Supplementary Materials for

Systematic humanization of yeast genes reveals conserved functions and genetic modularity

Aashiq H. Kachroo, Jon M. Laurent, Christopher M. Yellman, Austin G. Meyer, Claus O. Wilke, Edward M. Marcotte

correspondence to: marcotte@icmb.utexas.edu

This PDF file includes:

Other Supplementary Materials for this manuscript includes the following:

Materials and Methods

Constructing human gene expression clones in yeast destination vectors

Human clones in Gateway entry vectors were procured from the human ORFeome library (*15*, *16*) distributed by Open Biosystems (GE Dharmacon). 549 ORFeome clones corresponding to 469 human genes (and 80 additional variants of these genes) that have 1:1 essential orthologs in yeast [based on orthology assignments from InParanoid (*8*)] were sub-cloned into yeast destination vectors (*28*) using LR clonase II (Invitrogen) (*29*). Three sets of human gene expression clones were created:

1) pAG415GAL-ccdB (+Leu; CEN) clones; human genes in expression vectors under an inducible GAL promoter with a stop codon contributed by the vector backbone resulting in a 52 residue C-terminal tail. For a subset of human genes (e.g., **Fig. S18**), the C-terminal extension blocks complementation. We therefore additionally sub-cloned all non-replacing human genes from set 1 into a redesigned "short-tail" Gateway destination vector (**Table S5**) to make clone set 2.
2) pAG415GAL-ccdB +5 Stop (+Leu; CEN) clones; human genes in expression vectors under an inducible GAL promoter with a stop codon contributed by the vector backbone resulting in a short +5 residue C-terminal tail.
3) pAG416GPD-ccdB +6 Stop (+Ura; CEN) clones; human genes in expression vectors under a constitutive GPD promoter. The stop codon was positioned such that it resulted in a short +6 residue C-terminal tail.

Sequencing confirmed that ~90% of the subclones were correct, while ~10% carried either a substitution or deletion mutation within the ORF. Erroneous clones were re-cloned from cDNA obtained from Open Biosystems or omitted from subsequent analyses.

Screening for human genes that complement essential gene function in yeast

Yeast strains were inoculated in 96 well plates in selective medium (YPD + Geneticin [200μg/ml] for all three strain collections), transformed with human gene expression or control vectors as appropriate in 96-well format, and assayed as described below for each assay type:

A. **Yeast Tet-down collection** (*17*), comprising yeast strains with ~800 essential genes controlled by expression from a tetracycline-regulatable promoter. We obtained the yeast Tet-promoter strain set (Hughes collection) from Open Biosystems (GE Dharmacon). In principle, these yeast strains should not grow or grow poorly in the presence of doxycycline (15μg/ml), which represses expression from the Tet promoter. This allows for identification of human complementing clones that support growth on selective plates containing galactose + doxycycline. Medium containing dextrose + doxycycline served as negative control. Control plasmid transformations were carried out to verify the behavior of strains across all conditions tested; strains not showing the

confirmed lethality/slow growth phenotype using control plasmids were discarded from subsequent analyses.

After transformation with the pGAL:hORF plasmids, transformants were distributed equally on four 96 well test plates, comprised of:

1. −Leu dextrose medium, serving as a control for the transformation efficiency. This condition mimics the wild-type scenario in that only the yeast gene under investigation is expected to be transcribed while the human gene is not.
2. −Leu galactose medium. In this case both the yeast and the human gene are expected to be expressed, allowing us to the test for expression-induced toxicity of the human gene.
3. −Leu dextrose + doxycycline (15μg/ml), down-regulating the yeast gene and repressing the human gene. This condition should allow for no growth, except where leaky expression from the Gal promoter allowed for complementation, serving as a complementation assay at low expression levels. In these cases, empty vector transformants served as controls to verify a lack of growth.
4. −Leu galactose + doxycycline (15μg/ml), assaying for the human gene functional replacement when expressed at higher levels while the corresponding yeast gene is down regulated. In some cases, we observed complementation at low expression levels (dextrose + dox medium), but toxicity at higher levels (galactose + dox or galactose medium).

B. **Yeast temperature sensitive (TS) collection** (*18*), comprising yeast strains with ~1300 genes carrying a temperature sensitive mutation. The strains grow at permissive temperatures (22-26^{o}C) but not restrictive temperatures (35-37^{o}C). Growth at restrictive temperatures thus allows for the identification of human clones that complement the yeast defect on selective plates.

Assays were carried out essentially as above, using both pGAL:hORF (−Leu selection) & pGPD:hORF (−Ura selection) expression vectors to test for complementation. Cases were also identified in which a human gene expressed at low levels (dextrose medium, pGAL promoter) showed functional replacement. As was described above, we carried out empty vector plasmid transformations for all strains to test the strains in all four growth conditions.

For human genes under the GAL promoter, transformants were plated on the following conditions:

1. −Leu dextrose medium at the permissive temperature (25^{o}C), serving as a control for transformation efficiency. This condition mimics the wild-type

scenario since the yeast gene under investigation is expected to be functional while the human gene is not (or only minimally) expressed.

2. −Leu galactose medium at the permissive temperature (25^{o}C). In this case both the yeast and the human gene are expressed, allowing us to the test for toxicity of the human gene.
3. −Leu dextrose at the non-permissive temperature (37^{o}C), which renders the yeast gene non-functional in the absence of human gene expression. This condition was generally expected to allow for no growth. In several cases, the human gene was sufficiently expressed to allow for complementation (as compared to empty vector control transformants).
4. −Leu galactose at the non-permissive temperature (37^{o}C), testing for human gene functional replacement under conditions in which the corresponding yeast gene is non-functional. In several cases, we observed complementation at low expression levels (dextrose at 37^{o}C), but toxicity at high expression higher levels (galactose at 25^{o}C or galactose at 37^{o}C).

In the case of human genes expressed under GPD promoter, transformants were plated on the following conditions:

1. −Ura dextrose medium at the permissive temperature (25^{o}C), serving as a control for transformation efficiency and/or toxicity since both the yeast and the human gene are expressed.
2. −Ura dextrose medium at the non-permissive temperature (37^{o}C), testing for human gene functional replacement under conditions in which the corresponding yeast gene is non-functional.

C. **Yeast Magic Marker heterozygous diploid deletion collection** (*13, 19*), comprising diploid yeast strains with one allele deleted and replaced by a KanMX kanamycin-resistance cassette, allowing for selection on G418 (200µg/ml). We obtained the collection from ATCC. We transformed human expression clones or an empty control vector into appropriate strains and selected on −Ura G418 medium in 96 well format. (Toxicity was inferred as a repeated failure to obtain transformants.) Transformants were re-plated on GNA-rich pre-sporulation medium containing G418 & 50mg/L histidine. Individual colonies were then inoculated in liquid sporulation medium containing 0.1% potassium acetate, 0.005% Zinc acetate, and incubated with vigorous shaking at 25^{o}C for 3-5 days, after which sporulation efficiency was estimated by microscopy, and the mixture then re-suspended in water and equally plated on two assay conditions:

1. “G418 minus” magic marker dextrose medium (−His −Arg −Leu +Can −Ura), incubated at 30^{o}C. The haploid spores that carry the wild-type allele grow in this medium providing us with the control for sporulation efficiency. This condition also assays for toxicity if the haploid spores fail to grow.

2. "G418 plus" magic marker dextrose medium (−His −Arg −Leu +Can −Ura) containing 200μg/ml G418. In the absence of the human gene (as for control transformants), the resulting haploid knockout strain is expected not to grow, providing an assay of replaceability for strains carrying the human gene expression clone. Cases with approximately equal numbers of cells growing in the absence or presence of G418 were considered functional replacements. For cases with ambiguous growth (marked by moderate numbers of isolated colonies growing on the +G418 medium relative to −G418 medium), we screened varying quantities of the sporulation mixtures to test for leaky histidine selection (his3 is expressed under the sporulation specific promoter pMfa1) and confirmed complementation using tetrad analysis (**Table S1**), as described below.

Confirming complementation of yeast mutants by plasmid-expression of yeast genes

As positive controls for the assay pipeline, we tested whether plasmid-born yeast genes could rescue the corresponding chromosomally-deleted strains using the magic marker collection. For 29 randomly chosen genes, we amplified open reading frames from genomic DNA of the wild-type BY4741 yeast strain, sub-cloning them into the same vector (pAG416GPD-ccdB;URA;CEN) used for the expression of human cDNA clones. Genes were sequence-verified and tested for complementation (**Table S2**). In all cases tested, the yeast genes rescued the corresponding deletion (**Table S2**). Analysis was carried out in the magic marker heterozygous knockout background as described above.

Confirming individual segregant analysis by tetrad dissection

Individual tetrad dissection and replica-plating analysis (*30*) were used to further test the complementation results previously obtained by bulk spore analysis of Magic Marker strains (**Table S1**). Magic Marker strains with complementing plasmids were individually sporulated and at least ten tetrads of each strain were dissected and scored for the presence of all segregating markers and the complementing plasmid carrying a human gene. Marker segregation was used to determine whether the viability of spores with a given gene deletion was plasmid-dependent or independent. When necessary, the plasmid dependence of complementation was further tested by replica-plating to SC −Ura + 5-FOA plates (*31*) to determine whether forced plasmid loss was lethal.

Investigation of proteasome subunits across organisms

We obtained cDNAs of *C. elegans* proteasome alpha (*pas1-5, pas7*) & beta (*pbs1-3, pbs5-7*) subunits from GE Dharmacon. cDNAs for *C. elegans* pas6, pbs4 and *X. laevis PSMB5*, *PSMB6* and *PSMB7* were custom synthesized as g-blocks from IDT. Genes were PCR amplified to add attL1 and L2 to the 5' and 3' ends of the DNA, respectively, and the purified PCR product sub-cloned using the LR reaction into the pAG416GPD-ccdB URA CEN vector. All clones were sequence-verified before transformation into an appropriate yeast strain to test functional replacement. The *Saccharomyces kluyveri* orthologs of proteasome genes were cloned as genomic fragments under the regulation of the native *S. kluyveri* promoters and terminators. Complementation was assayed by the ability of the *S. kluyveri* gene on a CEN-ARS plasmid to confer viability on haploid *S. cerevisiae* cells with the corresponding gene deletion.

Replacing yeast proteasome alpha subunits with human orthologs vs. non-orthologous alpha subunits

To test whether yeast proteasome alpha subunit complementation was specific for cognate human orthologs versus non-orthologous alpha subunits, each of five human proteasome alpha subunits cloned under the GPD promoter (in pAG416GPD-CEN Ura+, as described above) and the empty vector control were transformed into five heterozygous diploid magic marker yeast strains, each harboring a deletion (replacement by the KanMX cassette) in a distinct alpha subunit (**Fig. S9**). Transformants were inoculated in sporulation medium (described above), and sporulation mix spotted (~3µl) on magic-marker medium (2% Dextrose −Ura −Arg −His −Leu +Can) with or without G418 (200µg/ml).

Confirmation of human protein expression via Western blotting

To verify human gene expression, we sub-cloned a test set of 8 human clones *via* Gateway cloning (*29*) from ORFeome entry vectors into plasmid pAG415GAL-ccdB-HA in order to obtain HA-tagged fusion proteins under control of the inducible GAL promoter. Clones were sequence-verified and transformed into wild-type BY4741 yeast for assay. Cells were grown in 2% Galactose medium at 30°C, and ~1 ml of cells at OD 1.0 were pelleted, washed with 250 µl of 0.2M NaOH, and whole cell lysates analyzed for protein expression by Western blotting with rabbit polyclonal anti-HA antibody (Abcam) and HRP conjugated goat anti-rabbit secondary antibody (**Fig. S5**). Bands were visualized using luminol (Santa Cruz Biotechnology).

Predictive features

Sequence features

Features for human genes obtained from the ORFeome collection were calculated based on the ORFeome 7.1 (*15*, *16*) fasta file, downloaded from http://horfdb.dfci.harvard.edu/hv7/docs/human_orfeome71.tar.gz. For clones not obtained from the human ORFeome collection, we calculated sequence features using the longest annotated transcript or its translation from Ensembl version 74, available at http://Dec2013.archive.ensembl.org/index.html. Yeast sequence features were calculated using Ensembl version 74.

Sequence length:

Sc_Length
Hs_Length
ScHs_LengthDifference
ScHs_AbsLengthDifference
HsORF-HsEns74_LengthDifference
HsORF-HsEns74_AbsLengthDifference

The length of each protein was calculated from the fasta files described above. Length difference was calculated as human protein length subtracted from yeast protein length, AbsLengthDifference is the absolute value of this difference. HsORF-HsEns74 was calculated as the length of the longest annotated human protein for a given gene in Ensembl version 74 subtracted from the ORFeome clone sequence given in the ORFeome 7.1 fasta.

Sequence similarity:

ScHs_PercentIDLongest
ScHs_PercentIDAligned
ScHs_PercentSimLongest
ScHs_PercentSimAligned

Identity and similarity were calculated from global alignments using NWalign (http://zhanglab.ccmb.med.umich.edu/NW-align/) with BLOSUM62 and gap open penalty of −11 and extension −1. Orthologous pairs were calculated by InParanoid (*32*). Longest refers to identity or similarity calculated as a fraction of the longer of the two orthologs. Aligned refers to calculating identity or similarity calculated as a fraction of the length of the aligned region of the sequences.

Codon usage:

Sc_CAI
Sc_CBI
Sc_FOP
Hs_CAI
Hs_CBI
Hs_FOP

Calculated from the above fasta entries using CodonW (http://sourceforge.net/projects/codonw/). The human features were calculated using yeast optimal codons as a measure of divergence from yeast optimality.

Human transcript features:

HsEns74_AverageLength
HsEns74_TranscriptCount
HsEns74_Longest
HsEns74_Shortest

AverageLength is the average length of all annotated transcripts for a given gene in human Ensembl v74. TranscriptCount is the number of transcripts annotated for a given gene in human Ensembl v74. Longest is the length of the longest transcript for a given gene in human Ensembl v74. Shortest is the length of the shortest transcript for a given gene in human Ensembl v74.

Sequence features:

Sc_UpstreamConservation
Sc_UpstreamNonCoding
Sc_Underwrapping
Sc_KaKs
Sc_Aromaticity
Sc_3'UTRLength
Sc_5'UTRLength
Sc_RecombinationRate

Values were obtained from Vavouri *et al.* (*33*)

Aggregation propensity:

Sc_TANGO
Hs_TANGO

Calculated using the TANGO algorithm (*34*), which estimates the inherent aggregation propensity of a given protein sequence.

Additional features

Network properties

When applicable (e.g. HumanNet, YeastNet), actual weights of interactions were taken into account when calculating these features. Otherwise, a default weight of 1.0 was used. Network features were defined as follows: Degree represents the count of interaction partners for a node in a given network. Betweenness represents betweenness centrality, a measure of how central in a network a given node is, calculated as the number of shortest paths between all node pairs in a network that pass through a given node. Clustering represents the node clustering coefficient, calculated as the fraction of edges that could possibly be present in a node's neighborhood that are actually present. FracComp represents the fraction complementing, and is the fraction of interaction partners observed to complement. TestedComp represents the count of interaction partners observed to complement. TestedDegree represents the count of interaction partners that were tested in our assays.

BIOGRID:

(Hs|Sc)_BIOGRID_(Degree|Betweenness|Clustering|FracComp|TestedComp|TestedDegree)
(Hs|Sc)_BIOGRID-LT_(Degree|Betweenness|Clustering|FracComp|TestedComp|TestedDegree)

Calculated from interactions present in BIOGRID 3.1.93 (*35*). The feature BIOGRID was calculated using only those interactions annotated as 'physical interactions', while BIOGRID-LT was calculated using the subset of physical interactions found by low-throughput experiments.

Complexes:

(Hs|Sc)_Complexes_(Degree|Betweenness|Clustering|FracComp|TestedComp|TestedDegree)

Human complex features were calculated using the CORUM database (*36*). Yeast complexes were calculated using protein complexes defined in Hart *et al.* (*37*)

KEGG:

(Hs|Sc)_KEGG_(FracComp|TestedComp|TestedDegree)

KEGG features were calculated using the KEGG database (*38*). To create the KEGG network, a given pathway was represented as a clique, such that all proteins annotated as belonging to the same pathway were connected to each other by pairwise edges.

Functional Networks:

(Hs|Sc)_*Net_(Degree|Betweenness|Clustering|FracComp|TestedComp|TestedDegree)

Human and yeast functional network features were calculated based on HumanNet (*39*) and YeastNet (*40*), respectively. The final sum log-likelihood score reported for each interaction was employed as an edge weight for calculations.

Genetic interactions:

Sc_SGA_(Degree|Betweenness|Clustering)

Genetic interaction features were calculated from the Synthetic Genetic Array data of Costanzo *et al*. (*41*) using the reported intermediate cutoff for interactions.

Abundance:

(Hs|Sc)_ProteinAbundance
(Hs|Sc)_TranscriptAbundance
(Hs|Sc)_RPFAbundance
(Hs|Sc)_TranslationEfficiency

Protein abundances were used as reported by Kulak *et al.* (*42*). Transcript abundance, RPF (Ribosome Protected Fragments) abundance and Translation were calculated from Guo *et al.* (Human) (*43*) and Ingolia *et al.* (Yeast) (*44*). Translation efficiency was calculated as the ratio of RPF reads to mRNA reads.

Yeast expression features:

Sc_ProteinHalfLife
Sc_TransriptionRate
Sc_TranslationRate
Sc_mRNAHalfLife
Sc_Noise
Sc_ExpressionDivergence
Sc_Responsiveness

All values were obtained from Vavouri *et al.* (*33*)

Calculating predictive strength of features

The predictive power of each feature was calculated as the area under the receiver-operator characteristic curve (AUC) while treating each feature as an individual classifier. Each feature was sorted in both ascending and descending directions, retaining the direction providing an AUC > 0.5. To assess significance, a shuffling procedure was performed as follows: For each feature, the replaceable/non replaceable status of each ortholog pair was shuffled (retaining the original ratio of replaceable to non-replaceable assignments), and the AUC was calculated. The shuffling procedure was carried out 1,000 times for each feature, and the mean AUC values and their standard deviations reported.

To construct the integrated classifier, a subset of informative features was selected using a greedy bi-directional hill-climbing algorithm (BestFirst), limited to 10 consecutive non-improving node additions. Subsets were evaluated by considering the predictive ability of individual features and the degree of redundancy between them (CfsSubsetEval; **Table S3**). A Bayes net classifier was constructed with the selected features and evaluated by 10-fold cross-validation. The network was constructed using the K2 network search algorithm, initializing with a random order of features and restricting the maximum number of parents allowed per node to two. Conditional probability tables of the network were learned using SimpleEstimator with alpha=0.5. All calculations were performed with the Weka data-mining tool (*45*). Supplemental **Files S1** and **S2** provide the Weka input (.arff format) files for the full gene set and withheld 10 literature cases, respectively; **File S3** provides the trained BayesNet, in Weka .xml format.

Isolation of complementing *PSMB7* mutants

We mutagenized the human *PSMB7* gene by error-prone PCR (Gene Morph II random mutagenesis kit, Agilent) to a rate of ~2-4 mutations per kbp and screened for complementing mutants (**Fig. S11A**). Briefly, PCR primers included a 15 bp 3' region complementary to the human gene, and a 100 bp 5' extension incorporating the attL1 or attL2 site. The human gene was PCR-mutagenized and the gel-purified PCR fragment sub-cloned into the pAG416GPD-ccdB URA CEN vector using the LR reaction. The resulting plasmid pool of *PSMB7* variants was amplified and electroporated into the yeast magic marker heterozygous knockout strain *Pup1/ΔPup1*::KanMX, selecting

transformants on −Ura G418 (200μg/ml). Transformants were replica-plated onto sporulation medium agar (0.1% potassium acetate, 0.005 zinc acetate, 2% agar) and incubated at 25^{o}C for 3 days. Sporulation plates were replica-plated onto magic marker medium (−Arg−His−Leu−Ura+Can+G418, 2% Dex, 2% agar) and incubated at 30^{o}C. Colonies were sub-cultured into magic marker medium (−Arg−His−Leu−Ura+Can+G418, 2% Dex) and grown in a shaking incubator at 30^{o}C. Plasmid DNA was isolated and re-transformed into the heterozygous diploid parent test strain to re-test for functional replacement of yeast *PUP1* and control for suppressor mutations at other locations. Two clones successfully complemented the *pup1* deletion strain growth defect (**Fig. S11B**); these were confirmed by tetrad analysis (**Fig. S11B**) and sequenced to identify mutations. One clone encoded three amino acid substitutions (S214G, G211C, and K127Q); the second clone exhibited a D71N mutation and a nonsense mutation resulting in a truncated protein with a 14-residue deletion at the C-terminus. Each individual mutation was subsequently independently regenerated (**Table S5**) and analyzed for rescue: The (S214G), (S214G, G211C), (D71N, premature-Stop) mutants were each able to rescue the *pup1* deletion (**Fig. S12**). Because several mutations clustered around the active site, we also created a catalytically dead mutant (T44A) to assay whether catalysis was required for replacement, introducing the T44A mutation into the wild type *PSMB7* background as well as into the other rescuing mutant backgrounds. The (S214G, T44A), (G211C, T44A), (S214G, G211C, T44A) and (D71N, pre-Stop, T44A) each rescued the *pup1* deletion whereas the T44A mutation alone did not (**Fig. S12**), indicating that complementation of *ΔPup1* by *PSMB7* did not require a catalytically active protein.

Analyzing SNPs in the human mevalonate kinase (*MVK*) gene

To examine the effects of natural human genetic variation on humanization, we considered in depth the case of mevalonate kinase (*MVK* in humans & *ERG12* in yeast), which phosphorylates a key intermediate, mevalonate, during sterol synthesis. In humans, mevalonate kinase deficiency (MKD) caused by mutations of this gene results in mevalonate aciduria (MVA) and hyperimmunoglobulinaemia D with periodic fever (HIDS).

We tested the effects both of disease-associated and common *MVK* variants as follows: We observed human *MVK* to functionally replace the yeast ortholog *ERG12* in both the Tet-repressible and heterozygous diploid deletion yeast strains; we employed the Tet-repressible strain for these experiments. *MVK* mutants were generated by site-specific oligonucleotide-directed mutagenesis using standard protocols and sequence verifying all clones. Wild type and mutant *MVK* genes were cloned into yeast expression vector pAG415GAL-CEN Leu+.

Humanized yeast strains were initially cultured in synthetic defined medium −Leu +2% dextrose. (Independent assays culturing cells in −Leu +2% raffinose showed similar results.) Cultures were grown to ~2.0 OD and diluted to ~1 x 10^{8} cells/ml before spotting (3μl) with serial dilutions on media containing doxycycline and incubating at 30°C. We surveyed a range of galactose:glucose ratios to sample different degrees of activation of the Gal promoter (*46*), allowing us to distinguish relative rates of complementation conferred by different human *MVK* mutants (**Fig. S7**). Independent assays culturing cells at 37°C showed no evidence for additional temperature sensitivity of the phenotypes.

Independent assays in media containing 2% raffinose and concentrations of galactose varying from 0.005-0.1% gave qualitatively similar results.

Yeast liquid growth assays

Liquid growth assays were carried out using a Biotek synergy HT incubating photospectrometer in 96-well format. All cultures of 150µl were seeded with an initial cell density of 2.5 - 5 x 10^5 cells/ml.

Molecular modeling of evolving interaction partners

We implemented an atomistic model of a pair of interacting proteins that evolve and diverge from their ancestral state. We used this model to systematically address under what conditions the evolved proteins could still bind to their respective ancestral partners.

Evolutionary model. We developed an efficient implementation of the evolutionary process described by Sella and Hirsh (*47*). This process models an evolving, monomorphic population, which can be represented by a single genotype at each point in time. A new mutation relative to this representative genotype is tested in each time step, and it is either accepted (in which it becomes the new representative genotype) or rejected (in which case the representative genotype remains unchanged). In this model, time is proportional to the number of mutations tested. The model accurately reflects the population genetics of an evolving population under the assumption that the product of population size and mutation rate is small, $N\mu \ll 1$.

In our model, genotypes corresponded to molecular complexes. For each complex i, we calculated a fitness x_i based on the complex's energetics (see next subsection). We accepted a mutation from complex i to j according to the Metropolis criterion, with probability

$$p_{\text{accept}} = \begin{cases} 1 & \text{for } x_j > x_i, \\ e^{-2N(x_i - x_j)} & \text{for } x_j \leq x_i. \end{cases}$$

Thus, a mutation that increases fitness is always accepted, and a mutation that decreases fitness may get accepted but is exponentially unlikely to do so. The Markov process defined in this way converges to a Boltzman distribution in x_i, which is the distribution expected by population genetics according to the theory of Sella and Hirsh.

Fitness function. The fitness of a complex was calculated from the thermodynamic stabilities of the two proteins and from their binding energy. The total fitness of complex i was calculated as a sum over individual components x_i^k, $x_i = \sum_k x_i^k$, where the index k runs over the three energy contributions we calculated for each complex, two protein stabilities and one binding energy. The corresponding energy ΔG_{i}^k was converted into a fitness component via a soft threshold potential, such that sufficiently stable proteins and/or sufficiently strongly bound complexes all had the same (maximal) fitness, while less stable proteins and/or less strongly bound complexes had increasingly lower fitness:

$$x_i^k = -\log\left[e^{\beta}\left(\Delta G_i^k - \Delta G_{\text{threshold}}^k\right) + 1\right].$$

Here, $\Delta G^{k}_{\text{threshold}}$ is the energy barrier on the sigmoidal curve and β is an arbitrary scaling constant that determines the softness of the threshold.

Simulation setup. We based our model of a protein complex on the PDB structure of a yeast complex, ubiquitin-like protein Smt3 bound to SUMO-conjugating enzyme Ubc9 (PDB id: 2EKE). Prior to simulation, we cleaned the model in the PDB file: We removed one of the two complexes in the asymmetric unit of the solved structure and removed all of the water molecules. The resulting complex had two chains, chain A (SUMO-conjugating enzyme Ubc9) and chain C (ubiquitin-like protein Smt3). From chain C, we further removed the N-terminal 18 amino acids. These residues appeared in an unstructured, extended conformation; they were not involved in the binding interface and did not appear to be in a physiologically relevant conformation.

Throughout all of the simulations, the FoldX energy function was used both to minimize the energy and calculate the energy of the complex. Each of the simulations was started using the same initial minimized structure. To test a new mutation in the evolutionary trajectory, a site was selected at random from the full sequence of amino acids in the minimized structure. That site was mutated to a random, non-synonymous amino acid, and minimized with FoldX. After minimization, the stability of each of the two chains and the binding interaction energy were calculated. We then calculated the probability of acceptance as described above, and randomly accepted or rejected the mutation based on that probability. In all simulations, we set $N = 1000$ and $\beta = 10$.

In total, we tested three simulation treatments; for each one, $\Delta G^{k}_{\text{threshold}}$ was the only modified parameter. We considered scenarios in which selection maintained protein stability, stable binding between the extant partners, or both. The corresponding treatments are called “Non-Bound”, “Low Stability”, and “Wild Type” (**Table S6**). We chose the $\Delta G^{k}_{\text{threshold}}$values for these treatments as follows. For “Wild Type”, we wanted all energies to remain near their starting values throughout evolution. Therefore, the threshold for the stability of each extant chain and for binding to the extant partner was set to its initial value minus $1/N$. In addition, the evolved energies were monitored to ensure minimal deviations from their starting (ancestral) values. For the treatment “Non-Bound”, the stabilities for each chain were the same as for wild type, binding was not enforced during evolution. For the treatment “Low Stability”, binding was enforced between the extant partners at the wild type level, but the stability threshold was set to zero for both chains.

For each treatment, we performed 80 replicates and we ran each replicate for 1000 time steps (equating to 1000 tested mutations).

Analysis of ancestral binding. For each replicate, we tested whether evolved chains could still bind their respective ancestral partner (for which binding was never selected in the simulations) by computationally building complexes between evolved and ancestral chains. To make the reconstruction computationally tractable, we only tested proteins yielded by every fifth accepted mutation rather than every accepted mutation. For each time point that we analyzed, we took the chain that had just been mutated and placed it into a complex with the ancestral version of its partner. Again, the energy was minimized using FoldX, and the binding energy value was recorded. In addition, to determine the identity between the evolved and ancestral strain, we aligned the sequences and

computed site-wise identities. From 1000 attempted mutations, the final analysis yielded approximately 100 ancestral energy calculations per evolutionary trajectory. For subsequent analysis each chain was analyzed individually.

To assess successful binding between evolved and ancestral proteins, we set a hard ancestral binding energy threshold of -7.5, above which we considered binding to be impossible. Our qualitative results did not depend on the specific choice of this cutoff. We calculated successful binding as a function of sequence divergence, and for each replicate we identified the largest divergence at which the evolved protein still bound its ancestor. The resulting data have the form of a survival data set, and we analyzed them using the R package "survival".

We identified interfacial and non-interfacial mutations using biopython's PDB package. To determine whether or not a mutation was in the interface, we computed the distance from each α-carbon on the test chain to the nearest atom in the (non-self) partner chain. If there was any atom of the partner within 8Å of the α-carbon, then the site of the α-carbon was called an interface site. To calculate the odds of non-interface vs. interface mutations, we first divided the number of mutations at interface sites by the total number of interface sites; likewise, the number of non-interface mutations was divided by the number of non-interface sites. Then, the odds of acceptance of a mutation at a non-interface versus an interface site was calculated as the ratio of these two quantities. These odds were calculated for all replicates for each treatment group, yielding an odds distribution for each treatment.

Analysis of humanizing vs. random mutations. We aligned human and yeast orthologs for both Ubc9 and Smt3 proteins using MAFFT (*48*), and identified 68 amino acid substitutions for Ubc9 and 35 for Smt3. We then simulated step-wise humanization of the yeast genes by directly introducing the mutational differences, one by one and in random order, into each chain one at a time. Thus, for a single trajectory of a chain, a mutation was randomly introduced from the pool of differences of that chain, the structure was minimized with FoldX, and the binding to the yeast partner was computationally assessed. The trajectory was continued until all of the differences in that chain had been introduced. Therefore, at the end of the humanizing trajectory, the mutating chain was composed of the human protein sequence and its binding partner remained as the yeast protein. To compare this analysis to the case of random mutations, we performed the same simulation, but picked random amino acid substitutions for each trajectory of the relevant chain; for the control, the number of random mutations introduced was the same as for the humanizing runs, 68 for Ubc9 and 35 for Smt3. For both chains, we performed 256 independent trajectories for each of the computational humanization and control runs.

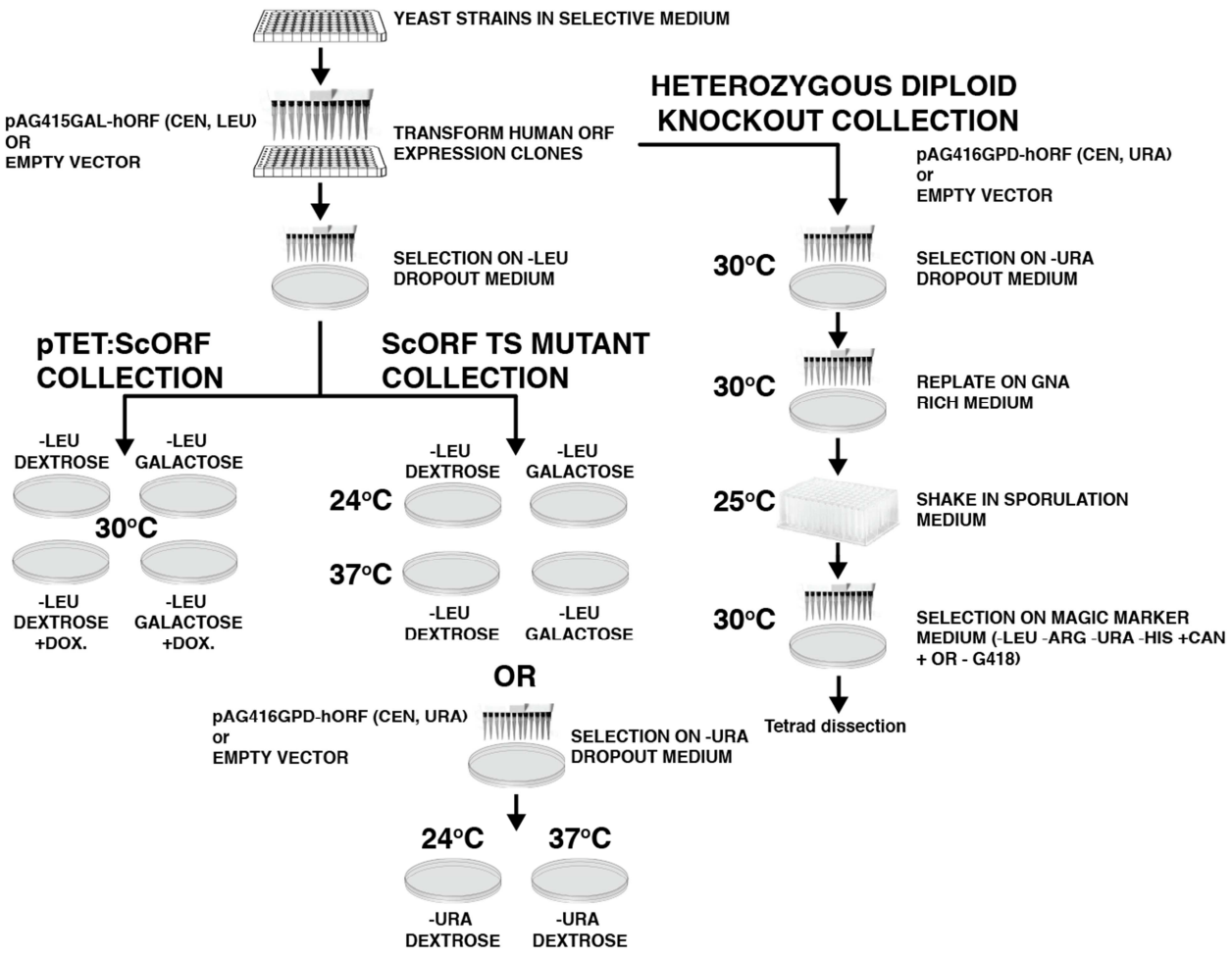

Fig. S1.

Flow chart of the experimental approach utilized to test functional complementation. Yeast strains were grown in 96-well plates in selective medium (YPD + 200 μg/ml G418). The matched orthologous human gene expression clones in 96-well plates were transformed into the appropriate yeast strains followed by selection on the appropriate dropout medium. The resulting transformants were spotted onto selective medium with appropriate markers to assay complementation. In the case of the yeast Magic Marker heterozygous knockout collection, replacement was further verified by carrying out tetrad dissection followed by testing for plasmid dependency.

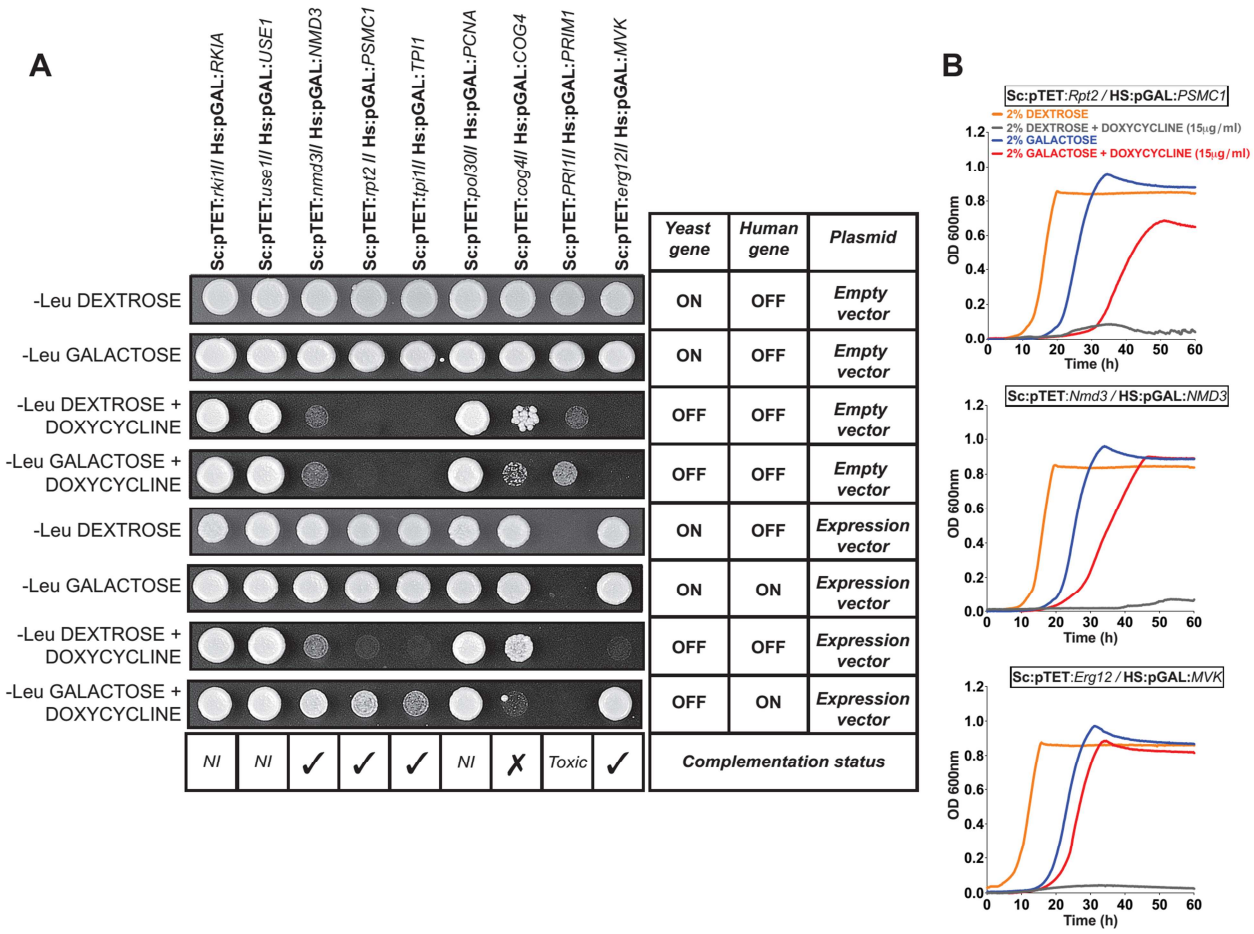

Fig. S2

Example complementation assays with Tet-repressible yeast strains. (A) Yeast strains carrying either a plasmid with a human gene under the GAL promoter or the empty vector control were initially grown in −Leu Dextrose medium without doxycycline (yeast gene ON). Cultures were diluted to ~1 x 10^8 cells/ml and spotted (3µl) on media with either dextrose (human gene OFF) or galactose (human gene ON) as a carbon source in the presence (yeast gene OFF) or absence (yeast gene ON) of doxycycline (15µg/ml). Complementation status was determined as follows: (NI) non-informative if the empty vector control strain grows in the presence of doxycycline (yeast gene OFF), (✓) complementing if the yeast grow better with the expression vector (either ON or OFF; certain clones expressed at low levels even in the presence of dextrose) than the corresponding empty vector control in the presence of doxycycline (yeast gene OFF), (✗) non-complementing if the yeast strain fails to grow better with the human gene expression plasmid (either ON or OFF, as described above) than the corresponding empty vector control in the presence of doxycycline (yeast gene OFF), and (Toxic) if the strains with the expression vector don't grow. **(B)** Qualitative plate-based growth was broadly consistent with quantitative growth curves (each curve mean of n = 3) in liquid media, shown for three examples. Strains were grown at 30 °C in four conditions: 2% dextrose without doxycycline (orange curve; yeast gene ON, human gene OFF), 2% dextrose with doxycycline (grey; yeast gene OFF, human gene OFF), 2% galactose without doxycycline (blue; yeast gene ON, human gene ON) and 2% galactose with doxycycline (red; yeast gene OFF, human gene ON).

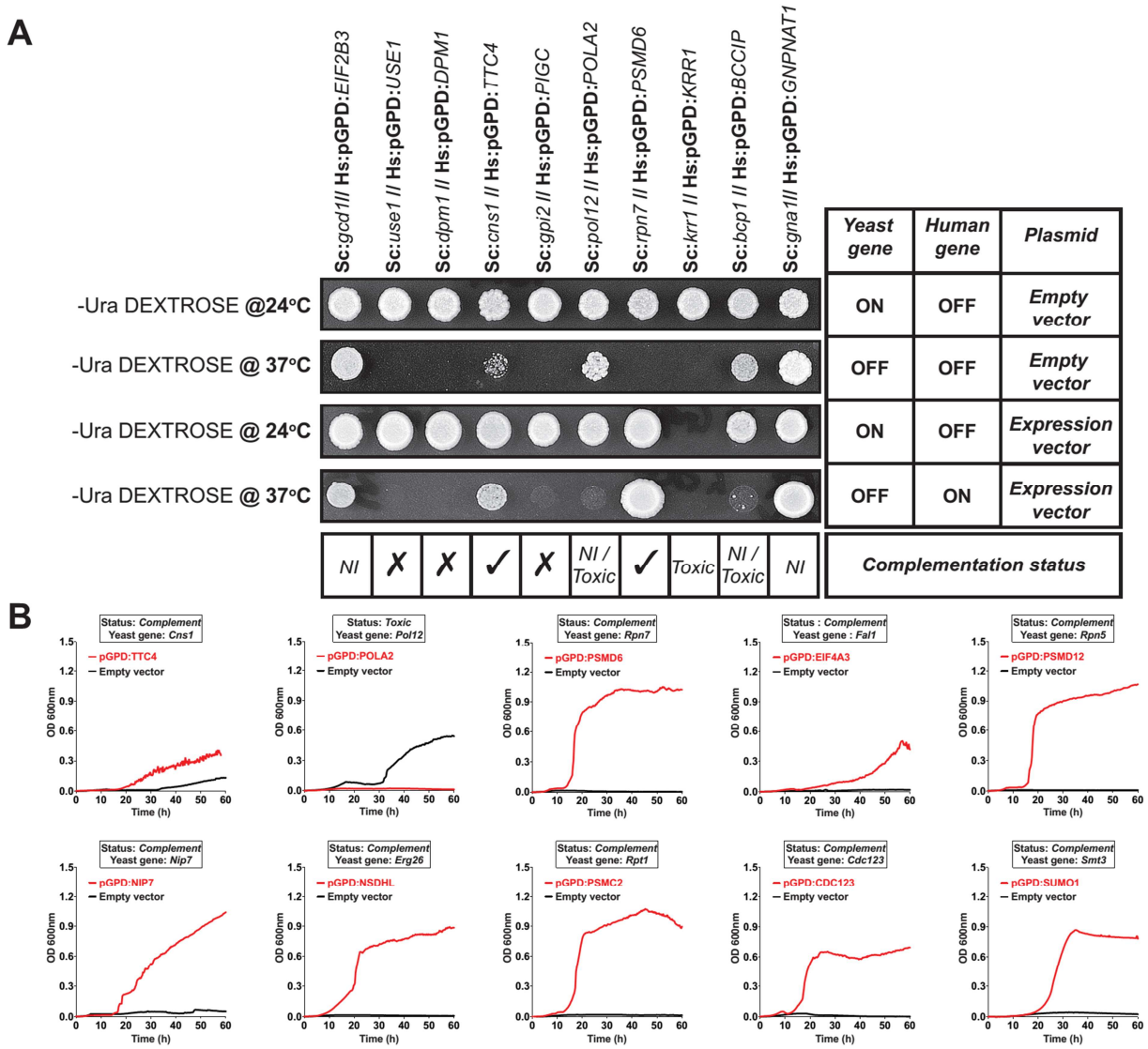

Fig. S3

Example complementation assays with temperature-sensitive yeast strains. (A) Yeast strains carrying either an empty vector (human gene OFF) or a plasmid harboring a human gene under the GPD promoter (human gene ON) were initially grown in −Ura Dextrose medium at a permissive temperature 24°C (yeast gene ON). The cultures were diluted to ~1 x 10^8 cells/ml and spotted (3µl) on SD −Ura dextrose medium. Plates were incubated either at the permissive temperature of 24°C (yeast gene ON) or the restrictive temperature of 37°C (yeast gene OFF). Complementation status was determined as follows: (NI) non-informative if the yeast strain carrying an empty vector grows at the restrictive temperature of 37°C (yeast gene OFF), (✓) complementing if the yeast grows at restrictive temperature of 37°C only when human gene is expressed, (✗) non-complementing when the particular yeast strain cannot grow or grows poorly at restrictive temperature of 37°C with either the empty vector or human gene containing plasmid, (Toxic) if the yeast strain doesn't grow at either the restrictive or permissive temperature in the presence of the human gene. **(B)** Yeast strains in liquid culture show qualitatively similar trends, e.g. as for 37°C growth curves of strains with either the empty vector (black line) or the human gene under the GPD promoter (red line) (each curve mean of *n*=3).

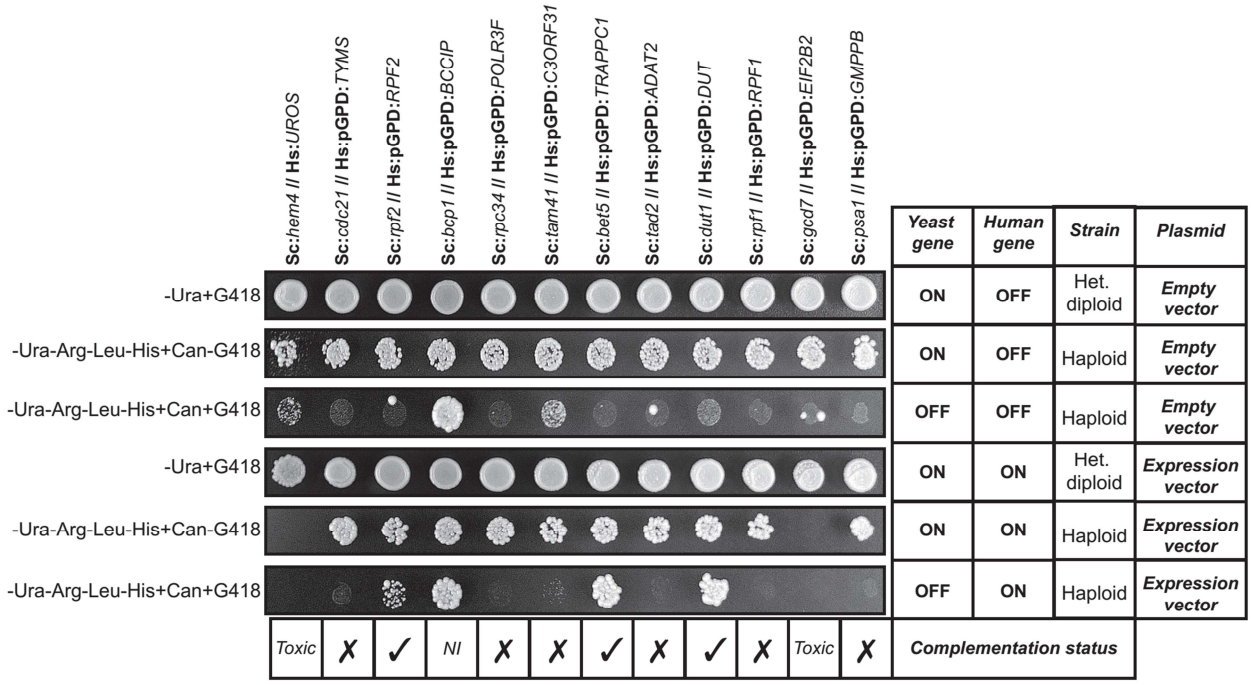

Fig. S4

Example complementation assays by sporulation of heterozygous diploid strains. Heterozygous diploid yeast strains carrying either an empty vector (human gene OFF) or a plasmid harboring a human gene under GPD promoter (human gene ON) were initially grown in −Ura Dextrose medium with G418 (200µg/ml) (yeast gene ON, human gene ON). After sporulation, the sporulation mix was spotted (3µl) on the magic marker medium (2% Dextrose −Ura −Arg −His −Leu +Can) in the presence (yeast gene OFF) or absence (yeast gene ON) of G418. Complementation status is determined as follows: (NI) non-informative when the yeast strain carrying an empty vector grows in the presence of G418 (yeast gene OFF), (✓) complementing when the yeast grows in the presence of G418 only when human gene is expressed, (✗) non-complementing when the particular yeast strain cannot grow in the presence of G418 with either empty vector or human gene containing plasmid, or (Toxic) when the yeast strain doesn't grow in the presence or absence of G418 when the human gene is expressed. Isolated colonies in the G418+ haploid selection condition (3rd row) generally corresponded to aneuploid cells escaping selection (as determined by tetrad analysis of 80 representative strains) and were ignored.

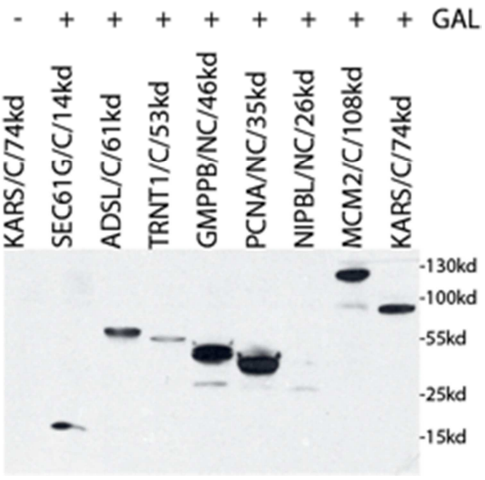

Fig. S5

Confirmation of human protein expression by Western-blotting for a random subset of ORFeome subclones. Human proteins were expressed as fusions with C-terminal HA epitope tags. Lanes are labeled with gene name, status (C=Complementing, NC=Non-complementing), and expected size including epitope tag. Serving as a negative control, the left-most gel lane contains protein extracted from cells grown without galactose induction of the GAL promoter, using the same strain as in the right-most lane.

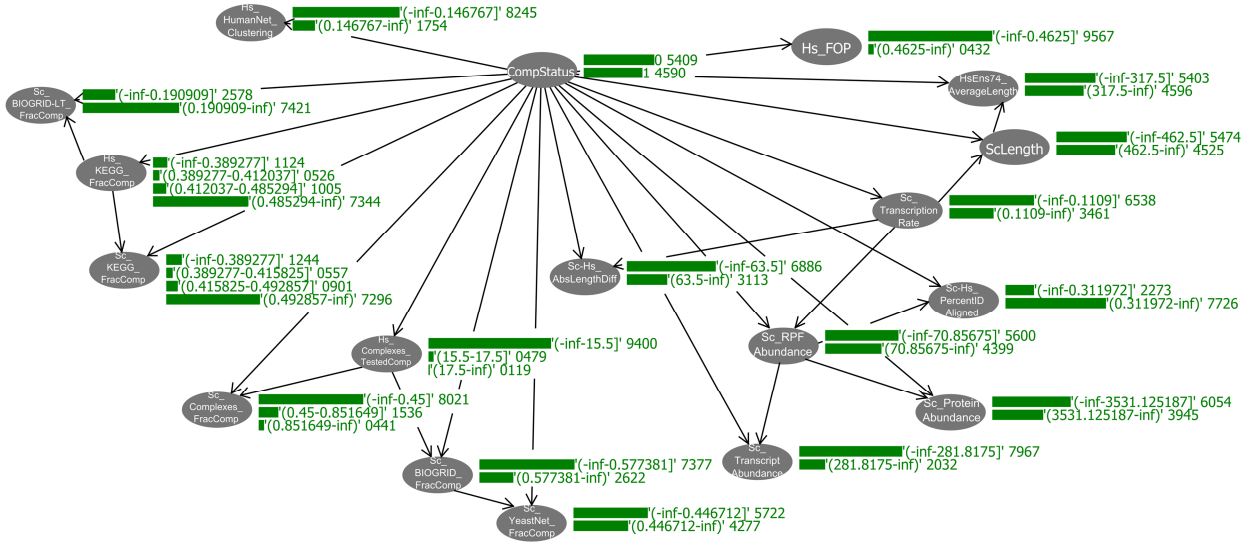

Fig. S6

Diagram of the BayesNet algorithm for predicting complementation. Nodes represent features; arrows represent dependencies. Each bar chart denotes the marginal distribution of feature values, represented as one or more numerical intervals of that feature with associated marginal probabilities (x 10^4). Conditional probability tables for each node are provided in **File S3**. Overall strengths and signs of feature association are listed in **Table S3**.

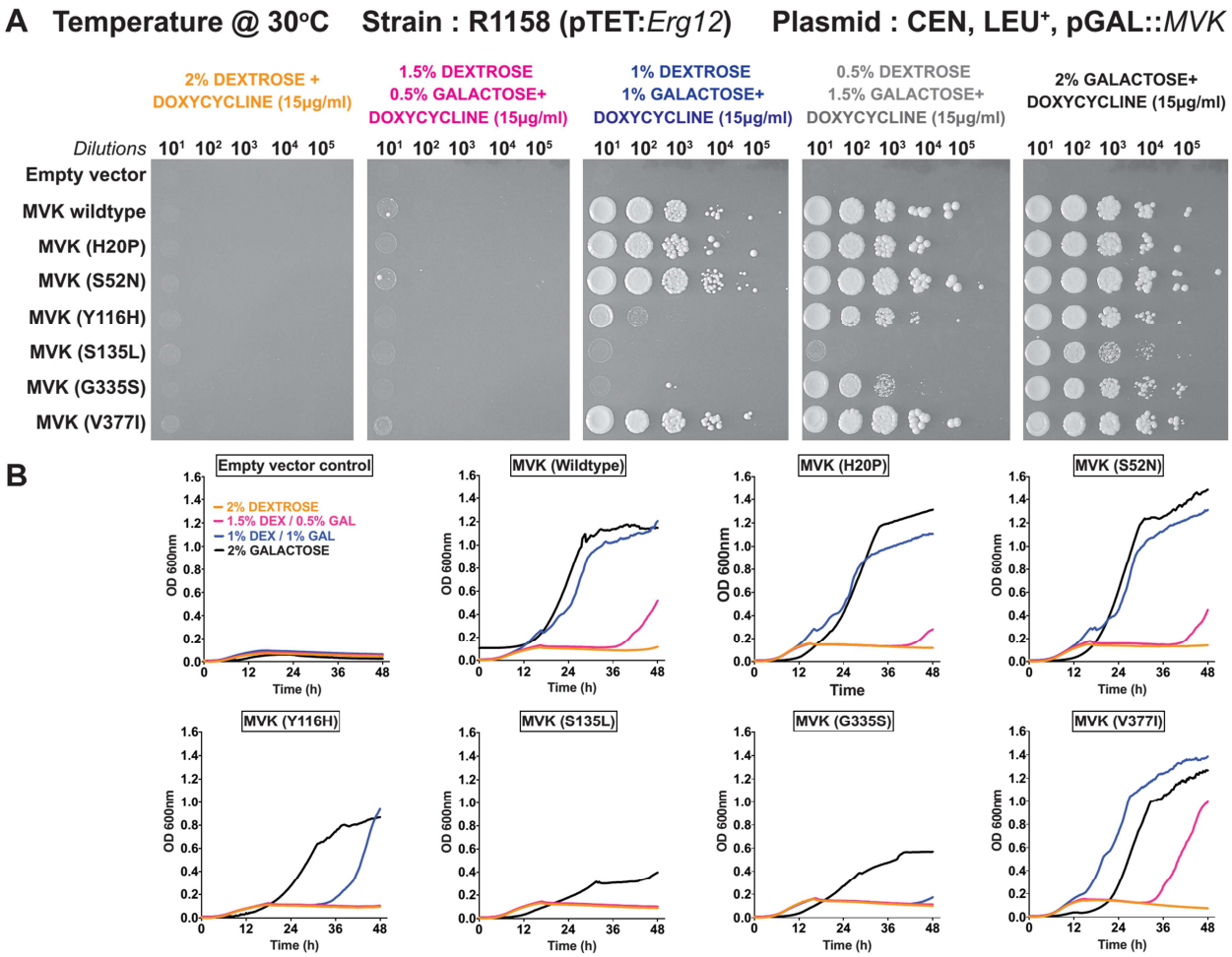

Fig. S7

Disease-specific variants of human mevalonate kinase (*MVK*) show reduced ability to replace the orthologous yeast gene. (A) Serial dilution of yeast strain (pTet:*ERG12*) carrying either an empty vector, wild type *MVK* or variants of *MVK* gene were spotted on a medium containing doxycycline (yeast gene OFF) with different ratios of dextrose and galactose. Wild type *MVK* and common *MVK* variants S52N, H20P and V377I (*49*, *50*) did not show any marked difference in their ability to complement the yeast gene function, while disease-specific variants Y116H (*51*), S135L (*52*) and G335S (*53*) did so at a lower efficiency. **(B)** Growth curves (mean of $n = 3$) of the yeast strain (pTet:*ERG12*) carrying different plasmids in SD −Leu medium with doxycycline. Varying ratios of dextrose and galactose were used to distinguish the ability to functionally replace yeast gene among variants of *MVK* gene. 2% dextrose (orange), 1.5% dextrose/0.5% galactose (pink), 1% dextrose/1% galactose (blue) and 2% galactose (black). All strains exhibited comparable growth rates in the absence of doxycycline.

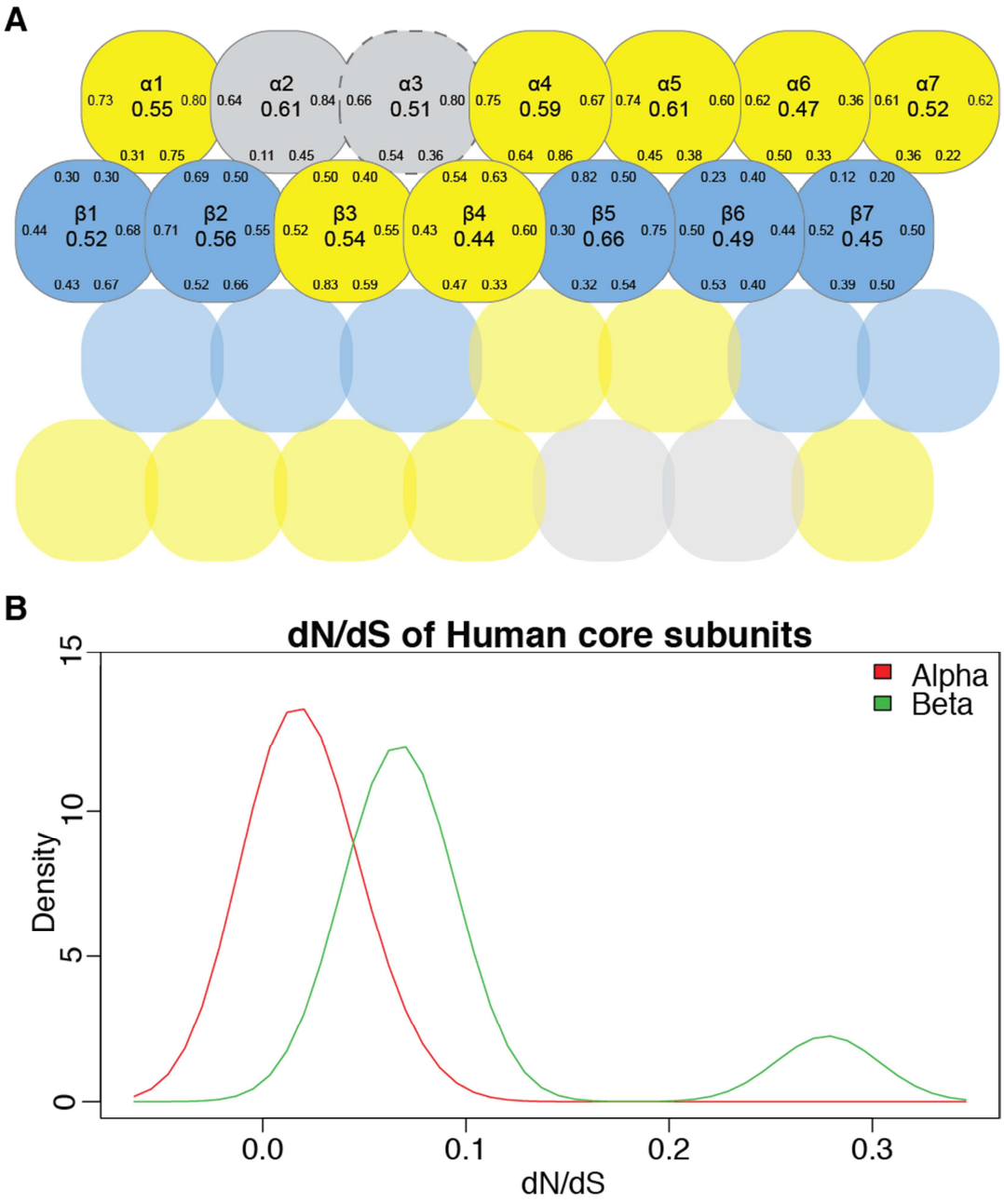

Fig. S8

Sequence identity between human and yeast 20S proteasome subunits for each subunit and subunit-subunit interface. (A) Each subunit is represented by a circle whose central value denotes that subunit's global fractional amino acid sequence identity. A value at the edge of a circle denotes the fractional amino acid identity for residues at the indicated interface, defined as residues found within 5Å of the interacting subunit in the yeast 20S proteasome structure (*21*). (**B**) Distributions of whole protein average *dN*/*dS* ratios for human alpha and beta proteasome subunits (calculated relative to mouse) as reported by Bayés *et al.* (*54*).

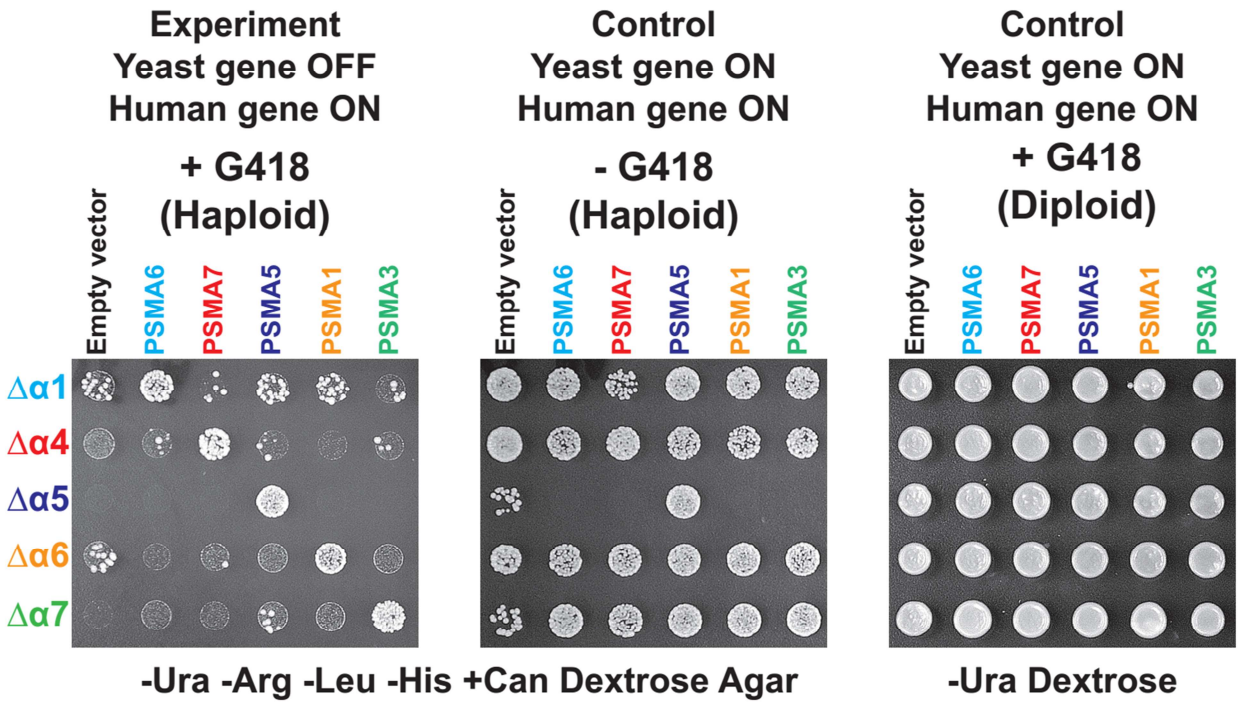

Fig. S9

Five human proteasome alpha subunits specifically replace their orthologs; non-orthologous subunits fail to replace. Five human proteasome alpha subunits were expressed in each of five heterozygous diploid yeast strains, harboring deletions of the corresponding five yeast proteasome alpha subunits (right panel). Out of all 25 tested combinations, only orthologous human sequences complemented the yeast growth defects upon sporulation (left panel). Orthologs are labeled with matching colors. As in **Fig. S4**, isolated colonies in the G418+ haploid selection condition (left panel) corresponded to cells escaping selection, most likely through aneuploidy, and were ignored. We observed the Δα1 strain to generally exhibit a higher rate of background colony formation, potentially reflecting higher aneuploidy rates. The Δα5 strain exhibited haploid-specific toxicity upon expression of the non-orthologous human genes *PSMA6*, *PSMA7*, *PSMA1*, and *PSMA3*, as indicated by strong growth defects even in the absence of G418 (middle panel).

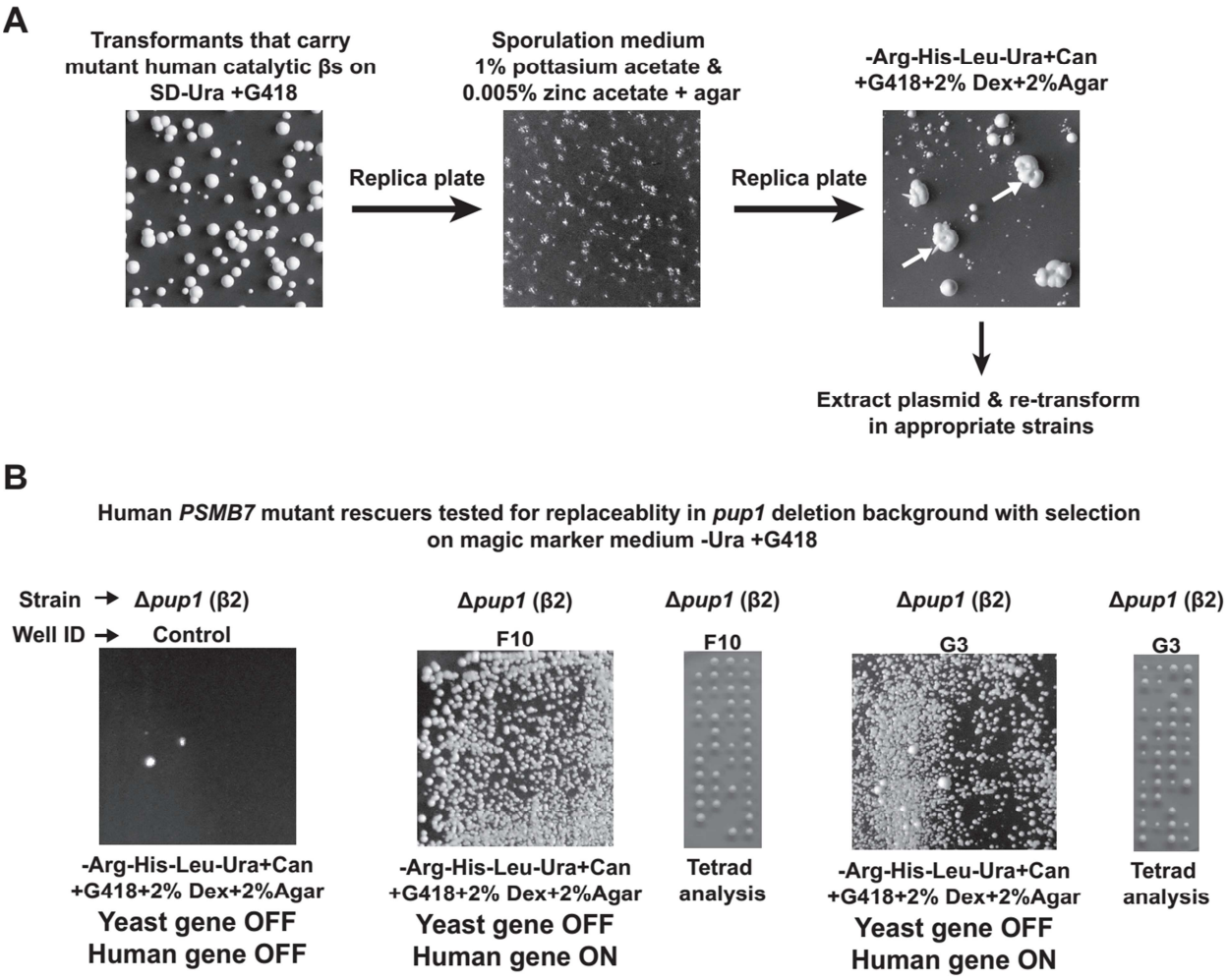

Fig. S10

Summary of mutational screen for human *PSMB7* variants complementing loss of the yeast ortholog *pup1*. (A) An error prone *PSMB7* expression clone library was transformed into the magic marker yeast heterozygous diploid knockout *Pup1*/Δ*pup1*::KanMX strain. Transformants were replica plated onto sporulation medium and plates incubated for 3-5 days at 25^{o}C. Spores were replica plated on magic marker medium with 200 μg/ml G418. Plasmids were isolated from surviving colonies and re-transformed into the Δ*pup1* strain to confirm functional replacement and control for non-plasmid based mutations. **(B)** Two clones were isolated capable of complementing the Δ*pup1* defect and were further verified by tetrad dissection. Strain F10 harbored S214G, G211C, and K127Q mutations; strain G3 harbored D71N and a premature stop codon truncating 14 C-terminal residues.

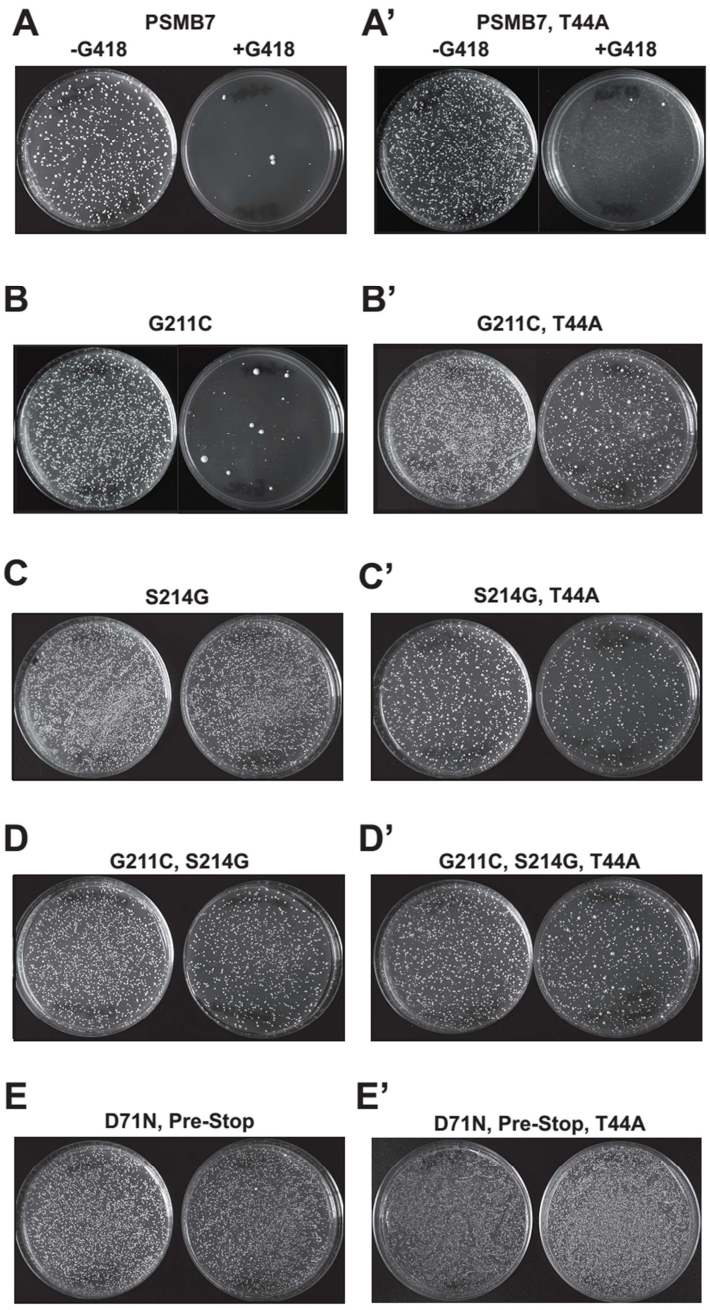

Fig. S11

Mutations in human *PSMB7* allow functional replaceability. Each individual mutation was tested for functional complementation in magic marker yeast heterozygous diploid knockout *Pup1*/Δ*pup1*::KanMX strain. **(A)** Wild-type human Psmb7 or **(A')** catalytically inactive Psmb7 (T44A) were unable to rescue the *pup1* deletion. **(B)** Psmb7 (G211C) on its own wasn't able to complement the *pup1* deletion but **(B')** with a T44A mutation rescued growth moderately. Other mutations in the *PSMB7* gene capable of complementing the yeast *PUP1* growth defect included **(C)** S214G, **(C')** S214G, T44A, **(D)** G211C, S214G, **(D')** G211C, S214G, T44A, **(E)** D71N, premature stop and **(E')** D71N, premature stop, T44A.

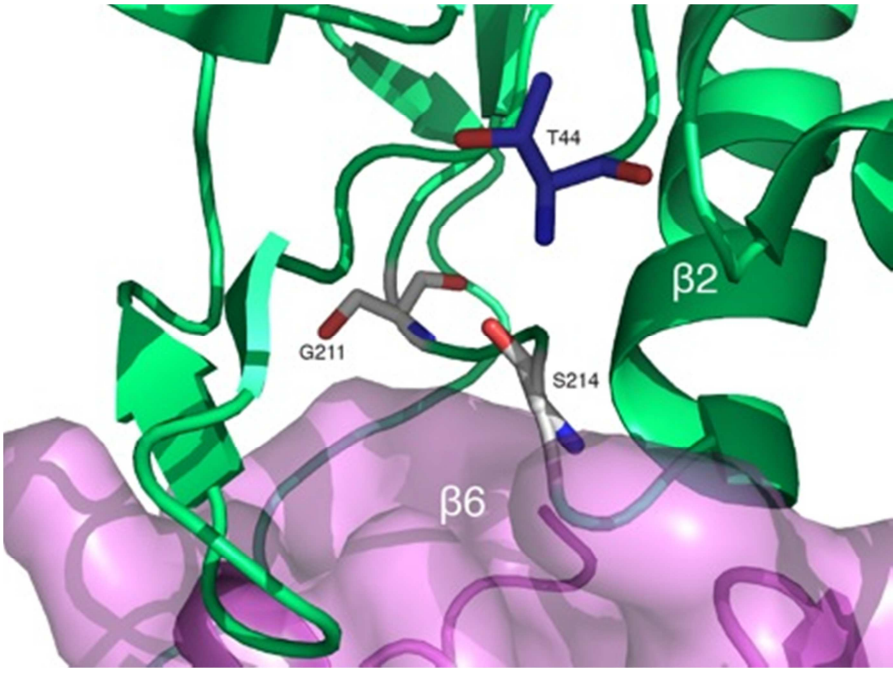

Fig. S12

Mutations enabling complementation with Psmb7, the human β2 proteasome subunit, fall near the active site threonine (T44) and the interface with the nearby yeast β6 subunit. Structure: PDB 1IRU (*55*).

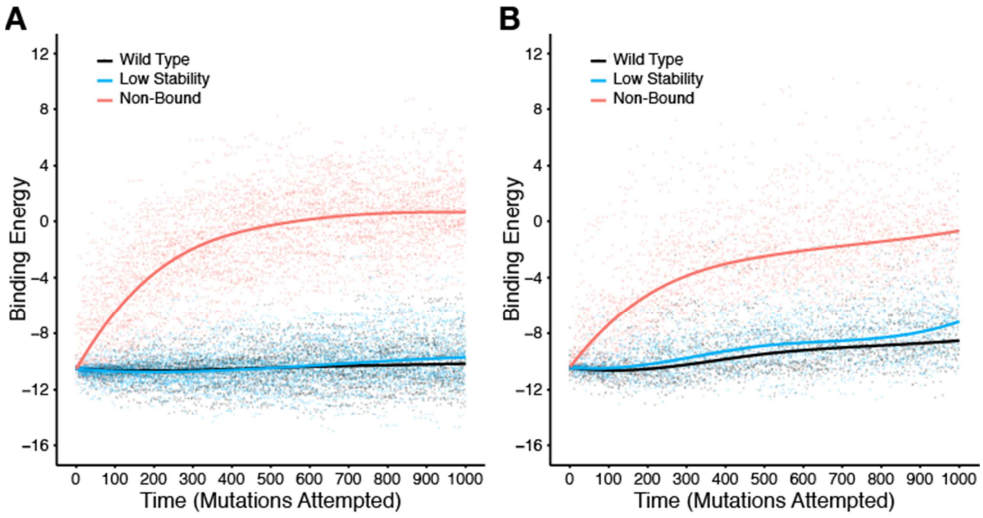

Fig. S13

Binding energy between evolved and ancestral chains vs. time, for the three simulation treatments we considered ("Wild Type", "Low Stability", "Non-Bound", see also **Table S6**). Time is proportional to the total number of mutations attempted during evolution. Dots represent individual accepted mutants from all 80 replicates for each treatment, and the solid lines show the median binding energy at that time. The ability to bind to the ancestral sequence decays rapidly when the proteins evolve without any selection pressure for successful binding (treatment "Non-Bound"), but it decays only very slowly when the proteins are selected for continued binding (treatments "Wild Type" and "Low Stability"). (**A**) Evolved chain A binding to ancestral chain C. (**B**) Evolved chain C binding to ancestral chain A.

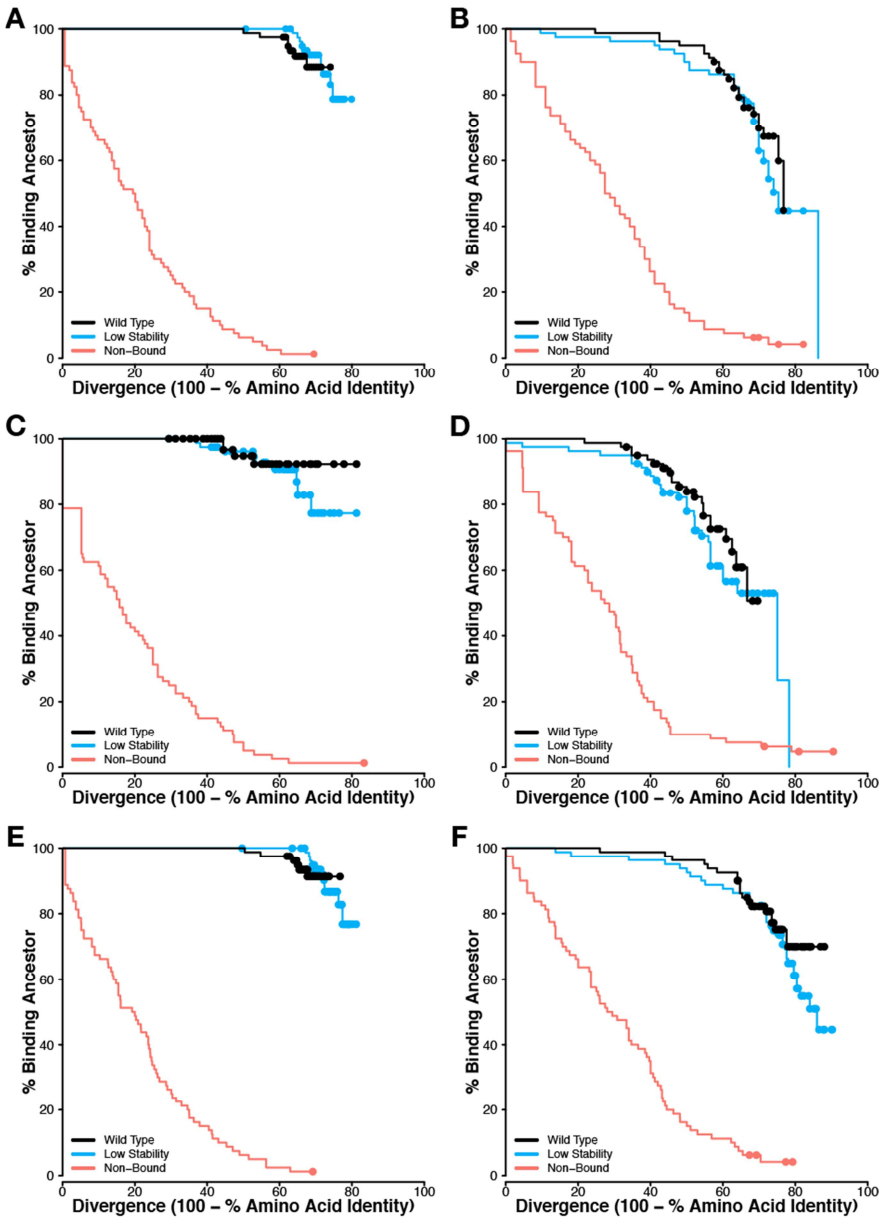

Fig. S14

Survival analysis of ancestral binding vs. sequence divergence. Solid lines indicate the percent of replicates for which the evolved chain retained the ability to bind to its ancestral binding partner up to the given amount of sequence divergence. We considered an evolved chain to be successfully bound to the ancestral partner if the binding energy fell below –7.5 (see **Fig. S13** for reference). Solid dots indicate right-censored replicates, for which the evolved chain could successfully bind to its ancestral binding partner up to the maximum observed sequence divergence. Sequence divergence was calculated for the entire sequence (**A, B**), for the binding interface only (**C, D**), and for all sites except those in the binding interface (**E, F**). In all cases, binding is typically lost by 40% divergence in the "Non-Bound" treatment but survives in the majority of replicates to 70% divergence or more in the other two treatments. (**A, C, E**) Evolved chain A binding to ancestral chain C. (**B, D, F**) Evolved chain C binding to ancestral chain A.

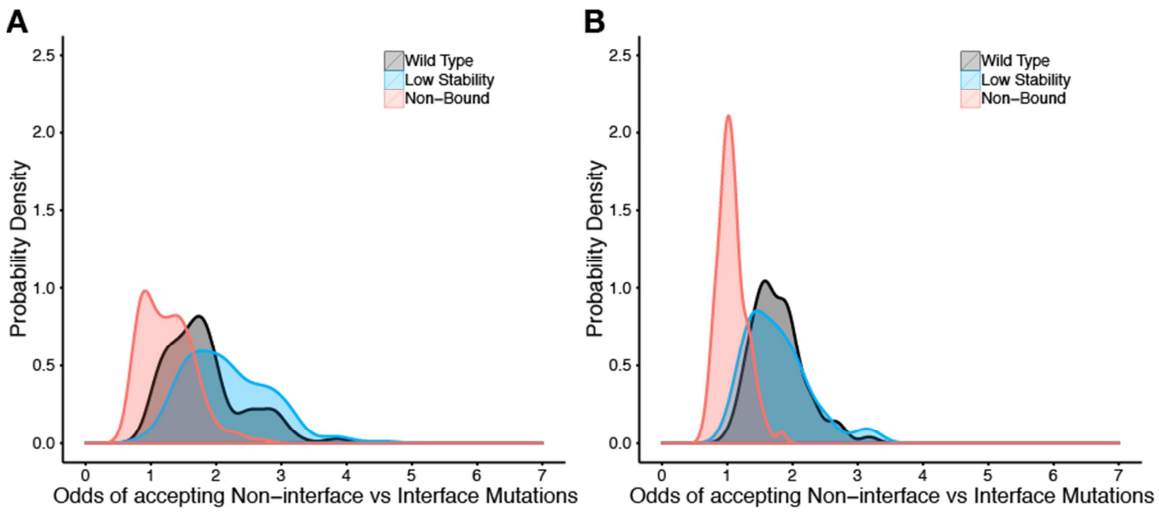

Fig. S15

Odds of accepting non-interface vs. interface mutations, over all replicates. The odds distribution has a mode near 1 for the "Non-Bound" treatment, i.e., interface and non-interface mutations are accepted equally frequently in that treatment. For the other two treatments, the odds distribution is significantly shifted towards the right, indicating that interface mutations are accepted less frequently than non-interface mutations in these treatments. (**A**) Chain A. (**B**) Chain C.

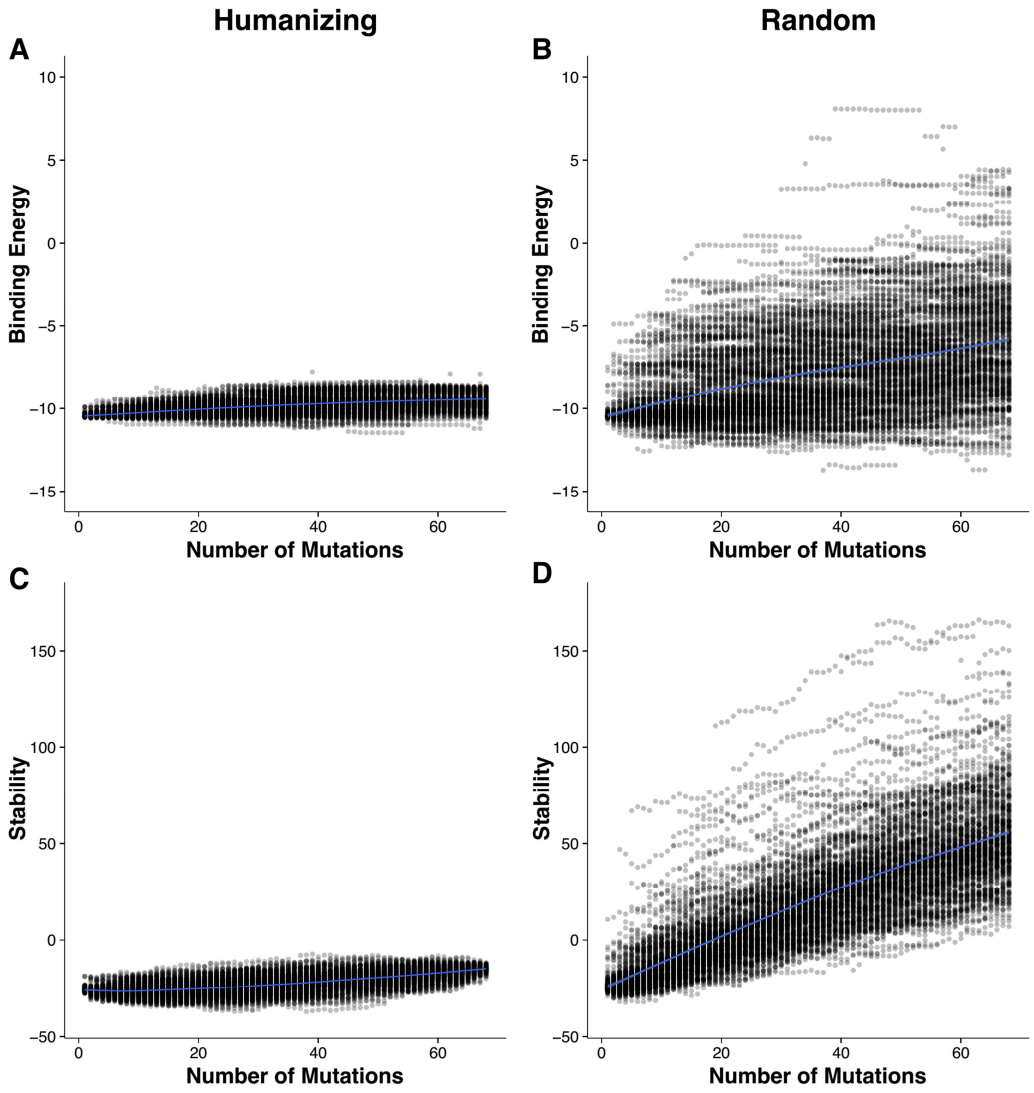

Fig. S16

Simulations suggest humanizing mutations are highly non-random within the Ubc9-Smt3 protein complex. Results are plotted for computationally humanizing (panels **A**, **C**) chain A (protein ScUbc9) of the yeast ScUbc9–ScSmt3 complex as compared to random mutations (panels **B**, **D**) in chain A, plotting the computed binding energy (panels **A**, **B**) and stability (panels **C**, **D**) of a mutated chain A bound to the yeast chain C. In the left column (panels **A**, **C**), we started from the yeast protein–protein complex and introduced, in random order, all 68 mutations which separate yeast chain A from its human ortholog. Chain C was left as the yeast protein. Dots represent mutants from 256 independent trajectories; the solid line shows the median of the distribution. In the right column (panels **B**, **D**), we repeated the analysis using 68 randomly chosen mutations. Introducing mutations from the human sequence preserves chain stability and binding to the orthologous partner, relative to random mutations.

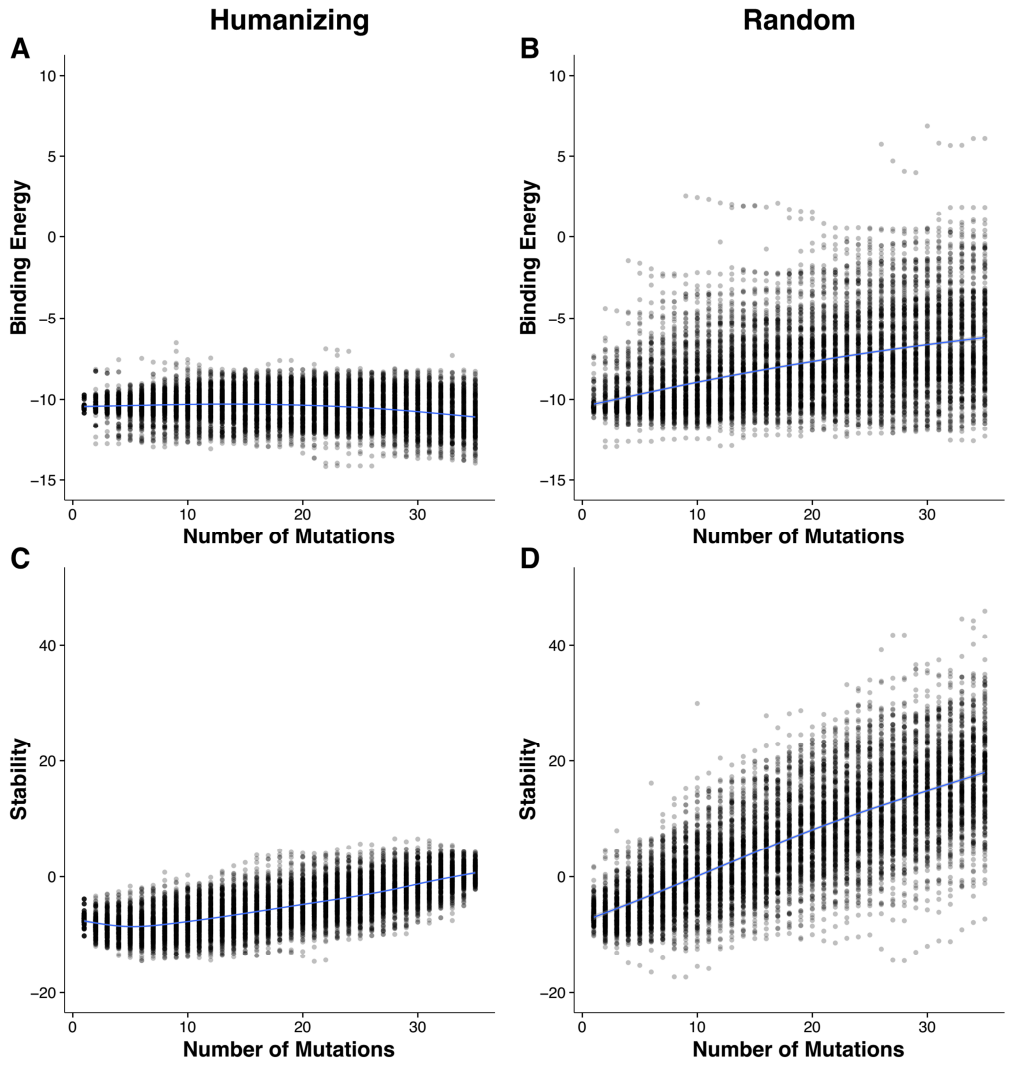

Fig. S17

Simulations suggest humanizing mutations are highly non-random within the Ubc9-Smt3 protein complex. Results are plotted for computationally humanizing (panels **A**, **C**) chain C (protein ScSmt3) of the yeast ScUbc9–ScSmt3 complex as compared to random mutations (panels **B**, **D**) in chain C, plotting the computed binding energy (panels **A**, **B**) and stability (panels **C**, **D**) of a mutated chain C bound to the yeast chain A. All labels and analyses are performed as in **Fig S16**, but considering the 35 mutations which separate yeast chain C from its human ortholog (panels **A**,**C**), versus 35 randomly chosen mutations (panels **B**, **D**). As in **Fig S16**, introducing mutations taken from the human *SMT3* ortholog preserves chain stability and binding to the orthologous partner (ScUbc9), relative to introducing random mutations.

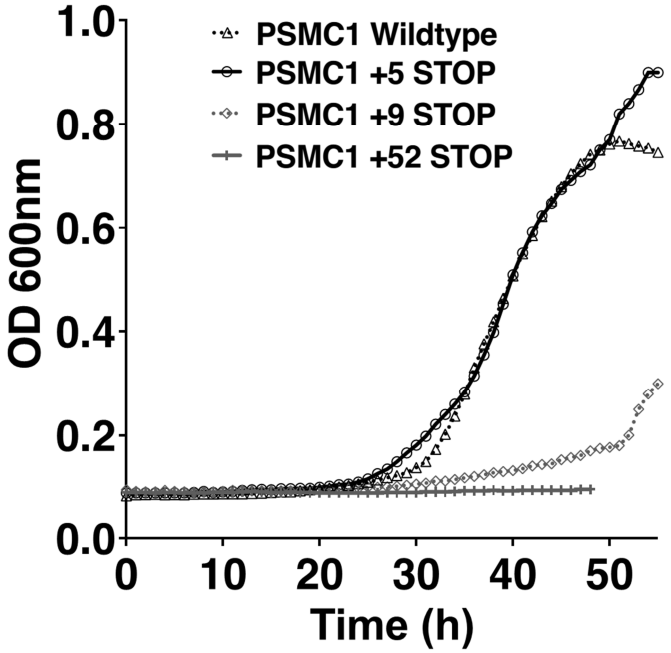

Fig. S18

Effects of Gateway-vector contributed C-terminal tails on complementation by *PSMC1*. Growth curves for the yeast strains expressing human *PSMC1* gene with native stop codon or with +5, +9, or +52 amino acid tails at the C-terminus indicate successful replacement by the wild-type and +5-aa tail construct, and poor or no complementation by the +9 or +52-aa constructs. Assays were carried out in the yeast strain pTET:*RPT2* in the presence of doxycycline (15μg/ml). Growth curves are the means of 3 independent experiments.

Table S2.

Yeast genes tested for replaceability of the corresponding deletion at the genomic locus. Yeast genes were cloned as in human expression plasmid collection #3.

Yeast systematic name	Yeast gene name	Human ORF status		Yeast ORF status
YPR180W	AOS1	Non complement		Complement
YBR155W	CNS1	Complement		Complement
YPR105C	COG4	Non complement	Toxic	Complement
YPL266W	DIM1	Complement	Toxic	Complement
YPR183W	DPM1	Complement	Toxic	Complement
YKR071C	DRE2	Complement		Complement
YDL205C	HEM3	Complement		Complement
YKR063C	LAS1	Non complement	Toxic	Complement
YOL135C	MED7	Non complement		Complement
YLR116W	MSL5	Non complement	Toxic	Complement
YER012W	PRE1	Complement		Complement
YOR362C	PRE10	Complement		Complement
YPR103W	PRE2	Non complement		Complement
YJL001W	PRE3	Non complement		Complement
YFR050C	PRE4	Non complement		Complement
YMR314W	PRE5	Complement		Complement
YOL038W	PRE6	Complement		Complement
YBL041W	PRE7	Non complement		Complement
YIR008C	PRI1	Non complement	Toxic	Complement
YDL030W	PRP9	Non complement		Complement
YOR157C	PUP1	Non complement		Complement
YGR253C	PUP2	Complement		Complement
YER094C	PUP3	Complement		Complement
YER021W	RPN3	Non complement		Complement
YGL011C	SCL1	Complement		Complement
YDR292C	SRP101	Complement	Toxic	Complement
YDR246W	TRS23	Non complement		Complement
YER093C	TSC11	Non complement		Complement
YGL098W	USE1	Non complement	Toxic	Complement

Table S4.
Summary of BayesNet performance on withheld literature test set.

Yeast gene	HsENSP74	HsEntrez	Literature Status	Predicted Status	Correct	P(Complement)
YHR143W-A	ENSP00000430106	5440	Complement	Complement	1	0.998
YJR007W	ENSP00000256383	1965	Complement	Complement	1	0.993
YGR267C	ENSP00000419045	2643	Complement	Complement	1	0.963
YGL040C	ENSP00000386284	210	Complement	Complement	1	0.942
YDL097C	ENSP00000261712	5717	Complement	Complement	1	0.857
YOR057W	ENSP00000367208	10910	Complement	Complement	1	0.732
YHR069C	ENSP00000361433	23404	Complement	Complement	1	0.731
YOR335C	ENSP00000261772	16	Complement	Complement	1	0.518
YAL035W	ENSP00000289371	9669	Complement	Non complement	0	0.377
YHR027C	ENSP00000310129	5708	Complement	Non complement	0	0.319

Table S6.

Values of $\Delta G^{k}_{\text{threshold}}$ for the three simulation treatments considered. Each treatment is defined by three $\Delta G^{k}_{\text{threshold}}$ values, one for chain A stability, one for chain C stability, and one for the binding energy between the two chains. The $\Delta G^{k}_{\text{threshold}}$ values for the "Wild Type" treatment were chosen such that the corresponding ΔG^{k}_{i} values remained stationary in that treatment.

	$\Delta G^{k}_{\text{threshold}}$		
Simulation treatment	Chain A stability	Chain C stability	Binding energy
Wild Type	–23.0	–5.0	–9.7
Low Stability	0.0	0.0	–9.7
Non-Bound	–23.0	–5.0	+∞

Additional Data Table S1 (separate file)
Summary of complementation results.

Additional Data Table S3 (separate file)
Summary of features and prediction performance.

Additional Data Table S5 (separate file)
Oligonucleotide primers used in this study.

Additional Data File S1 (separate file)
Weka machine learning feature file for the main gene set (.arff format).

Additional Data File S2 (separate file)
Weka machine learning feature file for the withheld literature test set (.arff format).

Additional Data File S3 (separate file)
BayesNet classifier (Weka .xml format).

References

28. S. Alberti, A. D. Gitler, S. Lindquist, *Yeast*. **24**, 913–919 (2007).
29. X. Liang, L. Peng, C.-H. Baek, F. Katzen, *BioTechniques*. **55**, 265–268 (2013).
30. D. Burke, Cold Spring Harbor Laboratory, *Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual* (Cold Spring Harbor Laboratory Press, Plainview, N.Y, 2000 ed., 2000).
31. C. Kaiser, Cold Spring Harbor Laboratory, *Methods in yeast genetics: a Cold Spring Harbor Laboratory course manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1994 ed., 1994).
32. M. Remm, C. E. Storm, E. L. Sonnhammer, *J. Mol. Biol.* **314**, 1041–1052 (2001).
33. T. Vavouri, J. I. Semple, R. Garcia-Verdugo, B. Lehner, *Cell*. **138**, 198–208 (2009).
34. A.-M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, *Nat. Biotechnol.* **22**, 1302–1306 (2004).
35. C. Stark *et al.*, *Nucleic Acids Res.* **34**, D535–539 (2006).
36. A. Ruepp *et al.*, *Nucleic Acids Res.* **36**, D646–650 (2008).
37. G. T. Hart, I. Lee, E. R. Marcotte, *BMC Bioinformatics*. **8**, 236 (2007).
38. H. Ogata *et al.*, *Nucleic Acids Res.* **27**, 29–34 (1999).
39. I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, E. M. Marcotte, *Genome Res.* **21**, 1109–1121 (2011).
40. I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte, *Science*. **306**, 1555–1558 (2004).
41. M. Costanzo *et al.*, *Science*. **327**, 425–431 (2010).
42. N. A. Kulak, G. Pichler, I. Paron, N. Nagaraj, M. Mann, *Nat. Methods*. **11**, 319–324 (2014).
43. H. Guo, N. T. Ingolia, J. S. Weissman, D. P. Bartel, *Nature*. **466**, 835–840 (2010).
44. N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, J. S. Weissman, *Science*. **324**, 218–223 (2009).
45. E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten, *Bioinforma. Oxf. Engl.* **20**, 2479–2481 (2004).
46. R. Escalante-Chong *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1636–1641 (2015).
47. G. Sella, A. E. Hirsh, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9541–9546 (2005).
48. K. Katoh, D. M. Standley, *Mol. Biol. Evol.* **30**, 772–780 (2013).
49. S. M. Houten, C. S. van Woerden, F. A. Wijburg, R. J. A. Wanders, H. R. Waterham, *Eur. J. Hum. Genet. EJHG*. **11**, 196–200 (2003).
50. A. D'Osualdo *et al.*, *Eur. J. Hum. Genet. EJHG*. **13**, 314–320 (2005).
51. M. Leyva-Vega *et al.*, *Am. J. Med. Genet. A.* **155A**, 1461–1464 (2011).
52. I. Koné-Paut, E. Sanchez, A. Le Quellec, R. Manna, I. Touitou, *Ann. Rheum. Dis.* **66**, 832–834 (2007).
53. M. Nevyjel *et al.*, *Pediatrics*. **119**, e523–527 (2007).
54. A. Bayés *et al.*, *Nat. Neurosci.* **14**, 19–21 (2011).
55. M. Unno *et al.*, *Struct. Lond. Engl. 1993*. **10**, 609–618 (2002).