bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Expanding the Atlas of Functional Missense Variation for Human Genes

Jochen Weile[1,2,3,4,*], Song Sun[1,2,3,4,5,*], Atina G. Cote[1,2,3], Jennifer Knapp[1,2,3], Marta Verby[1,2,3], Joseph Mellor[2,6], Yingzhou Wu[1,2,3,4], Carles Pons[7], Cassandra Wong[1,2], Natascha van Lieshout[1], Fan Yang[1,2,3,4], Murat Tasan[1,2,3,4], Guihong Tan[2,3], Shan Yang[8], Douglas M Fowler[9], Robert Nussbaum[8], Jesse D. Bloom[10], Marc Vidal[11,12], David E Hill[11], Patrick Aloy[7,13], Frederick P. Roth[1,2,3,4,14,†]

[1]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto Ontario M5G 1X5, Canada. [2]The Donnelly Centre and Departments of [3]Molecular Genetics and [4]Computer Science University of Toronto, Toronto, Ontario M5S 3E1, Canada. [5]Department of Medical Biochemistry and Microbiology, Uppsala University, SE-75123 Uppsala, Sweden. [6]SeqWell Inc, Boston, MA. [7]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain. .[8]Invitae Corp., San Francisco, CA. [9]Department of Genome Sciences, University of Washington, Seattle, WA. [10]Fred Hutchinson Research Center, Seattle, WA. [11]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA 02215, USA. [12]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [13]Institució Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. [14]Canadian Institute for Advanced Research, Toronto, ON M5G 1Z8, Canada.

[*]Contributed Equally; [†]Corresponding Author

Abstract

Although we now routinely sequence human genomes, we can confidently identify only a fraction of the sequence variants that have a functional impact. Here we developed a deep mutational scanning framework that produces exhaustive maps for human missense variants by combining random codon-mutagenesis and multiplexed functional variation assays with computational imputation and refinement. We applied this framework to four proteins corresponding to six human genes: UBE2I (encoding SUMO E2 conjugase), SUMO1 (small ubiquitin-like modifier), TPK1 (thiamin pyrophosphokinase), and CALM1/2/3 (three genes encoding the protein calmodulin). The resulting maps recapitulate known protein features, and confidently identify pathogenic variation. Assays potentially amenable to deep mutational scanning are already available for 57% of human disease genes, suggesting that DMS could ultimately map functional variation for all human disease genes.

Keywords: Genotype-Phenotype/Deep Mutational Scanning/VUS/Complementation

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Millions of people will soon have their genomes sequenced. Unfortunately, we have only a limited ability to interpret personal genomes, each carrying 100-400 rare missense variants[1] 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[2]. Functional variants can be predicted, but when high precision is required, computational tools[3,4] detect only one third as many pathogenic variants as experimental assays[5]. Unfortunately, validated experimental assays enabling rapid clinical interpretation of variants are not available for the vast majority of human disease genes.

Deep Mutational Scanning (DMS)[6,7], a strategy for large-scale functional testing of variants, can functionally annotate a large fraction of amino acid substitutions for a substantial subset of residue positions. Recent DMS studies, for example, covered the critical RING domain of BRCA1[8] associated with breast cancer risk, and the PPARγ protein associated with Mendelian lipodystrophy and increased risk of type 2 diabetes[9]. Such maps can accurately identify functionality of a clinical variant in advance of that variant's first clinical presentation. Diverse assays can be used for DMS (see Supplementary Table S1). Functional complementation assays test 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)[10,11]. Cell-based functional complementation assays can accurately identify disease variants across a diverse set of human disease genes[5].

Challenges to the DMS strategy include the need to establish robust assays measuring each variant's impact on the disease-relevant functions of a gene, and to generate maps that cover all possible amino acid changes. Also, published DMS maps have not typically controlled the overall quality of measurements nor estimated the quality of individual measurements. Thus, the use of DMS maps to confidently evaluate specific variants has been limited.

Here we describe a modular DMS framework to generate complete, high-fidelity maps of variant function based on functional complementation. This framework combines elements of previous DMS studies, uses machine learning to impute and improve the map with surprisingly high accuracy, and yields a confidence measure for each reported measurement. In the following sections, we give an overview of the overall framework for DMS, describe its initial application to the SUMO E2 conjugase UBE2I, present complete high-fidelity maps for three new disease-associated proteins and explore the potential for clinical relevance. Finally, we assemble information on functional assays for known human disease genes and conclude that DMS is already potentially extensible to the majority of human disease genes, suggesting the possibility of exhaustive maps of functional variation covering all human genes.

Results

We describe a framework for comprehensively mapping of functional missense variation, organized into six stages (see Figure 1A): 1) mutagenesis; 2) generation of a variant library; 3) selection of functional variants; 4) read-out of the selection results and analysis to produce an initial sequence-function map; 5) computational analysis to impute missing

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

values; and 6) computational analysis to refine measured values via machine learning. We describe and contrast two versions of this framework: DMS-BarSeq and DMS-TileSeq.

A barcode-based deep mutational scanning strategy

We first describe DMS-BarSeq and its application to map functional missense variation for the SUMO E2 conjugase UBE2I. In DMS-BarSeq, a heterogeneous pool of cells bearing a library of different barcoded expression plasmid is quantified via barcode-sequencing before and after selection. For Stage 1 of the DMS framework—mutagenesis—we sought a relatively even representation of all possible single amino acid substitutions. We wished to allow multiple mutations per clone, both because this allowed for greater mutational coverage for any given library size, and offered an opportunity to discover intragenic epistatic relationships. To this end, we scaled up a previous mutagenesis protocol[12] to develop Precision Oligo-Pool based Code Alteration (POPCode), which yields random codon replacements (see Online Methods).

For Stage 2 of the framework—generation of a variant library—we employed *en masse* recombinational cloning of mutagenized UBE2I amplicons into a pool of randomly-barcoded plasmids (see Online Methods). The full-length UBE2I sequence and barcode of each plasmid was established using a novel sequencing method called KiloSeq which combines plate-position-specific index sequences with Illumina sequencing to carry out full-length sequencing for thousands of samples (see Online Methods). We retained clones that carried at least one amino acid substitution to generate a final library comprised of 6,553 UBE2I variants, covering different combinations of 1,848 (61% of all possible) unique amino acid changes. Variant plasmids were pooled, together with empty vector and wild type control plasmids (see Online Methods).

For Stage 3—selection for clones encoding a functional protein—we employed a *S. cerevisiae* functional complementation assay[5,13], based on human *UBE2I*'s ability to rescue growth at an otherwise-lethal temperature in a yeast strain carrying a temperature sensitive (ts) allele of the *UBE2I* orthologue *UBC9*. Despite a billion years of divergence, yeast functional complementation assays can accurately discriminate pathogenic from non-pathogenic human variants[5]. The plasmid library from Stage 3 was transformed *en masse* into the appropriate ts strain. Pools were grown for 48 hours at the permissive (25°C) and selective (37°C) temperatures, respectively (see Online Methods).

To assess variant functions (Stage 4), barcodes were sequenced at multiple timepoints of the selection, enabling reconstruction of individual growth curves and normalized fitness quantification for each of the 6,553 barcoded strains. Functional complementation scores were calibrated so that 0 corresponds to the fitness of the null-allele and 1 to wild type complementation (see Online Methods). Using replicate agreement and extent of library representation, we estimated our uncertainty in each fitness value (see Online Methods).

Before further refinement in Stages 5 and 6, we wished to assess the quality of complementation scores. Based on both technical (Figure 1B, top) and biological replicates (different clones carrying the same mutation; Figure 1B, bottom), we found scores to be reproducible (Pearson's R of 0.97 and 0.78, respectively). Semi-quantitative manual complementation assays for a subset of mutants that spanned the range of fitness scores (see Online Methods) correlated well with DMS scores. Indeed, agreement between large-

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

scale and manual scores was on par with agreement between internal replicates of the large-scale scores (Figure 1B,C).

We also examined evolutionary conservation and computational predictors of deleteriousness, such as PolyPhen-2[3] and PROVEAN[14]. Although each is an imperfect measure of the functionality of amino acid changes[5], each should and did correlate with DMS results (Figure 1D top panel, Supplementary Figure S1). Finally, we 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 1D, bottom panel). Taken together, these observations support the biological relevance of the DMS-BarSeq approach.

A tiled-region strategy for mapping functional variation

While DMS-BarSeq has several advantages (see Discussion), its performance comes at the cost of producing an arrayed library of clones, each with known coding and barcode sequence. We therefore also evaluated an alternative approach, DMS-TileSeq in which each functional variant is detected via the effect of selection on the abundance of clones carrying that variant. The frequency of each variant in the pool is determined, before and after selection, by deep sequencing of short amplicons that tile the complete coding region.

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 (thereby skipping the step of arraying and sequencing individual clones). This plasmid pool was next transformed *en masse* into the *ubc9-ts* strain. As with DMS-BarSeq, DMS-TileSeq employs pooled strains grown competitively (Stage 3) at the permissive and selective temperatures. In Stage 4, like some previous DMS efforts[15], we directly sequenced the coding region from the clone population to determine variant frequency before and after selection. Use of tiled amplicons enables individual template molecules to be sequenced on both strands, allowing elimination of most base-calling errors[6] (see Online Methods for details).

To assess the reliability of DMS-TileSeq, we compared results with DMS-BarSeq for UBE2I. DMS-TileSeq and DMS-BarSeq correlation was similar to that observed between biological DMS-BarSeq replicates (Pearson's R = 0.75, Supplementary Figure S2). DMS-TileSeq and DMS-BarSeq also behaved similarly in their agreement with manual complementation assays (Supplementary Figure S3). 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 functional complementation scores.

After using regression to transform the DMS-TileSeq scores to the more intuitive scale of DMS-BarSeq (where 0 corresponds to the median score of null mutant controls and 1 corresponds to the median score of wildtype controls), we combined scores from the two methods, giving greater weight to more confident measurements (see Online Methods).

Machine learning to complete and refine maps

As with most previous DMS maps, our initial UBE2I map missed a number of entries (e.g. due to substitutions underrepresented in the input clone library). In total, 2563 of 3012 possible amino acid changes (85%) were measured. To complete the map (Stage 5 in the framework), we trained a random forest[16] regression model using the existing measurements

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the map. The model used four types of predictive feature: intrinsic (derived from other measurements in our map); conservation-based; chemicophysical; and structural. Particularly predictive features (Figure 2D) included the average score of observed substitutions at a given position, as weighted by measurement confidence. Conservation-based features included BLOSUM62[17], SIFT[18] and PROVEAN[14] scores, and position-specific AMAS[19] conservation. Chemicophysical features included mass and hydrophobicity of the original and wildtype amino acids, and the difference between them. Structural features included solvent accessibility and burial in interaction interfaces. For DMS-BarSeq, which scored many multi-mutant clones, we also used the confidence-weighted average score of all clones containing a particular substitution, and variant fitness expected from a multiplicative model[20] (see Online Methods).

We assessed imputation performance using cross-validation. Surprisingly, the error (root-mean squared deviation or RMSD) of imputed values (0.33) was on par with that of experimentally measured data (Figure 2A). 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 imputed values (Figure 2C), again finding strong agreement. Predictions showed the least error in positions with high mutation density and the most error for hypercomplementing variants, i.e. those yielding above-WT fitness levels in yeast (Figure 2B). Although hypercomplementation may indicate that a variant is adaptive in yeast, imputation generally predicted these variants to be deleterious, a hypothesis we explore further below.

To refine less-confident experimental measurements (Stage 6 of the framework), we combined experimental and imputed scores, weighting by confidence level. Manual complementation assays, applied to a set of variants that represented the full range of fitness scores (Supplementary Figure S3), served to validate the reliability of the complete, refined functional map of UBE2I after imputation and refinement. The map, as seen in Figure 3A, fulfills biochemical expectations, with the hydrophobic core, the active site and protein interaction interfaces being most strongly impacted by mutations (Figure 3B). Detailed observations with respect to structure, biochemistry and epistatic behaviour of double mutants can be found in supplementary text.

Hypercomplementing variants are likely to be deleterious in humans

We further investigated UBE2I variants exhibiting hypercomplementation (Figure 3A). Manual assays confirmed that complementation with these mutants allows greater yeast growth than does the wild type human protein (Supplementary Figure S4A). These hypercomplementing substitutions did not reliably correspond to ‘reversion’ substitutions that inserted the corresponding *S. cerevisiae* residue (Supplementary Figure 4B). Some substitutions could be adaptive by improving compatibility with yeast interaction partners. Indeed, a comparison with co-crystal structure data[21] shows that many of the hypercomplementing residues are on the surface proximal to the substrate, with some directly contacting the substrate’s sumoylation motif (Figure 2C). *In vitro* sumoylation assays performed previously for a small number of UBE2I mutants revealed increased sumoylation for some substrates[22]. Comparing our map with these sumoylation assay results, we saw that cases of hypercomplementation were enriched for substrate specificity shift (Supplementary Figure S4C). However, other cases of hypercomplementation hinted at different modes of adaptation (see supplementary text).

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To explore whether variants exhibiting hypercomplementation are more likely beneficial or deleterious in a human context, we used a quantitative phylogenetic approach[23,24] to compare three models relating complementation scores to evolutionary preference for an amino acid variant: (a) evolutionary preference is directly proportional to complementation score; (b) preference has a ceiling at the wildtype complementation score (values >1 were set to 1); or (c) preference is set to the reciprocal of complementation score for mutations with greater-than-wildtype scores, corresponding to a deleterious effect of hypercomplementing mutations. We used the phydms software[24] to test which of these three approaches best described the evolutionary constraint on a set of naturally occurring UBE2I homologs, using fitness scores that excluded conservation features from the refinement process, to avoid the circularity of using natural sequence data when deriving the scores. The best fit is achieved by treating variants with greater-than-wildtype complementation in yeast as deleterious in humans (Supplementary Table S2). We therefore reinterpreted cases of hyperactive complementation in our map as deleterious and repeated the imputation and refinement procedure. This also allowed for more reliable imputed values (reducing cross-validation RMSD from 0.33 to 0.24).

Variant impact maps for five additional disease-implicated genes

Having validated the framework, we sought to map functional variation for disease-relevant genes. We applied the higher-throughout TileSeq approach, coupled with yeast complementation, to a diverse set of genes: SUMO1, for which heterozygous null variants are associated with cleft palate[25]; Thiamine Pyrophosphokinase 1 (TPK1), associated with vitamin B1 metabolism dysfunction[26]; and CALM1, CALM2 and CALM3, associated with cardiac arrhythmias (long-QT syndrome[27] and catecholaminergic polymorphic ventricular tachycardia[28]). Because the three calmodulin genes encode the same polypeptide sequence, performing DMS for CALM1 also provided maps for CALM2 and CALM3.

Supporting the quality of the resulting four maps, each map showed clear differences in score between distributions of likely-neutral (synonymous) and likely-deleterious (nonsense) variants (Supplementary Figure S5). To assess the impact of the machine learning imputation and refinement on the different maps, we measured the completeness of each map before and after imputation, the cross-validation RMSD of the imputation, as well as the maximum standard error value for each map before and after refinement (Table 1). On average, 24.6% of scores were obtained purely by imputation, and 3.96% of scores were appreciably changed by >5% of the difference between null and wt controls as a result of refinement. Proteins for which map quality was initially lower were improved most by refinement, while others, like SUMO1, improved only modestly. Inspection of the maps yielded a number of interesting biochemical and structural observations (see supplementary text).

Phylogenetic analysis of SUMO1, as for UBE2I, showed that variants that complement yeast better than wild-type are best modeled as being deleterious in humans (Supplementary Table S2). We therefore transformed above-wild-type fitness scores to be deleterious (see Methods). Because hypercomplementing substitutions provide interesting clues about differences between yeast and human cellular contexts, we provide both transformed (Figure 4) and untransformed (Supplementary Figure S6) map versions.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DMS functional maps reflect clinical phenotypes.

To validate the utility of our maps in the context of human disease, we extracted known disease-associated variants from ClinVar[29], as well as rare and common polymorphisms observed independent of disease from GnomAD[30], and somatic variants previously observed in tumors from COSMIC[31].

While no germline disease-associated missense variants are known for UBE2I and SUMO1 in ClinVar, somatic cancer variants have been observed for both genes according to COSMIC. Somatic variants in these three genes exhibited higher functional impact in DMS maps than germline variants (Wilcoxon $P=2.6x10^{-5}$) (Figure 5A). This does not necessarily suggest that either of these genes are cancer drivers, as even passenger somatic variants should subject to less purifying selection than germline variants, but it does lend further credence to the biological relevance of our maps.

For TPK1, many very rare variants (minor allele frequency or MAF < 10^{-6}) are seen in GnomAD. The majority of these variants score as deleterious (Supplementary Figure S7A). Thiamine Metabolism Dysfunction Syndrome, reported to be caused by variants in TPK1, is a severe disease to which patients succumb in childhood[26]. Although GnomAD attempted to exclude subjects with severe pediatric disease, the abundance of rare predicted-deleterious variants may be understood by the disease's recessive inheritance pattern. Using phased sequence data from the 1000 Genomes Project[1] to determine diploid genotypes in TPK1, we assigned each subject a diploid score corresponding to the maximum score across each pair of alleles. This improved prediction performance markedly, leading to complete separation between disease and non-disease genotypes using DMS, PROVEAN or PolyPhen-2 scores (Supplementary Figure S7B). However, additional compound heterozygotes with known disease status will be required to compare DMS with computational methods in the task of identifying TPK1 disease variants.

Because the inheritance pattern of calmodulin disorders is typically dominant[27], we did not consider diploid genotypes but simply evaluated the ability of DMS scores to distinguish disease from non-disease variants (Figure 5B). DMS scores performed well according to precision-recall analysis, with an area under the precision-recall curve (AUC) of 0.72, exceeding both PROVEAN (AUC=0.48) and PolyPhen-2 (AUC=0.47) (Figure 5C). At a stringent precision threshold of 90%, DMS exceeded twice the sensitivity of PROVEAN and PolyPhen-2. We further investigated variants seen by Invitae, a clinical genetic testing company. Ten rare calmodulin variants had been observed, of which half were from tests ordered due to a cancer indication, the other half from tests ordered for a cardiac disease indication. Blinded to indication, we ranked the 10 Invitae variants by DMS score (Table 2). Setting DMS score thresholds based on disease and non-disease variants from ClinVar, we classified two Invitae variants as damaging, two as VUS, and six as benign. Based on the patient test indications subsequently revealed by Invitae, five out of the six variants we classified as benign were ordered due to a non-cardiac indication, while both variants with damaging predictions and both with VUS predictions corresponded to cardiac indications. Overall, DMS scores showed a significant association with cardiac indications (P=0.008; Mann-Whitney-U test).

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Potential for applying deep mutational scanning more widely

DMS mapping requires an *en masse* functional assay that can be applied at the scale of 10^4-10^5 variant clones. Among ~4000 disease genes, examination of four systematic screens and curated literature suggests that ~5% of human disease genes currently have a yeast complementation assay[5,32,33]. This number could grow dramatically via systematic complementation testing under different environments and genetic backgrounds. Moreover, complementation assays can also be carried out in other model systems including human cells[34], where current transfection efficiencies permit *en masse* screening at the required scale. Based on only three large-scale CRISPR studies[34–36], cellular growth phenotypes (which might serve as the basis for an *en masse* selection) have already been observed in at least one cell line for 29% of human disease genes. Beyond complementation, assays of protein interaction can, in addition to identifying variants directly impacting interaction, can detect variants ablating overall function through effects on protein folding or stability, In a recent study, approximately two thirds of disease-causing variants were found to impact at least one protein interaction[37]. Although only a minority of human protein interactions have been mapped[38], already 40% of human genes have at least one interaction partner detectable by yeast two-hybrid assay in a recent screen[38]. Taking the union of available assays, we estimate that 57% of known disease-associated genes (Supplementary Table S3) already have an assay that is potentially amenable to DMS .

Discussion

The framework for systematically mapping functional missense variation we describe here combines elements of previous DMS studies and introduces a new mutagenesis strategy and a machine learning-based imputation and refinement strategy. This framework enables DMS maps that are 'complete' in the sense that high-quality functional impact scores are provided for all missense variants to full-length proteins. Application to four proteins highlighted complex relationships between the biochemical functions of these proteins with phenotypes in the yeast model system. Analysis of pathogenic variation, especially for calmodulin, supported the potential clinical utility of DMS maps from this framework.

The two described versions of DMS, DMS-BarSeq and DMS-TileSeq, each have advantages and limitations. DMS-BarSeq permits study of the combined effects of variants at any distance along the clone, and therefore can reveal intramolecular genetic interactions. For DMS-BarSeq fully-sequenced variant clones are arrayed, enabling further investigation of individual variants. DMS-BarSeq can directly compare growth of any clone to null and wild type controls, resulting in an intuitive scoring scheme. However, despite the efficient KiloSeq strategy for sequencing arrayed clone sets we report for the first time here, DMS-BarSeq is more resource-intensive. Although the regional sequencing strategy of DMS-TileSeq can only analyze fitness of double mutant combinations falling within the same ~150bp tile, it is far less resource-intensive than DMS-BarSeq.

Given that most missense variants in individual human genes are single-nucleotide variants[30], and given that only ~30% of all possible amino acid substitutions are accessible by single nucleotide mutation, one might wonder why codon mutagenesis should be preferred over single-nucleotide mutagenesis. We see three arguments for codon-level mutagenesis: 1) knowing the functional impact of all 19 possible substitutions at each positions enables clearer understanding of the biochemical properties that are required at

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

each residue position; 2) an analysis of >60,000 unphased human exomes[30] found that each individual human harbors ~23 codons containing multiple nucleotide variants that together could encode an amino acid not encoded by either single variant; 3) it is not straightforward to generate balanced libraries in which every single-nucleotide variant has roughly equal representation, given that error-prone amplification methods strongly favor transition mutations over transversion mutations, while still avoiding frequent introduction of new stop codons; 4) the major cost of DMS will likely continue to be development and validation of the functional assay, so using codon-level mutagenesis instead of (or in addition to) nucleotide-level mutagenesis has a relatively small impact on overall cost.

This study yielded four DMS maps measuring functional impact of ~16,000 missense variants. The maps generated for sumoylation pathway members UBE2I and SUMO1, and disease-implicated genes CALM1/2/3 and TPK1 using our framework were consistent with biochemical expectations while providing new hypotheses. DMS maps based on functional complementation were highly predictive of disease-causing variants, outperforming popular computational prediction methods such as PolyPhen-2 or PROVEAN[5]. Given sufficient experimental data for training, our results show that imputation can 'fill the gaps' with scores that are nearly as reliable as experimental measurements, and that computational refinement can improve upon experimental measures.

Genome sequencing is likely to become common in clinical practice. Current estimates suggest that every human carries an average of 100-400 rare variants that have never before been seen in the clinic. DMS meets a critical 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 offers exhaustive maps of functional variation that enable interpretation immediately upon clinical presentation, even for rare and personal variation. Our survey of assays revealed that the majority (57%) of human disease genes are potentially already accessible to DMS analysis, so that we may begin to imagine an atlas of DMS maps to reveal pathogenic variation for all human disease proteins.

Acknowledgements

The authors thank Amy Caudy, Lincoln Stein, Igor Stagljar, Chris Lima and Brian Raught for their advice, and thank Brenda Andrews and Charles Boone for kindly providing temperature-sensitive yeast mutant strains.

The authors gratefully acknowledge funding by the National Human Genome Research Institute of the National Institutes of Health (NIH/NHGRI) Center of Excellence in Genomic Science (CEGS) Initiative, the Canadian Excellence Research Chairs (CERC) Program, and the Ontario Ministry of Research and Innovation (MRI).

Author contributions

FPR, JW, AGC and SS conceived the project; AGC,SS,JK,MV and CW performed the DMS experiments and manual validations; JM, MT and FR conceived the KiloSeq method, AGC, JK and JM performed KiloSeq, JW, SS and NL developed the analysis pipeline, YW and JW developed the machine learning imputation and refinement method with advice from DF; JW, CP and PA performed structural and epistasis analyses; SS and FY curated the list of assayable genes; JB performed the evolution analysis; SY and BN helped conduct the blind

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

test with Invitae variant data; GT constructed ts strains; DEH and MV provided human clones; and JW, FPR and SS wrote the manuscript. FPR supervised the project.

Conflicts of interest

FPR is a shareholder and scientific advisory board member of SeqWell Inc. and of Ranomics, Inc. RN and SY are employees of Invitae, Inc.

References

1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
2. Maxwell, K. N. *et al.* Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer. *Am. J. Hum. Genet.* **98,** 801–817 (2016).
3. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).
4. Choi, Y., Yongwook, C. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31,** 2745–2747 (2015).
5. Sun, S. *et al.* An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res.* **26,** 670–680 (2016).
6. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).
7. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11,** 801–807 (2014).
8. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200,** 413–422 (2015).
9. Majithia, A. R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat. Genet.* **48,** 1570–1575 (2016).
10. Lee, M. G. & Nurse, P. Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. *Nature* **327,** 31–35 (1987).
11. Osborn, M. J. & Miller, J. R. Rescuing yeast mutants with human genes. *Brief. Funct. Genomic. Proteomic.* **6,** 104–111 (2007).
12. Seyfang, A. & Jin, J. H. Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round. *Anal. Biochem.* **324,** 285–291 (2004).
13. Jiang, W. & Koltin, Y. Two-hybrid interaction of a human UBC9 homolog with centromere proteins ofSaccharomyces cerevisiae. *Mol. Gen. Genet.* **251,** 153–160 (1996).
14. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688 (2012).
15. Doud, M. B. & Bloom, J. D. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses* **8,** (2016).
16. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
17. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 10915–10919 (1992).
18. Ng, P. C. & Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **11,** 863–874 (2001).
19. Livingstone, C. D. & Barton, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9,** 745–756 (1993).
20. St Onge, R. P. *et al.* Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat. Genet.* **39,** 199–206 (2007).
21. Gareau, J. R., Reverter, D. & Lima, C. D. Determinants of small ubiquitin-like modifier 1 (SUMO1) protein specificity, E3 ligase, and SUMO-RanGAP1 binding activities of

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

nucleoporin RanBP2. *J. Biol. Chem.* **287,** 4740–4751 (2012).
22. Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108,** 345–356 (2002).
23. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. (2014). doi:10.1101/002899
24. Bloom, J. D. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. (2016). doi:10.1101/037689
25. Andreou, A. M. *et al.* TBX22 missense mutations found in patients with X-linked cleft palate affect DNA binding, sumoylation, and transcriptional repression. *Am. J. Hum. Genet.* **81,** 700–712 (2007).
26. Mayr, J. A. *et al.* Thiamine pyrophosphokinase deficiency in encephalopathic children with defects in the pyruvate oxidation pathway. *Am. J. Hum. Genet.* **89,** 806–812 (2011).
27. Crotti, L. *et al.* Calmodulin mutations associated with recurrent cardiac arrest in infants. *Circulation* **127,** 1009–1017 (2013).
28. Nyegaard, M. *et al.* Mutations in calmodulin cause ventricular tachycardia and sudden cardiac death. *Am. J. Hum. Genet.* **91,** 703–712 (2012).
29. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–8 (2016).
30. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
31. Forbes, S. A. *et al.* in *Current Protocols in Human Genetics* (2008).
32. Kachroo, A. H. *et al.* Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science* **348,** 921–925 (2015).
33. Hamza, A. *et al.* Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants. *Genetics* **201,** 1263–1274 (2015).
34. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163,** 1515–1526 (2015).
35. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–84 (2014).
36. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350,** 1092–1096 (2015).
37. Sahni, N. *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161,** 647–660 (2015).
38. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

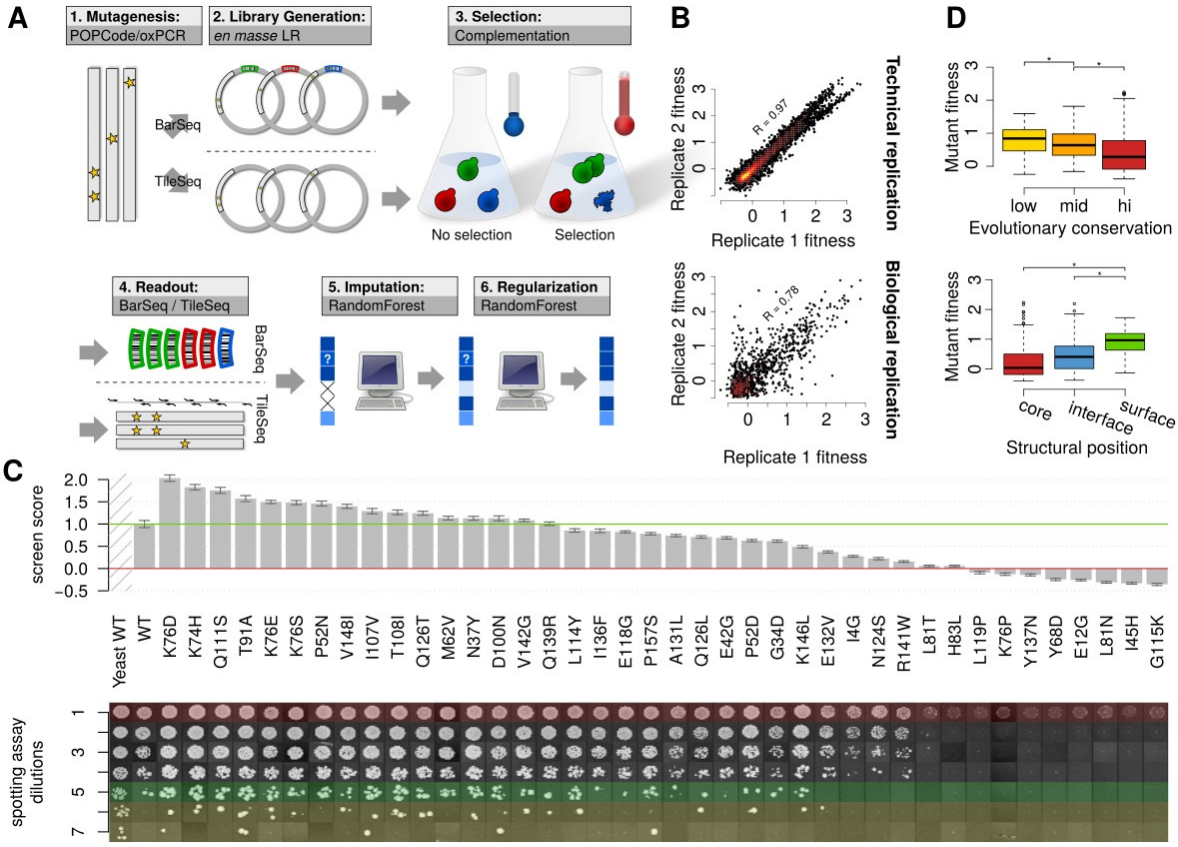

Figure 1: UBE2I screening and validation. (A) Modular structure of the screening framework. (B) Fitness scores in technical replicates (separate assays of the same pool) and biological replicates (separate substrains in the pool carrying the same variants). (C) Manual spotting assay validation of a representative set of variants. Each row represents a consecutive 5-fold dilution. Marked in red: Maximal dilution visible in empty vector control. Marked in green: Maximal dilution with visible human wt control. Marked in yellow: Dilution steps exceeding visible human wt control. Bar heights represent summary screen scores. Error bars indicate bayesian refined s.e.m. (D) Variants grouped by evolutionary conservation (AMAS score) of their respective sites (top) and grouped by structural context within the protein core, within protein-protein interaction interfaces or on remaining protein surface (bottom). See online methods for statistical details.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

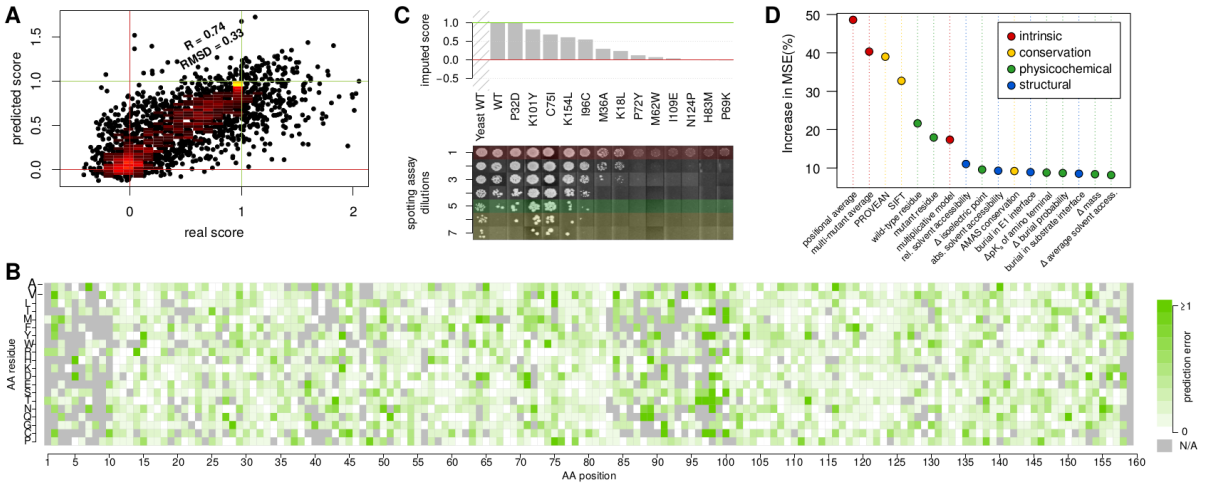

Figure 2: Validation of machine learning imputation for UBE2I. (A) Cross-validation evaluation: Scores measured in screen compared to machine learning prediction in 10x cross validation. The agreement is comparable to that between biological replicates in the screen itself (compare to Main Figure 1) (B) Error map, showing cross-validation results for each data point sorted by amino acid position and mutant residue. (C) Comparison of imputation predictions with individual spotting assays. (D) Most informative features in the RandomForest imputation, as measured in % increase in mean squared deviation upon randomization of a given feature.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

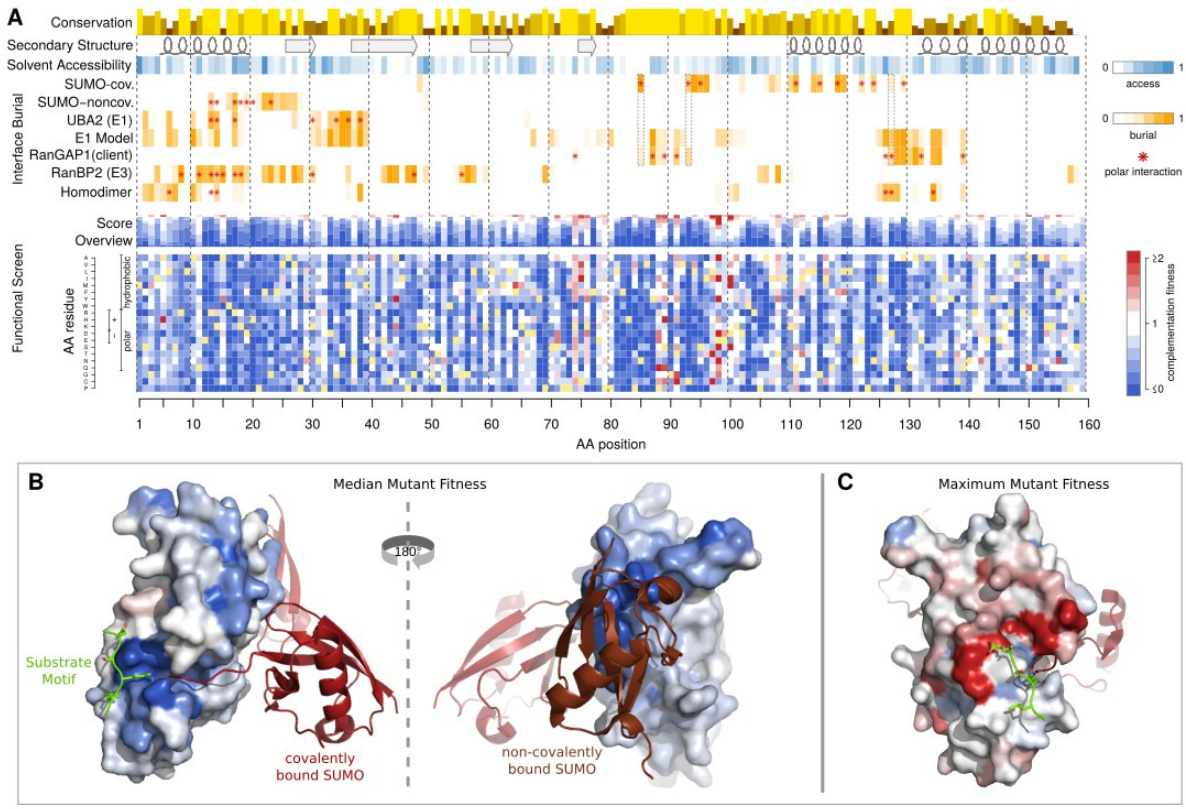

Figure 3: (A) A complete functional map of UBE2I as resulting from the combination of the complementation screen and machine learning imputation. An impact score of 0 (blue) corresponds to a fitness equivalent to the empty vector control. A score of 1 (white) corresponds to a fitness equivalent to the wildtype control. A score greater than 1 (red) corresponds to fitness above wildtype levels. Shown above, for comparison are sequence conservation, secondary structure, solvent accessibility, and burial of the respective amino acid in protein-protein interaction interfaces with covalently and non-covalently bound SUMO, the E1 UBA2, the sumoylation target RanGAP1, the E3 RanBP2 and UBE2I itself. Hydrogen bonds or salt bridges between residues and the respective interaction partner are marked with red asterisks. Residues buried in both the covalent SUMO and client interfaces are framed with dotted lines, marking the core members of the active site. (B) UBE2I crystal structure with residues colored according to the median mutant fitness. Colors as in A. The interacting substrate's ΨKxE motif is shown in green stick model; Covalently bound SUMO is shown as a red cartoon model; and non-covalently bound SUMO is shown in brown cartoon model. The structures shown were obtained by alignment of PDB entries 3UIP and 2PE6. (C) UBE2I crystal structure as in B, with residues colored according to maximum mutant fitness.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

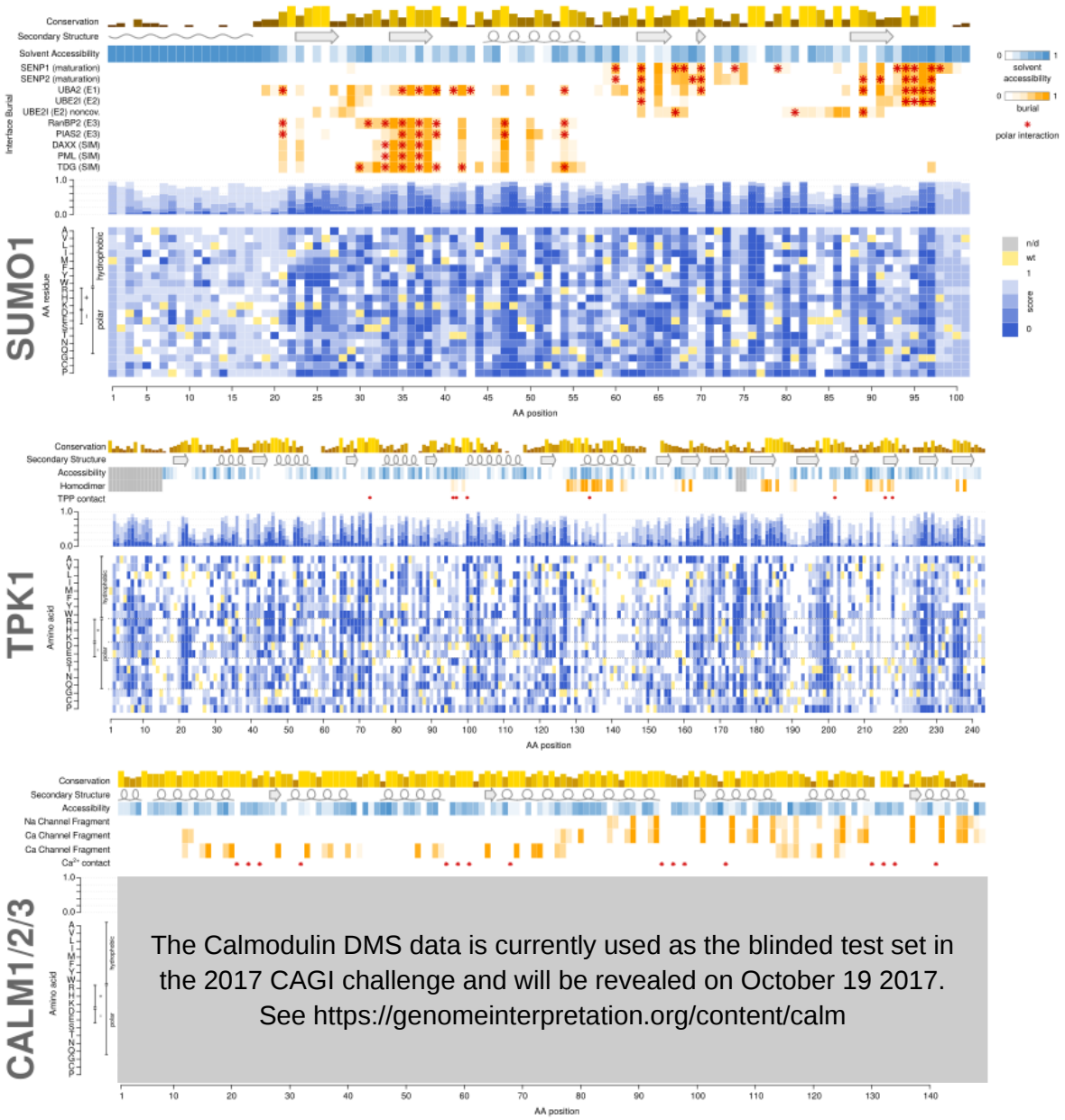

Figure 4: Functional maps of SUMO1, TPK1 and calmodulin (CALM1/2/3). Colors as in Figure 3.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

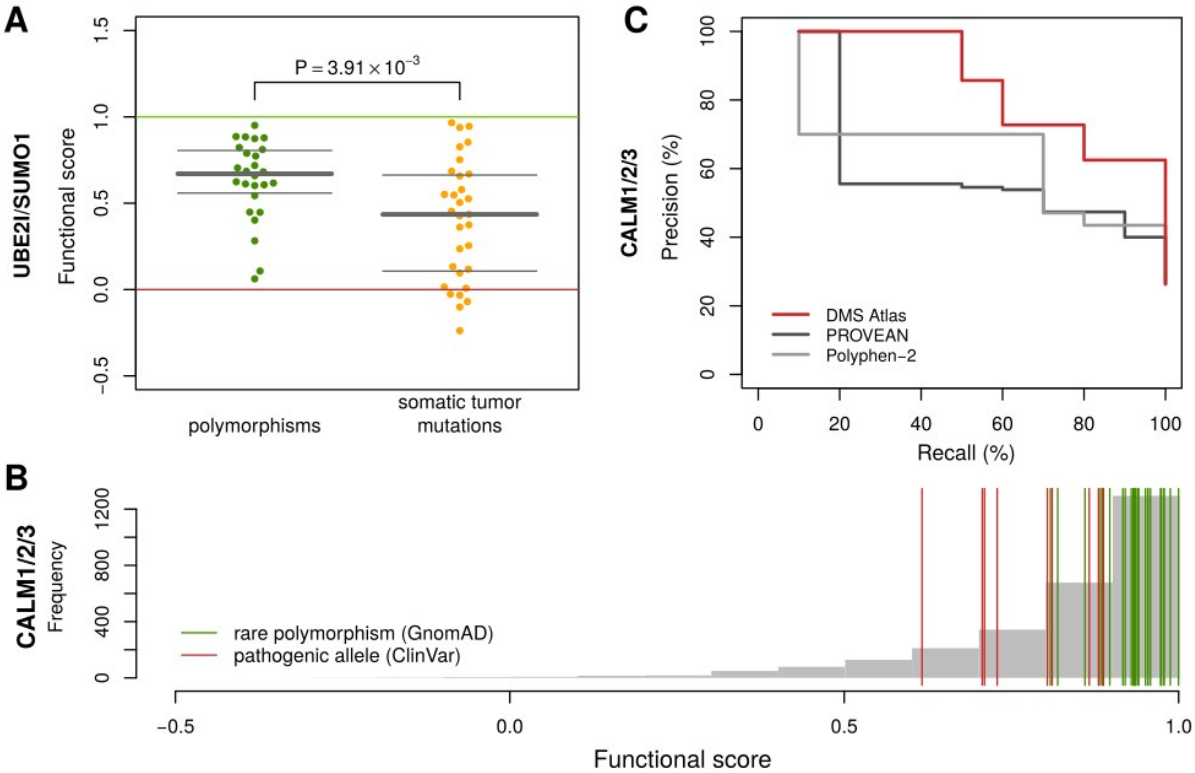

Figure 5: (A) Comparison of functional scores between rare polymorphisms (GnomAD) and somatic tumor mutations (COSMIC) in UBE2I and SUMO1. Bars show median and quartiles. One-sided Wilcoxon test, n={26,31} (unit:variants), W=570.5, P=3.73x10^{-3}. (B) Impact score distributions in calmodulin overlayed with previously observed alleles in CALM1, CALM2 and CALM3: Rare alleles from GnomAD are shown in green; ClinVar alleles classified as pathogenic are shown in red. (C) Precision-Recall Curves for our DMS atlas, PROVEAN, and PolyPhen-2 with respect to distinguishing Gnomad variants from pathogenic alleles from ClinVar.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Table 1: Map quality comparison

Gene	Possible AA changes	Achieved AA changes	Imputation RMSD	Experimental max(stderr)	Refined max(stderr)	Refinement > 0.05
UBE2I	3021	2563 (85%)	0.24	0.36	0.25	2.46%
SUMO1	1919	1700 (89%)	0.25	0.19	0.17	1.06%
TPK1	4617	3181 (69%)	0.34	0.49	0.37	5.51%
CALM1	2831	1813 (64%)	0.29	0.28	0.22	6.84%

Table 2: Invitae VUS classification

Variant	MAF	sd/rmsd	imp/reg	prereg	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	10^{-5}	0.05	mild	0.88	0.88	uncertain	Cardio
N98S	NA	0.05	mild	0.89	0.89	uncertain	Cardio
T35I	$4x10^{-6}$	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	$3x10^{-5}$	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

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Online Methods

POPCode Mutagenesis

The Precision Oligo-Pool based Code Alteration (POPCode) scales up a previous method[1]. to achieve coverage over the complete spectrum of possible amino acid changes at all protein positions. POPCode requires design of an oligonucleotide centered on each codon in the Open Reading Frame (ORF) of interest, such that the target codon is replaced with an NNK degenerate codon. This has been previously demonstrated to allow all amino acid changes while reducing the chance of generating stop codons[2]. Within each mutagenic oligonucleotide, the arm flanking the target codon is varied to achieve a predicted melting temperature that is as uniform as possible to facilitate an even mutation rate across the ORF sequence. We developed a web tool that automates this design step, available online at http://llama.mshri.on.ca/cgi/popcodeSuite/main. (See also: Code Availability section).

The POPCode mutagenesis experiment was performed via 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 degrees for 3 minutes and then cooled down to 4 degrees 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, 1X Taq buffer, 0.2mM dNTPs-dTTP, 0.2mM dUTP, 0.4uM 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 or Thermo Scientific. The phosphorylation reaction is as follows: a 50µl reaction containing 1X PNK buffer, 300 pmoles 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, 0.15uM phosphorylated oligo pool, and 1.5uM 5'-oligo was incubated at 95°C for 3 minutes followed by immediate cooling to 4°C. A 30µl reaction containing 1X 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 1X UDG buffer and 5U Uracil DNA Glycosylase (NEB) and incubated at 37°C for 2 hours.

Amplification of mutegenized DNA. 1µl UDG reaction was added to a 50µl reaction containing 1X Taq buffer, 0.2mM dNTPs, 0.4uM 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.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Single-nucleotide mutagenesis

Oxidized nucleotide PCR was performed as previously described by Mohan and colleagues[3]. Primers were designed to attach attB sites to the product in preparation for Gateway cloning.

Preparation of oxidized nucleotides. 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, 1X 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.

Library generation

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*), 1X 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 reported[4], but instead of being arrayed were maintained as pools with high complexity. Briefly, a linear PCR product containing two random 25 nucleotide "barcode" regions flanked by loxP and lox2272 sites along with common linker sequences for priming was combined with a gateway compatible vector at a SacI restriction site through *in vitro* DNA assembly[5]. 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 Clonase II Enzyme Mix, 1X 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 Clonase II Enzyme Mix, 1X 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.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

KiloSeq

For the BarSeq method, to establish the identity of each plasmid barcode and its associated set of mutations in the target ORF we used KiloSeq (either carried out in our laboratory or as a service from SeqWell Inc., Beverly, MA). 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. We then re-amplify those fragments 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.

To perform demultiplexing, barcode identification and insert resequencing, we developed a sequence analysis pipeline (see Code Availability section). In the first step Illumina bcl2fastq is used to demultiplex the reads at the plate level using the custom Nextera indices. The resulting 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 clustered by edit distance 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 software[6] is used to align reads to the ORF template, PCR duplicates are removed and nucleotide variants called using samtools pileup[7]. 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 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.

After successful genotyping with KiloSeq, we 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 (Biomatrix BM5-BC robot, S&P Robotics) into a condensed final library of 40 plates containing 6,548 clones.

High-throughput yeast based complementation screen

The yeast based functional assays were established and validated in our previous study[8]. The mutant alleles of the yeast temperature sensitive strains used in this study are u*bc9-2, smt3-331, thi80-ph, and cmd1-1*. The high-throughput screen was performed as follows: the POPCode generated mutant library was transferred to the expression vector pHYCDest[8] by en masse Gateway LR reactions followed by transformation into NEB5α competent E. coli cells (New England Biolabs) and selection for ampicillin resistance.

For the DMS-BarSeq approach, plasmids extracted from a 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 a temperature-sensitive (ts) allele which can be functionally complemented by the corresponding wild-type human gene[8]. Complexity for

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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. For each sample, plasmids were extracted from 10 ODU of cells and used as templates for the downstream barcode PCR amplification. The barcode loci were amplified for each library of plasmids with primers carrying sample-specific tags and then sequenced on an Illumina NextSeq 500.

For the DMS-TileSeq approach, plasmids extracted from a pool of ~100,000 clones were transformed into the corresponding *S. cerevisiae* temperature sensitive strain yielding around 1,000,000 total transformants. Plasmids were prepared from two of 10 ODU of cells and used as templates for the downstream tiling PCR (two replicates of non-selective condition). Two of 40 ODU of cells were inoculated into 200ml medium and grown to full density with shaking at 36°C and plasmids extracted from 10 ODU of each culture were used as templates for the downstream tiling PCR (two replicates of selective condition). In parallel, plasmid expressing the wild-type ORF was transformed to the corresponding S. cerevisiae ts strain and grown to full density under the selection. Plasmids were extracted from two of 10 ODU of cells and used as templates for the downstream tiling PCR (two replicates of wild-type control). For each plasmid library, the 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 first-step amplicon was indexed with an illumina sequencing adaptor in the second-step PCR. We perform paired end sequencing on the tiled regions across the ORF.

Fitness scoring and refinement

For DMS-BarSeq, a computational pipeline was implemented to identify and count individual sample tags and barcode combinations within each read (see Code Availability section). We first calculated the relative population size by dividing each clone's barcode count by the total number of barcodes in each condition. We then calculated the estimated absolute population size for each clone 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). We 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*, timepoint t_k at temperature τ, then $\forall i \in \{1 \leq i \leq N | i \in \mathbb{N}\}$, $\forall k \in \{1 \leq k \leq 5 | k \in \mathbb{N}\}$, $\forall \tau \in \{25°, 37°\}$

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

Where $r_{i,t_k}^{(\tau)}$ is the relative population size for clone *i*, timepoint t_k at temperature τ, $P_{i,t_k}^{(\tau)}$ is the absolute population size for clone *i*, timepoint t_k at temperature τ, $\rho_{i,t_k}^{(\tau)}$ is the measured hourly growth rate for clone *i*, timepoint t_k at temperature τ, $\phi_{i,t_k}^{(\tau)}$ is the fitness advantage relative to the pool growth for clone *i*, timepoint t_k at temperature τ, ϕ'_{i,t_k} is the normalized relative fitness advantage for clone *i*, timepoint 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 internall null and wildtype controls, this results in null-like mutants receiving a score of zero and wildtype-like mutants receiving a score of one.

Given limited amounts of replicates, the empirical standard deviations calculated for each clone or variant can be expected to be imprecise. Baldi and Long[9] have previously described a method for Bayesian regularization or refinement of the standard deviations which yield more robust estimates, leading to less classification error in statistical tests. Briefly, a prior estimate of the standard deviation is computed by linear regression based on the number of barcodes in the permissive condition and the fitness score. The prior is then combined with the empirical value 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 resulting according to the regression, n represents the degrees of freedom for the empirical data (i.e. the number of replicates) and s is the empirical standard deviation. The methods were implemented as part of a larger DMS analysis package (see Code Availability)

For DMS-TileSEQ, raw sequencing reads were aligned to the reference ORF cDNA sequences using Bowtie-2[6] and a custom Perl script was used to parse and compare the forward and reverse read alignment files to count the number of co-occurrences of a codon change in both paired reads. Mutational counts in each condition were normalized to sequencing depth at the respective position. Then, the normalized mutational counts from

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the wild type control libraries (control for sequencing errors) were subtracted from the normalized mutational counts from the non-selective and selective conditions respectively. Finally, the enrichment ratio was calculated for each variant based on the adjusted mutational counts before and after selection.

Re-scaling of fitness metrics

The results from the barcoded and regional sequencing screens do not scale linearly to each other. We 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 we can join corresponding data points using weighted means, where the weight is inversely proportional to the Bayesian regularized standard error. Output standard error was adjusted to account for differences in input fitness values and increased sample size:

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

where μ_0 is the 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. These steps were implemented as part of a larger DMS analysis package (see Code Availability)

Imputation of missing data

Next we aimed to find a machine learning method that would allow us to input the missing parts of the map. The first step towards this was to gather suitable features. We first evaluated the most promising features using linear regression and then applied a random forest model using all the available features.

The most important features were intrinsic, i.e. directly derived from unused information in the screen. These are: The average fitness across variants at the same position; The average fitness of multi-mutant clones that contain the variant of interest; the estimated fitness according to a multiplicative model to infer mutant fitness A using a double mutant AB and single mutant B. Another set of features was computed from differences between various chemical properties of the wildtype and mutant amino acids. These properties include size, volume, polarity, charge, hydropathy.A third set of features is derived from the structural context of each amino acid position. This includes 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[10]. Solvent accessibility and interface burial were calculated using the GETAREA tool[11] on the following PDB entries: For UBE2I: 3UIP[12] ; 4W5V (Boucher *et al.* unpublished) ; 3KYD[13] ; 2UYZ[14] ; 4Y1L[15]. For SUMO1: 2G4D[16]; 2IO2[17]; 3KYD[13]; 3UIP[12]; 2ASQ[18]; 4WJO[19]; 4WJQ[19]; 1WYW[20]. For calmodulin: 3G43[21]; 4DJC[22]. And for TPK1: 3S4Y[23]

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hydrogen bond and salt bridge candidates were predicted using OpenPyMol and evaluated for validity by manual inspection. Additional features used are the BLOSUM score for a given amino acid change, the PROVEAN score, 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[24], which was used as input for AMAS[25]. We then applied the complete set of features in a random forest model using the R package randomForest[26]. These procedures were implemented as part of a larger DMS analysis package (see Code Availability section).

Refinement of low-confidence measurements

The machine-learning predictions resulting generated above can also be used to refine experimental measurements of lower confidence. To this end, the corrected standard error associated with each datapoint can be used to determine the weight of assigned to the measurement.

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

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

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

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

Where μ_0 is the 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 methods were implemented as part of a larger DMS analysis package (see Code Availability section)

Experimental validation by yeast spotting assays

To validate the reliability of the fitness scores obtained during the screen, we 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. After genotype verification using Sanger sequencing, each variant was transferred to the yeast expression plasmid pHYCDest by Gateway technology and individually transformed into the yeast ts mutant 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-LEU plates as appropriate to maintain the plasmid and incubated at either the permissive or nonpermissive 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.

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Assessing relationship of hyperactive complementation to reversion

To examine whether changing amino acid residues into those residues naturally occur in yeast were more likely to show hyperactive complementation we 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[24]. 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. The methods were implemented as part of a larger DMS analysis package (see Code Availability section).

In vitro sumoylation comparison

Images from in vitro sumoylation assays performed for UBE2I variants by Bernier-Villamor et al.[27] 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 analysis package provided and also available online at https://bitbucket.org/rothlabto/dmspipeline.

Phylogenetic comparison of different models for hypercomplementation

We used the phydms software package[28] to test three different models relating the effect of complementation-enhancing substitutions in SUMO1 and UBE2I to actual preference for the substituted amino acid in a real biological context. Specifically, using the substitution models described in Bloom 2016[28], we tested 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 this variant. In the first model, $\pi_{r,a} = f_{r,a}$. In the second model, $\pi_{r,a} = \min(f_{r,a}, f_{r,wt})$ where $f_{r,wt}$ is the fitness score for the wildtype amino-acid at site r. In the third model, $\pi_{r,a} = f_{r,a}$, if $f_{r,a} <= f_{r,wt}$ and $1/f_{r,a}$ otherwise. We fit each of these models to the set of Ensembl homologs with at least 75% sequence identity to the human protein. As shown in Supplementary Table S2, in all cases the last model (which assigns low preference to variants that strongly enhance activity) best fits the sequences. The computer code that performs this analysis is available on GitHub at https://github.com/jbloomlab/AtlasPaper_SUMO1_UBE2I_ExpCM

Statistical details

Figure 1C: Error bars show Bayesian regularized standard error based on three technical replicates and a prior based on pre-selection counts and final score (see subsection on score calculation for details).

Figure 1D: As normality cannot be assumed for the distributions of fitness scores, one-sided two-sample Wilcoxon-Mann-Whitney tests were used. Low conservation (n=60 clones) vs Medium Conservation (n=105 clones) W = 3789, P = 0.015; Medium Conservation (n=105 clones) vs High Conservation (n=404 clones) W = 28043, P = 1.8×10^{-7}; Core (n=208 clones) vs Surface (n=42 clones) W = 1649, P = 1.01×10^{-10}; Interface (n=215 clones) vs Surface (n=42 clones) W = 2461, P = 1.58×10^{-6}.

Figure 5A: As normality cannot be assumed for the distributions of fitness scores, a one-sided two-sample Wilcoxon-Mann-Whitney test was used: n={26,31} variants, W=570.5, P=3.73×10^{-3}.

Code and data availability

All code associated with this work can be checked out using mercurial from the following repositories: (1) For the KiloSeq analysis pipeline: https://bitbucket.org/rothlabto/kiloseq; (2)

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for the popcode oligo design tool: https://bitbucket.org/rothlabto/popcodesuite; (3) For the BarSeq sequence analysis pipeline: https://bitbucket.org/rothlabto/screenpipeline; (4) For the TileSeq sequence analysis pipeline: https://bitbucket.org/rothlabto/tileseq_package For all raw data and downstream analyses: https://bitbucket.org/rothlabto/dmspipeline. All final variant maps and associated data tables can be downloaded at http://dalai.mshri.on.ca/~jweile/projects/dmsData/

Online References

1. Seyfang, A. & Jin, J. H. Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round. *Anal. Biochem.* **324,** 285–291 (2004).
2. Pal, G. & Fellouse, F. in *Drug Discovery Series* 111–142 (2005).
3. Mohan, U., Kaushik, S. & Banerjee, U. C. PCR Based Random Mutagenesis Approach for a Defined DNA Sequence Using the Mutagenic Potential of Oxidized Nucleotide Products. *Open Biotechnol. J.* **5,** 21–27 (2011).
4. Yachie, N. *et al.* Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* **12,** 863 (2016).
5. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).
6. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
8. Sun, S. *et al.* An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res.* **26,** 670–680 (2016).
9. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17,** 509–519 (2001).
10. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins* **23,** 566–579 (1995).
11. Fraczkiewicz, R. & Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.* **19,** 319 (1998).
12. Gareau, J. R., Reverter, D. & Lima, C. D. Determinants of small ubiquitin-like modifier 1 (SUMO1) protein specificity, E3 ligase, and SUMO-RanGAP1 binding activities of nucleoporin RanBP2. *J. Biol. Chem.* **287,** 4740–4751 (2012).
13. Olsen, S. K., Capili, A. D., Lu, X., Tan, D. S. & Lima, C. D. Active site remodelling accompanies thioester bond formation in the SUMO E1. *Nature* **463,** 906–912 (2010).
14. Knipscheer, P., van Dijk, W. J., Olsen, J. V., Mann, M. & Sixma, T. K. Noncovalent interaction between Ubc9 and SUMO promotes SUMO chain formation. *EMBO J.* **26,** 2797–2807 (2007).
15. Alontaga, A. Y. *et al.* RWD Domain as an E2 (Ubc9)-Interaction Module. *J. Biol. Chem.* **290,** 16550–16559 (2015).
16. Xu, Z. *et al.* Crystal structure of the SENP1 mutant C603S-SUMO complex reveals the hydrolytic mechanism of SUMO-specific protease. *Biochem. J* **398,** 345–352 (2006).
17. Reverter, D. & Lima, C. D. Structural basis for SENP2 protease interactions with SUMO precursors and conjugated substrates. *Nat. Struct. Mol. Biol.* **13,** 1060–1068 (2006).
18. Song, J., Zhang, Z., Hu, W. & Chen, Y. Small ubiquitin-like modifier (SUMO) recognition of a SUMO binding motif: a reversal of the bound orientation. *J. Biol. Chem.* **280,** 40122–40129 (2005).
19. Cappadocia, L. *et al.* Structural and Functional Characterization of the Phosphorylation-Dependent Interaction between PML and SUMO1. *Structure* **23,** 126–138 (2015).
20. Baba, D. *et al.* Crystal structure of thymine DNA glycosylase conjugated to SUMO-1. *Nature* **435,** 979–982 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/166595; this version posted July 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

21. Fallon, J. L. *et al.* Crystal structure of dimeric cardiac L-type calcium channel regulatory domains bridged by Ca2-calmodulins. *Proceedings of the National Academy of Sciences* **106,** 5135–5140 (2009).
22. Sarhan, M. F., Tung, C.-C., Van Petegem, F. & Ahern, C. A. Crystallographic basis for calcium regulation of sodium channels. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 3558–3563 (2012).
23. Baker, L.-J., Dorocke, J. A., Harris, R. A. & Timm, D. E. The Crystal Structure of Yeast Thiamin Pyrophosphokinase. *Structure* **9,** 539–546 (2001).
24. Sievers, F. & Higgins, D. G. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* **1079,** 105–116 (2014).
25. Livingstone, C. D. & Barton, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9,** 745–756 (1993).
26. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
27. Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108,** 345–356 (2002).
28. Bloom, J. D. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. (2016). doi:10.1101/037689