map as a result of regularization. B) Six variants were that were well measured in DMS-BarSeq but less well measured in DMS-TileSeq. For evaluation, regularization was performed only on the DMS-TileSeq data and compared to the DMS-BarSeq gold standard. Base and tip of arrows indicate pre- and post-regularization values, gray circles indicate the machine learning predictions used. Red targets indicate BarSeq gold standard. C) Comparison of values in the regularized map against manual complementation spotting assay values. D) Completeness of the map (in terms of coverage of possible amino acid changes) at different stages of the framework.

2.3. Discussion

Here I have demonstrated the capabilities of a new improved Deep Mutational Scanning framework that uses functional complementation in yeast to map the impact of mutations on the overall ability of a protein to function. I integrated a machine learning-based imputation and regularization strategy into the deep mutational scanning process, to create the first DMS map that is complete with respect to high-quality functional impact scores over the full length of a protein.

The two versions of DMS described, DMS-BarSeq and DMS-TileSeq, each have advantages and limitations. DMS-BarSeq permits study of the combined effects of mutations located at any distance along the clone, and therefore can reveal intramolecular genetic interactions (as will be explored further in the next chapter). Futhermore, mutant clones produced for DMS-BarSeq are arrayed, sequenced and indexed which enables potential follow up investigation of individual variants. DMS-BarSeq also allows for the direct comparison of growth of any clone to null and wild type controls, resulting in an intuitive scoring scheme. However, the cost of arraying and sequencing clones for DMS-BarSeq renders it more costly and labour intensive, even given the efficient KiloSeq strategy. By contrast, the regional sequencing strategy of DMS-TileSeq is substantially more efficient, but can only analyze fitness of those double mutant combinations that fall within the same tile.

The use of codon-replacement mutagenesis allows for the observation of a fuller repertoire of amino-acid substitutions than single-nucleotide mutagenesis would have allowed (only $\sim$ 30% of all possible amino acid substitutions are accessible by single nucleotide mutation). However, given that the majority of missense variants observed in individual genomes are single-nucleotide variants [105], one might reasonably wonder whether codon mutagenesis is worth carrying out in addition to single-nucleotide mutagenesis. There are three arguments for using codon-level mutagenesis to reveal the impact of all 19 possible amino acid substitutions at each position: 1) a full picture of functional missense variation enables a clearer understanding of what biochemical properties are required of each functionally important residue; 2) an analysis of over 60,000 unphased human exomes [105] found that each individual human harbors approximately 23 codons containing multiple nucleotide variants that collectively encode an amino acid not encoded by either single variant; 3) it seems likely that, going forward, the dominant cost of DMS will be development and validation of the functional assay, so that carrying out codon-level mutagenesis

instead of (or in addition to) nucleotide-level mutagenesis has a relatively small impact on overall cost.

2.4. Methods

2.4.1. Mutagenesis and library construction

Oxidized nucleotide PCR: Oxidized nucleotide PCR was performed by Jennifer Knapp as previously described by Mohan and colleagues [63]. A 100µM dNTP mixture was incubated at 37°C with 5mM $FeSO_4$ for 10 minutes. Addition of 0.5M Mannitol was used to stop the reaction. Oxidized nucleotides were prepared fresh for every PCR reaction. PCR in presence of oxidized nucleotides. PCR reaction containing: 1-5ng template DNA, 1× Thermopol Buffer (Invitrogen), 1.5mM $MgCl_2$, 0.2mM dNTP, 0.33µM forward and reverse primers containing attB sites, 1U Taq polymerase was set up during the nucleotide oxidation reaction. Oxidized nucleotides were the last component added to the PCR reaction at a concentration of 0.1mM (half the amount of regular dNTP). Thermal cycler program: 95°C for 10 min, 30 cycles of 95°C for 1 min, 50°C for 1 min, 72°C for 1 min, final extension at 72°C for 10 min. Mutagenized PCR product was visualised on a 1% agarose gel, and gel-extracted using a gel extraction kit (Qiagen). The gel extracted PCR product is the pooled mutagenesis product carrying *attB* sites that is carried through to the KiloSeq stage.

POPCode mutagenesis

Oligonucleotide design: POPCode oligos are generated using the POPCodeSuite webtool I created. Given a target oligo length and a maximum length offset, the tool calculates for every codon in the target gene the set of possible oligos conforming to the length and offset parameters. Then, melting temperatures for the 5' and 3' halves of each oligo are calculated. For each codon, the oligo that most closely matches the median 5' and 3' melting temperatures is chosen. Based on parameters derived from previous observations, the expected mutation frequency is calculated for each oligo and used to simulate variant coverage rates at different library sizes. The source code is provided on the attached storage media and can also be found online[2].

[2] `http://dalai.mshri.on.ca/~jweile/projects/popcodeSuite/`

The POPCode mutagenesis protocol was performed by Atina Coté, Jennifer Knapp and Marta Verby in the following steps: (i) the uracil-containing wild type template was generated by PCR-amplifying the ORF with dNTP/dUTP mix and HotTaq DNA polymerase, (ii) the mixture of phosphorylated oligonucleotide pool and uracil-containing template was denatured by heating it to 95°C for 3 minutes and then cooled down to 4°C to allow the oligos hybridize to the template, (iii) gaps between hybridized oligonucleotides were filled with the non-strand-displacing Sulpholobus Polymerase IV (NEB) and sealed with T4 DNA ligase (NEB), (iv) after degradation of the uracil-doped wild-type strand using Uracil-DNA-Glycosylase (UDG) (NEB), the mutant strand was amplified with attB-sites-containing primers and subsequently transferred en masse to a donor vector by Gateway BP reaction to generate a library of entry clones.

Synthesis of uracil-containing template: A 50µl PCR reaction contained the following: 1ng template DNA, 1× Taq buffer, 0.2mM dNTPs-dTTP, 0.2mM dUTP, 0.4µM forward and reverse oligos, and 1U Hot Taq Polymerase. Thermal cycler conditions are as follows: 98°C for 30s, 25 cycles of 98°C for 15s, 60°C for 30s, and 72°C for 1min. A final extension was performed at 72°C for 5 min. Uracilated amplicon was gel-purified using the Minelute gel purification kit (Qiagen).

Phosphorylation of mutagenic oligos: Desalted oligos were purchased from Eurofins and Thermo Scientific. The phosphorylation reaction is as follows: a 50µl reaction containing 1× PNK buffer, 300 pmol oligos, 1mM ATP, and 10U Polynucleotide Kinase (NEB) was incubated at 37°C for 2 hours. The reaction was used directly in the subsequent POPCode reaction.

POPCode oligo annealing and fill-in: A 20µl reaction containing 20ng uracilated DNA, 0.15µM phosphorylated oligo pool, and 1.5µM 5'-oligo was incubated at 95°C for 3 minutes followed by immediate cooling to 4°C. A 30µl reaction containing 1× Taq DNA Ligase buffer, 0.2mM dNTPs, 2U Sulfolobus DNA Polymerase IV (NEB), and 40U Taq DNA Ligase (NEB) was added to the DNA and was incubated at 37°C for 2 hours.

Degradation of wild-type template: 1µl fill-in reaction was added to a 20µl reaction containing 1× UDG buffer and 5U Uracil DNA Glycosylase (NEB)

and incubated at 37°C for 2 hours.

Amplification of mutagenized DNA: 1µl UDG reaction was added to a 50µl reaction containing 1× Taq buffer, 0.2mM dNTPs, 0.4µM forward and reverse oligos, and 1U Hot Taq Polymerase. Thermal cycler conditions are as follows: 98°C for 30s, 25 cycles of 98°C for 15s, 60°C for 30s, and 72°C for 1min. A final extension was performed at 72°C for 5 min.

Library construction

Library construction was performed by Atina Coté, Jennifer Knapp and Marta Verby following the *en masse* LR cloning protocol previously described in Yachie *et al.* [106].

Generation of mutagenised pool of Entries: An *en masse* Gateway BP reaction containing 150ng of pooled mutagenesis PCR product carrying *attB* sites, 150ng of pDONR223, 1µL Gateway BP Clonase II Enzyme Mix (Invitrogen), 1× TE Buffer is prepared. This reaction is incubated overnight at room temperature and then transformed into *E. coli* aiming for the maximum number of transformants (at least 100,000 CFUs) to keep complexity high. Several colonies are picked at this stage for a quality control check by Sanger sequencing, and the rest are put through a pooled DNA extraction. The result is a pool of mutagenised PCR product inserted into the entry vector pDONR223.

Generation of Barcoded Destination Pools: Barcoded destination plasmids were generated as previously described in Yachie *et al.* [106], but instead of being arrayed were maintained as pools with high complexity. Briefly, a linear PCR product containing two random 25 nucleotide barcode regions along with common linker sequences for priming was combined with a Gateway-compatible vector at a SacI restriction site through *in vitro* DNA assembly [107]. This barcoded destination vector pool was transformed into One Shot ccdB Survival T1R Competent Cells (Invitrogen). The transformations were spread onto large round LB+ampicillin petri plates for increased selection capacity and pool complexity was estimated from CFU counts. The plates were combined into a single pool for plasmid DNA extraction by maxiprep.

En masse Gateway LR reaction: An *en masse* Gateway LR reaction was used to transfer the mutagenised pool of entries into the barcoded destination pool.

This reaction takes place over five days. On Day 1, a 5μL reaction containing 150ng of mutagenised ORF pool in pDONR223 backbone, 150ng barcoded pHYC expression vector pool, 1μL LR ClonaseII Enzyme Mix, 1× TE buffer is prepared. The reaction is incubated at room temperature overnight. On each of days 2-5, add in a 5μL volume consisting of 150ng barcoded pHYC expression vector, 1μL LR ClonaseII Enzyme Mix, 1× TE Buffer, incubating at room temperature overnight each day. On day 5, the final volume is 25μL.

Transformations and colony picking: LR reactions were transformed into *E. coli* and plated to achieve a density of 400-600 individual colonies per plate. A Biomatrix robot (Biomatrix BM5-BC robot, S&P Robotics) was then used to automatically pick and array 384 colonies per plate for a total of ∼20,000 clones in ∼52 plates per ORF of interest. Each colony at this stage should contain a pHYC expression vector harbouring a variant of the ORF of interest and a unique barcode.

2.4.2. KiloSeq and library condensation

Experimental procedures: KiloSeq library preparation was performed by Atina Coté, Jennifer Knapp and Marta Verby. The first step is to PCR-amplify a segment of the plasmid containing both ORF and barcode locus. PCRs were carried out using the Hydrocycler 16 (LGC Group, Ltd.), using primers with well-specific index sequences. Amplicons from each plate were pooled, and subjected to Nextera tagmentation using Tn5 transposase to generate a library of amplicons with random breaks to which the adapters have been ligated. The fragments are then re-amplified to generate a library of amplicons such that one end of each amplicon bears the well-specific tag and the other (ladder) end bears the Nextera adapter. These libraries can be re-amplified to introduce Illumina TruSeq adaptors, allowing multiple plates of amplicons to be sequenced together. Paired-end sequencing was carried out using Illumina NextSEQ 500. In each pair of reads, one read will reveal the well tag and the barcode locus, whereas the other will contain a fragment of the mutant ORF, and these fragments can be assembled into a contiguous sequence.

Computational procedures: I developed a sequence analysis pipeline to process all KiloSeq data. The pipeline runs on a high-performance computing cluster (Figure 2.11). In the first step, Illumina `bcl2fastq` is used to demultiplex the reads at the plate level using the custom Nextera indices. The resulting

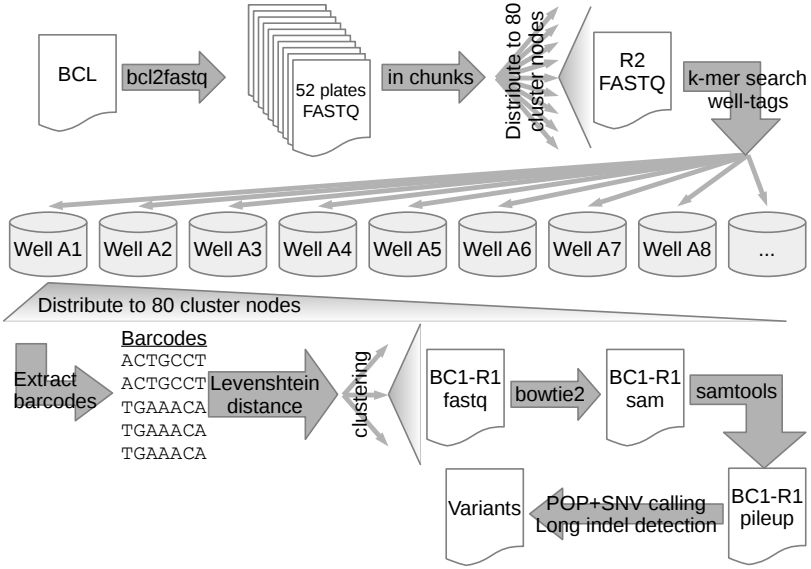

Figure 2.11.: KiloSeq analysis pipeline: `bcl2fastq` is used to demultiplex by plate. The resulting FASTQ files are broken up in to chunks and feed to worker nodes on the fly. Each worker identifies well-tags in the R2 reads and demultiplexes by well accordingly. After demultiplexing is complete, jobs for each well in each plate are distributed across worker nodes. There, barcode sequences are extracted and clustered based on Levenshein distance. R1 reads from each cluster are aligned to the ORF reference and pileups are generated, which are used for variant calling and long indel detection (via recognition of sudden changes in read depth using a modified Sobel filter.)

`FASTQ` files are then further demultiplexed using the well-tags in a highly parallel fashion. This results in a folder structure containing tens of thousands of individual `FASTQ` files sorted by plate and well location. These are then further processed in parallel to identify barcodes. Wells can sometimes contain more than one clone (e.g., due to incomplete washing in the robotic pinning process). Thus barcode sequences are extracted from each read and then clus-

tered by edit distance [97] to determine the set of barcodes in each well. The associated paired reads for each barcodes are then further split by barcode. Each barcode-specific set of ORF reads can then be analyzed with respect to mutations. `Bowtie2` [108] is used to align reads to the ORF template, PCR duplicates are removed and nucleotide variants called using samtools pileup [109]. Given limited read lengths, identification of longer indels is not straightforward. A solution was found by extracting depth of coverage tracks for each clone and normalizing them with respect to average positional coverage across each 384-well plate, applying an edge-detection algorithm [98] to find sudden increases or decreases within normalized coverage, indicating the presence under-covered regions that can arise as a result of insertions or deletions. The source code is provided on the attached storage media and can also be found online[3].

After successful genotyping with KiloSeq, I determined the subset of clones that (i) contained a minimum of one missense mutation, (ii) did not contain any insertions or deletions, (iii) did not contain mutations outside of the ORF, (iii) had unique barcodes, (iv) had sufficient read coverage during KiloSeq to allow for confident genotyping. We re-arrayed this filtered subset of clones (using the Biomatrix BM5-BC robot, S&P Robotics) into a condensed final library of 40 plates containing 6,548 clones. I created a custom software library to automatically program the Biomatrix robot's picking protocol. The software is provided on the attached storage media and can also be found online[4].

2.4.3. DMS-BarSeq

Complementation competition experiment: Complementation experiments were performed by Jennifer Knapp, Song Sun and Marta Verby. Plasmids extracted from the pool of 6,548 barcoded and KiloSeq-validated mutant clones, together with barcoded null and wildtype controls, were transformed into a *S. cerevisiae* strain carrying the temperature-sensitive (ts) *ubc9-2* allele which can be functionally complemented by the corresponding wild-type human gene [13, 79]. Complexity for this transformation was 100,000 CFU. For the time series BarSeq screen, the pools were grown separately at both non-selective (25°C) and selective (38°C) temperatures in triplicates to be examined at 5 different timepoints (0h, 6h, 12h, 24h, 48h) yielding 30 samples. At their respective time points, plates were scraped, OD quantified, and their barcode loci amplified with primers carrying sample-specific tags. The amplified product is then

[3] `http://dalai.mshri.on.ca/~jweile/projects/kiloseq/`
[4] `http://dalai.mshri.on.ca/~jweile/projects/biomatrix/`

sequenced on an Illumina NextSeq 500.

Sequence analysis: I created a custom sequence analysis pipeline, which was used to identify and count individual sample tags and barcode combinations within each read. The pipeline uses a k-mer search algorithm in a highly parallelized fashion on a SunGridEngine HPC cluster. Barcodes are counted and the counts aggregated across cluster nodes. The pipeline source code is provided on the attached storage media and can also be found online[5].

Scoring: I developed a custom software to perform scoring and statistical analysis. First, the relative population size for each clone is calculated by dividing each clone's barcode count by the total number of barcodes in each condition. Then the estimated absolute population size for each clone is calculated by multiplying the relative population size with the estimated total number of cells on the respective plate at the corresponding time point (obtained from OD measurements). I then treat the amount of growth between each individual time point compared to the pool average as an individual estimate of fitness, all of which act cumulatively. This is calculated as follows: Let c^{τ}_{i,t_k} be the barcode count for clone i, time point t_k at temperature τ, then $\forall i \in \{1 \leq i \leq N | i \in \mathbb{N}\}$, $\forall k \in \{1 \leq k \leq 5 | k \in \mathbb{N}\}$, $\forall \tau \in \{25°, 37°\}$

[5] `http://dalai.mshri.on.ca/~jweile/projects/screen_pipeline/`

$$
\begin{aligned}
r_{i,t_k}^{(\tau)} &= \frac{c_{i,t_k}^{(\tau)}}{\sum_j c_{j,t_k}^{(\tau)}} \\
P_{i,t_k}^{(\tau)} &= r_{i,t_k}^{(\tau)} \cdot P_{*,t_k}^{(\tau)} \\
\rho_{i,t_k}^{(\tau)} &= \sqrt[(t_k - t_{k-1})]{\frac{P_{i,t_k}^{(\tau)}}{P_{i,t_{k-1}}^{(\tau)}}} \\
\phi_{i,t_k}^{(\tau)} &= \frac{\rho_{i,t_k}^{(\tau)}}{\rho_{*,t_k}^{(\tau)}} \\
\phi'_{i,t_k} &= \frac{\phi_{i,t_k}^{(37^\circ)}}{\phi_{*,t_k}^{(25^\circ)}} \\
s_i &= \prod_k \phi'_{i,t_k} \\
s'_i &= \frac{s_i - s_{\text{null}}}{s_{\text{wt}} - s_{\text{null}}},
\end{aligned}
$$

where $r_{i,t_k}^{(\tau)}$ is the relative population size for clone i and time point t_k at temperature τ, $P_{i,t_k}^{(\tau)}$ is the absolute population size for clone i, time point t_k at temperature τ, $\rho_{i,t_k}^{(\tau)}$ is the measured hourly growth rate for clone i, time point t_k at temperature τ, $\phi_{i,t_k}^{(\tau)}$ is the fitness advantage relative to the pool growth for clone i, time point t_k at temperature τ, ϕ'_{i,t_k} is the normalized relative fitness advantage for clone i at time point t_k, and s_i is the cumulative normalized relative fitness advantage for clone i. Finally, s'_i is the fitness score relative to the internal null and wild type controls. This results in null-like mutants receiving a score of zero and wild type-like mutants receiving a score of one.

The scoring software is part of a larger DMS analysis package provided on the attached storage media. It is also available online[6].

Error regularization: I regularized the standard error measurements for each clone using a Bayesian method published by Baldi and Long [99]. A prior estimate for each measurement was obtained via linear regression over permis-

[6] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

sive read counts and fitness values. The prior is combined with the empirical standard deviation obtained from technical replication using Baldi and Long's original formula

$$\sigma^2 = \frac{v_n \sigma_n^2}{v_n - 2} = \frac{v_0 \sigma_0^2 + (n-1)s^2}{v_0 + n - 2},$$

where v_0 represents the degrees of freedom assigned to the prior estimate, σ_0 is the prior estimate, n represents the degrees of freedom for the empirical data (i.e. the number of replicates) and s is the empirical standard deviation.

The error regularization procedure is part of a larger DMS analysis package provided on the attached storage media. It is also available online[7].

2.4.4. DMS-TileSeq

Complementation competition experiment The TileSEQ experiment was performed by Song Sun and Marta Verby. Plasmids extracted from a pool of $\sim 10^5$ PopCode-generated clones were transformed into the *S. cerevisiae ubc9-2* ts strain yielding around 10^6 total transformants. Plasmids were prepared from two replicates of each 10 ODU of cells and used as templates for the downstream tiling PCR. These serve as the two replicates in the non-selective condition. A further two replicates of 40 ODU of cells were inoculated into 200ml medium and grown under continuous shaking to full density at 36°C. Plasmids were extracted from 10 ODU of each culture and were used as templates for the downstream tiling PCR. These serve as the two replicates in the selective condition. Finally, plasmid expressing the wild-type ORF was transformed into the *S. cerevisiae ubc9-2* ts strain and grown to full density under selection. Plasmids were extracted from two replicates of 10 ODU of cells and used as templates for the downstream tiling PCR. These serve as the two replicates of wild-type control. For each plasmid library, a tiling PCR was performed in two steps: (i) the targeted region of the ORF was amplified with primers carrying a binding site for Illumina sequencing adaptors, (ii) each amplicon was indexed with an Illumina sequencing adaptor. Finally, paired end sequencing is performed on the tiled regions across the ORF using an Illumina NextSeq 500.

TileSeq Analysis pipeline: Sequencing data is demultiplexed using Illumina `bcl2fastq`. Reads are the aligned to the UBE2I template using Bowtie2 [108]

[7] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

and variants called where both reads in each pair agree. Variants are counted and aggregated for each condition and replicate. Counts in each condition are normalized to sequencing depth at the respective position. Then, wildtype control counts are subtracted from the selective and permissive condition counts. Finally, the log ratio between adjusted selective and permissive counts is calculated. Error regularization was performed the same way as in DMS-BarSeq using the Baldi and Long method [99]. The scoring procedure is implemented as part of a larger DMS analysis package provided on the attached storage media. It is also available online[8].

2.4.5. Joining of maps, imputation and regularization

While DMS-TileSeq produces only one fitness score per variant, DMS-BarSeq in many cases contains multiple biological replicates of the same variant associated with different barcodes. To provide summary fitness values on a per-variant basis, I combined scores from biological replicates using weighted means, where the weight is inversely proportional to the Bayesian regularized standard error. The standard error associated with the joint score is also adjusted to account for differences in input fitness measurements and increased sample size.

The results from the barcoded and regional sequencing screens do not scale linearly with each other. I used regression to find a monotonic transformation function

$$f(x) = a \cdot e^x + b \cdot x + c$$

between the two screens' respective scales. The standard deviation is transformed accordingly using a Taylor series-based approximation.

$$\sigma' = \sigma \cdot (a \cdot e^{\mu} + b)$$

After both datasets have been brought to the same scale I can join corresponding data points using weighted means, where the weight is again inversely proportional to the Bayesian regularized standard error. Output standard error was adjusted again to account for differences in input fitness values and

[8] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

increased sample size.

$$w_0 = \frac{1}{1 + \frac{\sigma_{\bar{x}}^{(0)}}{\sigma_{\bar{x}}^{(1)}}};\ w_1 = \frac{1}{1 + \frac{\sigma_{\bar{x}}^{(1)}}{\sigma_{\bar{x}}^{(0)}}}$$

$$\mu_{\text{joint}} = w_0 \cdot \mu_0 + w_1 \cdot \mu_1$$

$$\sigma_{\text{joint}}^2 = w_0 \cdot (\sigma_0^2 + \mu_0^2) + w_1 \cdot (\sigma_1^2 + \mu_1^2) - \mu_{\text{joint}}^2$$

$$\sigma_{\bar{x}}^{(\text{joint})} = \frac{\sigma_{\text{joint}}}{\sqrt{df_0 + df_1}}$$

where μ_0 is the DMS-BarSeq value, σ_0 the associated standard deviation, $\sigma_{\bar{x}}^{(0)}$ the associated standard error, df_0 the associated degrees of freedom, μ_1 is the DMS-TileSeq value, σ_1 the associated standard deviation, $\sigma_{\bar{x}}^{(1)}$ the associated standard error, and df_1 the associated degrees of freedom.

Imputation of missing values was performed using `RandomForest` Regression [100]. The following intrinsic features were generated: the confidence-weighted average fitness across mutations at the same position; the average fitness of multi-mutant clones that contain the mutation of interest; and the estimated fitness according to a multiplicative model to infer mutant fitness A using a double mutant AB and single mutant B. A second set of features was computed from differences between various chemical properties of the wild-type and mutant amino acids. These properties include size, volume, polarity, charge, and hydropathy. A third set of features is derived from the structural context of each amino acid position. These include secondary structure, solvent accessibility, burial in interfaces with different interaction partners, and involvement in hydrogen bonds or salt bridges with interaction partners. Secondary structures were calculated using `Stride` [110]. Solvent accessibility and interface burial were calculated using the `GETAREA` tool [111] on the following PDB entries: `3UIP` [85]; `4W5V` [84]; `3KYD` [83]; `2UYZ` [112]; `4Y1L` [90]. Hydrogen bonds and salt bridges candidates were predicted using `OpenPyMol` [113] and evaluated for validity by manual inspection. Additional features used are the PROVEAN [12] and BLOSUM [103] scores for a given amino acid change and the evolutionary conservation of the amino acid position. Conservation was obtained by generating a multiple alignment of direct functional orthologues across many eukaryotic species using `CLUSTAL` [114], which was used as input for `AMAS` [104].

The machine learning predictions generated above were also used to regularize experimental measurements of lower confidence. To this end, the corrected

standard error associated with each data point can be used to determine the weight assigned to the measurement, as follows:

$$
\begin{aligned}
w_0 = \frac{1}{1 + \frac{\sigma_{\bar{x}}^{(0)}}{\sigma_{\bar{x}}^{(1)}}};\ w_1 &= \frac{1}{1 + \frac{\sigma_{\bar{x}}^{(1)}}{\sigma_{\bar{x}}^{(0)}}} \\
\mu_{\text{joint}} &= w_0 \cdot \mu_0 + w_1 \cdot \mu_1 \\
\sigma_{\text{joint}}^2 &= w_0 \cdot (\sigma_0^2 + \mu_0^2) + w_1 \cdot (\sigma_1^2 + \mu_1^2) - \mu_{\text{joint}}^2 \\
\sigma_{\bar{x}}^{(\text{joint})} &= \frac{\sigma_{\text{joint}}}{\sqrt{df_0 + df_1}}
\end{aligned}
$$

where μ_0 is the measured value, σ_0 the associated standard deviation, $\sigma_{\bar{x}}^{(0)}$ the associated standard error, df_0 the associated degrees of freedom, μ_1 is the RandomForest predicted value, σ_1 the associated standard deviation as approximated by cross-validation RMSD, $\sigma_{\bar{x}}^{(1)}$ the associated standard error, and df_1 the associated virtual degrees of freedom.

The joining, imputation, and regularization procedures are implemented as part of a larger DMS analysis package provided on the attached storage media, and also available online[9].

2.4.6. Complementation spotting assays

To validate the reliability of the fitness scores obtained during the screen, I selected three subsets of clones from our original UBE2I variant library: (1) A set of clones carrying variants with functional scores representing the full spectrum in the screen; (2) A set of clones carrying hypercomplementing variants in the screen; and (3) A set of clones carrying variants not present in the imputation training data set. Jennifer Knapp and I performed genotype verification using Sanger sequencing. The spotting assay was then performed by Jennifer Knapp as follows. Each verified variant was transferred to the yeast expression plasmid pHYCDest by Gateway cloning and individually transformed into the *S. cerevisiae* ubc9-2 ts strain. Cells were grown to saturation in 96-well cell culture plates at room temperature. Each culture was then adjusted to an OD600 of 1.0 and serially diluted to $5^{-1}, 5^{-2}, 5^{-3}, 5^{-4}$, and 5^{-5}. These cultures (5µl of each) were then spotted on SC-Leucine plates as appropriate to maintain

[9] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

the plasmid and incubated at either the permissive (25°C) or non-permissive (37°C) temperatures for two days. Each variant was assayed alongside negative and positive controls for loss of complementation (expression of either the wild type human protein or a GFP control). Results were interpreted by comparing the growth difference between the yeast strains expressing human genes and the corresponding control strain expressing the GFP gene.

I developed a custom software, PlateOrganizer, to organize and analyze image data from spotting assays. It is provided on the attached storage media and can also be found online[10].

[10] `http://dalai.mshri.on.ca/~jweile/projects/PlateOrganizer/`

3. Expanding the atlas of variant effects in human disease genes

As in the previous chapter, the work described here is the result of a team effort including multiple members of the Roth Lab as well as other collaborators. Wet lab procedures were executed by Marta Verby, Song Sun, Atina Coté and Jennifer Knapp, while computational aspects of the work were developed and implemented by myself, except where indicated otherwise.

3.1. Introduction

Within coming decades, millions of people will have their genome sequenced. Unfortunately, we have limited ability to interpret personal genomes, each carrying 100-400 rare missense variants [9] of which many must currently be classified as Variants of Uncertain Significance (VUS). For example, gene panel sequencing aimed at identifying germline cancer risk variants in families yielded VUS for the majority of missense variants [8]. While functional variants can be predicted via computational tools such as PolyPhen-2 [10] and PROVEAN [12], these methods can confidently detect only one third as many disease variants as are detectable by experimental assays [13]. Unfortunately, experimental assays are either unavailable or economically inviable for most human disease genes.

Recent DMS studies have provided individual maps for the critical RING domain of BRCA1 [57] associated with breast cancer risk, and the PPARγ protein associated with Mendelian lipodystrophy and increased risk of type 2 diabetes [61]. Such maps can not only identify functionality of a clinical variant accurately, but also potentially do so in advance of that variant's first clinical presentation.

In the previous chapter, a framework for comprehensive high-quality screening of functional effects across all possible missense mutations in human genes was established. The functional complementation assay used in the assay allows for the generation of maps that not only represent the overall functional con-

sequences of mutations, but also serves as a common basis to make maps more directly comparable. In addition, the statistical analysis and machine learning component introduced allows for high overall map quality and completeness. Using this framework a complete functional map for the SUMO E2 conjugase *UBE2I* was created. Here I describe the creation of a map of a second member of the Sumoylation pathway, *SUMO1*. I examine both map in detail before discussing the interpretation of yeast complementation phenotypes in terms of humans.

To demonstrate the value of the DMS framework in terms of clinical interpretation of variants, a diverse set of six new disease gene maps was added to the atlas: *TPK1* encoding Thiamin Pyrophosphokinase 1, *NCS1* encoding Neuronal Calcium Sensor 1, as well as the paralogues *CALM1*, *CALM2* and *CALM3*, which each encode the protein Calmodulin. The maps are evaluated in terms of pathogenicity prediction and VUS reclassification.

3.2. Results

3.2.1. A functional map of SUMO E2 recapitulates known biology and poses new questions

The DMS map of *UBE2I* produced in the previous chapter paints a comprehensive picture of variant effects on protein function. The complete, refined functional map of UBE2I after imputation and regularization can be seen in Figure 3.1. For comparison, additional tracks showing position-specific evolutionary conservation, secondary structure, relative solvent accessibility and burial in protein-protein interaction interfaces are also shown. Based on the map, several observations can be made. Consistent with the results of smaller-scale biochemical studies of the SUMO E2 conjugase [87, 94], the areas most sensitive to mutation are those proximal to the active site (particularly residues 81-88, 90, 92-96, and 127-130), and the N-terminal α-helix which mediates four protein interactions including the critical interaction with the E1 SUMO-activating complex. Within the active site, particularly strong sensitivity to mutation can be observed at the cysteine residue at position 93. This is consistent with its central role in E2 function, as it forms a thioester bond with the SUMO C-terminus [87].

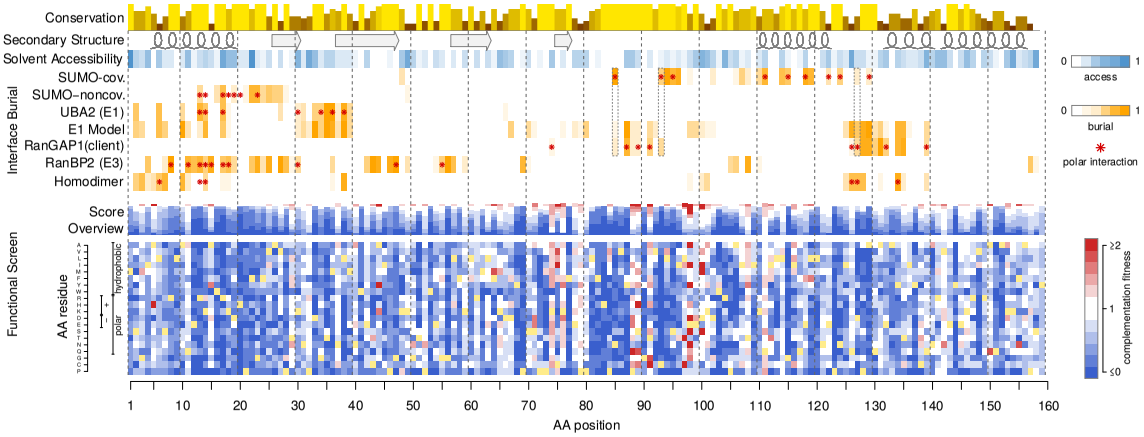

Figure 3.1.: Functional map of UBE2I. From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in protein-protein interaction interfaces with covalently bound SUMO, non-covalently bound SUMO, the SUMO E1 complex at two different stages of activation, the sumoylation substrate RanGAP1, the E3 RanBP2, and the UBE2I homodimer; A summary track showing the relative number of amino acid changes resulting in different fitness effects; and finally the individual amino acid change effects sorted by physicochemical groups.

An interesting feature of the map is the alternating tendency towards damaging and benign substitutions across positions 55-65. A comparison with solvent accessibility reveals this to be caused by alternating externally and internally-oriented residues, with the latter positions constrained to be hydrophobic. This alternating tendency is also reflected in evolutionary conservation across these positions.

All protein-protein interaction interfaces previously captured in co-crystal structures show increased sensitivity to mutation when compared to other surface residues (Figure 3.2). When comparing individual protein interaction interfaces, the most substantial fitness defects are observed in those for the E1 activating complex binding interface and the covalent and non-covalent SUMO binding interfaces (Figure 3.2A). While the homodimerization interface also shows significant sensitivity (Wilcoxon $P = 6.87 \cdot 10^{-21}$), the effects are not as severe as those at the E1 interface (Wilcoxon $P = 4.28 \cdot 10^{-8}$) (Figure 3.2B). This is consistent with the Alontaga and colleagues' hypothesis regarding its involvement in SUMO chaining [90], as in yeast SUMO chain formation has so far only been observed to be involved in meiosis [93], which is not a mechanism vital to fitness in a complementation assay. Alontaga *et al.* also postulate however, that non-covalent SUMO binding is necessary for SUMO chain formation. In contrast to the homodimerization interface, the non-covalent SUMO binding interface shows a much stronger sensitivity to mutation (Wilcoxon $P = 4.73 \cdot 10^{-8}$). This may be due to two different reasons: (i) there is a 27% overlap between the interface for non-covalent SUMO binding interface and the interface for E1-E2 binding, which is among the most sensitive surfaces of UBE2I; and (ii) non-covalent SUMO binding also plays an important role as an adapter for many E3 proteins [115].

Another interesting observation can be made with respect to a known phosphorylation site on the surface of UBE2I. Su and colleagues previously discovered that phosphorylation of Serine 71 via the Cyclin-dependent Kinase CDK1 results in sumoylation hyperactivity [116]. The map shows that substitutions with phosphomimetic residues at this position lead to hyperactive complementation, consistent with Su *et al.*'s observations. Furthermore, other residues amenable to phosphorylation are also tolerated, while hydrophobic replacements are generally deleterious (Figure 3.3).

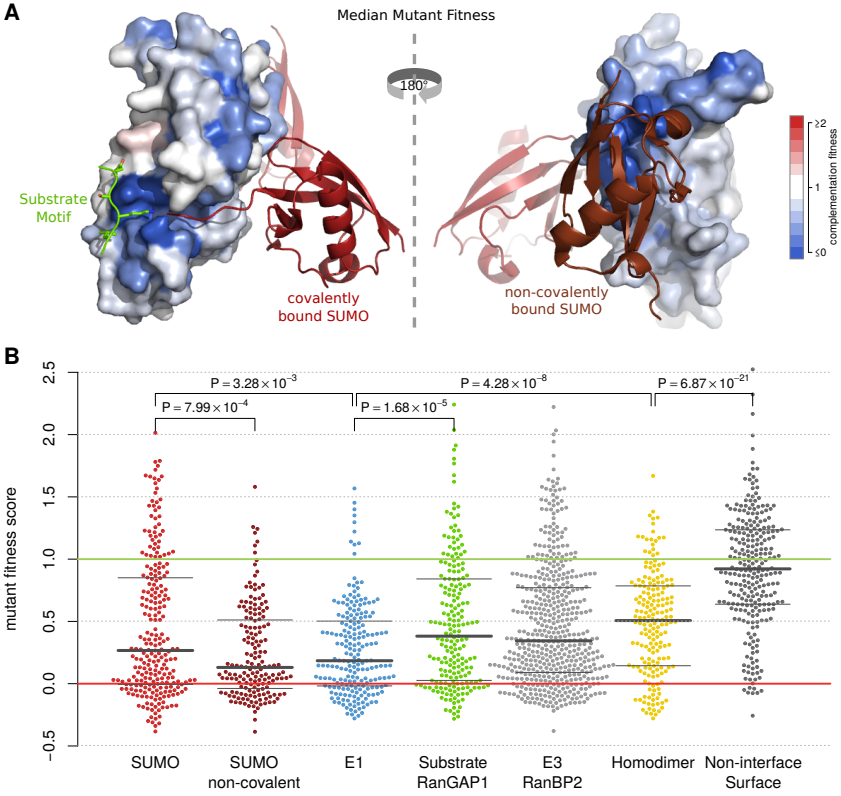

Figure 3.2.: Complementation fitness of mutations at interaction interfaces. A) Median mutant fitness mapped to the crystal structure of UBE2I. The ΨKxE substrate recognition motif is shown as green stick model, covalently and non-covalently bound SUMO are shown as crimson and brown cartoon model, respectively. B) Mutant fitness scores distributions for residues at different interaction interfaces (and non-interface surfaces as control). P-values from one-sided Wilcoxon tests. Bold bars represent medians, thin bars indicate 25% and 75% quartiles.

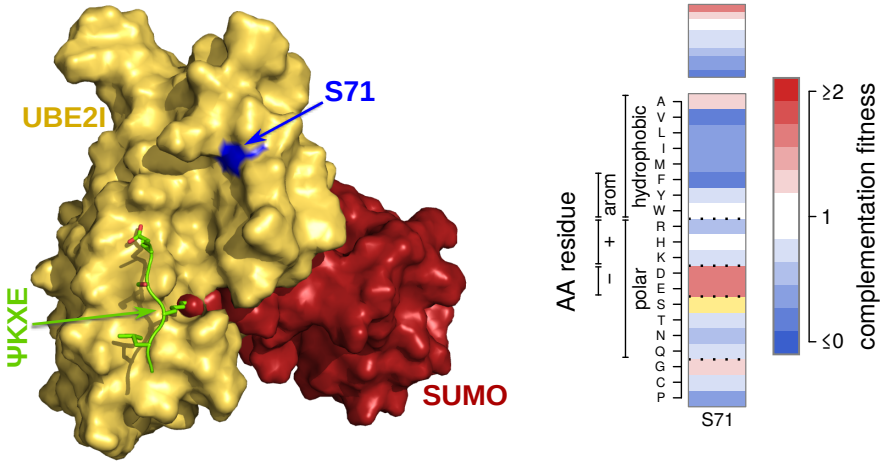

Figure 3.3.: Phosphorylation site of UBE2I shows hyperactive complementation when mutated to phosphomimetic residues.

Substrate specificity shifts and E2 hyperactivity

Intriguingly, many sites show fitness that is better than wildtype (e.g., positions 74, 76, 88, 89, 91 and 98). Manual functional complementation spotting assays confirmed that complementation with these mutants allows greater growth than does the wild type human protein, but resemble more closely the growth at the permissive temperature for the *ubc9-ts* strain (Figure 3.4A). One might be tempted to interpret these cases as reversions to residues present in the yeast protein. However, a comparison of fitness score distributions between changes to *S. cerevisiae* residues and those occurring in the distant species *Dictyostelium discoideum* (amoeba) or *Drosophila melanogaster* (fly) showed no significant difference (Figure 3.4B). Recognizing that in this assay, human UBE2I must function with the yeast versions of other sumoylation pathway members, it stands to reason that some substitutions could be adaptive by improving compatibility with yeast interaction partners. A comparison with co-crystal structure data [85] shows that many of the hypercomplementing residues are located on the surface facing the general direction of the substrate, with some being in direct contact with the substrate's sumoylation mo-

tif (Figure 3.4C). This suggests a possible adaptation via improved recognition of substrates for which sumoylation is most important for yeast growth. Indeed, *in vitro* sumoylation assays performed previously for a small number of UBE2I mutants revealed increased sumoylation for some substrates [87]. Comparing the map with these sumoylation assay results, I observed many cases of substrate specificity shift (Figure 3.4D). Of the three cases tested in the sumoylation assay that showed hyperactive behaviour in the map (E98A, T91A and K74A), one displayed hyperactive sumoylation of P53, while two displayed decreased sumoylation of P53 and IκBα. However, similar behaviour was seen for 3 other variants (D100, P128A and S89A), which scored as wildtype-like in the map. Most cases for which P53 saw wild-type level sumoylation showed either wildtype-like or slightly below complementation in the map. Finally, the four cases that disrupted sumoylation of all substrates were strongly deleterious in the map. In conclusion, variants that either positively or negatively or negatively affect P53 sumoylation levels *in vitro* appear show either wildtype-like or hyperactive complementation in yeast. This may indicate that differential sumoylation of one or more yeast proteins with a P53-like interface positively affects yeast growth.

As Figure 3.4D shows, substrate specificity does not paint a complete picture of the mechanisms potentially underlying hyperactive complementation. A particularly interesting exception can be observed at residues A15 and T108. Both residues harbor hyperactive mutations but do not face towards the substrate. Instead, they form part of the interface with the E3 SUMO ligase RanBP2, and flank a small cavity on UBE2I's surface into which RanBP2 inserts a phenylalanine residue upon binding [85]. Changing either A15 or T108 into aromatic residues results in a large fitness increase (Figure 3.5). This may be the result from the emergence of a π-stack interaction that strengthens E2-E3 binding.

It is unclear how to interpret the effect of mutations that enhance growth in the yeast complementation assay. If fitness measured in the assay is directly proportional to fitness in the real biological context, then these enhancing mutations would be beneficial. However one can also imagine an alternative scenario in which activity-enhancing mutations are deleterious in the real biological context. To objectively distinguish between these possibilities, we collaborated with Jesse Bloom to employ a method he recently published that leverages likelihood-based phylogenetics to quantitatively compare how well different experimental measurements represent actual evolutionary constraints in nature [52, 117]. We compared three models relating the experimental fitness to the evolutionary preference for a mutated amino-acid sequence: (a) the

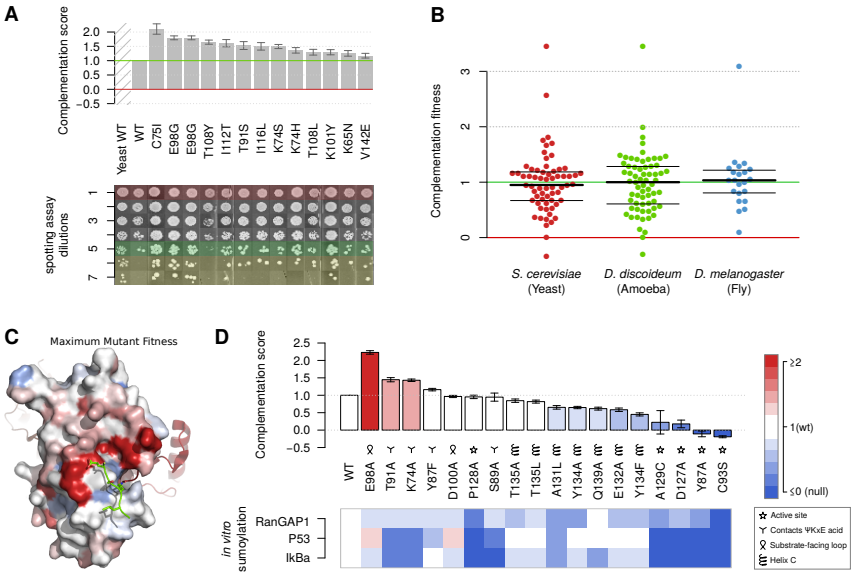

Figure 3.4.: Hyperactive complementation in UBE2I. A) Variants scoring higher than the wildtype controls show stronger growth in manual complementation spotting assays and resemble the WT yeast. B) Distribution of scores for changes to residues naturally occurring in yeast, amoeba and fly are not significantly different from each other. C) Maximum mutant score mapped to amino acid positions on UBE2I structure. Hyperactive mutations are clustered at the substrate recognition site. Structure data from `PDB:3UIP` [85] D) *In vitro* sumoylation assay data from Bernier-Villamor *et al.* [87] in comparison to the complementation fitness scores.

evolutionary preference was directly proportional to the untransformed experimental fitness; (b) the preference had a ceiling at the wildtype experimental fitness (values greater than 1 were set to 1); or (c) the preference was set to the reciprocal of fitness for mutations with greater-than-wildtype scores, corresponding to a deleterious effect of enhancing mutations. Dr. Bloom kindly

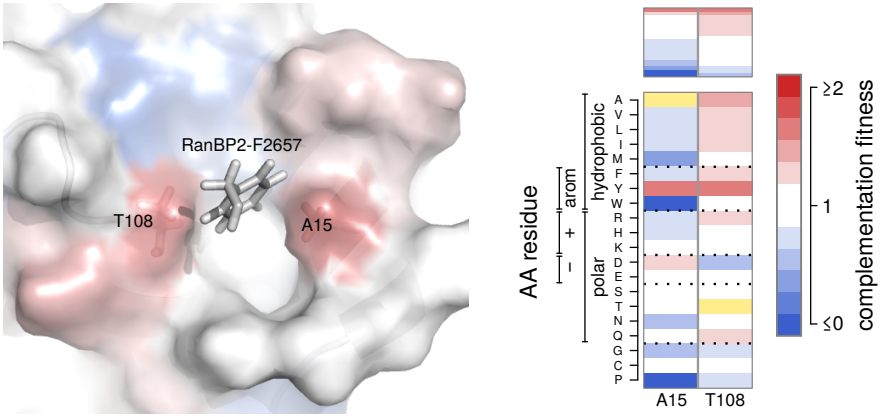

Figure 3.5.: Potential de-novo pi-stack interaction between UBE2I and the E3 RanBP2. Structure data from `PDB:3UIP` [85]

provided the phydms software [117] to test which of these three approaches best described the evolutionary constraint on a set of naturally occurring *UBE2I* homologs. The analysis was performed using fitness scores that excluded conservation features from the regularization process, to avoid the circularity of using natural sequence data when deriving the scores. As shown in Table 3.1, the best fit is achieved using the model that assumes that enhancing mutations are deleterious. This result provides objective support for the idea that mutations that enhance activity above wildtype levels in the complementation assay are actually deleterious in a real biological context.

Based on these observations I reinterpreted cases of hyperactive complementation in the map as deleterious. I repeated the imputation and regularization procedure on the transformed map, which resulted in substantially improved cross-validation performance (Root-Mean-Squared-Deviation, RMSD, decreased from 0.33 to 0.24).

Intragenic epistasis and compensatory mutations

Full-length *UBE2I* clones generated for DMS-BarSeq analysis often encoded more than one amino acid change. Multi-mutant clones offer the opportunity

Table 3.1.: Comparison of different models for the effects of hyperactivating mutations. AIC: Akaike Information Criterion

Model	ΔAIC relative to best model
Hyperactive mutations as deleterious	0
Hyperactive mutations as WT	27.7
Hyperactive mutations as beneficial	60.6

to search for intragenic genetic interactions. Genetic interaction is defined as the case of a combination of mutations that yields an unexpected phenotypic effect. Therefore, identifying genetic interactions requires modeling the phenotype that is expected from a combination of mutations, given the single-mutant effects. Here I used a previously-described multiplicative model [101, 102] in which genetic interaction is measured as $\varepsilon_{ij} = f_i \cdot f_j - f_{ij}$, where f_i and f_j represent single mutant fitness and f_{ij} represents double mutant fitness scores. Most double mutants (71%) did not show a significant deviation from $\varepsilon_{ij} = 0$ under this model, while 328 position pairs did show significant genetic interaction (Figure 3.6).

Of particular interest are compensatory interactions, i.e. cases where a double mutation is more fit than either of the component single mutations. Where compensatory residues are proximal in the protein structure, the combination of two mutant residues may be able to re-establish a physical interaction that was lost in each of the single mutants. Although the majority of genetically interacting sites were not proximal in the structure (Figure 3.6B), there were interesting exceptions. For example, the I4T-P69S double mutant appears to exhibit compensatory behaviour: In the wild type structure, the van-der-Waals radii of the two residues are in direct contact (Figure 3.6C). Either mutation alone would be expected to destabilize the hydrophobic interaction between isoleucine and proline. However, in the double mutant, hydroxyl groups on the two residues could adopt a hydrogen bond that re-establishes interaction and re-stabilizes the fold (Figure 3.6D).

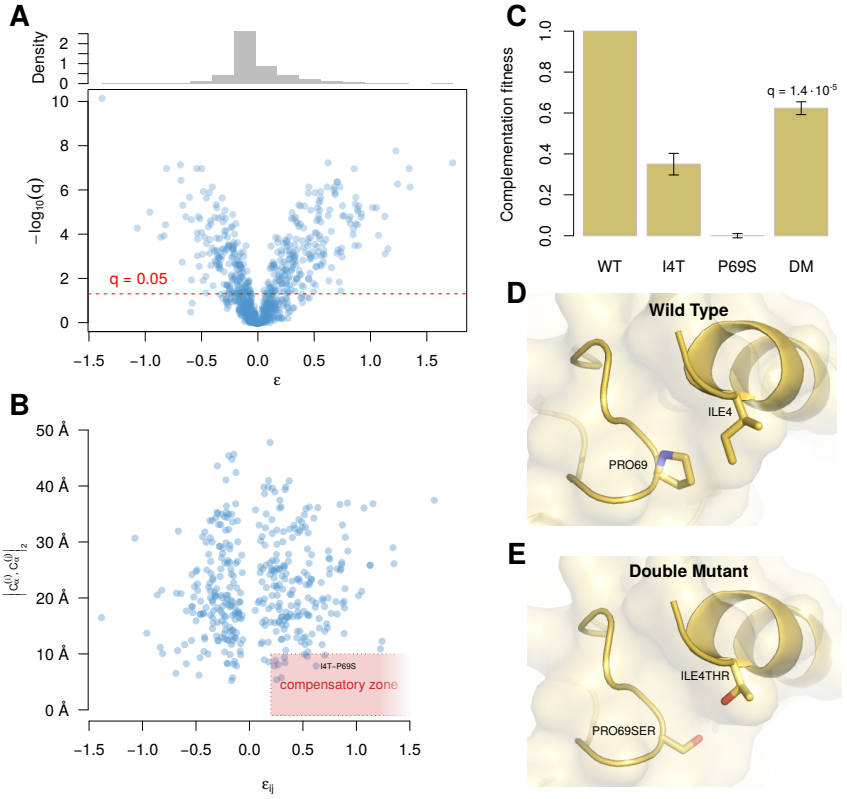

Figure 3.6.: Intragenic epistasis in UBE2I. A) Volcano plot of epsilon values (deviation from expectation) vs FDR-corrected significance levels. Above: Histogram of epsilon values. B) Comparison of epsilon values with inter-alpha-carbon distance in the 3D structure. C) Positive genetic interaction between I4T and P69S. The double mutant fitness (DM) significantly exceeds that of both single mutants and their product. D) Wiltype structural context for I4 and P69. E) Simulated double mutant structural context.

3.2.2. A comparison of complementation and Y2H reveals a interaction interface

An important factor behind the choice of UBE2I as a testing ground for the DMS framework was the mechanistic complexity of the Sumoylation pathway, in which the central component UBE2I engages in many different protein-protein interactions. Having examined the relative importance of its known interaction interfaces we wished to evaluate the possibility of detecting new interfaces. To this end, we adapted the DMS framework to use a Y2H assay in the selection step.

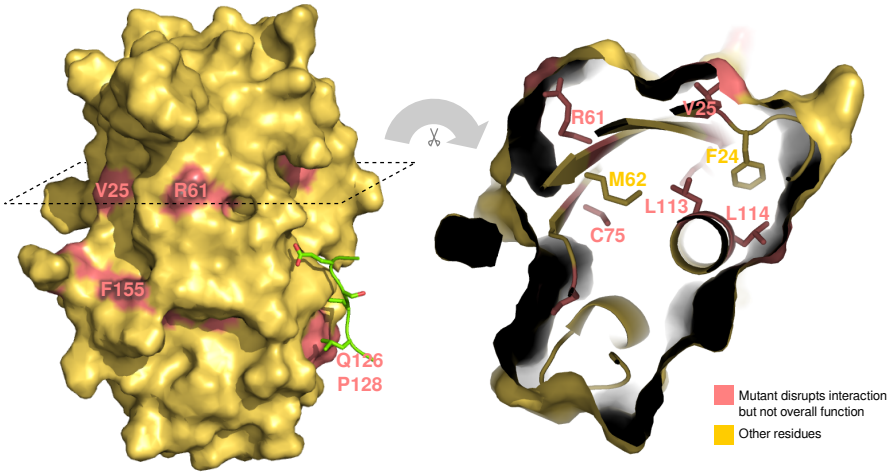

Figure 3.7.: Potential interfacial residues for UBE2I's interaction with SATB1. Highlighted residues disrupt Y2H interaction without disrupting overall function as measured by complementation. The dotted frame in the left panel indicates the plane across which the structure was cut to produce the panel on the right.

We explored the set of previously identified Y2H interactions of UBE2I and found its interaction with the Special AT-rich sequence Binding protein SATB1, a sumoylation target [118], to be the strongest interaction signal. We used DMS-BarSeq to map the effects of UBE2I variants on the UBE2I-SATB1 in-

teraction and compared the results to those of the complementation assay. Although too few variants in the Y2H screen were measured with high enough confidence to perform reliable imputation, I was able to identify 15 variants that specifically disrupted the UBE2I-SATB1 binding without affecting its overall function as measured by the complementation assay. Interestingly, three amino acid positions (V25, C75 and F155) were represented with multiple variants in this list, highlighting their importance. Figure 3.7 marks the affected residues on the surface of UBE2I, which may determine the specificity of the UBE2I-SATB1 interaction. Consistent with SATB1's known role as a sumoylation target [118], the residues are clustered near the known substrate recognition and binding surface. Intriguingly, I also found a number of residues within UBE2I's hydrophobic core, that upon mutation to alternative hydrophobic residues resulted in a disruption of UBE2I-SATB1 binding (Figure 3.7). The fact that these residues are physically close to the locations of surface residues with similar behaviour may indicate that mutations at these positions could result in subtle shifts of UBE2I's fold that disrupt the SATB1 binding interface without affecting other functions.

3.2.3. A functional map for SUMO1

Using the DMS-TileSeq version of the framework established in the previous chapter we also created a complete functional map for SUMO1 (Figure 3.8A). Out of the 1919 possible amino acid changes, fitness effects for 1700 (89%) were measured directly in the complementation competition experiment. The remaining 11% were obtained through imputation, which achieved a cross-validation RMSD of 0.25, a performance very similar to that of the UBE2I map.

The most immediately apparent feature of the SUMO1 map was the strong enrichment for neutral substitutions within the first 20 amino acid positions, which is consistent both with the low level of evolutionary conservation for this region and its annotation as a disordered region. The last four amino acid positions appeared similarly insensitive to mutation, consistent with the cleavage of this region by SENP proteases during SUMO maturation. By contrast, other residue positions were strongly sensitive to mutation, including many inward-facing residues that are apparently constrained to be hydrophobic. As expected, the C-terminal diglycine, directly preceding the last four cleaved residues, is also very sensitive to mutation, as it is required for the covalent binding of SUMO to the E1, the E2 and to the sumoylation target protein.

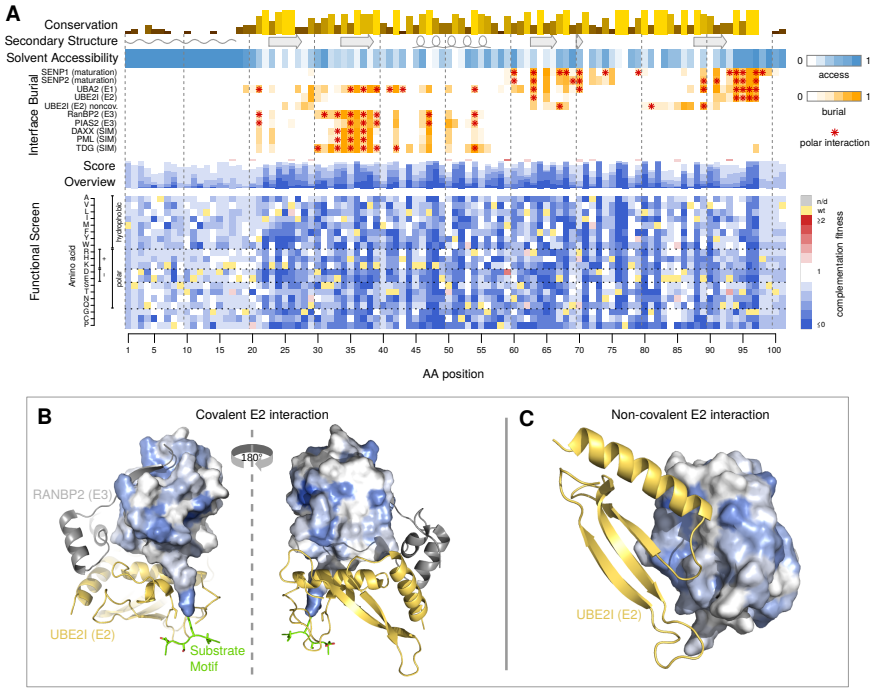

Figure 3.8.: Functional map of SUMO1. A) From top to bottom: Position-wise evolutionary conservation; Secondary structure; Relative solvent accessibility; Relative burial in interaction interfaces with SENP1, SNEP2, UBA2, UBE2I (covalent binding), UBE2I (noncovalent binding), RanBP2, PIAS2, DAXX, PML, and TDG; breakdown of relative amounts of variants with different complementation scores; Heatmap of individual amino acid change effects on complementation fitness. B) SUMO1 structure coloured according to median complementation fitness score. Partial structure of covalently bound UBE2I shown in golden cartoon representation, ΨKxE substrate recognition motif shown in green stick representation C) Colors as in B, but partial structure of UBE2I is shown in noncovalent binding mode.

Interestingly, except for the C-terminal diglycine, the residues that directly touch the E2 during covalent binding are not as sensitive (Figure 3.8B). This may be due to SUMO being force-fed to the E2 by the E1 activating complex and the thioester bond it forms with the E2's cysteine 93 being sufficient to maintain the complex. By contrast, residues in the interface for non-covalent E2 binding are much more sensitive (Figure 3.8C), especially leucine 80 and methionine 82.

Other strongly constrained residues are core members of interaction interfaces. These include the central phenylalanine 36 in the SUMO recognition motif (SRM) interface; glycine 68, which forms the apex of a tight turn within the interface with de-sumoylation enzymes, as well as the E1 and E2 proteins; and leucine 80, which is part of the interface with non-covalently bound E2.

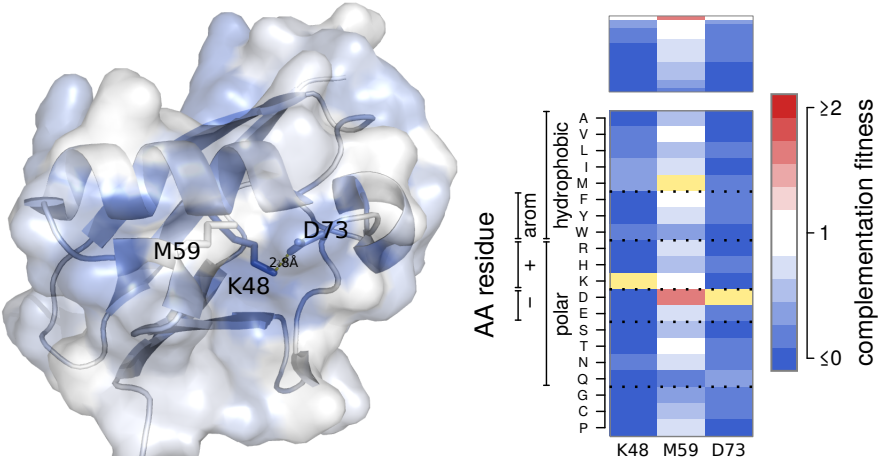

Figure 3.9.: A salt bridge within SUMO1 between Asp73 and Lys38 appears important for stability. Met59Asp may increase stability even further.

The proximity and orientation of aspartate 73 and lysine 48 suggests that they are able to form a salt bridge with one another. The importance of each residue according to the DMS map supports a model in which this salt bridge is important for SUMO folding and/or stability. Interestingly, substituting aspartate for methionine 59, which points towards lysine 48 from an angle similar to that of aspartate 73, enhances the complementation fitness of SUMO1

Table 3.2.: Map quality comparison. RMSD: Root-Mean-Squared-Deviation in 10× cross validation. max($\sigma_{\bar{x}}$): maximal standard error across non-imputed values in the map.

Gene	Possible AA changes	Achieved AA changes	Imputation RMSD	Experimental max($\sigma_{\bar{x}}$)	Regularized max($\sigma_{\bar{x}}$)
UBE2I	3021	2563 (85%)	0.24	0.36	0.25
SUMO1	1919	1700 (89%)	0.25	0.19	0.17
TPK1	4617	3181 (69%)	0.34	0.49	0.37
CALM1	2831	1813 (64%)	0.29	0.28	0.22
NCS1	3610	2542 (70%)	0.63	1.84	0.97

beyond wild type levels. This further underlines the potential importance of a polar interaction involving lysine 48 (Figure 3.9).

3.2.4. Functional maps of three human disease genes

Having established and evaluated the Deep Mutational Scanning framework on two members of the sumoylation pathway, we aimed to create maps for a diverse set of genes that have been associated with disease with varying degrees of confidence. While heterozygous null mutations in SUMO1 have previously been associated with cleft palate [119], we wished to create maps that could be tested in the context of variant classification in terms of disease. Based on the availability of robust complementation assays, we applied DMS-TileSeq to the following protein targets: Thiamine Pyrophosphokinase 1 (TPK1), associated with vitamin B1 metabolism dysfunction [120]; Neuronal Calcium Sensor 1 (NCS1), which has been implicated in autism based on a single *de novo* mutation [121]; and CALM1, CALM2 and CALM3 associated with the heart conditions long-QT syndrome [122] and catecholaminergic polymorphic ventricular tachycardia [123]. Although the three calmodulin genes differ in nucleotide sequence, each encodes the same polypeptide sequence. Thus, we performed a deep mutational scan only for CALM1, which enabled us to also map missense variant effects in CALM2 and CALM3. In each case, we used the TileSeq approach coupled with complementation to generate a map of missense variant functions.

As was shown above for UBE2I, phylogenetic analysis of SUMO1 similarly showed that variants with ability to complement yeast better than wild-type are likely deleterious in humans. I therefore transformed fitness scores so that such

hypercomplementing mutations are considered to be deleterious (see Methods). The transformed disease gene maps can be seen in Figures 3.10 and 3.12. However, since hypercomplementing substitutions may provide interesting clues about differences between yeast and human cellular contexts, I also provide untransformed versions of each map (see Appendix A).

A thiamine pyrophosphokinase map reflects a recessive phenotype

Thiamine pyrophosphokinase (TPK1) is a protein that forms a dimer to perform its biochemical function. Its substrate, thiamine diphosphate, is bound within two active sites formed by the dimerization interface [124]. That is, each monomer contributes half of the residues making up each of the two active sites. Each monomer in turn is made up of an N-terminal globular domain and a C-terminal β-sandwich domain (Figure 3.11A). The residues most sensitive to mutation in the protein make up the hydrophobic cores of the two domains: L21, V22, W36, G48, Y53, P65, G70, Y83, L108, I122, T124, and G127 for the N-terminal domain; and L161, G168, G199, L200, V227, V229, L236, and W237 for the C-terminal domain (Figure 3.11B).

As might have been expected, mutation-sensitive residues include those closely involved in forming the active sites: D46, G70, D71, D73, D100, and K103 in the N-terminal half of the active site, contacting the diphosphate portion of the substrate (Figure 3.11C). In the C-terminal half of the active site, K203, L209, G212, L214, S216, T217, and N219 show similar sensitivity. Interestingly, the tryptophan residue at position 202 appears to be insensitive to mutation despite its close and extensive contact with the thiamine ligand. By contrast, a neighbouring lysine at position 201 is surprisingly sensitive suggesting potential importance in coordinating the ligand. The remainder of the dimerization interface also features a number of sensitive residues, such as M136, G184, V188, G189 and G211. Finally, residues 1-12, which form a β-strand anchoring the N-terminal domain back to the C-terminal domain were also found to be sensitive.

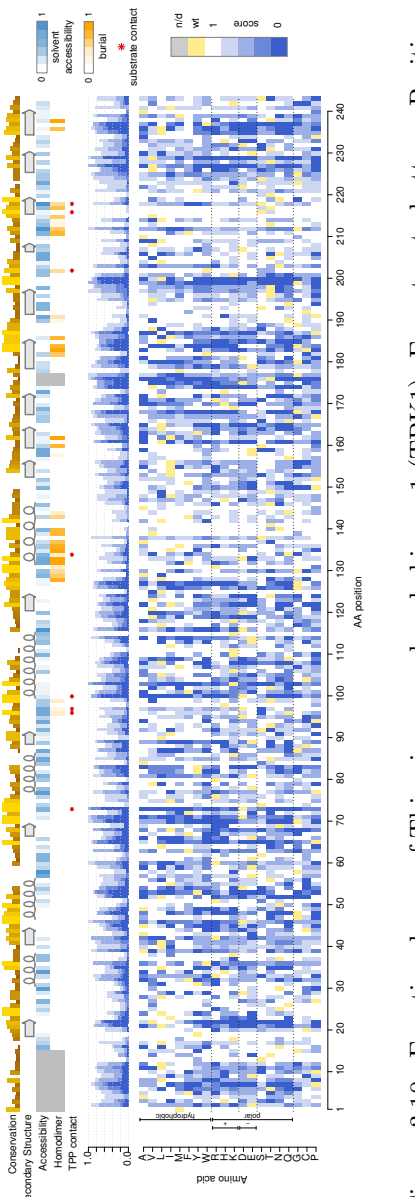

Figure 3.10.: Functional map of Thiamine pyrophosphokinase 1 (TPK1). From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in homodimerization interface; Contacts with Thiamine diphosphate; Summary track showing the shares of amino acid changes resulting in varying degrees of fitness effects; Detailed heatmap showing individual amino acid change effects.

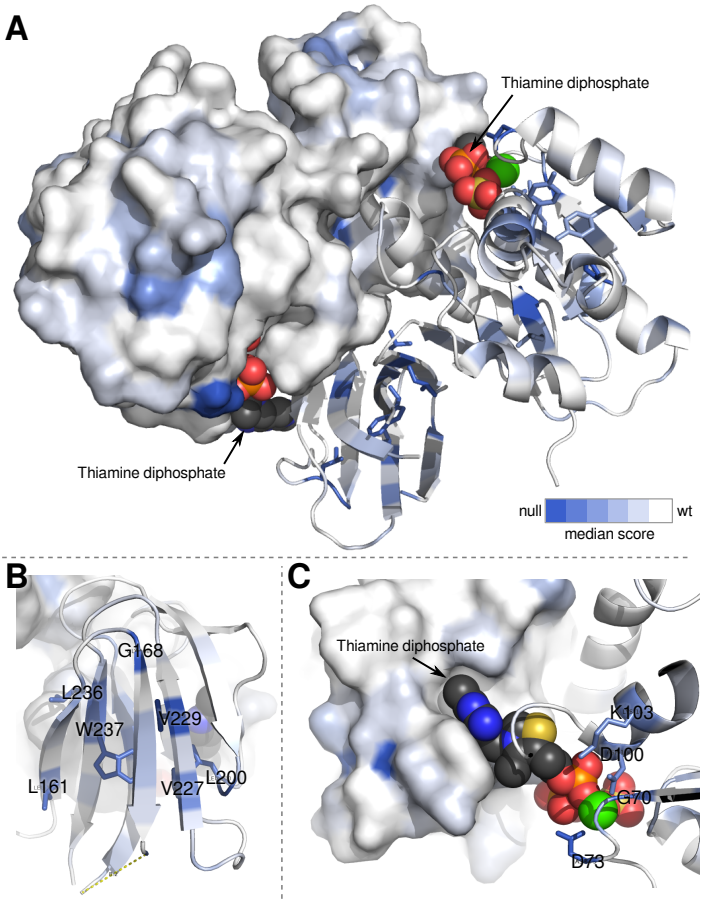

Figure 3.11.: Thiamine pyrophosphokinase 1 coloured by median complementation score. A) TPK1 homodimer structure showing one monomer as surface model, the other monomer as cartoon model. B) Hydrophobic residues facing the inside of the C-terminal β-sandwich domain are sensitive to mutation. C) Active site residues in contact with the substrate are sensitive to mutation. Structure data from `PDB:3S4Y` [124]

The two calcium sensors NCS1 and Calmodulin show different profiles

Calmodulin (CALM1/2/3) and the Neuronal Calcium Sensor protein (NCS1) are homologs (E-value $4 \cdot 10^{-5}$ when searched against the human proteome [125, 126]) with 24% sequence identity and 48.5% similarity [127]. However, they display different impact patterns despite their similar domain structure and similar molecular roles as calcium sensing proteins. Both are comprised of four Calcium-binding EF-hands, with NCS1 containing additional sequences upstream and downstream of the four hands. A comparison of previously published NMR structures reveals that the overall folds of the two proteins differ substantially [128, 129]. In its active (Ca^{2+}-bound) form, Calmodulin features a long central helix that separates two globular domains, called the N-lobe and the C-lobe, each comprised of two EF hands. Two hydrophobic pockets serving as a binding interface for interacting proteins are formed within the lobes. By contrast, NCS1's active form takes a single shell-like shape, centered around a large hydrophobic crevice. This crevice acts as a binding interface for interacting proteins. Thus, the divergent DMS profiles observed for CALM1/2/3 and NCS1 are not surprising given these substantial structural differences.

The Neuronal Calcium Sensor NCS1 displays the greatest sensitivity to mutation within the N-terminal region containing the myristoylation site. This myristoylation site is essential for anchoring NCS1 into the plasma membrane. One other residue that stands out is the tryptophan at position 30, which results in complete loss of function when replaced with any other amino acid. Like most other sensitive residues W30 is found among those contributing to the hydrophobic crevice acting as an interaction interface. Other cases include F55, F56, A104, M121, I152, and A182. An interesting observation can be made with respect to the two helices that separate the two N-terminal EF hands from the two C-terminal EF hands. A kink between the two helices brings them into an angle that allows the globular shape of the overall protein to form. Without this kink it is conceivable that NCS1's fold would much more resemble that of active Calmodulin. A glycine residue (G95) is likely responsible for forming that kink due to its helix breaking properties. This residue is also found to be quite sensitive to mutation.

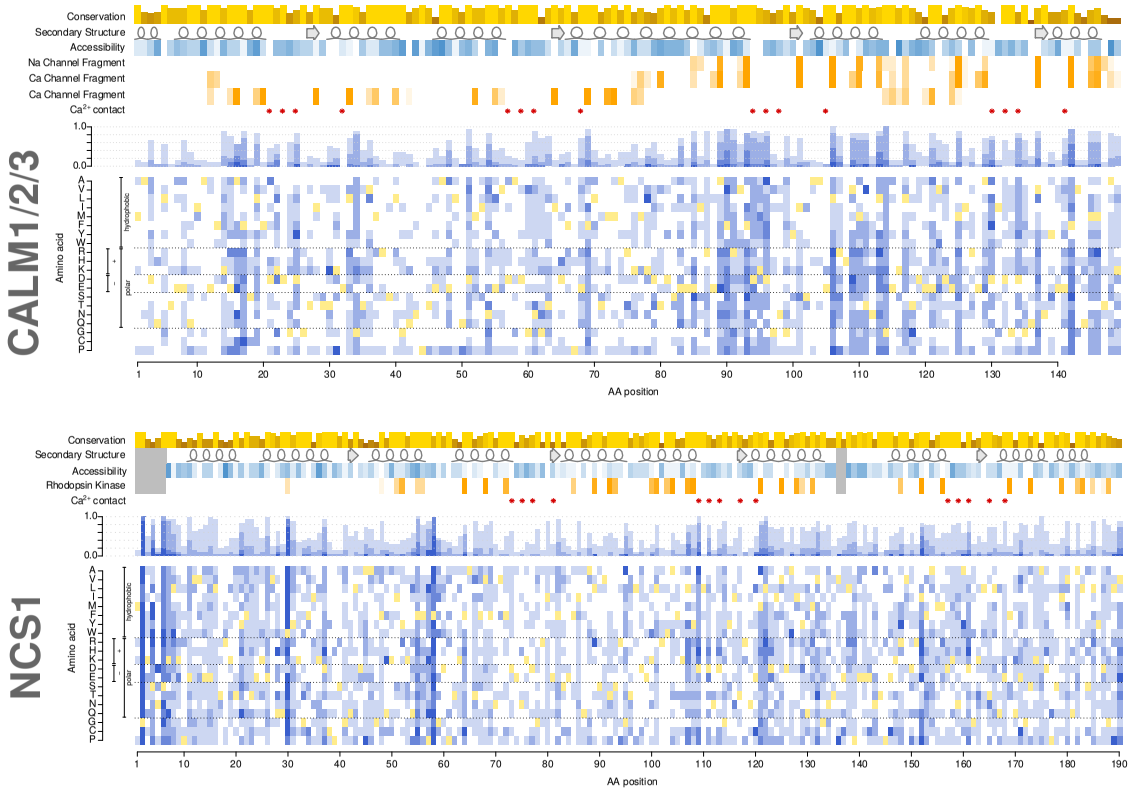

Figure 3.12.: Functional map of Thiamine pyrophosphokinase 1 (TPK1). From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibililty; Relative burial in homodimerization interface; Contacts with Thiamine diphosphate; Summary track showing the shares of amino acid changes resulting in varying degrees of fitness effects; Detailed heatmap showing individual amino acid change effects.

Within Calmodulin, the regions most sensitive to mutation are: 1) the hydrophobic cores of the two globular domains; 2) interfacial residues for protein-protein interactions, and 3) a subset of the negatively charged residues in EF hands that contact Ca^{2+} ions. Within the hydrophobic cores of the two lobes, five mutually interacting phenylalanine residues at positions 17, 69, 90, 93, and 142 stand out in particular, as all of them are found in the top 9 most sensitive residues on the map. Within the interaction interface, the residues D85, A89, F93, M100, L106, V109, L113, G114, L117, M125, V137, F142, M145, M146 are the most strongly sensitive to mutation. Regarding the four Calcium-binding EF-hand loops, it was interesting to find that only a subset of the negatively-charged residues contacting Ca^{2+} are even moderately sensitive. Within EF1, only D25 appears to be important, in EF2 only N61, in EF3 only D94 and D96, and in EF4 only D130 and D134. Overall, the EF3/4 in the C-lobe also appear to be more important than their N-lobe counterparts. This is in agreement with previous observations made by Sarhan and colleagues [128], who described the C-lobe as displaying a higher Ca^{2+} affinity. A number of unexplained sensitivities exist as well: Arginines at positions 54 and 91 show strong phenotypes despite extending from seemingly unused surfaces of the protein, offering the possibility that these residues are functionally relevant sites of interaction or modification.

3.2.5. Functional maps recapitulate known disease cases

To validate the utility of the maps in the context of human disease, I extracted known disease-associated variants from ClinVar [130], as well as rare and common polymorphisms observed independent of disease from GnomAD [105], and somatic variants previously observed in tumors from COSMIC [131].

For TPK1, a large number of very rare variants (minor allele frequency or MAF $< 10^{-6}$) is known from GnomAD. At first look, it appears the majority of these variants are shown to be deleterious (Figure 3.13). This seems unlikely, given that Thiamine Metabolism Dysfunction Syndrome, reported to be caused by mutations in this gene, is a very severe disease to which patients succumb in childhood [120], and given that GnomAD attempts to filter out subjects with severe pediatric disease. However, the disease is also known to follow a recessive inheritance pattern, with only homozygous or compound heterozygous individuals being affected. I thus used phased sequence data from the 1000 Genomes Project [9] to determine the diploid genotypes in the TPK1 locus for all listed individuals. Using these diploid phenotypes, I based the phenotype predictions

on the maximum fitness score of either (i.e. paternal and maternal) allele. This improved prediction performance markedly, leading to complete separation between disease and non-disease genotypes. However, both PROVEAN and PolyPhen-2 were also able to perfectly separate the two groups when using diploid genotypes (Figure 3.13B). Additional compound heterozygotes with known disease status will be required to determine whether this DMS map is more useful than computational methods for classifying pathogenic TPK1 variants.

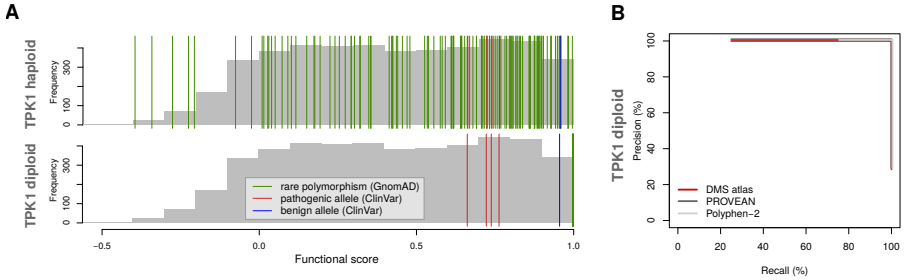

Figure 3.13.: Variant classificationin TPK1. A) Distribution of functional scores for rare polymorphisms (GnomAD) (green) and pathogenic and benign variants (ClinVar) (red, blue) in TPK1 overlaid on a histogram of functional scores for all missense variant. Top panel: Haploid scores considering the phenotype based only on a single allele. Bottom panel: Diploid scores, considering the phenotype based on both alleles in each individual. B) Precision-Recall curve for disease variant classification using diploid scores from the atlas presented here, PROVEAN and PolyPhen-2, based on the data from (A).

Evaluating the utility of the NCS1 map was similarly difficult. NCS1 does not have any entries in ClinVar. However, a previous publication identified the variant R102Q as a *de novo* variant in a single patient with autism spectrum disorder [121]. While the variant did not affect overall protein folding and localization, the authors did observe that the dynamics of cytosol-membrane cycling were altered. The complementation map did not show any functional impact for this variant. As is the case for TPK1, the emergence of more patient data in the future may enable a more useful evaluation of this map.

While no disease-associated missense alleles are recorded for UBE2I and SUMO1 in ClinVar, a number of somatic mutations for these genes have been observed in cancer according to COSMIC. While these can be expected to be passenger mutations, one may still hypothesize that somatic variants are likely not subject to the same selection pressures as germline variants, as interference with developmental processes is not necessarily detrimental to a tumour. I thus tested whether germline polymorphisms in these three genes were enriched for being functional compared to their somatic counterparts in the maps. Indeed, I observed a significant difference between the two sets (Wilcoxon $P = 2.6 \cdot 10^{-5}$) (Figure 3.14A).

Finally, I examined the functional map of Calmodulin. Here a sufficient number of disease-associated alleles were recorded in ClinVar. I found that the map was able to distinguish the disease variants from non-disease variants visibly well (Figure 3.14B). In contrast to TPK1, the Calmodulin map did not need to be corrected for diploid genotypes, as previously reported disease variants have been described as following a dominant inheritance pattern [122]. A precision-recall (PRC) plot reveals a superior performance (AUC = 0.74) compared to PROVEAN (AUC = 0.47) and PolyPhen-2 (AUC = 0.47) (Figure 3.14C). Remarkably, at 100% precision, the DMS map still achieves a recall of 50%, while PROVEAN and PolyPhen-2 only reach 20% and 15%, respectively.

To further put the Calmodulin map to the test in a clinical scenario, we inquired with Invitae, a company offering gene panel sequencing services for Long QT syndrome, including CALM1/2/3. In a blind test, we requested a list of Calmodulin variants they observed in patients but were unable to classify. After calibrating the map with respect to the above ClinVar and GnomAD datasets, I classified these 10 new variants (Table 3.3). Two were classified as damaging, six as benign, and two were too close to the threshold to be called either. In the next phase, Invitae revealed the associated patient cardiovascular phenotypes. Five out of the six patients with benign predictions were revealed to be unrelated to cardiovascular phenotypes, while both patients with damaging predictions did show a positive phenotype. The two uncertain cases were revealed to be affected as well. A Wilcoxon test showed these results to be statistically significant (P = 0.008).

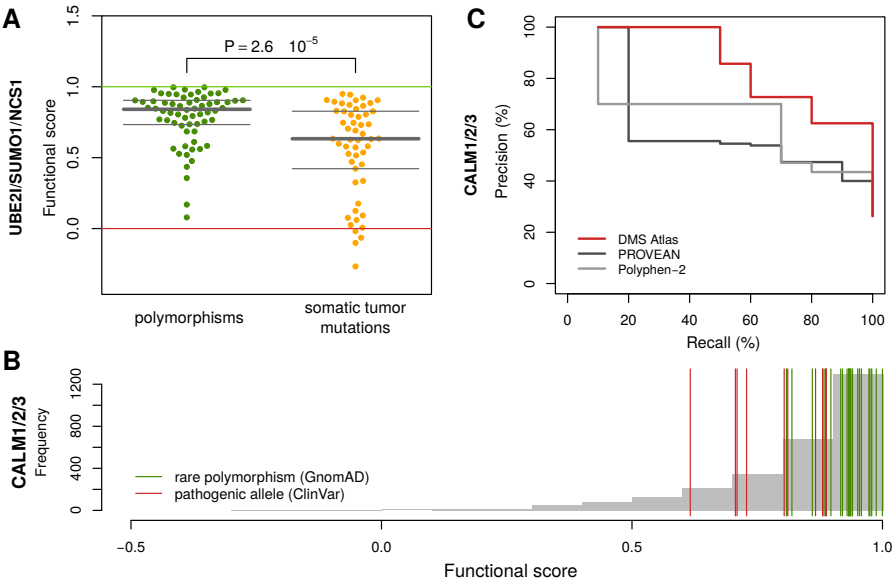

Figure 3.14.: Detection of disease-associated variants. A) Distributions of functional scores for rare polymorphisms (GnomAD) (green) and for somatic in cancer (COSMIC) (gold) in UBE2I, SUMO1 and NCS1. B) Distribution of functional scores for rare polymorphisms (GnomAD) (green) and known pathogenic alleles (ClinVar) (red) in CALM1, CALM2 and CALM3, overlaid on a histogram of all missense variant scores (gray). C) Precision-Recall curves for classification of disease variants using the variant atlas presented here, PROVEAN and PolyPhen-2 in CALM1, CALM2 and CALM3 based on rare polymorphisms from GnomAD and pathogenic variants from ClinVar.

3.3. Discussion

In total, this study has produced five maps with functional impacts for 15,998 possible missense variants. The functional maps generated for sumoylation pathway members UBE2I and SUMO1 and disease-implicated genes NCS1,

Table 3.3.: Re-classification attempt for variants of uncertain significance found in Invitae gene panel sequencing. MAF: Minor allele frequency in GnomAD, if known; sd/rmsd: standard error or RMSD of observation in map. imp/reg: imputed or degree of regularization; DMS score pre-regularization; DMS score post-regularization; DMS call: Classification according to DMS score; Indication: Type of sequencing panel ordered.

Variant	MAF	sd/ rmsd	imp/ reg	pre-reg	DMS	DMS call	indication
D94A	NA	0.26	imputed	NA	0.46	likely damaging	Cardio
D96H	NA	0.26	imputed	NA	0.72	likely damaging	Cardio
I28V	$1 \cdot 10^{-5}$	0.05	mild	0.88	0.88	uncertain	Cardio
N98S	NA	0.05	mild	0.89	0.89	uncertain	Cardio
T35I	$4 \cdot 10^{-6}$	0.04	mild	0.93	0.93	likely benign	Non-Cardio
E48G	NA	0.05	mild	0.93	0.93	likely benign	Cardio
G26D	NA	0.06	mild	0.94	0.94	likely benign	Non-Cardio
T27S	$3 \cdot 10^{-5}$	0.05	mild	0.96	0.96	likely benign	Non-Cardio
V122A	NA	0.05	mild	0.98	0.98	likely benign	Non-Cardio
A104G	NA	0.08	mild	1.00	1.00	likely benign	Non-Cardio

CALM1/2/3 and TPK1 using the DMS framework were consistent with biochemical expectations while providing new hypotheses. DMS maps based on functional complementation were highly predictive of disease causing mutations, outperforming computational prediction methods such as PolyPhen-2 or PROVEAN. The imputation method I employed allows me to generate complete functional maps while maintaining the reliability on par with the experimental results.

Given the prospect of personalized and precision medicine, genome sequencing is expected to become increasingly common in everyday medical practice. Current estimates suggest that every human carries an average of 200-300 rare missense mutations that have never before been seen in the clinic [9]. This creates a need for fast, reliable interpretation of variant effects. Instead of generating clones and functionally testing variants of unknown significance after they are first observed, DMS technology offers to generate exhaustive maps of functional variation that enable interpretation immediately upon clinical presentation, even for rare and personal variation.

A key requirement for DMS mapping is an *en masse* functional assay that can be applied at the scale of $10^4 - 10^5$ variant clones, such as complementation in yeast. However, among $\sim$ 4000 disease genes, examination of four systematic screens and curated literature suggests that only $\sim$ 5% of human disease genes have a yeast complementation assay [13, 132, 133]. Complementation assays can also be carried out in human cells [48], and *en masse* transfection is achievable at the required scale [29]. Based on only three large-scale CRISPR studies [28–30], cellular growth phenotypes have already been observed in at least one cell line for 29% of human disease genes. Beyond complementation, sub-functional assays, e.g. of protein interaction, can not only reveal variation that impacts the specifically assayed sub-function but also folding/stability mutations that ablate overall function. In a recent study, approximately two thirds of disease-causing variants were found to impact at least one protein interaction [25]. Although only a minority of human protein interactions have been mapped [134], already 40% of human genes have at least one interaction partner detectable by yeast two-hybrid assay in a recent screen [134]. Taking the union of available assays, one may estimate that 57% of known disease-associated genes already have an assay potentially amenable to DMS. Emerging protein interaction data and CRISPR screens suggests that the proportion of DMS-accessible disease genes will continue to rise.

3.4. Methods

3.4.1. DMS-TileSeq

The DMS-TileSeq experiment for *SUMO1*, *TPK1*, *NCS1*, and *CALM1* was performed by Song Sun and Marta Verby as described in chapter 2. The mutant alleles for the yeast temperature sensitive strains used were *smt3-331*, *thi80-ph*, *frq1-1* and *cmd1-1*. The downstream sequencing data analysis, scoring, imputation and regularization was performed by me as described in chapter 2.

3.4.2. DMS-BarSeq Y2H

The DMS-BarSeq Y2H experiment was performed by Jennifer Knapp as described in chapter 2, except for the following differences: (1) The *en masse* LR cloning of the UBE2I variant library was targeted into barcoded Y2H AD vectors; (2) Following KiloSeq and re-arraying, the library was pooled and transformed into haploid MATa yeast strains and mated with MATα strains

carrying SATB1-DB plasmids. Following diploid selection, the pool was grown for 48h in triplicates on Histidine-supplemented (permissive) and Histidine-deficient (selective) media, respectively. Plates were scraped and barcode loci amplified for BarSEQ.

Downstream sequencing data analysis and scoring was performed by me. Barcode counts were divided by the overall number of barcodes reads in each condition to obtain relative barcode frequency. Barcode frequencies in the -HIS condition were divided by frequencies in +HIS condition. Scores were then scaled to the null and WT controls, such that 0 corresponds to the average fitness of null controls and 1 to the average fitness of WT controls. Standard deviation based on technical replicates were regularized using the Baldi & Long method as described in chapter 2. The underlying code is part of a larger DMS analysis package provided on the attached storage media, and also available online[1].

3.4.3. UBE2I-SATB1 analysis

I integrated the complementation and Y2H data and filtered out low-quality measurements (s.d. > 0.3). To find interface candidates, I then selected the set of variants for which (i) the complementation score was greater than 0.5, (ii) the Y2H score was less than 0.5, and (iii) the Y2H score is at least 0.5 units below the complementation score. I then mapped the resulting variants on the UBE2I crystal structure.

3.4.4. UBE2I interface analysis

Co-crystal structure data for UBE2I was obtained from the PDB (Entries: `3UIP` [85]; `4W5V` [84]; `3KYD` [83]; `2UYZ` [112]; `4Y1L` [90]). A custom script was developed to obtain solvent accessibility using `GETAREA` [111] for monomers and complexes, allowing for the calculation of relative burial of interfacial residues. Complementation fitness distributions for each interaction's interfacial residues were tallied and tested for statistically significant differences using Wilcoxon tests. Distributions were plotted using the R package `beeswarm` [135]. The methods were implemented as part of a larger DMS analysis package provided on the attached storage media, and also available online[2].

[1] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`
[2] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

3.4.5. Structure coloration

A custom script was developed to calculate median and maximum complementation fitness values for each residue and autogenerate coloration commands for OpenPyMol [113]. The methods were implemented as part of a larger DMS analysis package provided on the attached storage media, and also available online[3].

3.4.6. Complementation spotting assays

Complementation spotting assays were performed by Jennifer Knapp as described in chapter 2. Image data was processed using PlateOrganizer and integrated and compared to the high-throughput results using custom scripts.

3.4.7. Hypercomplementing mutation analysis

Hypercomplementation and reversion to yeast residues: To examine whether changing amino acid residues into those residues naturally occur in yeast were more likely to show hyperactive complementation I compared these cases to changes into residues occurring in other species. The UBE2I amino acid sequence was aligned to that of its orthologues in *S. cerevisiae*, *D. discoideum* and *D. melanogaster* using `CLUSTAL` [114]. A custom script was used to extract inter-species amino acid changes and lookup the corresponding complementation fitness values in the UBE2I map. Distributions were plotted using the R package `beeswarm` [135]. Wilcoxon tests revealed no significant differences between the distributions. The methods were implemented as part of a larger DMS analysis package provided on the attached storage media, and also available online[4].

In vitro sumoylation comparison Images from *in vitro* sumoylation assays performed for UBE2I variants by Bernier-Villamor *et al.* [87] were scored by visual inspection while blinded to the underlying variant information. Scores were then represented as a heatmap and compared complementation scores from the UBE2I map. The methods were implemented as part of a larger DMS

[3] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`
[4] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

analysis package provided on the attached storage media, and also available online[5].

Phylogenetic comparison of different models for hyperactive mutations Jesse Bloom at the Fred Hutchinson Research Center in Seattle kindly provided the `phydms` software package [117] and applied it to test three different models relating the effect of activity-enhancing mutations in SUMO1 and UBE2I to the actual evolutionary preference for that amino acid in a real biological context. Specifically, using the substitution models described in [117], three different ways of relating the evolutionary preference $\pi_{r,a}$ for amino-acid a at site r to the fitness score $f_{r,a}$ for a given mutation were tested.

In the first model,

$$\pi_{r,a} = f_{r,a}.$$

In the second model,

$$\pi_{r,a} = \min(f_{r,a}, f_{r,\mathrm{wt}}),$$

where $f_{r,\mathrm{wt}}$ is the fitness score for the wildtype amino-acid at site r.

Finally, in the third model,

$$\pi_{r,a} = \begin{cases} f_{r,a} & \text{if } f_{r,a} \leq f_{r,\mathrm{wt}} \\ \frac{1}{f_{r,a}} & \text{otherwise} \end{cases}.$$

Each of these models were fit to the set of Ensembl homologues with at least 75% sequence identity to the human protein.

3.4.8. Transformation of maps for human phenotypes

Having established the third substitution model to provide the best fit for evolutionary preference (see above), I applied the corresponding transformation function underlying the model to the complementation data for each tested gene and repeated the imputation and regularization steps described in the previous chapter on the transformed data.

3.4.9. Intragenic epistasis analysis

Genetic interactions were determined based on a previously described multiplicative model [101,102], that expects double mutant fitness to conform to the

[5] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

product of single mutant fitness effects in the absence of interaction between the two. Under this model, the strength of genetic interaction is defined as

$$\varepsilon_{ij} = f_i \cdot f_j - f_{ij},$$

where f_i and f_j represent single mutant fitness and f_{ij} represents double mutant fitness scores. To test for deviation from this model, all cases where double mutant and both corresponding single mutants were known in the data were extracted. The standard deviation for the expected double mutant fitness $f_i \cdot f_j$ was estimated using

$$\mathbb{V}(XY) = \mathbb{E}(X^2Y^2) - (\mathbb{E}(XY))^2 = \mathbb{V}(X)\mathbb{V}(Y) + \mathbb{V}(X)(\mathbb{E}(Y))^2 + \mathbb{V}(Y)(\mathbb{E}(X))^2$$

Using these estimates, Student t-tests were performed between the measured and expected double mutant fitnesses and corrected for multiple hypothesis testing using the Benjamini-Hochberg [136] method at a 5% FDR threshold.

To detect potential direct compensatory relationships, the genetic interactions were compared with physical distance in the protein's 3D structure. The Euclidean distance between the C_α atoms in of each pair of residues was calculated using a custom script using structural data from the PDB (`3UIP` [85]). The methods were implemented as part of a larger DMS analysis package provided on the attached storage media, and also available online[6].

3.4.10. Structural analysis of disease gene maps

Co-crystal and NMR structure data for SUMO1, TPK1, NCS1 and CALM1 was obtained from the PDB (Entries: `2G4D` [82]; `2IO2` [137]; `3KYD` [83]; `3UIP` [85]; `2ASQ` [138]; `4WJO` [139]; `4WJQ` [139]; `1WYW` [140]; `2L2E` (Ames *et al.* unpublished); `4GUK` (Chengpeng *et al.* unpublished); `5AFP` [141]; `3G43` [142]; `4DJC` [128]; `3S4Y` [124];). Structures were colorized using the same method described above for UBE2I and analyzed using OpenPyMol [113].

3.4.11. Disease variant analysis

Missense variant tables for *UBE2I*, *SUMO1*, *TPK1*, *NCS1*, *CALM1*, *CALM2* and *CALM3* were integrated ClinVar, COSMIC, and GnomAD and compared with complementation scores. To calculate diploid scores for TPK1, phased

[6] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

variant call files (VCF) for the TPK1 gene obtained from the 1000 genomes project database to identify homozygous, heterozygous and compound heterozygous cases for all present variants using a custom script. For each case, the diploid score was calculated as $s_{\text{diploid}} = \max(s_1, s_2)$, where s_1 and s_2 are the variant scores for the paternal and maternal allele. The methods were implemented as part of a larger DMS analysis package provided on the attached storage media, and also available online[7].

[7] `http://dalai.mshri.on.ca/~jweile/projects/popcodePipeline/doc`

4. Conclusion

4.1. Summary

Here we have presented a complete framework for the construction of comprehensive, high-fidelity functional maps. We have demonstrated two versions of this framework: DMS-BarSeq, a barcode-based approach that allows for high-confidence measurement of individual clones including double- and higher-order multi-mutants; and DMS-TileSeq, a fast and efficient framework that generalizes fitness effects over many different clones sharing variants of interest. Both versions use a new mutagenesis protocol, POPCode, which thanks to its accompanying webtool makes it easer than before to generate variant libraries covering the complete space of amino acid changes. At its core, the framework relies on a functional complementation assay in yeast, which can measure the overall effect of variants on protein function and has been shown to be highly predictive of variant pathogenicity in humans, outperforming common *in silico* methods, despite the $\sim$ 1 billion year divergence between the two organisms. The DMS analysis software developed here introduces novel advances to deep mutational scanning: (i) The degree of confidence behind each measurement is carefully assessed and recorded in order to help variant classification; and (ii) variants that were missing in the complementation library or measured with low confidence were supplemented using a RandomForest-based machine learning method, yielding predictions that were found to be surprisingly reliable.

We have evaluated the technical features of the framework on the two sumoylation pathway members UBE2I and SUMO1. We found that the functional maps generated with our method were able to successfully recapitulate known features of the proteins' biology and biochemistry and even hint at novel features that warrant further investigation. We found a large number of genetic interactions between variants in UBE2I, some of which may be due to direct compensatory relationships of amino acid replacements. Most interactions however were found to involve residue pairs separated by larger physical distances.

Having validated the framework, we demonstrated its power to detect pathogenic

variants in the disease genes *TPK1*, *NCS1*, *CALM1*, *CALM2*, and *CALM3*. We found that our Calmodulin map excelled at distinguishing disease-associated variants from benign polymorphisms and greatly outperformed the common prediction algorithms PolyPhen-2 and PROVEAN. We subsequently applied our functional map for *CALM1*, *CALM2*, and *CALM3* to classify VUS observed in patients during gene panel sequencing and found our predictions to correlate significantly with patient indications.

Limitations of the DMS framework

Despite these successes, there are a number of limitations to the current form of our DMS framework. A fairly simple problem is the current restriction to scan relatively short genes. This is due to three reasons: (1) Longer genes would require a re-formulation of the mutagenesis protocol, as the number of mutations introduced per clone can be expected to increase linearly with gene length. This would need to be addressed by varying the concentration of mutant oligos in the amplification step. This solution could be tested systematically for templates of different lengths to determine the exact relationship between the factors involved. The results can then be added to the POPCode oligo design web tool to automatically report the most suitable protocol for each case to the experimenter. (2) Variant clone pools for longer genes must be kept at larger population sizes at all times to avoid bottlenecking the complexity of the pools. (3) Finally, larger libraries also require more sequencing reads to cover all variants at adequate depth. Thus they either require the use of higher-throughput instruments or would have to be processed in batches. A possible solution to all three problems would be to mutagenize only sections of longer genes that would be scanned separately from each other, although this would be more time consuming and costly.

A more difficult problem is that currently, the number of genes amenable to functional complementation in yeast is very limited. Song Sun and other members of the Roth lab have previously determined that only ~ 200 human disease genes can currently be examined using this assay [13]. In addition, we found that some of these genes suffer from mapping quality issues. We observed this in the *NCS1* map, which was of lower quality compared to other genes due to its relatively weak wildtype complementation fitness resulting in a less favourable signal-to-noise ratio. However it is possible that these assays might be improved by using different yeast strains with different backgrounds or by using different growth selection conditions. Moreover, as mentioned in

section 3.3 of the previous chapter, we have determined that 57% of disease genes could potentially be assayed using DMS variants based on Y2H or human cell lines instead, as will be discussed in further detail in the next section.

4.2. Outlook

4.2.1. Using DMS data in a clinical context

As introduced in chapter 1, a major motivating factor behind the development of our framework is to address the growing problem of variants of uncertain significance observed in the clinic. While our results show that functional maps as produced by our framework can be helpful in the effort of VUS reclassification, a single line of evidence is not usually sufficient. Even though the ACMG considers functional assays among the strongest classification criteria, they require at least one additional criterium of moderate strength, such as enrichment in cases over controls, or negligible allele frequency in the general population [6]. While most of the data informing the required criteria cannot be generated *en masse*, other information, such as allele frequencies in the general population are available from the 1000 genomes project [9] and the genome and exome aggregation database (GnomAD) [105]. Thus an important goal for the future would be the construction of a public database with an underlying automatic data integration and classification system that obtains information from available sources and automatically applies the ACMG's recommended decision-making process towards variant classification. Classification results should be presented transparently, revealing the individual underlying evidence, confidence levels, and reasoning structure. Alternatively, it is conceivable that future iterations of DMS maps can be validated to be sufficiently rigorous to allow for a change in ACMG guidelines.

Another factor to consider with respect of the presentation of DMS maps for clinical use is the reporting of imputation and regularization. Even though this work has shown that imputed values are equivalent in quality to their experimental counterparts and that regularization leads to improved performance, a sociological bias against computational predictions may lead to users dismissing these data. While full disclosure of data provenance is necessary, it may also lead to the misinterpretation of data if its presentation is not handled carefully.

The commitment towards the construction of a resource is only warranted if its primary source of information, functional maps generated using Deep

Mutational Scanning, can continue to be provided. The Roth lab is planning to continue building functional maps of disease genes and to expand the list of genes amenable to deep mutational scanning. A shortlist of ∼100 genes is planned to be addressed in the coming years. However, this undertaking is a costly one. Per 500 amino acid positions scanned, approximately $5500 need to be spent on consumables, primarily for sequencing and oligos for POPCode mutagenesis. Assuming six genes being scanned in parallel, approximately 45 full-time employee hours need to be invested per gene. Ultimately, this undertaking cannot be shouldered by one lab alone and will require outreach to other groups. As shown in chapter 1 section 1.4, a fair number of groups are already performing deep mutational scans and may be interested in collaboration. As a first step, the Roth and Fowler labs are already collaborating with respect to mapping a number of heart disease associated genes.

4.2.2. Adaptation and extentions to DMS technology

DMS in human cell lines

As mentioned above, an important future direction is the adaptation of the deep mutational scanning framework toward directly using human cell lines in competition assays. Recent genome-wide CRISPR screens have revealed a sizable number of genes with growth phenotypes in different human cell lines [28–30]. While a number of DMS efforts have already been performed using human cells [47, 48, 55, 61], the underlying assays were not generalizable, for example, the most recent effort by Majithia and colleagues [61] for PPARγ was only possible due to the fortuitous circumstances of having found a surface marker whose expression level directly reflects PPARGγ activity. Atina Cote in the Roth Lab is currently working on establishing a generalizable growth-based complementation assay using CRISPR in human cell lines.

Screening of other functional elements

Another important future direction is to expand the capability of Deep Mutational Scanning to enable assaying variants outside of protein-coding regions of the genome. However, since the space of the human genome is simply too large to be tested in its entirety the logical choice is to concentrate on elements most likely to be functionally relevant, such as splice sites, promoters, or transcription factor binding sites. Hanane Ennajdaoui in the Roth Lab is currently

working on adapting our DMS framework to scan intronic regions (shortened to exclude medial sequences).

4.2.3. Other uses of DMS functional map data

Screening for viral suppressors

Deep Mutational Scanning has many other potential uses beyond disease variant classification. As we have demonstrated in chapter 3 for *UBE2I*, the method helps shed light on the biophysical mechanisms underlying the function of a gene. We have also shown that using a combination of Y2H and complementation, DMS can point to potential new protein interaction interfaces.

UBE2I is known to be directly targeted by many viruses, such as HIV and EBV, through specific protein-protein interactions to subvert host defenses [143]. Using Y2H as the selection assay in our DMS framework and using our existing functional map of UBE2I as a reference, it would be possible to scan for variants that specifically disrupt interactions with viral proteins while not affecting overall UBE2I function. At the same time, this approach could help finding the specific interface for the interactions in question and could inform future drug development.

Advances in computational prediction of disease variants

As the number of functional maps produced via DMS grows, so does their value as training data for *in silico* prediction methods. Currently the number of genes scanned is not yet representative enough to cover the functional diversity of the proteome. However, Yingzhou Wu in the Roth Lab has already begun to explore its potential value for extrapolation. In an initial experiment, he was able to show that a machine learning method trained on the functional data obtained for UBE2I was able to make better predictions towards the effects of mutations in SUMO1 than if trained on the data set underlying PolyPhen-2 (HumDiv). Thus, with each new functional map added to our variant atlas, computational prediction method have the potential to become more powerful.

Functional classification of amino acid positions

The same wealth of functional data that may serve as training data for future computational prediction methods may also help us learn more about the set of roles played by different residues within proteins. In an initial experiment, I

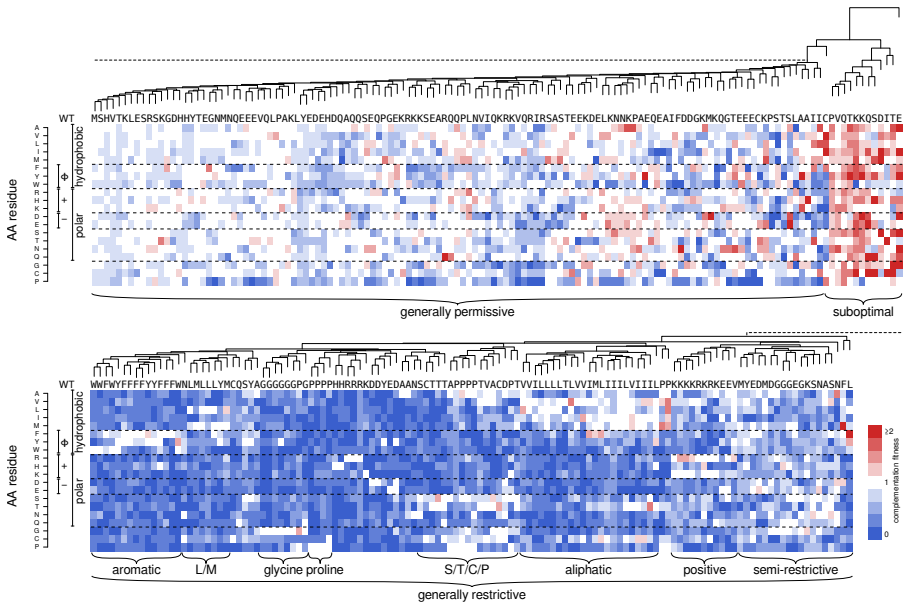

Figure 4.1.: Hierarchical clustering of amino acid positions in UBE2I and SUMO1 based on mutation profile similarity

have generated a hierarchical cluster map across amino acid positions in UBE2I and SUMO1 (Figure 4.1). The clustering hints at distinct functional classes occupied by different positions. There are three broad groups: (1) Positions that are generally unrestricted and can be occupied by almost any amino acid; (2) Positions that are generally constrained to a certain small number of amino acids and; (3) Positions that show hyperactivity for many possible amino acids. Within these groups there are a number of subclusters visible. For example, within the second group, certain positions only tolerate aliphatic residues, while others only tolerate aromatic residues. Evolution only samples a subset of the possible amino acids at a given position. By growing the set of proteins with complete functional maps we can potentially collect a catalog of possible functional 'archetypes' for positions within proteins. Using multiple alignments we can then make predictions as to the archetype of any given position.

Appendices

A. Variant maps with hypercomplementation

As was shown in chapter 3 section 3.2.1, phylogenetic analysis of *UBE2I* and *SUMO1* both showed that variants with ability to complement yeast better than wild-type are likely deleterious in humans. Thus, fitness scores were transformed so that such hypercomplementing mutations are considered to be deleterious. However, since hypercomplementing substitutions may provide interesting clues about differences between yeast and human cellular contexts, the untransformed versions the maps are provided below.

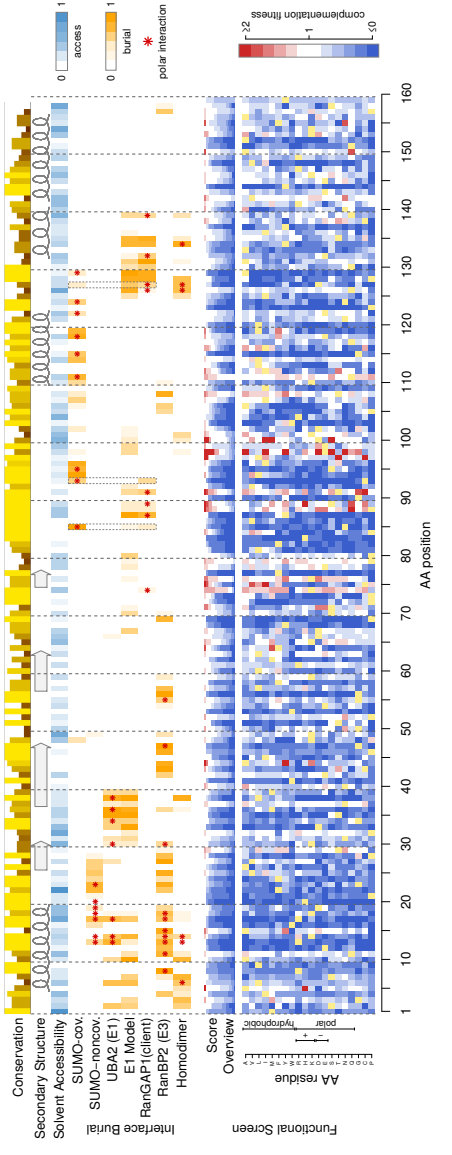

Figure A.1.: Functional map of UBE2I. From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in protein-protein interaction interfaces with covalently bound SUMO, non-covalently bound SUMO, the SUMO E1 complex at two different stages of activation, the sumoylation substrate RanGAP1, the E3 RanBP2, and the UBE2I homodimer; A summary track showing the relative number of amino acid changes resulting in different fitness effects; and finally the individual amino acid change effects sorted by physicochemical groups.

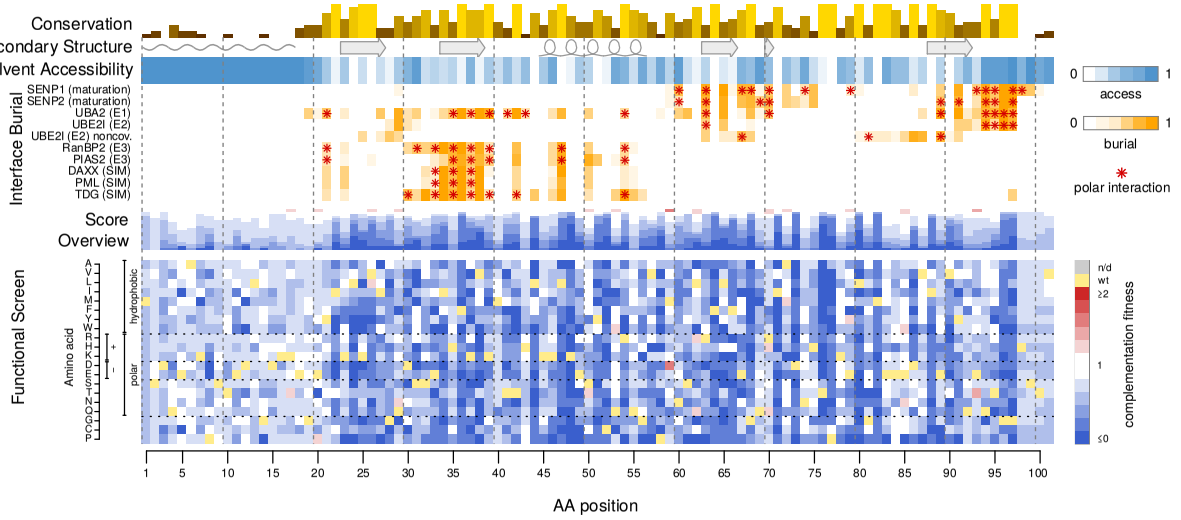

Figure A.2.: Functional map of SUMO1. From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in protein-protein interaction interfaces with SENP proteases, E1 complex, covalent and noncovalent E2 binding, E3s and three different SIM motifs; A summary track showing the relative number of amino acid changes resulting in different fitness effects; and finally the individual amino acid change effects sorted by physicochemical groups.

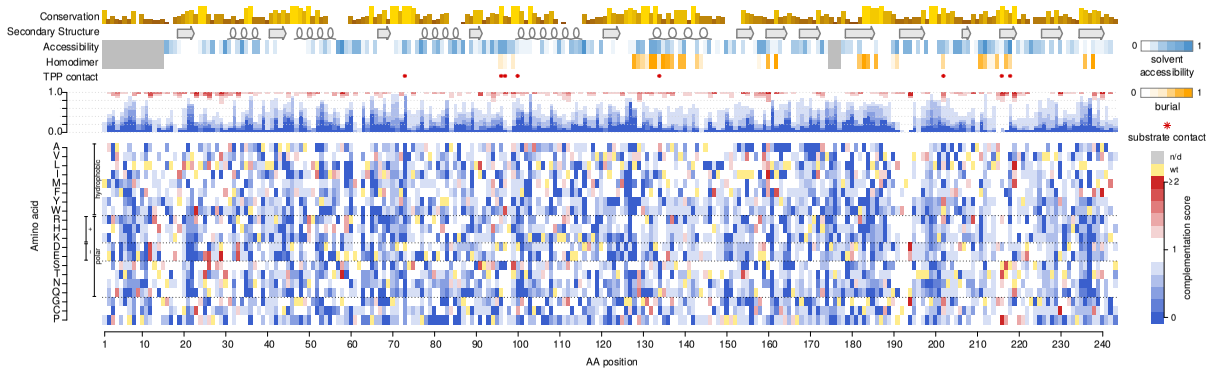

Figure A.3.: Functional map of TPK1. From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in homodimerization interfaces; Positions contacting the Thiamine pyrophosphate (TPP) molecule; A summary track showing the relative number of amino acid changes resulting in different fitness effects; and finally the individual amino acid change effects sorted by physicochemical groups.

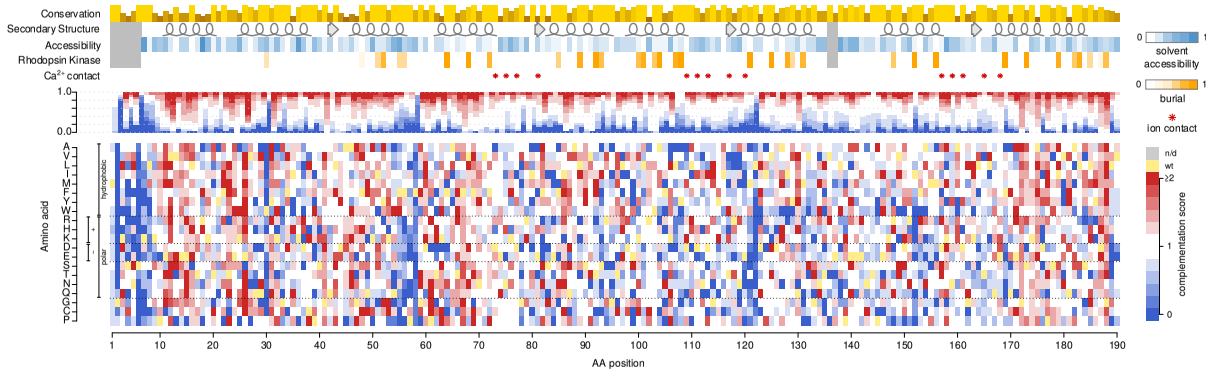

Figure A.4.: Functional map of NCS1. From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in interaction interface with rhodopsin kinase; Positions contacting Calcium ions; A summary track showing the relative number of amino acid changes resulting in different fitness effects; and finally the individual amino acid change effects sorted by physicochemical groups.

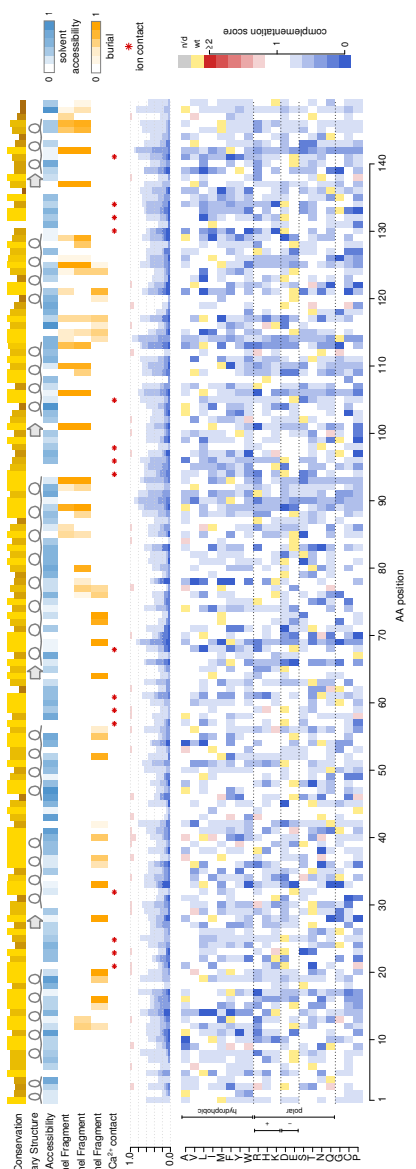

Figure A.5.: Functional map of Calmodulin. From top to bottom: Position-wise evolutionary conservation (AMAS); Secondary structure; Relative solvent accessibility; Relative burial in interaction interface with ion channel fragments; Positions contacting Calcium ions; A summary track showing the relative number of amino acid changes resulting in different fitness effects; and finally the individual amino acid change effects sorted by physicochemical groups.

Bibliography

[1] Douglas M. Fowler, Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. High-resolution mapping of protein sequence-function relationships. *Nature Methods*, 7(9):741–746, 09 2010.

[2] Andreas Ernst, David Gfeller, Zhengyan Kan, Somasekar Seshagiri, Philip M. Kim, Gary D. Bader, and Sachdev S. Sidhu. Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing. 6(10):1782–1790, 10 2010.

[3] Ryan T. Hietpas, Jeffrey D. Jensen, and Daniel N. A. Bolon. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences*, 108(19):7896–7901, 05 2011.

[4] Stacey L. Edwards, Jonathan Beesley, Juliet D. French, and Alison M. Dunning. Beyond GWASs: Illuminating the Dark Road from Association to Function. *American Journal of Human Genetics*, 93(5):779–797, 11 2013.

[5] Murat Taşan, Gabriel Musso, Tong Hao, Marc Vidal, Calum A. MacRae, and Frederick P. Roth. Selecting causal genes from genome-wide association studies via functionally coherent subnetworks. *Nature Methods*, 12(2):154–159, 02 2015.

[6] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, Heidi L. Rehm, and on behalf of the ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, 17(5):405–423, 05 2015.

[7] Jae Yeon Cheon, Jessica Mozersky, and Robert Cook-Deegan. Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? *Genome Medicine*, 6:121, 2014.

[8] Kara N. Maxwell, Steven N. Hart, Joseph Vijai, Kasmintan A. Schrader, Thomas P. Slavin, Tinu Thomas, Bradley Wubbenhorst, Vignesh Ravichandran, Raymond M. Moore, Chunling Hu, Lucia Guidugli, Brandon Wenz, Susan M. Domchek, Mark E. Robson, Csilla Szabo, Susan L. Neuhausen, Jeffrey N.

Weitzel, Kenneth Offit, Fergus J. Couch, and Katherine L. Nathanson. Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer. *The American Journal of Human Genetics*, 98(5):801–817, 05 2016.

[9] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 10 2015.

[10] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. In *Current Protocols in Human Genetics*. John Wiley & Sons, Inc., 2001. DOI: 10.1002/0471142905.hg0720s76.

[11] Pauline C. Ng and Steven Henikoff. Predicting Deleterious Amino Acid Substitutions. *Genome Research*, 11(5):863–874, 05 2001.

[12] Yongwook Choi, Gregory E. Sims, Sean Murphy, Jason R. Miller, and Agnes P. Chan. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE*, 7(10):e46688, 10 2012.

[13] Song Sun, Fan Yang, Guihong Tan, Michael Costanzo, Rose Oughtred, Jodi Hirschman, Chandra L. Theesfeld, Pritpal Bansal, Nidhi Sahni, Song Yi, Analyn Yu, Tanya Tyagi, Cathy Tie, David E. Hill, Marc Vidal, Brenda J. Andrews, Charles Boone, Kara Dolinski, and Frederick P. Roth. An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Research*, 26(5):670–680, 05 2016.

[14] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 07 1989.

[15] Meijia Yang, Zining Wu, and Stanley Fields. Protein-peptide interactions analyzed with the yeast two-hybrid system. *Nucleic Acids Research*, 23(7):1152–1156, 04 1995.

[16] Wim Van Criekinge and Rudi Beyaert. Yeast Two-Hybrid: State of the Art. *Biological Procedures Online*, 2:1–38, 10 1999.

[17] Paul L. Bartel and Stanley Fields. Analyzing protein-protein interactions using two-hybrid system. In *Methods in enzymology*, volume 254, pages 241–263. 1995.

[18] Igor Stagljar, Chantal Korostensky, Nils Johnsson, and Stephan te Heesen. A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proceedings of the National Academy of Sciences*, 95(9):5187–5192, 04 1998.

[19] Kavitha Iyer, Lukas Bürkle, Daniel Auerbach, Safia Thaminy, Martin Dinkel, Kim Engels, and Igor Stagljar. Utilizing the Split-Ubiquitin Membrane Yeast Two-Hybrid System to Identify Protein-Protein Interactions of Integral Membrane Proteins. *Sci. STKE*, 2005(275):pl3–pl3, 03 2005.

[20] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America*, 98(8):4569–4574, 04 2001.

[21] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 02 2000.

[22] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amelie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-Laszlo Barabasi, and Marc Vidal. An empirical framework for binary interactome mapping. *Nat Meth*, 6(1):83–90, 01 2009.

[23] Kirill Tarassov, Vincent Messier, Christian R. Landry, Stevo Radinovic, Mercedes M. Serna Molina, Igor Shames, Yelena Malitskaya, Jackie Vogel, Howard Bussey, and Stephen W. Michnick. An in Vivo Map of the Yeast Protein Interactome. *Science*, 320(5882):1465–1470, 06 2008.

[24] Sven Eyckerman, Annick Verhee, José Van der Heyden, Irma Lemmens, Xaveer Van Ostade, Joël Vandekerckhove, and Jan Tavernier. Design and application of a cytokine-receptor-based interaction trap. *Nature Cell Biology*, 3(12):1114–1119, 12 2001.

[25] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I. Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I. Karras, Yang Wang, István A. Kovács, Atanas Kamburov, Irina Krykbaeva, Mandy H. Lam, George Tucker, Vikram Khurana, Amitabh Sharma, Yang-Yu Liu, Nozomu Yachie, Quan Zhong, Yun Shen, Alexandre Palagi, Adriana San-Miguel, Changyu Fan, Dawit Balcha, Amelie Dricot, Daniel M. Jordan, Jennifer M. Walsh, Akash A. Shah, Xinping Yang, Ani K. Stoyanova, Alex Leighton, Michael A. Calderwood, Yves Jacob, Michael E. Cusick, Kourosh Salehi-Ashtiani, Luke J. Whitesell, Shamil Sunyaev, Bonnie Berger, Albert-László Barabási, Benoit Charloteaux, David E. Hill, Tong Hao, Frederick P. Roth, Yu Xia, Albertha J. M. Walhout, Susan Lindquist, and Marc Vidal. Widespread

Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, 161(3):647–660, 04 2015.

[26] Melanie G. Lee and Paul Nurse. Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. *Nature*, 327(6117):31–35, 05 1987.

[27] Michael J. Osborn and J. Ross Miller. Rescuing yeast mutants with human genes. *Briefings in Functional Genomics*, 6(2):104–111, 06 2007.

[28] Traver Hart, Megha Chandrashekhar, Michael Aregger, Zachary Steinhart, Kevin R. Brown, Graham MacLeod, Monika Mis, Michal Zimmermann, Amelie Fradet-Turcotte, Song Sun, Patricia Mero, Peter Dirks, Sachdev Sidhu, Frederick P. Roth, Olivia S. Rissland, Daniel Durocher, Stephane Angers, and Jason Moffat. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell*, 163(6):1515–1526, 12 2015.

[29] Vincent A. Blomen, Peter Májek, Lucas T. Jae, Johannes W. Bigenzahn, Joppe Nieuwenhuis, Jacqueline Staring, Roberto Sacco, Ferdy R. van Diemen, Nadine Olk, Alexey Stukalov, Caleb Marceau, Hans Janssen, Jan E. Carette, Keiryn L. Bennett, Jacques Colinge, Giulio Superti-Furga, and Thijn R. Brummelkamp. Gene essentiality and synthetic lethality in haploid human cells. *Science*, 350(6264):1092–1096, 11 2015.

[30] Tim Wang, Jenny J. Wei, David M. Sabatini, and Eric S. Lander. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science*, 343(6166):80–84, 01 2014.

[31] Rocio Acuna-Hidalgo, Joris A. Veltman, and Alexander Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, 17:241, 2016.

[32] B. C. Cunningham and J. A. Wells. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908):1081–1085, 06 1989.

[33] Yasuhiro Fujino, Risako Fujita, Kouichi Wada, Kotomi Fujishige, Takashi Kanamori, Lindsey Hunt, Yoshihiro Shimizu, and Takuya Ueda. Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochemical and Biophysical Research Communications*, 428(3):395–400, 11 2012.

[34] Bharat V. Adkar, Arti Tripathi, Anusmita Sahoo, Kanika Bajaj, Devrishi Goswami, Purbani Chakrabarti, Mohit K. Swarnkar, Rajesh S. Gokhale, and Raghavan Varadarajan. Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure*, 20(2):371–381, 02 2012.

[35] Richard N. McLaughlin Jr, Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, 11 2012.

[36] K. M. Schlinkmann, A. Honegger, E. Tureci, K. E. Robison, D. Lipovsek, and A. Pluckthun. Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *Proceedings of the National Academy of Sciences*, 109(25):9810–9815, 06 2012.

[37] Timothy A. Whitehead, Aaron Chevalier, Yifan Song, Cyrille Dreyfus, Sarel J. Fleishman, Cecilia De Mattos, Chris A. Myers, Hetunandan Kamisetty, Patrick Blair, Ian A. Wilson, and David Baker. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology*, 30(6):543–548, 06 2012.

[38] Michael W. Traxlmayr, Christoph Hasenhindl, Matthias Hackl, Gerhard Stadlmayr, Jakub D. Rybka, Nicole Borth, Johannes Grillari, Florian Rüker, and Christian Obinger. Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by Combining Directed Evolution with High Throughput Sequencing. *Journal of Molecular Biology*, 423(3):397–412, 10 2012.

[39] Nicholas C. Wu, Arthur P. Young, Sugandha Dandekar, Hemani Wijersuriya, Laith Q. Al-Mawsawi, Ting-Ting Wu, and Ren Sun. Systematic Identification of H274y Compensatory Mutations in Influenza A Virus Neuraminidase by High-Throughput Screening. *Journal of Virology*, 87(2):1193–1199, 01 2013.

[40] Benjamin P. Roscoe, Kelly M. Thayer, Konstantin B. Zeldovich, David Fushman, and Daniel N. A. Bolon. Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *Journal of Molecular Biology*, 425(8):1363–1377, 04 2013.

[41] L. M. Starita, J. N. Pruneda, R. S. Lo, D. M. Fowler, H. J. Kim, J. B. Hiatt, J. Shendure, P. S. Brzovic, S. Fields, and R. E. Klevit. Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 04 2013.

[42] Erik Procko, Rickard Hedman, Keith Hamilton, Jayaraman Seetharaman, Sarel J. Fleishman, Min Su, James Aramini, Gregory Kornhaber, John F. Hunt, Liang Tong, Gaetano T. Montelione, and David Baker. Computational Design of a Protein-Based Enzyme Inhibitor. *Journal of Molecular Biology*, 425(18):3563–3575, 09 2013.

[43] Christine E. Tinberg, Sagar D. Khare, Jiayi Dou, Lindsey Doyle, Jorgen W. Nelson, Alberto Schena, Wojciech Jankowski, Charalampos G. Kalodimos, Kai Johnsson, Barry L. Stoddard, and David Baker. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, 501(7466):212–216, 09 2013.

[44] Li Jiang, Parul Mishra, Ryan T. Hietpas, Konstantin B. Zeldovich, and Daniel N. A. Bolon. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLOS Genetics*, 9(6):e1003600, 06 2013.

[45] Ikjin Kim, Christina R. Miller, David L. Young, and Stanley Fields. High-throughput Analysis of in vivo Protein Stability. *Molecular & Cellular Proteomics*, 12(11):3370–3378, 11 2013.

[46] Daniel Melamed, David L. Young, Caitlin E. Gamble, Christina R. Miller, and Stanley Fields. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA*, 19(11):1537–1551, 11 2013.

[47] Charles M. Forsyth, Veronica Juan, Yoshiko Akamatsu, Robert B. DuBridge, Minhtam Doan, Alexander V. Ivanov, Zhiyuan Ma, Dixie Polakoff, Jennifer Razo, Keith Wilson, and David B. Powers. Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *mAbs*, 5(4):523–532, 07 2013.

[48] Timothy R. Wagenaar, Leyuan Ma, Benjamin Roscoe, Sung Mi Park, Daniel N. Bolon, and Michael R. Green. Resistance to vemurafenib resulting from a novel mutation in the BRAFV600e kinase domain. *Pigment Cell & Melanoma Research*, 27(1):124–133, 01 2014.

[49] Elad Firnberg, Jason W. Labonte, Jeffrey J. Gray, and Marc Ostermeier. A Comprehensive, High-Resolution Map of a Gene′s Fitness Landscape. *Molecular Biology and Evolution*, 31(6):1581–1592, 06 2014.

[50] C. Anders Olson, Nicholas C. Wu, and Ren Sun. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology*, 24(22):2643–2651, 11 2014.

[51] Alexandre Melnikov, Peter Rogov, Li Wang, Andreas Gnirke, and Tarjei S. Mikkelsen. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Research*, 42(14):e112–e112, 08 2014.

[52] Jesse D. Bloom. An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit. *Molecular Biology and Evolution*, 31(8):1956–1978, 08 2014.

[53] Bargavi Thyagarajan and Jesse D. Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *eLife*, 3:e03300, 07 2014.

[54] Michael A. Stiffler, Doeke R. Hekstra, and Rama Ranganathan. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell*, 160(5):882–892, 02 2015.

[55] Michael B. Doud, Orr Ashenberg, and Jesse D. Bloom. Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Molecular Biology and Evolution*, 32(11):2944–2960, 11 2015.

[56] Jacob O. Kitzman, Lea M. Starita, Russell S. Lo, Stanley Fields, and Jay Shendure. Massively parallel single-amino-acid mutagenesis. *Nature Methods*, 12(3):203–206, 03 2015.

[57] Lea M. Starita, David L. Young, Muhtadi Islam, Jacob O. Kitzman, Justin Gullingsrud, Ronald J. Hause, Douglas M. Fowler, Jeffrey D. Parvin, Jay Shendure, and Stanley Fields. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*, page genetics.115.175802, 03 2015.

[58] Parul Mishra, Julia M. Flynn, Tyler N. Starr, and Daniel N. A. Bolon. Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. *Cell Reports*, 15(3):588–598, 04 2016.

[59] Michael B. Doud and Jesse D. Bloom. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*, 8(6):155, 06 2016.

[60] David Mavor, Kyle Barlow, Samuel Thompson, Benjamin A. Barad, Alain R. Bonny, Clinton L. Cario, Garrett Gaskins, Zairan Liu, Laura Deming, Seth D. Axen, Elena Caceres, Weilin Chen, Adolfo Cuesta, Rachel E. Gate, Evan M. Green, Kaitlin R. Hulce, Weiyue Ji, Lillian R. Kenner, Bruk Mensa, Leanna S. Morinishi, Steven M. Moss, Marco Mravic, Ryan K. Muir, Stefan Niekamp, Chimno I. Nnadi, Eugene Palovcak, Erin M. Poss, Tyler D. Ross, Eugenia C. Salcedo, Stephanie K. See, Meena Subramaniam, Allison W. Wong, Jennifer Li, Kurt S. Thorn, Shane Ó Conchúir, Benjamin P. Roscoe, Eric D. Chow, Joseph L. DeRisi, Tanja Kortemme, Daniel N. Bolon, and James S. Fraser. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, 5:e15802, 04 2016.

[61] Amit R. Majithia, Ben Tsuda, Maura Agostini, Keerthana Gnanapradeepan, Robert Rice, Gina Peloso, Kashyap A. Patel, Xiaolan Zhang, Marjoleine F. Broekema, Nick Patterson, Marc Duby, Ted Sharpe, Eric Kalkhoven, Evan D. Rosen, Inês Barroso, Sian Ellard, UK Monogenic Diabetes Consortium, Sekar Kathiresan, Myocardial Infarction Genetics Consortium, Stephen O'Rahilly, UK Congenital Lipodystrophy Consortium, Krishna Chatterjee, Jose C. Florez, Tarjei Mikkelsen, David B. Savage, and David Altshuler. Prospective functional classification of all possible missense variants in PPARG. *Nature Genetics*, 48(12):1570–1575, 12 2016.

[62] R. C. Cadwell and G. F. Joyce. Mutagenic PCR. *Genome Research*, 3(6):S136–S140, 06 1994.

[63] Utpal Mohan, Shubhangi Kaushik, and Uttam Chand Banerjee. PCR Based Random Mutagenesis Approach for a Defined DNA Sequence Using the Mutagenic Potential of Oxidized Nucleotide Products. *The Open Biotechnology Journal*, 5(1):21–27, 07 2011.

[64] C. A. Hutchison, S. Phillips, M. H. Edgell, S. Gillam, P. Jahnke, and M. Smith. Mutagenesis at a specific position in a DNA sequence. *The Journal of Biological Chemistry*, 253(18):6551–6560, 09 1978.

[65] Andreas Seyfang and Jean Huaqian Jin. Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round. *Analytical Biochemistry*, 324(2):285–291, 01 2004.

[66] Elad Firnberg and Marc Ostermeier. PFunkel: Efficient, Expansive, User-Defined Mutagenesis. *PLOS ONE*, 7(12):e52031, 12 2012.

[67] T. A. Kunkel. Rapid and efficient site-specific mutagenesis without phenotypic selection. *Proceedings of the National Academy of Sciences*, 82(2):488–492, 01 1985.

[68] Gabor Pal and Frederic A Fellouse. Methods for the Construction of Phage-Displayed Libraries. In *Phage Display In Biotechnology and Drug Discovery*, Drug Discovery Series, pages 111–142. CRC Press, 07 2005. DOI: 10.1201/9780849359125.ch3.

[69] J. K. Scott and G. P. Smith. Searching for peptide ligands with an epitope library. *Science*, 249(4967):386–390, 07 1990.

[70] C. F. Barbas, J. D. Bain, D. M. Hoekstra, and R. A. Lerner. Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proceedings of the National Academy of Sciences*, 89(10):4457–4461, 05 1992.

[71] Sriram Kosuri, Nikolai Eroshenko, Emily M. LeProust, Michael Super, Jeffrey Way, Jin Billy Li, and George M. Church. Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature Biotechnology*, 28(12):1295–1299, 12 2010.

[72] G. P. Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science (New York, N.Y.)*, 228(4705):1315–1317, 06 1985.

[73] L. C. Mattheakis, R. R. Bhatt, and W. J. Dower. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proceedings of the National Academy of Sciences*, 91(19):9022–9026, 09 1994.

[74] M. H. Julius, T. Masuda, and L. A. Herzenberg. Demonstration That Antigen-Binding Cells Are Precursors of Antibody-Producing Cells After Purification with a Fluorescence-Activated Cell Sorter. *Proceedings of the National Academy of Sciences*, 69(7):1934–1938, 07 1972.

[75] Joseph B. Hiatt, Rupali P. Patwardhan, Emily H. Turner, Choli Lee, and Jay Shendure. Parallel, tag-directed assembly of locally derived short sequence reads. *Nature Methods*, 7(2):119–122, 02 2010.

[76] Jesse D. Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16:168, 2015.

[77] Douglas M. Fowler, Carlos L. Araya, Wayne Gerard, and Stanley Fields. Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics*, 27(24):3430–3431, 12 2011.

[78] Alan F. Rubin, Nathan Lucas, Sandra M. Bajjalieh, Anthony T. Papenfuss, Terence P. Speed, and Douglas M. Fowler. Enrich2: a statistical framework for analyzing deep mutational scanning data. *bioRxiv*, page 075150, 09 2016.

[79] W. Jiang and Y. Koltin. Two-hybrid interaction of a human UBC9 homolog with centromere proteins ofSaccharomyces cerevisiae. *Molecular and General Genetics MGG*, 251(2):153–160, 05 1996.

[80] Ruth Geiss-Friedlander and Frauke Melchior. Concepts in sumoylation: a decade on. *Nature Reviews Molecular Cell Biology*, 8(12):947–956, 12 2007.

[81] Karim Nacerddine, François Lehembre, Mantu Bhaumik, Jérôme Artus, Michel Cohen-Tannoudji, Charles Babinet, Pier Paolo Pandolfi, and Anne Dejean. The SUMO Pathway Is Essential for Nuclear Integrity and Chromosome Segregation in Mice. *Developmental Cell*, 9(6):769–779, 12 2005.

[82] Zheng Xu, So Fun Chau, Kwok Ho Lam, Ho Yin Chan, Tzi Bun Ng, and Shannon W. N. Au. Crystal structure of the SENP1 mutant C603s–SUMO complex reveals the hydrolytic mechanism of SUMO-specific protease. *Biochemical Journal*, 398(3):345–352, 09 2006.

[83] Shaun K. Olsen, Allan D. Capili, Xuequan Lu, Derek S. Tan, and Christopher D. Lima. Active site remodelling accompanies thioester bond formation in the SUMO E1. *Nature*, 463(7283):906–912, 02 2010.

[84] Katherine H. Reiter, Anita Ramachandran, Xue Xia, Lauren E. Boucher, Jürgen Bosch, and Michael J. Matunis. Characterization and Structural Insights into Selective E1-E2 Interactions in the Human and Plasmodium falciparum SUMO Conjugation Systems. *Journal of Biological Chemistry*, 291(8):3860–3870, 02 2016.

[85] Jaclyn R. Gareau, David Reverter, and Christopher D. Lima. Determinants of Small Ubiquitin-like Modifier 1 (SUMO1) Protein Specificity, E3 Ligase, and SUMO-RanGAP1 Binding Activities of Nucleoporin RanBP2. *Journal of Biological Chemistry*, 287(7):4740–4751, 02 2012.

[86] Deborah A. Sampson, Min Wang, and Michael J. Matunis. The Small Ubiquitin-like Modifier-1 (SUMO-1) Consensus Sequence Mediates Ubc9 Binding and Is Essential for SUMO-1 Modification. *Journal of Biological Chemistry*, 276(24):21664–21669, 06 2001.

[87] Victor Bernier-Villamor, Deborah A. Sampson, Michael J. Matunis, and Christopher D. Lima. Structural Basis for E2-Mediated SUMO Conjugation Revealed by a Complex between Ubiquitin-Conjugating Enzyme Ubc9 and RanGAP1. *Cell*, 108(3):345–356, 02 2002.

[88] Matthew S. Macauley, Wesley J. Errington, Manuela Schärpf, Cameron D. Mackereth, Adam G. Blaszczak, Barbara J. Graves, and Lawrence P. McIntosh. Beads-on-a-String, Characterization of Ets-1 Sumoylated within Its Flexible N-terminal Sequence. *Journal of Biological Chemistry*, 281(7):4164–4172, 02 2006.

[89] Frederick C. Streich Jr and Christopher D. Lima. Capturing a substrate in an activated RING E3/E2–SUMO complex. *Nature*, 536(7616):304–308, 08 2016.

[90] Aileen Y. Alontaga, Nigus D. Ambaye, Yi-Jia Li, Ramir Vega, Chih-Hong Chen, Krzysztof P. Bzymek, John C. Williams, Weidong Hu, and Yuan Chen. RWD Domain as an E2 (Ubc9)-Interaction Module. *Journal of Biological Chemistry*, 290(27):16550–16559, 07 2015.

[91] Michael H. Tatham, Ellis Jaffray, Owen A. Vaughan, Joana M. P. Desterro, Catherine H. Botting, James H. Naismith, and Ronald T. Hay. Polymeric Chains of SUMO-2 and SUMO-3 Are Conjugated to Protein Substrates by SAE1/SAE2 and Ubc9. *Journal of Biological Chemistry*, 276(38):35368–35374, 09 2001.

[92] Allan D. Capili and Christopher D. Lima. Structure and Analysis of a Complex between SUMO and Ubc9 Illustrates Features of a Conserved E2-Ubl Interaction. *Journal of Molecular Biology*, 369(3):608–618, 06 2007.

[93] Chung-Hsu Cheng, Yu-Hui Lo, Shu-Shan Liang, Shih-Chieh Ti, Feng-Ming Lin, Chia-Hui Yeh, Han-Yi Huang, and Ting-Fang Wang. SUMO modifications control assembly of synaptonemal complex and polycomplex in meiosis of Saccharomyces cerevisiae. *Genes & Development*, 20(15):2067–2081, 08 2006.

[94] Kalman P. Bencsath, Michael S. Podgorski, Vishwajeeth R. Pagala, Clive A. Slaughter, and Brenda A. Schulman. Identification of a Multifunctional Binding Site on Ubc9p Required for Smt3p Conjugation. *Journal of Biological Chemistry*, 277(49):47938–47945, 12 2002.

[95] Michael Kossin. Thermometer icon, Wikimedia Commons, 2004.

[96] Hylke Bons, Lapo Calamandrei, Olivier Charavel, Ryan Collier, Rodney Dawes, Vinicius Depizzol, Steven Garrity, Tuomas Kuosmanen, Garret LeSage, Niko Mirthes, Jesus D Navarro, Andreas Nilsson, Kalle Persson, Jakub Steiner, and Josef Vybiral. Computer icon, The Tango Project, 2005.

[97] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 02 1966.

[98] Irwin Sobel and Gary Feldman. A 3x3 isotropic gradient operator for image processing. In *The Stanford Artificial Intelligence Project*, pages 271–272, 1968.

[99] Pierre Baldi and Anthony D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 06 2001.

[100] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 10 2001.

[101] Patrick C. Phillips. The Language of Gene Interaction. *Genetics*, 149(3):1167–1171, 07 1998.

[102] Robert P. St Onge, Ramamurthy Mani, Julia Oh, Michael Proctor, Eula Fung, Ronald W. Davis, Corey Nislow, Frederick P. Roth, and Guri Giaever. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics*, 39(2):199–206, 02 2007.

[103] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 11 1992.

[104] Craig D. Livingstone and Geoffrey J. Barton. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Bioinformatics*, 9(6):745–756, 12 1993.

[105] Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O′Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, Daniel P. Birnbaum, Jack A. Kosmicki, Laramie E. Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M. Ruderfer, Khalid Shakir, Peter D. Stenson, Christine Stevens, Brett P. Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M. Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M. Neale, Aarno Palotie, Shaun M. Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J. Daly, Daniel G. MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 08 2016.

[106] Nozomu Yachie, Evangelia Petsalaki, Joseph C. Mellor, Jochen Weile, Yves Jacob, Marta Verby, Sedide B. Ozturk, Siyang Li, Atina G. Cote, Roberto Mosca, Jennifer J. Knapp, Minjeong Ko, Analyn Yu, Marinella Gebbia, Nidhi Sahni, Song Yi, Tanya Tyagi, Dayag Sheykhkarimli, Jonathan F. Roth, Cassandra Wong, Louai Musa, Jamie Snider, Yi-Chun Liu, Haiyuan Yu, Pascal Braun, Igor Stagljar, Tong Hao, Michael A. Calderwood, Laurence Pelletier, Patrick Aloy, David E. Hill, Marc Vidal, and Frederick P. Roth. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Molecular Systems Biology*, 12(4):863, 04 2016.

[107] Daniel G. Gibson, Lei Young, Ray-Yuan Chuang, J. Craig Venter, Clyde A. Hutchison, and Hamilton O. Smith. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5):343–345, 05 2009.

[108] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 04 2012.

[109] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 08 2009.

[110] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 12 1995.

[111] Robert Fraczkiewicz and Werner Braun. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry*, 19(3):319–333, 02 1998.

[112] Puck Knipscheer, Willem J. van Dijk, Jesper V. Olsen, Matthias Mann, and Titia K. Sixma. Noncovalent interaction between Ubc9 and SUMO promotes SUMO chain formation. *The EMBO Journal*, 26(11):2797–2807, 06 2007.

[113] Schrödinger. The PyMOL Molecular Graphics System, 2016.

[114] Fabian Sievers and DesmondG. Higgins. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In David J Russell, editor, *Multiple Sequence Alignment Methods*, number 1079 in Methods in Molecular Biology, pages 105–116. Humana Press, 01 2014. DOI: 10.1007/978-1-62703-646-7_6.

[115] Laurent Cappadocia, Andrea Pichler, and Christopher D. Lima. Structural basis for catalytic activation by the human ZNF451 SUMO E3 ligase. *Nature Structural & Molecular Biology*, 22(12):968–975, 12 2015.

[116] Yee-Fun Su, Tsunghan Yang, Hoting Huang, Leroy F. Liu, and Jaulang Hwang. Phosphorylation of Ubc9 by Cdk1 Enhances SUMOylation Activity. *PLOS ONE*, 7(4):e34250, 04 2012.

[117] Jesse D. Bloom. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biology Direct*, 12:1, 2017.

[118] Joseph-Anthony T. Tan, Yujie Sun, Jing Song, Yuan Chen, Theodore G. Krontiris, and Linda K. Durrin. SUMO conjugation to the matrix attachment region-binding protein, special AT-rich sequence-binding protein-1 (SATB1), targets SATB1 to promyelocytic nuclear bodies where it undergoes caspase cleavage. *The Journal of Biological Chemistry*, 283(26):18124–18134, 06 2008.

[119] Artemisia M. Andreou, Erwin Pauws, Marius C. Jones, Manvendra K. Singh, Markus Bussen, Kit Doudney, Gudrun E. Moore, Andreas Kispert, Jan J. Brosens, and Philip Stanier. TBX22 missense mutations found in patients with X-linked cleft palate affect DNA binding, sumoylation, and transcriptional repression. *American Journal of Human Genetics*, 81(4):700–712, 10 2007.

[120] Johannes A. Mayr, Peter Freisinger, Kurt Schlachter, Boris Rolinski, Franz A. Zimmermann, Thomas Scheffner, Tobias B. Haack, Johannes Koch, Uwe Ahting, Holger Prokisch, and Wolfgang Sperl. Thiamine Pyrophosphokinase Deficiency in Encephalopathic Children with Defects in the Pyruvate Oxidation Pathway. *The American Journal of Human Genetics*, 89(6):806–812, 12 2011.

[121] Mark T. W. Handley, Lu-Yun Lian, Lee P. Haynes, and Robert D. Burgoyne. Structural and Functional Deficits in a Neuronal Calcium Sensor-1 Mutant Identified in a Case of Autistic Spectrum Disorder. *PLOS ONE*, 5(5):e10534, 05 2010.

[122] Lia Crotti, Christopher N. Johnson, Elisabeth Graf, Gaetano M. De Ferrari, Bettina F. Cuneo, Marc Ovadia, John Papagiannis, Michael D. Feldkamp, Subodh G. Rathi, Jennifer D. Kunic, Matteo Pedrazzini, Thomas Wieland, Peter Lichtner, Britt-Maria Beckmann, Travis Clark, Christian Shaffer, D. Woodrow Benson, Stefan Kääb, Thomas Meitinger, Tim M. Strom, Walter J. Chazin, Peter J. Schwartz, and Alfred L. George. Calmodulin Mutations Associated with Recurrent Cardiac Arrest in Infants. *Circulation*, 127(9):1009–1017, 02 2013.

[123] Mette Nyegaard, Michael T. Overgaard, Mads T. Søndergaard, Marta Vranas, Elijah R. Behr, Lasse L. Hildebrandt, Jacob Lund, Paula L. Hedley, A. John Camm, Göran Wettrell, Inger Fosdal, Michael Christiansen, and Anders D. Børglum. Mutations in calmodulin cause ventricular tachycardia and sudden cardiac death. *American Journal of Human Genetics*, 91(4):703–712, 10 2012.

[124] David E Timm, Jingyuan Liu, L. J Baker, and Robert A Harris. Crystal structure of thiamin pyrophosphokinase1. *Journal of Molecular Biology*, 310(1):195–204, 06 2001.

[125] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 10 1990.

[126] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 01 2015.

[127] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–277, 06 2000.

[128] Maen F. Sarhan, Ching-Chieh Tung, Filip Van Petegem, and Christopher A. Ahern. Crystallographic basis for calcium regulation of sodium channels. *Proceedings of the National Academy of Sciences*, 109(9):3558–3563, 02 2012.

[129] Pétur O. Heidarsson, Ida J. Bjerrum-Bohr, Gitte A. Jensen, Olaf Pongs, Bryan E. Finn, Flemming M. Poulsen, and Birthe B. Kragelund. The C-Terminal Tail of Human Neuronal Calcium Sensor 1 Regulates the Conformational Stability of the Ca2+-Activated State. *Journal of Molecular Biology*, 417(1–2):51–64, 03 2012.

[130] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R. Maglott. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, 01 2016.

[131] S.a. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C.g. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, and P.j. Campbell. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. In *Current Protocols in Human Genetics*. John Wiley & Sons, Inc., 2001. DOI: 10.1002/cphg.21.

[132] Akil Hamza, Erik Tammpere, Megan Kofoed, Christelle Keong, Jennifer Chiang, Guri Giaever, Corey Nislow, and Philip Hieter. Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants. *Genetics*, 201(3):1263–1274, 11 2015.

[133] Aashiq H. Kachroo, Jon M. Laurent, Christopher M. Yellman, Austin G. Meyer, Claus O. Wilke, and Edward M. Marcotte. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, 348(6237):921–925, 05 2015.

[134] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, Atanas Kamburov, Susan D. Ghiassian, Xinping Yang, Lila Ghamsari, Dawit

Balcha, Bridget E. Begg, Pascal Braun, Marc Brehme, Martin P. Broly, Anne-Ruxandra Carvunis, Dan Convery-Zupan, Roser Corominas, Jasmin Coulombe-Huntington, Elizabeth Dann, Matija Dreze, Amélie Dricot, Changyu Fan, Eric Franzosa, Fana Gebreab, Bryan J. Gutierrez, Madeleine F. Hardy, Mike Jin, Shuli Kang, Ruth Kiros, Guan Ning Lin, Katja Luck, Andrew MacWilliams, Jörg Menche, Ryan R. Murray, Alexandre Palagi, Matthew M. Poulin, Xavier Rambout, John Rasla, Patrick Reichert, Viviana Romero, Elien Ruyssinck, Julie M. Sahalie, Annemarie Scholz, Akash A. Shah, Amitabh Sharma, Yun Shen, Kerstin Spirohn, Stanley Tam, Alexander O. Tejeda, Shelly A. Trigg, Jean-Claude Twizere, Kerwin Vega, Jennifer Walsh, Michael E. Cusick, Yu Xia, Albert-László Barabási, Lilia M. Iakoucheva, Patrick Aloy, Javier De Las Rivas, Jan Tavernier, Michael A. Calderwood, David E. Hill, Tong Hao, Frederick P. Roth, and Marc Vidal. A Proteome-Scale Map of the Human Interactome Network. *Cell*, 159(5):1212–1226, 11 2014.

[135] Aron Eklund. The Bee Swarm Plot, an Alternative to Stripchart, 2016.

[136] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[137] David Reverter and Christopher D. Lima. Structural basis for SENP2 protease interactions with SUMO precursors and conjugated substrates. *Nature Structural & Molecular Biology*, 13(12):1060–1068, 12 2006.

[138] Jing Song, Ziming Zhang, Weidong Hu, and Yuan Chen. Small Ubiquitin-like Modifier (SUMO) Recognition of a SUMO Binding Motif — A reversal of the bound orientation. *Journal of Biological Chemistry*, 280(48):40122–40129, 12 2005.

[139] Laurent Cappadocia, Xavier H. Mascle, Véronique Bourdeau, Samuel Tremblay-Belzile, Malik Chaker-Margot, Mathieu Lussier-Price, Junya Wada, Kazuyasu Sakaguchi, Muriel Aubry, Gerardo Ferbeyre, and James G. Omichinski. Structural and Functional Characterization of the Phosphorylation-Dependent Interaction between PML and SUMO1. *Structure*, 23(1):126–138, 01 2015.

[140] Daichi Baba, Nobuo Maita, Jun-Goo Jee, Yasuhiro Uchimura, Hisato Saitoh, Kaoru Sugasawa, Fumio Hanaoka, Hidehito Tochio, Hidekazu Hiroaki, and Masahiro Shirakawa. Crystal structure of thymine DNA glycosylase conjugated to SUMO-1. *Nature*, 435(7044):979–982, 06 2005.

[141] Sravan Pandalaneni, Vijaykumar Karuppiah, Muhammad Saleem, Lee P. Haynes, Robert D. Burgoyne, Olga Mayans, Jeremy P. Derrick, and Lu-Yun Lian. Neuronal Calcium Sensor-1 Binds the D2 Dopamine Receptor and G-protein-coupled Receptor Kinase 1 (GRK1) Peptides Using Different Modes of Interactions. *Journal of Biological Chemistry*, 290(30):18744–18756, 07 2015.

[142] Jennifer L. Fallon, Mariah R. Baker, Liangwen Xiong, Ryan E. Loy, Guojun Yang, Robert T. Dirksen, Susan L. Hamilton, and Florante A. Quiocho. Crystal structure of dimeric cardiac L-type calcium channel regulatory domains bridged by Ca2+·calmodulins. *Proceedings of the National Academy of Sciences*, 106(13):5135–5140, 03 2009.

[143] Archana Varadaraj, Domenico Mattoscio, and Susanna Chiocca. SUMO Ubc9 enzyme as a viral target. *IUBMB Life*, 66(1):27–33, 01 2014.