PLOS GENETICS

RESEARCH ARTICLE

Identifying pathogenicity of human variants via paralog-based yeast complementation

Fan Yang[1,2,3,4], Song Sun[1,2,3,4,5], Guihong Tan[1,2], Michael Costanzo[1,2], David E. Hill[6,7], Marc Vidal[6,7], Brenda J. Andrews[1,2], Charles Boone[1,2,8], Frederick P. Roth[1,2,3,4,6,8]*

1 Donnelly Centre, Toronto, Ontario, Canada, **2** Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada, **3** Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, **4** Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada, **5** Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, **6** Center for Cancer Systems Biology (CCSB), Dana- Farber Cancer Institute, Boston, Massachusetts, United States of America, **7** Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America, **8** Canadian Institute for Advanced Research, Toronto, Ontario, Canada

* fritz.roth@utoronto.ca

Abstract

To better understand the health implications of personal genomes, we now face a largely unmet challenge to identify functional variants within disease-associated genes. Functional variants can be identified by trans-species complementation, e.g., by failure to rescue a yeast strain bearing a mutation in an orthologous human gene. Although orthologous complementation assays are powerful predictors of pathogenic variation, they are available for only a few percent of human disease genes. Here we systematically examine the question of whether complementation assays based on paralogy relationships can expand the number of human disease genes with functional variant detection assays. We tested over 1,000 paralogous human-yeast gene pairs for complementation, yielding 34 complementation relationships, of which 33 (97%) were novel. We found that paralog-based assays identified disease variants with success on par with that of orthology-based assays. Combining all homology-based assay results, we found that complementation can often identify pathogenic variants outside the homologous sequence region, presumably because of global effects on protein folding or stability. Within our search space, paralogy-based complementation more than doubled the number of human disease genes with a yeast-based complementation assay for disease variation.

Author summary

Functional complementation assays of human disease-associated gene variants can reveal many more human disease variants at high confidence than current computational approaches, even using highly-diverged model organisms. However, this has generally only been possible for a minority of human disease genes for which orthologous complementation is known in the relevant model organism, so that alternative assays are urgently needed. Here we show that complementation relationships can be found for many additional human disease genes by exploiting paralogous human-yeast gene relationships, and

Citation: Yang F, Sun S, Tan G, Costanzo M, Hill DE, Vidal M, et al. (2017) Identifying pathogenicity of human variants via paralog-based yeast complementation. PLoS Genet 13(5): e1006779. https://doi.org/10.1371/journal.pgen.1006779

Editor: Gregory M. Cooper, HudsonAlpha Institute for Biotechnology, UNITED STATES

Received: November 25, 2016

Accepted: April 25, 2017

Published: May 25, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files, excepting image files for negative results which can be found at http://dx.doi.org/10.5061/dryad.j05n0.

Funding: We acknowledge support from National Institutes of Health/National Human Genome Research Institute (NIH/NHGRI) Centers of Excellence in Genomic Sciences grant HG004233 (MV and FPR), NIH/NHGRI grant HG001715 (MV, DEH, and FPR), the Canada Excellence Research Chairs Program (FPR), a Swedish Research

Council International Postdoc Grant (SS), National Institutes of Health grant HG005853 (BJA and CB), and Canadian Institutes of Health Research grants FDN-143264 and FDN-143265 (BJA and CB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

that disease variant identification using paralogy-based assays performs on par with orthology-based assays.

Introduction

As a result of rapid developments in sequencing technology, we are identifying many rare variants in individual human genomes [1]. To fully exploit this resource, we must be able to rapidly identify which of the many variants in each individual are most likely to be functional and disease-causing.

Yeast remains an extremely useful model organism for studying gene functions [2, 3], genetic interactions [4], protein-protein interactions [5–7], and genotype-phenotype relationships [8, 9]. The scale of experiments in yeast ranges from individual assays to high-throughput genome-wide experiments [10–12]. Of the ~6000 genes in yeast only about 15% are completely un-annotated with a function, and even for these genes there are clues from a wide range of large-scale experiments. Core cellular biology is well conserved between yeast and humans, with ~60% of yeast genes having human homologs and 87% of yeast protein domains being present in a human protein [13]. Functional complementation assays using model organisms can allow us to, for example, assess the functions of all possible missense variants of a gene in advance of their first appearance in the human population [14–16].

Assays of functional variation using complementation are constructed via two steps. First, a complementation relationship is identified, such that expression of a wild-type human gene product rescues phenotypic defects in a yeast strain lacking the cognate function. Second, the pathogenicity of genetic variants is assessed by comparing their ability to complement with that of the wild-type allele. Previously, we developed yeast-based functional complementation assays to evaluate the functional effects of missense variants in human disease-associated genes [17]. We and others have shown that yeast-based functional complementation assays can efficiently reveal the functionality of human genetic variants [15]. Indeed, our previous work showed that yeast-based functional complementation assays achieved three times the sensitivity of computational methods for detecting disease variants at the same high threshold of precision [17]. Although complementation assays have been largely restricted to orthologous human-yeast gene pairs, a few examples of paralogous complementation are known. For example, the human gene *RAC1* can complement the yeast *ras1*$^{-}$ strain, suggesting functional similarity between these genes [18]. In another example, a recent systematic screen found that the human gene *SEC61A1*, implicated as a host factor for influenza, HIV and dengue viruses [19], can complement loss of the yeast gene *RFT1* [15]. Thus, the set of human disease genes with complementation-based functional variation assays could potentially be expanded via paralog relationships [18].

There are over 130,700 'disease-causing' variants according to the most stringent annotation in the Human Gene Mutation Database (HGMD; as of November 2015), corresponding to 3535 unique disease genes [20]. Of these disease genes, 972 have an annotated ortholog in *Saccharomyces cerevisiae*, while another 762 genes have at least one paralog. (Except where noted, we follow the practical operating definition of "paralog" as any homolog not annotated as an ortholog.) If we include less stringent HGMD disease gene annotations (see Methods), the number of disease genes with a yeast ortholog rises to 1869, with an additional 1087 having a paralog.

Orthologs are (by definition) diverged by speciation rather than by duplication within a species, and it is generally believed that they are more likely to serve in the same biological role across species. In contrast, it is generally thought that paralogs—homologs that diverged by duplication within the genome of a species—are more likely to have evolved a distinct or specialized function. However, it is quite possible for orthologs to acquire different properties and for paralogs to retain the same function [21]. Although complementation assays based on human-yeast ortholog pairs can accurately predict pathogenic variants [17], it is unclear whether similar assays based on paralogs are as useful in predicting pathogenic variants. Therefore, we assessed the ability of paralogous complementation assays to detect pathogenic variation using an objective panel of disease and non-disease variants.

Results

Many complementation relationships exist for human-yeast paralogs

To expand the set of human disease genes with a functional complementation assay, we identified human disease genes, each having one or more essential yeast paralogs for which a conditional mutant was available. Because protein domains are distinct functional and structural units in a protein, because variants within a particular domain have a heightened chance of affecting structural and functional properties of the proteins in which they appear [22–24], and because domain-based mutational studies have proven useful in elucidating the functional and disease effects of variants [22, 25, 26], we also used protein domain annotations to select human-yeast paralogs for which all domains in the yeast protein could be found in the human protein. This yielded 314 human disease genes with a suitable yeast paralog to test. Given that a human gene may have multiple yeast paralogs, this resulted in a larger search space of 1060 human-yeast paralog pairs (S1 Table).

For each of the 314 human genes in our search space, we obtained an open reading frame (ORF) from the hORFeome 8.1 collection [7, 27], and generated a 'humanized' yeast expression plasmid via recombinational cloning [17]. To assess complementation for each human-yeast pair, the human protein was expressed in yeast strains bearing temperature-sensitive mutations [28] in the corresponding yeast gene, and growth was assessed at multiple temperatures (Fig 1; see Methods for detail).

In addition to functional complementation tests for 1060 human-yeast paralog pairs (including one positive-control paralog pair previously to complement), we assessed 7 additional known-complementing orthologous pairs as positive control. All complementation tests were performed twice, and 42 pairs yielded complementation in at least one replicate. These 42 included all 7 positive-control orthologous pairs, and 35 paralogous pairs. The 35 complementing paralog pairs included the positive control and 34 novel pairs, of which 33 were subsequently confirmed. (S1 Fig, see Methods for a complete description of complementation testing procedures.) Images showing negative functional assay results are available via the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.j05n0.

Thus, within a test space of 1060 human-yeast paralog pairs, we recovered 34 complementing pairs of which 33 (97%) were novel. Of the 314 human disease-associated genes tested, 33 (10.4%) yielded a complementation relationship with at least one yeast paralog.

Some essential yeast genes are complemented by multiple human paralogs sharing only a single domain

Among the 33 novel human-yeast paralog complementation assays established here, there were four yeast genes that could each be complemented by multiple human genes. For each of these yeast genes, the corresponding set of complementing human genes shared a common protein domain. For example, the function of yeast serine/threonine protein kinase Kin28

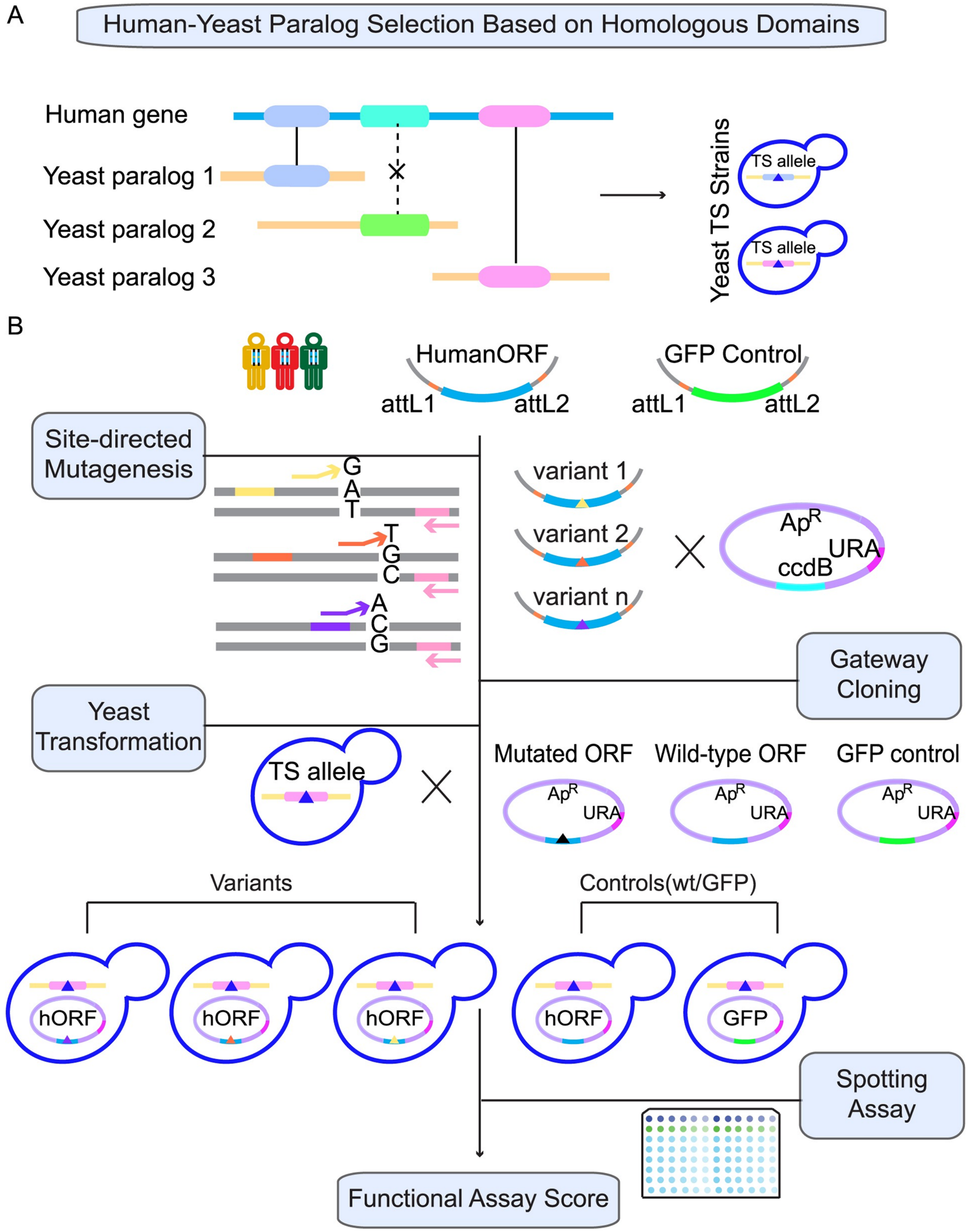


Fig 1. Schematic overview of process for assessing the functional effect of human disease-associated variants via complementation testing. A. We selected paralog pairs where a human disease protein has a yeast paralog for which all protein domains are also found in the human protein. Homologous pairs of domains are connected by solid lines, while non-homologous domain pairs are connected by a dashed line. B. For a subset of those paralog pairs for which we identified complementation relationships, we used these relationships to assess whether the functionality of variants in these assays predicted variant pathogenicity.

https://doi.org/10.1371/journal.pgen.1006779.g001

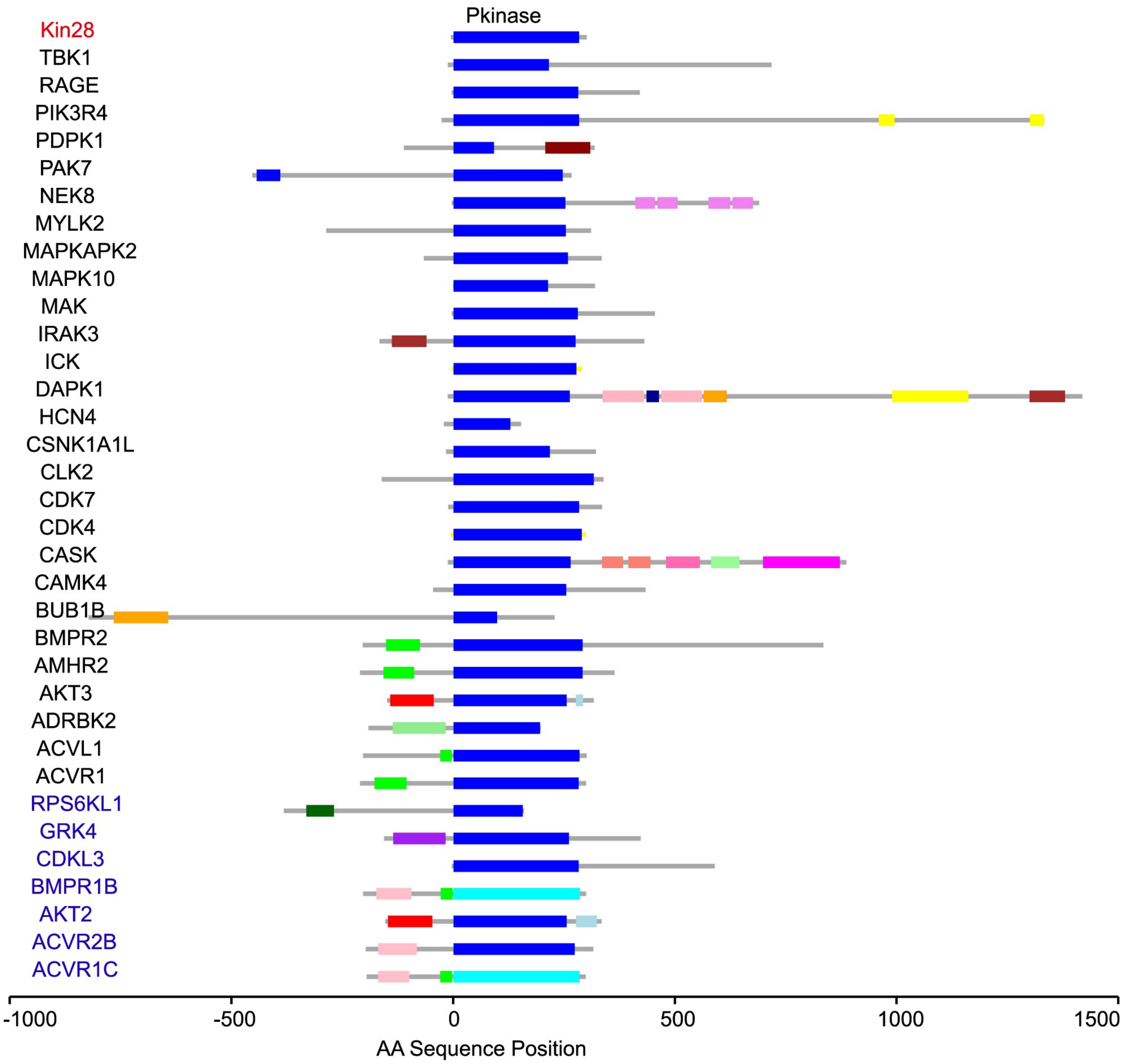


Fig 2. Protein domain architecture of yeast Kin28 and human paralogs. Shown are yeast Kin28 (red text), and human paralogs tested for complementation (in blue text if we found complementation and black text otherwise). Protein domain patterns Pkinase_Tyr (PFAM pattern PF07714) and Pkinase (PFAM pattern PF00069) are indicated in light and dark blue, respectively.

https://doi.org/10.1371/journal.pgen.1006779.g002

(ORF ID: YDL108W) could be complemented by expression of seven different human proteins (Fig 2): Ribosomal Protein S6 Kinase-Like 1 (RPS6KL1), G Protein-Coupled Receptor Kinase 4 (GRK4), Cyclin-Dependent Kinase-Like 3 (CDKL3), Bone Morphogenetic Protein Receptor, type IB (BMPR1B), V-Akt Murine Thymoma Viral Oncogene Homolog 2 (AKT2), Activin Receptor Type-2B (ACVR2B) and Activin A Receptor, Type 1C (ACVR1C), each sharing the same Pkinase protein domain found within yeast Kin28 (Table 1). However, each of these seven human proteins contain one or more additional protein domains and have different functions in different pathways. Indeed, the only apparent common thread among Kin28-complementing human proteins is the Pkinase protein domain.

Table 1. Seven human genes can complement yeast Kin28.

Human Gene Name	Human Gene Symbol	Protein Domain
Ribosomal Protein S6 Kinase-Like 1	RPS6KL1	PF00069,PF04212
G Protein-Coupled Receptor Kinase 4	GRK4	PF00069
Cyclin-Dependent Kinase-Like 3	CDKL3	PF00069
Bone Morphogenetic Protein Receptor	BMPR1B	PF00069,PF01064,PF08515
V-Akt Murine Thymoma Viral Oncogene Homolog 2	AKT2	PF00069,PF00169,PF00433
Activin Receptor Type-2B	ACVR2B	PF00069,PF01064
Activin A Receptor	ACVR1C	PF00069,PF01064,PF08515

https://doi.org/10.1371/journal.pgen.1006779.t001

The other three examples of yeast genes complemented by multiple human genes were *CAK1* (encoding Cdk-activating kinase Cak1), *SEC12* (encoding guanine nucleotide exchange factor Sec12), and *NAN1* (encoding Net1-Associated Nuclear protein Nan1). Complementing the loss of yeast Cak1 were two human genes encoding Serine/threonine-Protein Kinase (TBK1) and Cyclin-Dependent Kinase 7 (CDK7) (Fig 3), both of which contain a Pkinase domain. Complementing loss of yeast Sec12 were human genes *IFT122*, *ELP2*, and *GNB1L*, each sharing the WD40 repeat domain (PF00400). Loss of yeast Nan1 was rescued by human genes *PAFAH1B1* and *RFWD2*, also sharing the WD40 repeat domain. Thus, protein domain function, even when encoded by otherwise highly-diverged gene pairs, can be sufficiently conserved to allow functional rescue of a yeast protein and thus a potential assay for functional human variants.

Our search for complementation involved many kinases. Indeed, of the 1060 human/yeast gene pairs we tested, 480 (45%) of those pairs contained a yeast or human kinase-encoding gene. To understand this prevalence, we note that we only sought complementation where the human gene was a disease gene, where the yeast gene was essential, and where all domains in the yeast protein could be found in the human proteins. Although kinases do not seem to be enriched amongst human disease genes (they represent ~23% of annotated disease genes as compared with ~30% of non-disease-annotated genes), they are abundant. In yeast, there are 230 kinase-coding genes, of which 29 are essential. Moreover, 39% of human disease genes that had a yeast homolog are kinases, so that kinases are enriched for conservation in yeast. Thus, enrichment for kinases in our search space seems due to the fact that the kinase domain is ancient and found frequently in both yeast and human proteins. There were 14 yeast kinase-encoding genes and 56 human kinase-encoding genes amongst these 480 yeast/human kinase pairs, so that each gene appears in many pairs.

Paralog complementation is only weakly predicted by sequence similarity

We examined the extent of sequence identity between human disease-associated genes and their yeast paralogs. For each human and yeast gene pair, we calculated the pairwise sequence identity (PID; the percentage of aligned positions with identical residues). For a yeast gene with multiple human paralogs tested, we examined PID for complementing and non-complementing human-yeast paralog pairs. As expected, complementing pairs had higher PID than non-complementing pairs (Fig 4A, *P*-value = 0.007, Wilcoxon test). Similarly, for human genes that had multiple yeast paralogs tested, complementing pairs had relatively higher average PID (Fig 4B, *P*-value = 0.003, Wilcoxon test). A similar analysis performed for three additional sequence-identity calculation methods reached similar conclusions, except for one method which calculates a substantially lower percent identity in cases where the length of the

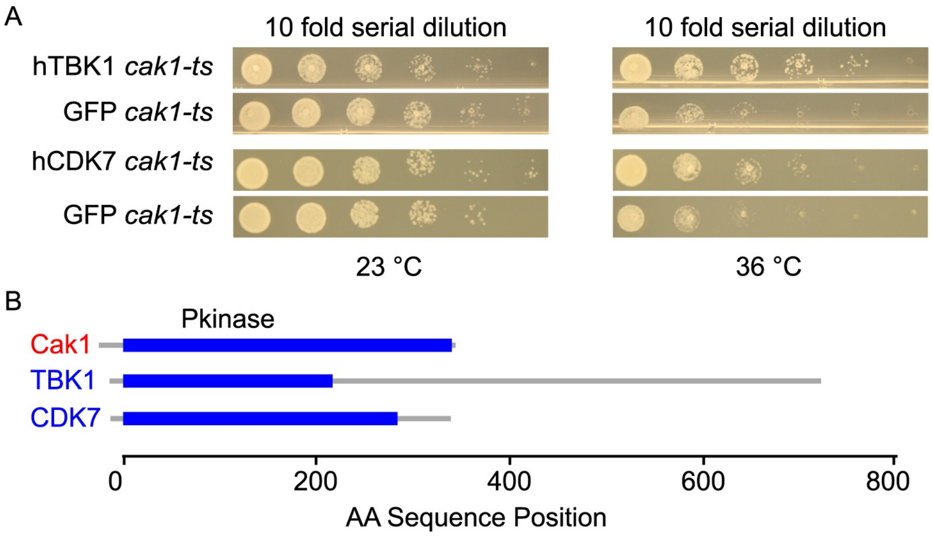


Fig 3. Functional assay and protein domain architecture of yeast Cak1 and its complementing human paralogs. (A) Functional complementation assay results showing that expression of human proteins TBK1 and CDK7 complements defects in a strain (YFL029C_tsa650) that encodes a temperature sensitive variant of Cak1 (described as "*cak1-ts*" above). (B) Pkinase domains are shown in dark blue. Complementing paralogs indicated in blue text.

https://doi.org/10.1371/journal.pgen.1006779.g003

aligned region differs greatly between two aligned proteins (see S1 File). Our results show that, as with human-yeast orthologs [12, 15], sequence similarity between human-yeast paralogs is correlated with—but only weakly predictive of—functional complementation. For example, a 30% PID threshold captured 60% of the complementing pairs, but 30% of non-complementing pairs also exceeded this threshold. Thus, systematic experimental testing will continue to be required for discovery of complementing paralog pairs.

Assessing the pathogenicity of missense variants

Having established functional complementation relationships between human-yeast paralogs, we wondered whether these relationships could be exploited to assess the pathogenicity of human genetic variants. Of the 33 disease-associated genes for which we could identify a novel complementation relationship, there were 17 with known pathogenic missense variants according to HGMD DM annotation. To assess the ability of human/yeast paralog complementation assays to identify pathogenic variants (identified as those with high confidence "DM" annotation, indicating disease causality, from the HGMD database), we selected a subset of seven human disease-associated genes with multiple annotated disease-causing missense variants [29–31] (Table 2). Non-disease-annotated missense variants were present in the dbSNP database [32–34] for five of these seven genes. In total, we tested 19 disease-causing missense variants, each qualifying as causal according to the most stringent "DM" annotation in HGMD and the most stringent "pathogenic" annotation in ClinVar [35]. We also tested 16 non-disease-associated variants from dbSNP, selecting lower allele frequency variants where possible to better control for the generally low allele frequency of disease-causing variants.

For each of these 35 human variants, we generated an expression clone by site-directed mutagenesis and recombinational cloning, transformed it into the appropriate temperature-sensitive (TS) yeast strain, and assessed functional complementation (Fig 1; see Methods). For each genetic variant, this yielded a semi-quantitative Failure-to-Complement (FC) score, corresponding to the previously described "FCS score" [17]. FC scores were calibrated so that the

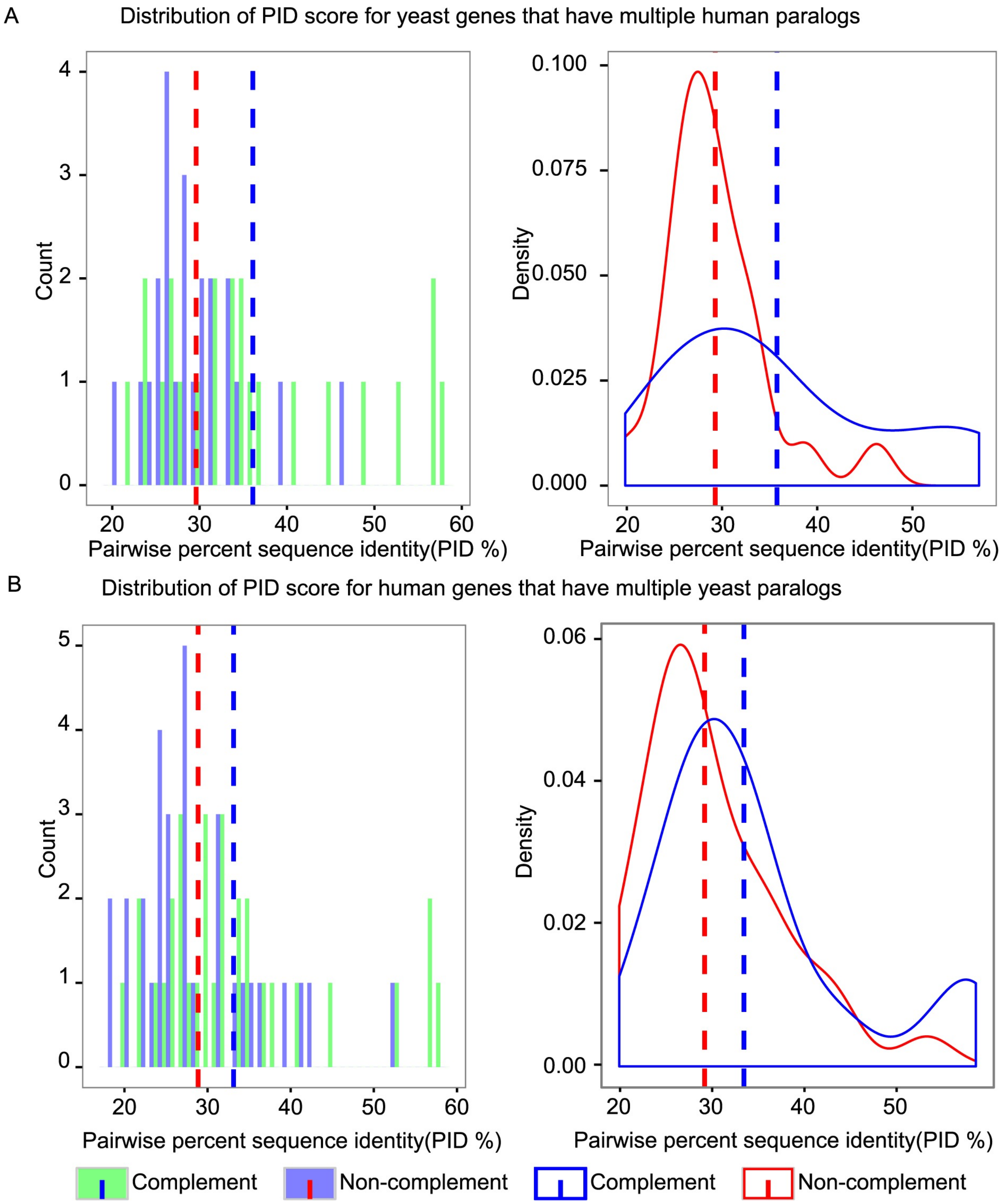


Fig 4. Relating sequence similarity and ability of a paralog to complement. The average percent identity (PID) score distribution is shown for human-yeast pairs such that multiple human paralogs were tested for a given yeast protein (A), and for human-yeast pairs such that multiple yeast paralogs were tested for a given human protein (B). In each case, the distribution is shown separately for complementing and non-complementing pairs. Each bin height is the count of human or yeast genes having a PID within the appropriate range for that bin. That complementing and non-complementing distributions are both shifted in positon relative to one another and highly overlapping suggests that sequence similarity is an informative but imperfect predictor of complementation.

https://doi.org/10.1371/journal.pgen.1006779.g004

Table 2. Deleteriousness predictions from functional complementation (FC), Polyphen-2 (PPH2) and PROVEAN.

Gene Symbol	Entrez	Variant	Disease Assoc?	FC Score	FC Prediction	FC Correct?	PPH2 Score	PPH2 Prediction	PPH2 Correct?	Provean Score	Provean Prediction	ProveanCorrect?	Within Aligned Region?
CASK	8573	T573I	No	0.6	Damaging	No	0.021	Neutral	Yes	-2.35	Neutral	Yes	No
CASK	8573	D471N	No	0.4	Neutral	**Yes**	0.005	Neutral	Yes	-1.48	Neutral	Yes	No
CASK	8573	M438L	No	0.4	Neutral	**Yes**	0	Neutral	Yes	-1.24	Neutral	Yes	Yes
CASK	8573	R430C	No	0.4	Neutral	**Yes**	0.035	Neutral	Yes	-2.51	Damaging	No	Yes
CASK	8573	R28L	Yes	0.8	Damaging	**Yes**	1	Damaging	Yes	-3.59	Damaging	Yes	Yes
CYP19A1	1588	M21T	No	0.6	Damaging	No	0.01	Neutral	Yes	-0.65	Neutral	Yes	No
CYP19A1	1588	M85R	Yes	0.8	Damaging	**Yes**	0.128	Neutral	No	-2.77	Damaging	Yes	Yes
CYP19A1	1588	W39R	No	0.4	Neutral	**Yes**	0.343	Neutral	Yes	-5.16	Damaging	No	Yes
CYP19A1	1588	M127R	Yes	0.8	Damaging	**Yes**	1	Damaging	Yes	-4.87	Damaging	Yes	Yes
CYP19A1	1588	Y81C	Yes	0.8	Damaging	**Yes**	1	Damaging	Yes	-6.87	Damaging	Yes	Yes
DHDDS	79947	K42E	Yes	0	Neutral	No	0.786	Damaging	Yes	-3.65	Damaging	Yes	Yes
EMG1	10436	D86G	Yes	0.6	Damaging	**Yes**	1	Damaging	Yes	-6.99	Damaging	Yes	Yes
IFT122	55764	G51A	No	0.2	Neutral	**Yes**	0.016	Neutral	Yes	-4.11	Damaging	No	No
IFT122	55764	T91I	No	0.2	Neutral	**Yes**	0.953	Damaging	No	-3.99	Damaging	No	No
IFT122	55764	S373F	Yes	0.6	Damaging	**Yes**	0.951	Damaging	Yes	-5.038	Damaging	Yes	No
IFT122	55764	L99W	No	0.4	Neutral	**Yes**	0.861	Damaging	No	-0.178	Neutral	Yes	No
IFT122	55764	R328W	No	0.2	Neutral	**Yes**	0.994	Damaging	No	-6.168	Damaging	No	No
RAB33B	83452	N148K	Yes	0.8	Damaging	**Yes**	0.005	Neutral	No	0.6	Neutral	No	No
RAB33B	83452	K46Q	Yes	0.8	Damaging	**Yes**	1	Damaging	Yes	-3.55	Damaging	Yes	Yes
RAB33B	83452	P142L	No	0.6	Damaging	No	1	Damaging	No	-9.99	Damaging	No	Yes
RAB33B	83452	T177M	No	0.6	Damaging	No	1	Damaging	No	-5.21	Damaging	No	Yes
VCP	7415	A232G	Yes	0.6	Damaging	**Yes**	0.005	Neutral	No	-1.87	Neutral	No	No
VCP	7415	I151V	Yes	0.4	Neutral	No	0	Neutral	No	-0.51	Neutral	No	Yes
VCP	7415	I27V	No	0.2	Neutral	**Yes**	0	Neutral	Yes	-0.43	Neutral	Yes	Yes
VCP	7415	Q19R	No	0.4	Neutral	**Yes**	0	Neutral	Yes	0.61	Neutral	Yes	Yes
VCP	7415	S171N	No	0	Neutral	**Yes**	0.004	Neutral	Yes	-1.18	Neutral	Yes	No
VCP	7415	T436I	No	0.4	Neutral	**Yes**	0.236	Neutral	Yes	-3.76	Damaging	No	No
VCP	7415	I206F	Yes	0.6	Damaging	**Yes**	0.983	Damaging	Yes	-3.7	Damaging	Yes	Yes
VCP	7415	L198W	Yes	0.6	Damaging	**Yes**	1	Damaging	Yes	-4.71	Damaging	Yes	Yes
VCP	7415	R159G	Yes	0.6	Damaging	**Yes**	1	Damaging	Yes	-6.56	Damaging	Yes	No
VCP	7415	R159C	Yes	0.8	Damaging	**Yes**	1	Damaging	Yes	-6.31	Damaging	Yes	No
VCP	7415	R159H	Yes	0.8	Damaging	**Yes**	0.517	Damaging	Yes	-2.97	Damaging	Yes	No
VCP	7415	R191G	Yes	0.6	Damaging	**Yes**	0.999	Damaging	Yes	-6.49	Damaging	Yes	Yes
VCP	7415	P137L	Yes	0.4	Neutral	No	1	Damaging	Yes	-9.31	Damaging	Yes	Yes
VCP	7415	R155G	Yes	0.4	Neutral	No	0.998	Damaging	Yes	-5.18	Damaging	Yes	No

The annotation of "FC correct?", "PPH2 Correct?", "Provean Correct?" is based on whether deleteriousness annotations from FC, PPH2 or Provean agree with current pathogenicity (HGMD "DM") annotations. FC predictions that were correct according to HGMD "DM" are emphasized using a bold-text "Yes".

https://doi.org/10.1371/journal.pgen.1006779.t002

Table 3. Pathogenicity prediction performance for the human disease gene paralog test set.

Method	MCC	AUPRC	AUROC	REC90
PolyPhen-2	0.48	0.76	0.55	0.74
PROVEAN	0.37	0.7	0.52	0.71
Paralog-based FC	0.59	0.83	0.55	0.78

(MCC) Matthews correlation coefficient;
(AUPRC) area under the precision-recall curve;
(AUROC) area under the receiver-operating characteristic curve;
(REC90) recall at 90% precision.
Performance estimates for best-performing methods are indicated by underline

https://doi.org/10.1371/journal.pgen.1006779.t003

positive (complementing) control wild-type human plasmid achieves a FC score of 0, and a Green Fluorescent Protein (GFP) negative (non-complementing) control achieves an FC score of 1. Following previous conventions, only variants with a score greater than 0.5 were considered deleterious [17, 36].

Functional complementation assays predicted 15 (79%) of 19 disease variants and 4 (25%) of the 16 non-disease-associated variants we tested to be deleterious (S2 Fig). Our observation that 25% of non-disease-annotated variants failed to complement raises the possibility that many non-disease-annotated genetic variants may in fact impact gene function, so that our estimates of recall and precision may be conservatively low. Nevertheless, functional complementation assays clearly distinguish disease and non-disease-associated genetic variants: For the five genes that have both disease-associated and non-disease-associated variants, disease-associated variants exhibited significantly higher FC scores (*P*-value = 0.001, Wilcoxon test; Table 3, Fig 5A).

To put performance of functional complementation assays in the context of computational alternative methods, we applied PolyPhen-2 [36] and Protein Variation Effect Analyzer (PROVEAN) [37], two widely used computational methods for predicting pathogenic variants. At the 0.5 threshold, paralog-based functional complementation assays achieved 83% precision (fraction of predicted-deleterious variants that are annotated as pathogenic; 95% CI 58% - 96%) at 79% recall (fraction of pathogenic variants predicted to be deleterious). At a threshold score (0.5) which achieves the same 79% recall value, PolyPhen-2 achieved precision 75% (95% CI 51% - 90%). Different performance tradeoffs could be achieved at different thresholds. At the 0.7 threshold, paralog-based functional complementation assays achieved 100% precision (95% CI 60%-100%) at 42% recall. At the same 42% recall, PolyPhen-2 achieved 84% precision (95% CI 68%-100%). Although paralog-based variant testing numerically outperforms PolyPhen-2 in terms of precision at matched-recall thresholds, the limited sample sizes do not allow us conclude that this increase is significant. However, multiple performance measures—Matthews correlation coefficient (MCC), area under the precision-recall curve (AUPRC), area under the receiver operating characteristic (AUROC) curve, and recall at 90% precision (REC90)—suggest that paralog-based functional complementation assays are at least on par with computational methods in predicting pathogenicity (Table 3).

To more generally assess the performance of complementation-based pathogenicity assays against computational tests, we combined paralog-based and previous ortholog-based complementation pathogenicity tests [17]. At score thresholds where FC score and PolyPhen-2 both achieve a recall of 90%, the FC precision is 81% while PolyPhen-2 precision is 72%. Using the previously described performance threshold value of 0.5 for the FC score [17] achieves a recall of 78% and precision of 89% for the FC score. At a matched 78%, recall threshold, PolyPhen-2

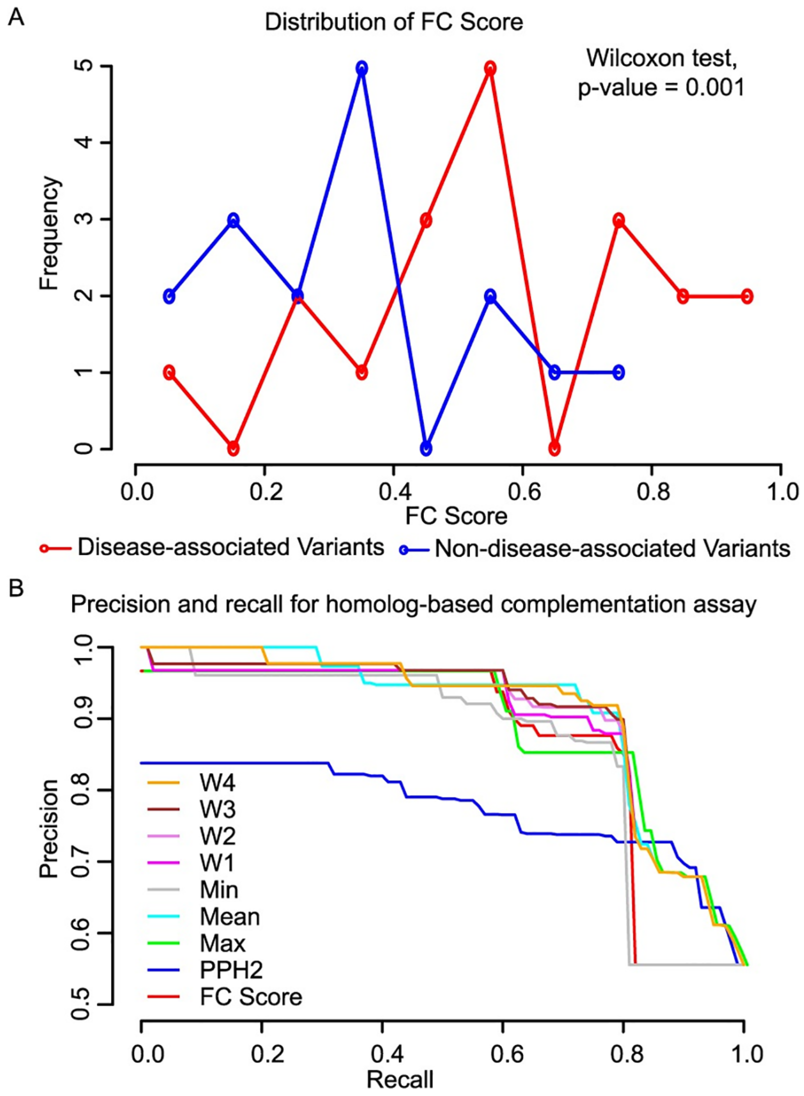


Fig 5. Ability of functional complementation to predict pathogenicity. (A) Distribution of FC scores for disease associated (red line) or non-disease-associated variants (blue line). FC scores from paralog-based complementation assays are significantly higher for disease-associated variants than non-disease-associated variants (*P*-value, Wilcoxon test). (B) Precision vs. recall performance for functional complementation scores (both paralog- and ortholog-based), PolyPhen-2 scores, and various options for combining the two approaches (see Methods).

https://doi.org/10.1371/journal.pgen.1006779.g005

yields a lower precision of 73% (Fisher's exact test *P*-value = 0.003). A similar comparison using only ortholog-based assays yielded the same conclusion, albeit with a less significant *P*-value of 0.008 [17]. Thus, inclusion of paralog-based complementation strengthens previous conclusions that complementation-based identification of functional variation outperforms current computational approaches.

We next investigated whether the combination of FC with PolyPhen-2 scores could yield performance that exceeds either approach alone. We used seven alternative ways to combine these scores: minimum, maximum, mean, and four alternative weighted means (w1 through w4) (Fig 5B). The results confirmed our previous conclusion that combining FC and PolyPhen-2 scores can improve the performance in the high precision/low recall region.

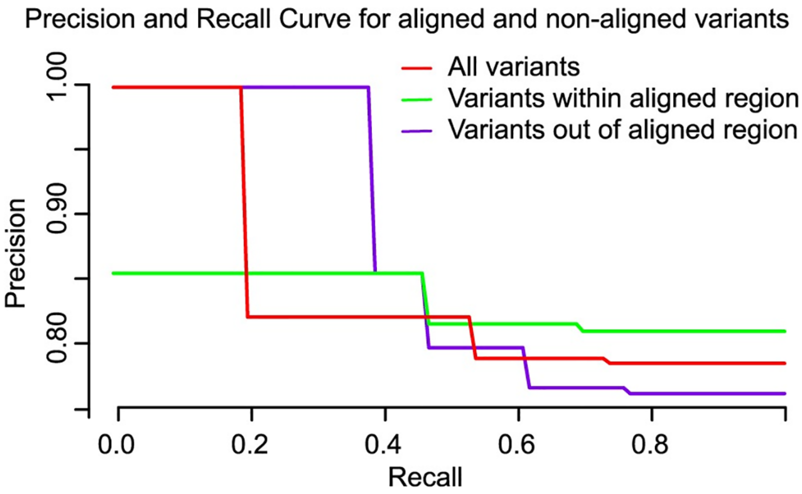


Fig 6. Performance of pathogenic variant identification does not strongly depend on whether the variant is in the aligned region. Here we show precision vs recall performance for varants that either do ('aligned') or do not (non-aligned) fall within the sequence region that can be aligned between human and yeast homologs.

https://doi.org/10.1371/journal.pgen.1006779.g006

We wondered whether complementation assays are capable of detecting pathogenic variants when these variants fall outside of the aligned homology region. It is possible that variants will affect additional human gene functions that are not needed for complementation, so that such pathogenic variants will be missed. However, variants which alter protein folding, or stability in a human cell may often do the same in a yeast cell. Interestingly, the ability of complementation to identify disease variation did not depend strongly on whether or not the variation falls within the aligned region of homology between yeast and human genes. As shown in Fig 6. at a score threshold achieving 90% recall, the likelihood of detecting a disease variant was comparable: 0.76 and 0.87 respectively for variants inside and outside of the aligned region of human and yeast paralogous pairs. Taking ortholog- and paralog-based complementation assay data together, the distributions of FC scores for variants inside and outside of the aligned region were statistically indistinguishable (*P*-value = 0.37, Wilcoxon test). All Wilcoxon tests are unaffected by our somewhat arbitrary assignment of numeric FC scores to different qualitative classes of observed complementation, because these tests only use the ranking order of quantitative values. At score thresholds yielding a recall of 90%, pathogenic variant detection variation achieved a precision rate of 92% and 88% respectively for variants inside and outside of the aligned region. Thus, functional complementation assays are capable of accurately detecting pathogenic variants, even outside of the aligned homology region.

Discussion

Considerable effort has been made to understand how genetic changes give rise to the molecular effects that cause diseases [38–40]. There are many databases and tools for prioritizing candidate single nucleotide polymorphisms (SNPs) or hypothesizing the molecular causes of genetic disease. Functional complementation assays enable identification of pathogenic disease variants with substantially greater sensitivity than computational methods [17]. Although previous trans-species functional complementation assays have been almost exclusively based on orthology relationships, our systematic search yielded novel paralogy-based functional complementation assays for 33 human disease genes.

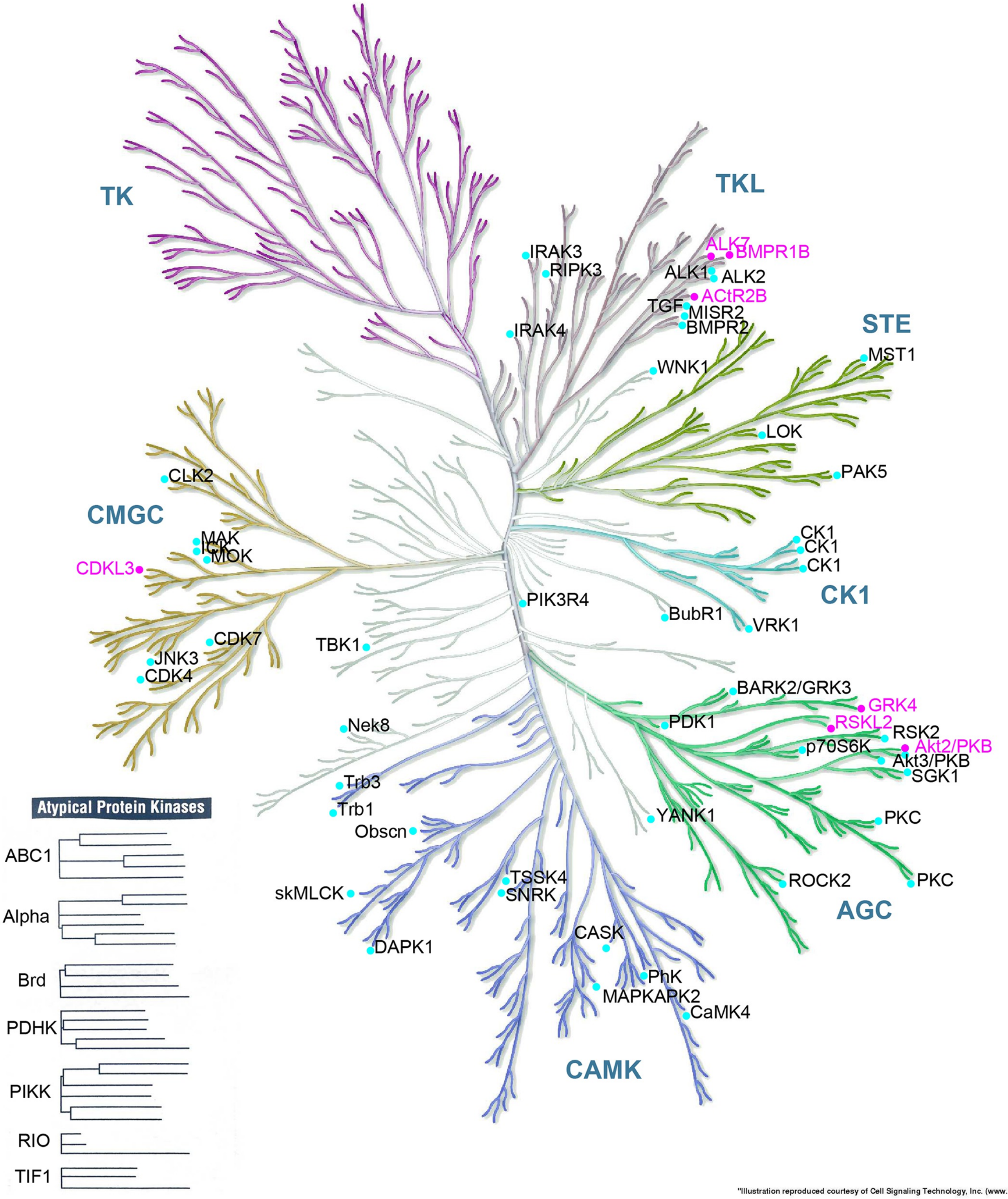
The gene *RAB33B*, which encodes a small GTP-binding protein of the RAB family and is associated with Smith-McCort Dysplasia, can illustrate paralog-based functional complementation. We successfully observed failure to complement for the two disease associated variants, P219S and K46Q [41, 42]. Interestingly, both non-disease-annotated variants, P142L (rs369719131) and T177M (rs140381459), also showed loss of complementation. Our findings agreed with PolyPhen-2 and PROVEAN which each also predicted them to be deleterious. All four variants tested are within the Ras domain. Thus, even though variants P142L and T177M are not known to be associated with disease, they appear to affect protein function.

Another example is the human *CASK* gene, which encodes calcium/calmodulin-dependent serine protein kinase. *CASK* encodes a 921-amino acid polypeptide with an N-terminal calcium/calmodulin-dependent protein kinase-like domain, PDZ and SH3 domains, a potential protein-binding motif, and a domain homologous to guanylate kinase [43]. Sequence variants in CASK cause intellectual disability [44]. The only annotated disease variant we tested in CASK was the kinase domain variant R28L causing FG Syndrome [45], an X-linked disorder causing intellectual disability, physical anomalies and developmental delays. This variant exhibited loss of complementation. We also tested several non-disease-associated CASK variants (D471N, M438L, R430C, and T573I). Three of the four non-disease variants tested retained the ability to complement. By contrast, the variant T573I (rs141840001), despite not being annotated as associated with Mendelian disease [30, 35, 46] or via any GWA study [47], showed reduced complementation. This variant was originally identified in a clinical genetics laboratory (Emory Genetics Laboratory, ClinVar accession RCV000175306.1) in an autistic male, so that the evidence of functionality we found for T573I may spur further investigation.

In addition to yielding a direct benefit in the form of novel functional assays, our systematic search for paralogous complementation enabled some general observations about complementation relationships. First, as with orthologs, sequence similarity is only a very weak predictor of complementation relationships (Fig 4), necessitating experimentation to identify complementation relationships.

Second, despite the idea that paralogs often have divergent functions, we found that multiple human genes (having in common a single protein domain) can sometimes complement the same yeast gene. For example, the seven human disease-associated genes that can complement yeast *kin28* all encode a protein kinase domain. Interestingly, the seven complementing genes fall into three different major kinase groups, including TKL kinases, CMGC kinases and AGC kinases (Fig 7). An additional 31 human disease-associated genes that encode the same protein domain (many of which fall into the same three major kinase groups) did not complement yeast *kin28*. Using the multiple sequence alignment tool Clustal [48] to examine the phylogenetic tree of tested human protein homologs of yeast Kin28), we also found no evident clustering of the yeast Kin28-complementing human homologs that could distinguish them from non-complementing human kinases. This result highlights the idea that closer evolutionary relationships do not guarantee complementation. When we mapped the 38 kinases to KEGG and REACTOME pathways, 5 of 7 complementing kinases mapped to signal transduction pathways, but this was not significantly different from rate at which 22 tested non-complementing kinases of 31 mapped to the same pathway (P = 0.6, Fisher's exact test). Thus, we found no obvious predictors of which kinases were more likely to complement (S4 Table).

We were surprised to find that the success of paralog-based complementation assays is on par with ortholog-based complementation in terms of identifying disease variation. Examining a test set of disease- and non-disease-associated variation, we found that paralog-based complementation could detect ~78% of pathogenic variants at 90% precision, which was statistically indistinguishable from the performance of ortholog-based complementation. The



"Illustration reproduced courtesy of Cell Signaling Technology, Inc. (www.cellsignal.com)"

Fig 7. The kinome tree of yeast Kin28 and its kinase paralogs tested here. Kinases that can complement yeast Kin 28 were colored in pink, other kinases tested for ability to complement yeast Kin28 were colored in cyan. (The image was generated from the Kinome-Render Tool [49] hosted at Cell Signaling, Inc.).

https://doi.org/10.1371/journal.pgen.1006779.g007

combination of paralog and ortholog-based complementation continues to outperform current computational approaches.

Paralog-based assays have high potential to extend the arsenal of assays to assess the functionality of human coding variation. This is despite the fact that paralogs yielded a complementation relationship for a smaller fraction of human genes than had been observed for ortholog pairs: this fraction was 10% in the current study as compared with 19% for ortholog pairs [17]. Indeed, Kachroo *et al* [12] achieved an even higher 47% rate of complementation for a subset of orthologous pairs that are "one to one", i.e., for which there are no close paralogs in either human or yeast. According to the YeastMine database [50] there are 773 additional human disease-associated genes with yeast paralogs, suggesting that a functional assay could potentially be developed for at least ~70 additional human disease-associated genes through further examination of paralog complementation. According to HGMD, about 3019 human disease-associated genes have paralogs in either *S. cerevisiae* or *Schizosaccharomyces pombe*. Simple extrapolation suggests that a more exhaustive search for complementation relationships in these two yeast species could yield complementation assays for assessing functional variation in 300 human disease genes. Considering multicellular model organisms, the number of potential complementation assays increases further (see Table 4 for a summary of human disease-associated genes with either an ortholog or paralog in five model animal systems). Given that complementation tests work as well as they do for identifying pathogenic variation in the billion-year diverged model organism *S. cerevisiae*, it stands to reason that other model systems (including complementation in human cells where cell-autonomous selectable phenotypes are known) should also be explored.

Our results combining paralog- and ortholog-based complementation tests show that these assays can be used to accurately identify pathogenic variants even when those variants fall outside of the aligned region. This is consistent with the idea that many deleterious variants affect protein folding or stability and disrupt the function of the entire protein. Thus, even where only a single domain is required for a human protein to complement its yeast paralog, that relationship can be exploited to detect a substantial subset of functional variation throughout the length of the human protein.

It is worth revisiting our working definition of paralogy (homology without annotated orthology). Paralogs under this definition may be previously unrecognized orthologs, and gene pairs with complementation relationships may be enriched in such cases. However, for the practical purpose of identifying pathogenic variants using a complementation assay, it seems that the distinction between paralogy and cryptic orthology is essentially irrelevant. In

Table 4. Numbers of human disease-associated genes with orthologs and paralogs in five model species.

Organism	Human disease-associated genes	
	Orthologs	**Paralogs**
S. cerevisiae or *Schizosaccharomyces pombe*	6648	3019
Mus musculus	5547	256
Rattus norvegicus	5492	265
Danio rerio	4619	231
Drosophila melanogaster	3021	384*
Caenorhabditis elegans	2665	169

*This figure is conservative, in that the HGMD source for this information used a more stringent criterion for paralogy (elsewhere in this study homologs without annotated orthology are referred to as paralogs).

https://doi.org/10.1371/journal.pgen.1006779.t004 either case, complementation relationships between human genes and their homologs in other species beyond *S. cerevisiae* provide substantial further opportunities to study the functional properties of human disease-associated variants.

One potential limitation of complementation testing is that, while it may accurately detect many loss-of-function variants, we expect that it is less likely to identify gain of function variants. We reviewed the primary literature for the 19 disease variants we tested via paralog-based complementation. Of these 19, the literature suggested "loss of function" for 16 and "gain of function" for only one (S3 Table). The putative "gain of function" variant retained its ability to complement, and was thus, as expected, not detected as damaging by our complementation assay.

Given that computational approaches are faster, cheaper and available for a wider range of genes than are functional complementation assays, it is worth asking whether systematic experimental variant assessment is worth pursuing. Where variant assessment is critical for diagnosis and therapy, and where computational methods cannot return a sufficiently confident call for a large fraction of disease variants, alternatives are clearly needed. Moreover, new advances in "deep mutational scanning" have enabled the *en masse* application of a cell-based functional assay to essentially all missense variants for a given protein [51]. For a fixed initial cost, deep mutational scans can provide a comprehensive 'look-up' table allowing instantaneous interpretation of missense variants as they appear in the clinic.

Materials and methods

Selecting human-yeast homologs for testing

To systematically test the ability of wild-type human disease-associated genes to rescue mutations in paralogous yeast genes, we defined the search space to be human genes for which HGMD [29–31] has annotated one or more alleles as being 'DM' (disease-causing) and for which a clone was available in ORFeome version 8.1 [27].

Because protein domains are distinct functional and structural units in a protein, because variants within a particular domain have a heightened chance of affecting structural and functional properties of the proteins in which they appear [22–24], and because domain-based mutational studies have proven useful in elucidating the functional and disease effects of variants [22, 25, 26], we also used protein domain annotations as a criterion for selecting human-yeast paralogs. We searched both yeast and human genes against the Pfam domain types from the Pfam protein domain family database (version 27) [52], using an *E*-value cutoff of 0.001 [53], and identified cases where all protein domains encoded by a yeast gene were fully 'covered' by a human gene. In our previous study [17], we used the InParanoid database [54] to select yeast/human orthologous pairs for which the human gene had at least one disease-associated variant according to either HGMD or OMIM databases. Here yeast/human pairs were chosen similarly, except that we accepted all homologs reported by InParanoid except those annotated as orthologs. The InParanoid program uses NCBI-BlastP pairwise similarity scores for constructing orthology groups. An orthology group is initially composed of two so-called seed orthologs that are found by mutual best hits between two proteomes.

Considering only paralog pairs where the yeast gene was essential and had an available temperature sensitive mutation, where the human gene had an available expression clone, and where all protein domains in the yeast gene were covered in the corresponding human gene, we selected 1060 human-yeast paralog pairs corresponding to 314 human genes and 162 yeast genes. We note that a single gene in one species can have multiple paralogs in another species, and thus appear in multiple tested paralog pairs.

Constructing wild-type human ORFs and human ORFs with disease-associated variants

Wild-type human disease-associated ORFs were selected from the human ORFeome version 8.1 [27]. As described in Sun *et al* [17], human ORFs with disease-associated variants were constructed by site-directed mutagenesis using the Thermo Scientific Phusion Site-Directed Mutagenesis Kit. The Gateway Donor plasmid was amplified using phosphorylated primers that introduce the desired changes followed by a 5-minute, room-temperature ligation reaction. The resulting plasmid was then transformed into NEB5α competent *E. coli* cells (New England Biolabs).

Constructing the *S. cerevisiae* expression plasmid pHYC-URA-ORF/GFP

All expressed ORFs used in these studies—including wild-type human disease-associated ORFs, human ORFs with constructed alleles, and the GFP control—were transferred into the destination vector pCM188- URA [55] by Gateway LR reactions using the All Gateway LR Clonase enzyme kit from Life Technologies. The destination vector pCM188-URA was obtained from ATCC, and subsequently altered to be Gateway compatible following the procedure applied in Sun et al (Genome Research 2016) to vectors pHYCDest-LEU2 and pHYCDest-NatMX. Plasmids generated by Gateway LR cloning were transformed into NEB5α competent *E. coli* cells (New England Biolabs) and selected on LB Agar plates with 100μg/mL Ampicillin. All plasmid DNA samples were isolated and purified using the NucleoSpin 96 Plasmid toolkit (Ref: 740625.24) and confirmed by Sanger sequencing. Plasmids carrying expressed ORFs were then transformed into the corresponding yeast temperature-sensitive strains.

Yeast-based functional complementation assay

Yeast temperature-sensitive (TS) strains carrying human ORFs or GFP control were spotted in a 10-fold dilution series and grown at a range of temperatures (room temperature of ~24˚C, and 28, 30, 32, 33, 34, 35, 36 and 38˚C). Results were interpreted by comparing the growth difference between the yeast strains expressing human genes and the corresponding control strain expressing the GFP gene. Each test was initially performed twice and pairs were found in at least one replicate were considered complementation candidates. For confirmation experiments, we went back to the glycerol stock of the relevant yeast TS strain, and re-transformed the expression plasmid for the candidate complementing human gene (and negative GFP control) into this fresh isolate. We further considered only those candidates passing a third replicate functional complementation assay.

Predicting functional effects for missense variants

To predict functional effects for each missense genetic variant, we assessed complementation with the above-described yeast spotting assays and assigned a semi-quantitative Failure-to-Complement (FC) score (corresponding to the previously-described FCS score [17]). Semi-quantitative FC scores were assigned to each variant: 0 (wild-type-like complementation), 0.6 (reduced complementation), 0.8 (severely reduced complementation) and 1 (complete loss of complementation). The predicted functional impact score for disease-associated variants were generated by the two best-performing computational methods in our previous study [17]: Polymorphism Phenotyping v2 (PolyPhen-2 [36, 56]) and PROVEAN [37].

As a pre-processing step before combining computational and FC scores, we followed the same method introduced by Sun *et al*, to calibrate each scoring system. To calculate the calibrated score for each disease variant, we combined the variants tested in both paralog-based and ortholog-based complementation assays, and randomly separated them into 10 groups. Precision within the FC (or PolyPhen-2) training data was calculated at different thresholds of each scoring method. precision and recall performance was then evaluated for seven methods of combining the two scores: minimum, maximum, mean, and four alternative weighted mean methods, where each method takes the form of α × calibrated-FC-score + (1-α) × calibrated-PolyPhen2-score. Specifically, methods w1, w2, w3 and w4 corresponded to α values of 0.9, 0.8, 0.7 and 0.6.

The area under the precision-recall curve (AUPRC) was calculated using R package "PRROC". When comparing the performance of functional complementation assays in predicting disease associated variants in either aligned or not aligned regions, we wished to account for the fact that changing the prior probability of pathogenicity can alter precision estimates. Therefore, performance was estimated using the ratio of AUPRC relative to the prior probability (designated as AUPRC_norm) instead of AUPRC.

Supporting information

S1 Table. Human-yeast complementing paralogous pairs.
(XLSX)

S2 Table. 1060 Human-yeast homologous pairs tested.
(XLSX)

S3 Table. 35 human disease variants tested with FC Score. (LOF: loss of function, GOF: gain of function).
(XLSX)

S4 Table. Pathway information of human kinases which are paralogs of yeast Kin28.
(XLSX)

S1 Fig. Functional assay result of complementing paralogous pairs at permissive, semi-permissive and non-permissive temperatures.
(PNG)

S2 Fig. Functional assay result and FC score of human variants at permissive, semi-permissive and non-permissive temperatures.
(PNG)

S1 File. Testing association between sequence similarity and complementation relationships for human-yeast paralogs.
(DOCX)

Author Contributions

Conceptualization: FY SS FPR.

Formal analysis: FY.

Investigation: FY GT MC BJA CB DEH MV.

Writing – original draft: FY.

Writing – review & editing: FPR FY SS GT MC DEH MV BJA CB.

References

1. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. https://doi.org/10.1038/nature11632 PMID: 23128226.
2. Kruger WD, Cox DR. A yeast assay for functional detection of mutations in the human cystathionine beta-synthase gene. Hum Mol Genet. 1995; 4(7):1155–61. PMID: 8528202.
3. Trevisson E, Burlina A, Doimo M, Pertegato V, Casarin A, Cesaro L, et al. Functional complementation in yeast allows molecular characterization of missense argininosuccinate lyase mutations. The Journal of biological chemistry. 2009; 284(42):28926–34. https://doi.org/10.1074/jbc.M109.050195 PMID: 19703900.
4. Kuzmin E, Sharifpoor S, Baryshnikova A, Costanzo M, Myers CL, Andrews BJ, et al. Synthetic genetic array analysis for global mapping of genetic networks in yeast. Methods Mol Biol. 2014; 1205:143–68. https://doi.org/10.1007/978-1-4939-1363-3_10 PMID: 25213244.
5. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. Cell. 2005; 122(6):957–68. https://doi.org/10.1016/j.cell.2005.08.029 PMID: 16169070.
6. Rolland J, Condamine FL, Jiguet F, Morlon H. Faster speciation and reduced extinction in the tropics contribute to the Mammalian latitudinal diversity gradient. PLoS Biol. 2014; 12(1):e1001775. https://doi.org/10.1371/journal.pbio.1001775 PMID: 24492316.
7. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005; 437(7062):1173–8. https://doi.org/10.1038/nature04209 PMID: 16189514.
8. Botstein D, Fink GR. Yeast: an experimental organism for 21st Century biology. Genetics. 2011; 189 (3):695–704. https://doi.org/10.1534/genetics.111.130765 PMID: 22084421.
9. Botstein D, Fink GR. Yeast: an experimental organism for modern biology. Science. 1988; 240 (4858):1439–43. PMID: 3287619.
10. Osborn MJ, Miller JR. Rescuing yeast mutants with human genes. Brief Funct Genomic Proteomic. 2007; 6(2):104–11. https://doi.org/10.1093/bfgp/elm017 PMID: 17698857.
11. Lee MG, Nurse P. Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. Nature. 1987; 327(6117):31–5. https://doi.org/10.1038/327031a0 PMID: 3553962.
12. Kachroo AH, Laurent JM, Yellman CM, Meyer AG, Wilke CO, Marcotte EM. Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science. 2015; 348 (6237):921–5. https://doi.org/10.1126/science.aaa0769 PMID: 25999509.
13. Peterson TA, Park D, Kann MG. A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. BMC genomics. 2013; 14 Suppl 3:S5. PMID: 23819456.
14. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014; 11 (8):801–7. https://doi.org/10.1038/nmeth.3027 PMID: 25075907.
15. Hamza A, Tammpere E, Kofoed M, Keong C, Chiang J, Giaever G, et al. Complementation of Yeast Genes with Human Genes as an Experimental Platform for Functional Testing of Human Genetic Variants. Genetics. 2015; 201(3):1263–74. https://doi.org/10.1534/genetics.115.181099 PMID: 26354769.
16. Sun S, Yang F, Tan G, Costanzo M, Oughtred R, Hirschman J, et al. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 2016; 26(5):670–80. https://doi.org/10.1101/gr.192526.115 PMID: 26975778.
17. Sun S, Yang F, Tan G, Costanzo M, Oughtred R, Hirschman J, et al. An extended set of yeast-based functional assays accurately identifies human disease mutations. Genome Res. 2016. https://doi.org/10.1101/gr.192526.115 PMID: 26975778.
18. Tamanoi F. Ras signaling in yeast. Genes Cancer. 2011; 2(3):210–5. https://doi.org/10.1177/1947601911407322 PMID: 21779494.
19. Heaton NS, Moshkina N, Fenouil R, Gardner TJ, Aguirre S, Shah PS, et al. Targeting Viral Proteostasis Limits Influenza Virus, HIV, and Dengue Virus Infection. Immunity. 2016; 44(1):46–58. https://doi.org/10.1016/j.immuni.2015.12.017 PMID: 26789921.
20. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics. 2014; 133(1):1–9. https://doi.org/10.1007/s00439-013-1358-4 PMID: 24077912.
21. Jensen RA. Orthologs and paralogs—we need to get it right. Genome biology. 2001; 2(8):INTERACTIONS1002. PMID: 11532207.

22. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol. 2004; 14(2):208–16. https://doi.org/10.1016/j.sbi.2004.03.011 PMID: 15093836.

23. Ponting CP, Russell RR. The natural history of protein domains. Annu Rev Biophys Biomol Struct. 2002; 31:45–71. https://doi.org/10.1146/annurev.biophys.31.082901.134314 PMID: 11988462.

24. Yang F, Petsalaki E, Rolland T, Hill DE, Vidal M, Roth FP. Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS Comput Biol. 2015; 11(3):e1004147. https://doi.org/10.1371/journal.pcbi.1004147 PMID: 25794154.

25. Bashton M, Chothia C. The generation of new protein functions by the combination of domains. Structure. 2007; 15(1):85–99. https://doi.org/10.1016/j.str.2006.11.009 PMID: 17223535.

26. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics. 2015; 200(2):413–22. https://doi.org/10.1534/genetics.115.175802 PMID: 25823446.

27. Team MGCP, Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, et al. The completion of the Mammalian Gene Collection (MGC). Genome Res. 2009; 19(12):2324–33. https://doi.org/10.1101/gr.095976.109 PMID: 19767417.

28. Li Z, Vizeacoumar FJ, Bahr S, Li J, Warringer J, Vizeacoumar FS, et al. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. Nature biotechnology. 2011; 29 (4):361–7. https://doi.org/10.1038/nbt.1832 PMID: 21441928.

29. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics. 2012; Chapter 1:Unit1 13. https://doi.org/10.1002/0471250953.bi0113s39 PMID: 22948725.

30. Cooper DN, Stenson PD, Chuzhanova NA. The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. Curr Protoc Bioinformatics. 2006; Chapter 1:Unit 1 13. https://doi.org/10.1002/0471250953.bi0113s12 PMID: 18428754.

31. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003; 21(6):577–81. https://doi.org/10.1002/humu.10212 PMID: 12754702.

32. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2014; 42(Database issue):D7–17. https://doi.org/10.1093/nar/gkt1146 PMID: 24259429.

33. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2009; 37(Database issue): D5–15. https://doi.org/10.1093/nar/gkn741 PMID: 18940862.

34. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2008; 36(Database issue): D13–21. https://doi.org/10.1093/nar/gkm1000 PMID: 18045790.

35. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42(Database issue):D980–5. https://doi.org/10.1093/nar/gkt1113 PMID: 24234437.

36. Sokic G, Dukanovic D. [Polyphen in the treatment of some diseases]. Stomatol Glas Srb. 1971; 18 (3):159–62. PMID: 5288161.

37. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7(10):e46688. https://doi.org/10.1371/journal.pone.0046688 PMID: 23056405.

38. Mooney S. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Briefings in bioinformatics. 2005; 6(1):44–56. PMID: 15826356.

39. Steward RE, MacArthur MW, Laskowski RA, Thornton JM. Molecular basis of inherited diseases: a structural perspective. Trends in genetics: TIG. 2003; 19(9):505–13. https://doi.org/10.1016/S0168-9525(03)00195-1 PMID: 12957544.

40. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006; 7:61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630 PMID: 16824020.

41. Dupuis N, Lebon S, Kumar M, Drunat S, Graul-Neumann LM, Gressens P, et al. A novel RAB33B mutation in Smith-McCort dysplasia. Hum Mutat. 2013; 34(2):283–6. https://doi.org/10.1002/humu.22235 PMID: 23042644.

42. Alshammari MJ, Al-Otaibi L, Alkuraya FS. Mutation in RAB33B, which encodes a regulator of retrograde Golgi transport, defines a second Dyggve—Melchior—Clausen locus. Journal of medical genetics. 2012; 49(7):455–61. https://doi.org/10.1136/jmedgenet-2011-100666 PMID: 22652534.

43. Cohen AR, Woods DF, Marfatia SM, Walther Z, Chishti AH, Anderson JM. Human CASK/LIN-2 binds syndecan-2 and protein 4.1 and localizes to the basolateral membrane of epithelial cells. J Cell Biol. 1998; 142(1):129–38. PMID: 9660868.

44. Atasoy D, Schoch S, Ho A, Nadasy KA, Liu X, Zhang W, et al. Deletion of CASK in mice is lethal and impairs synaptic function. Proc Natl Acad Sci U S A. 2007; 104(7):2525–30. https://doi.org/10.1073/pnas.0611003104 PMID: 17287346.

45. Piluso G, D'Amico F, Saccone V, Bismuto E, Rotundo IL, Di Domenico M, et al. A missense mutation in CASK causes FG syndrome in an Italian family. American journal of human genetics. 2009; 84(2):162–77. https://doi.org/10.1016/j.ajhg.2008.12.018 PMID: 19200522.

46. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005; 33(Database issue):D514–7. https://doi.org/10.1093/nar/gki033 PMID: 15608251.

47. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42(Database issue):D1001–6. https://doi.org/10.1093/nar/gkt1229 PMID: 24316577.

48. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology. 2011; 7:539. https://doi.org/10.1038/msb.2011.75 PMID: 21988835.

49. Chartier M, Chenard T, Barker J, Najmanovich R. Kinome Render: a stand-alone and web-accessible tool to annotate the human protein kinome tree. PeerJ. 2013; 1:e126. https://doi.org/10.7717/peerj.126 PMID: 23940838.

50. Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database (Oxford). 2012; 2012:bar062. https://doi.org/10.1093/database/bar062 PMID: 22434830.

51. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet. 2016; 48(12):1570–5. https://doi.org/10.1038/ng.3700 PMID: 27749844.

52. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2010; 38(Database issue):D211–22. https://doi.org/10.1093/nar/gkp985 PMID: 19920124.

53. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, et al. Pfam: clans, web tools and services. Nucleic acids research. 2006; 34(Database issue):D247–51. Epub 2005/12/31. https://doi.org/10.1093/nar/gkj149 PMID: 16381856.

54. Sonnhammer EL, Ostlund G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic acids research. 2015; 43(Database issue):D234–9. https://doi.org/10.1093/nar/gku1203 PMID: 25429972.

55. Gari E, Piedrafita L, Aldea M, Herrero E. A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in Saccharomyces cerevisiae. Yeast. 1997; 13(9):837–48. https://doi.org/10.1002/(SICI)1097-0061(199707)13:9<837::AID-YEA145>3.0.CO;2-T PMID: 9234672.

56. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512.