Comprehensive prediction of synthetic lethality between paralog pairs in cancer cell lines

Barbara De Kegel[1], Niall Quinn[1], Nicola A. Thompson[2], David J. Adams[2], Colm J. Ryan[1]*

[1] School of Computer Science and Systems Biology Ireland, University College Dublin, Dublin, Ireland
[2] Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

* Corresponding author: colm.ryan@ucd.ie

Abstract

Pairs of paralogs may share common functionality and hence display synthetic lethal interactions. As the majority of human genes have an identifiable paralog, exploiting synthetic lethality between paralogs may be a broadly applicable approach for targeting gene loss in cancer. However, only a minority of human paralog pairs have been tested for synthetic lethality to date and the factors predictive of synthetic lethality are largely unknown. Here we derive a set of synthetic lethal paralog pairs by integrating molecular profiling data with CRISPR screens in cancer cell lines. Using this resource we identify a number of features that are individually predictive of synthetic lethality between paralogs, including shared protein-protein interactions and evolutionary conservation. We develop a machine-learning classifier to predict, with explanations, which paralog pairs are most likely to be synthetic lethal. We show that our classifier accurately predicts the results of combinatorial CRISPR screens in cancer cell lines and furthermore can distinguish pairs that are synthetic lethal in multiple cell lines from those that are cell-line specific.

Introduction

Gene duplication is the primary means by which new genes are created (Zhang, 2003) and consequently the majority of human genes are duplicates (paralogs) (Zerbino *et al*, 2018). Although paralog pairs may functionally diverge over time, many of them maintain at least some functional overlap, even after long evolutionary periods (Conant & Wolfe, 2008; Vavouri *et al*, 2009). This functional overlap may allow paralog pairs to buffer each other's loss, contributing to the overall ability of cells and organisms to tolerate genetic perturbations. Indirect evidence for the contribution of paralog buffering to genetic robustness is provided by the finding that, in both model organisms and cancer cell lines, paralog genes are less likely to be essential than genes with no identifiable paralogs (De Kegel & Ryan, 2019; Gu *et al*, 2003; Wang *et al*, 2015; Blomen *et al*, 2015; Dandage & Landry, 2019). More direct evidence for functional buffering is provided by observations of synthetic lethality between paralog pairs (DeLuna *et al*, 2008; Dean *et al*, 2008; VanderSluis *et al*, 2010; Dede *et al*, 2020; De Kegel & Ryan, 2019). Synthetic lethality occurs when the perturbation of two genes individually is well tolerated but their combined perturbation is lethal. Synthetic lethal relationships appear to be much more frequent among paralog pairs than other gene pairs – in the budding yeast *Saccharomyces cerevisiae*, where comprehensive double perturbation screens have been performed, 25-35% of paralog pairs are synthetic lethal, compared to <5% of random gene pairs. It is not yet established how frequent synthetic lethality is among paralog pairs in human cells.

As gene loss is a common feature of tumour cells, synthetic lethality has long been proposed as a means of selectively killing tumour cells while limiting toxicity to healthy, non-transformed cells (Kaelin, 2005; O'Neil *et al*, 2017; Hartwell *et al*, 1997). The first therapies that exploit synthetic lethality - the use of PARP inhibitors in patients with loss-of-function *BRCA1* or *BRCA2* mutations - have now been approved for clinical use in select breast, ovarian, and prostate tumours (Lord & Ashworth, 2017). Many large-scale efforts are now focussed on the identification of new synthetic lethal treatments, in particular for cancers with recurrent genetic alterations that currently have no effective targeted therapies associated with them (Huang *et al*, 2020; Behan *et al*, 2019; Han *et al*, 2017). However, it remains challenging to identify new, robust synthetic lethal interactions in cancer due to a lack of understanding of what makes a gene pair likely to be synthetic lethal.

As >60% of human genes have an identifiable paralog, and as paralog pairs are more likely to be synthetic lethal than random gene pairs, exploiting synthetic lethality among paralog pairs may be a widely applicable approach for targeting a variety of gene loss events in cancer. Indeed, several groups have already identified synthetic lethal interactions involving paralogs that are recurrently lost in cancer, resulting in about a dozen validated paralog synthetic lethalities (Hoffman *et al*, 2014; Oike *et al*, 2013; Ehrenhöfer-Wölfer *et al*, 2019; Helming *et al*, 2014; van der Lelij *et al*, 2020; Benedetti *et al*, 2017; Szymańska *et al*, 2020; Lord *et al*, 2020; Muller *et al*, 2012; Schick *et al*, 2019; Ogiwara *et al*, 2016; Tsherniak *et al*, 2017; Viswanathan *et al*, 2018; Dey *et al*, 2017). These include approaches targeting recurrently altered driver genes, such as the synthetic lethal interaction between the tumour suppressor *SMARCA4* and its paralog *SMARCA2*, and approaches targeting passenger gene loss events, such as the synthetic lethal interaction between *ME2* and its paralog *ME3. ME2* is not believed to be a cancer driver gene, but is located in the same chromosomal

region as the tumour suppressor *SMAD4* and is frequently homozygously deleted in tumours that have lost *SMAD4* (Dey *et al*, 2017).

To date, synthetic lethal interactions between paralog pairs have been primarily discovered through the analysis of large scale loss-of-function screens (e.g. *ARID1A* mutant cell lines were found to be sensitive to loss of its paralog *ARID1B* through analysis of genome-wide shRNA screens of 165 tumour cell lines (Helming *et al*, 2014)) or through the rational testing of individual paralog pairs (*ME3* was tested for synthetic lethality with *ME2* because of an understanding of their role in metabolism (Dey *et al*, 2017)). Recently, combinatorial screening using CRISPR approaches has emerged as an alternative approach for determining synthetic lethality between paralog pairs, allowing researchers to quantify synthetic lethal relationships between hundreds of pairs in pooled screens (Gonatopoulos-Pournatzis *et al*, 2020; Dede *et al*, 2020; Thompson *et al*, in press). However, there are tens of thousands of paralog pairs in the human genome (Zerbino *et al*, 2018) and consequently there is a need for approaches to prioritize those pairs most likely to be synthetic lethal. While a number of statistical and machine learning approaches have been developed for genome-wide prediction of synthetic lethality (Benstead-Hume *et al*, 2019; Lee *et al*, 2018), there are as of yet no approaches that are tailored specifically to paralogs. We, and others, have previously identified a small number of features (sequence identity, duplication mode and protein complex membership) that are somewhat predictive of synthetic lethality between paralog pairs (De Kegel & Ryan, 2019; DeLuna *et al*, 2008; Guan *et al*, 2007) but a more comprehensive analysis of what makes a paralog pair likely to be synthetic lethal has yet to be performed.

Here we develop an ensemble classifier to predict synthetic lethality among paralog pairs in cancer cell lines, and attempt to explain why specific paralog pairs are likely (or unlikely) to be synthetic lethal. We first derive a ‘bronze standard’ set of synthetic lethal (SL) paralog pairs by integrating molecular profiling data with genome-wide CRISPR screens in 762 cancer cell lines from the DepMap project (Meyers *et al*, 2017). We then assess the ability of 22 different features of paralog pairs to distinguish synthetic lethal from non-synthetic lethal pairs in this dataset. We identify a number of individual features, such as shared protein-protein interactions and evolutionary conservation, that are predictive of synthetic lethal relationships. We then integrate all features using a random forest classifier and show that, for our ‘bronze standard’ dataset, this classifier can significantly outperform predictions based on the best individual features. We further demonstrate that our classifier can accurately predict the results of two independent combinatorial CRISPR screens. Finally, we make predictions for all paralog pairs, accompanied by local explanations that indicate why specific pairs are likely to be SL.

Results

Identification of a ‘bronze standard’ set of synthetic lethal paralog pairs

The number of experimentally validated synthetic lethal interactions between paralogs in human cells is extremely small - from the literature we identified only 12 SL paralog pairs with low throughput experimental validation (Table S1). This precludes any analysis of what

features make a paralog pair more or less likely to be SL. In the absence of an experimentally determined 'gold standard', we used a computational approach to identify a 'bronze standard' set of putative SL and non-SL paralog pairs from existing CRISPR screen data (Fig. 1A). Genome-wide CRISPR screens report on the fitness cost of perturbing each protein-coding gene in a given cell line and, when combined with molecular profiling data, can provide insight into which genes are essential in cell lines with specific alterations. We integrated genome-wide CRISPR screens of 762 cancer cell lines (Meyers *et al*, 2017) (DepMap 20Q2) with molecular profiling data for the same cell lines (Ghandi *et al*, 2019) and searched for cases where loss or silencing of one member of a paralog pair (termed A2) results in increased dependency on the other member (termed A1) (Fig. 1A). As in previous work (De Kegel & Ryan, 2019) we first reprocessed the DepMap CRISPR screen results to remove guides that might target multiple genes, an issue that disproportionately affects sequence-similar paralogs and which could confound our results (Fortin *et al*, 2019) (Table S2, see Methods). We started our analysis with a comprehensive list of protein-coding paralog pairs from Ensembl (Zerbino *et al*, 2018) (see Methods) and filtered it according to a number of criteria (Fig 1A). Briefly, we removed pairs with low (<20%) sequence identity, pairs from very large paralog families, pairs with tissue specific-expression, and pairs that were not present in our reprocessed DepMap CRISPR screen data (see Methods). Finally, we required that at least one gene in the pair is lost - i.e. subject to a homozygous deletion, loss-of-function mutation, or loss of mRNA expression - in at least 10 (and max. 90%) of the 762 cancer cell lines (Table S3, see Methods). The application of all five filtering steps culminated in a set of 3,810 candidate paralog pairs for further analysis (Fig. 1A).

To identify associations between loss of one paralog (A2) and altered dependency upon the other paralog (A1) among these candidate pairs we used a multiple linear regression model with A1 dependency as the dependent variable and A2 status (wildtype or loss) and cell line lineage as the independent variables (see Methods). Cell line lineage was included to account for variation in A1 dependency that is due to tissue type. For 126 of the 3,810 candidate pairs we found that A2 status was significantly associated with A1 dependency (10% FDR), i.e. A2 loss resulted in increased A1 dependency, indicative of a synthetic lethal relationship (Table S4). These putative SL pairs include six of the previously validated synthetic lethalities that we curated from the literature (Table S1), including *ARID1A/ARID1B (Helming et al, 2014), STAG1/STAG2* (van der Lelij *et al*, 2020) and *SMARCA2/SMARCA4* (Oike *et al*, 2013), as well as several pairs with strong associations that have not, to our knowledge, been previously reported, such as *SEC24C/SEC24D*, *VPS26A/VPS26B* and *HSPA4/HSPH1* (Fig. 1B).

For our bronze standard dataset the 126 putative SL pairs were labelled as true positives, while the 3,508 pairs for which there was no association between A2 loss and A1 dependency (uncorrected $p>0.05$) were labelled as true negatives (Table S5). The remaining tested pairs (with uncorrected $p\leq0.05$ but FDR>10%) were discarded. The final dataset thus contains 3,634 pairs, of which ~3.5% are considered SL.

Identification of features predictive of synthetic lethality between paralogs

We next sought to identify features that are predictive of synthetic lethality between paralogs. To that end we computed 22 features of paralog pairs (Table S6) including properties of the individual genes in the paralog pair (e.g. mean gene expression of the more

highly expressed paralog), and comparative properties that quantify some relationship between the two genes (e.g. mRNA expression correlation). The features we assembled are either Boolean (True/False) or quantitative properties and fall into three broad categories: sequence, neighborhood and expression features (Table S6). Features in the 'sequence' category are derived from analyses of the protein-coding sequences of the pairs and include sequence identity, measures of conservation in other species, gene age, and the ratio of the lengths of the two proteins. 'Neighborhood' features pertain mainly to the protein-protein interactions (PPI) and the protein complex membership of the genes in each paralog pair. The 'expression' features are calculated from measurements of mRNA expression in non-diseased tissues (GTEx Consortium, 2020) and comprise the average expression and expression correlation of the genes in each paralog pair. Three of the features - sequence identity, duplication mode, and protein complex membership - are ones for which we previously observed an association with putative synthetic lethality in human cells (De Kegel & Ryan, 2019). Sequence identity and duplication mode have also been associated with paralog synthetic lethality in budding yeast (Guan *et al*, 2007; DeLuna *et al*, 2008).

We assessed each feature's ability to predict synthetic lethality in our bronze standard dataset by treating each feature as an individual classifier and computing area under the receiver operating characteristic curve (ROC AUC) and average precision (Fig. 2A, Table S6, see Methods). Average precision is a measure of the area under the precision-recall (PR) curve and is often a more sensitive measure of performance when a classification challenge has high class imbalance, i.e. when the positive examples are much rarer than the negative examples (Lever *et al*, 2016). As our set of paralog pairs has high class imbalance (many more pairs are not SL than are SL) we consider both metrics for evaluation. It should be noted that by chance the average precision is only ~3.5%, and thus average precision is expected to be small (in absolute terms) even for well-performing classifiers.

The two individual features that best predict synthetic lethality in our dataset are sequence identity and shared protein-protein interactors (measured as the -log10 *p*-value from a Fisher's exact test of the overlap of A1 and A2's interactors) (Fig. 2A). These features both have a ROC AUC of ~0.73 and an average precision of ~10%. Both features can be considered measurements of 'functional similarity' - with sequence identity measuring how similar the paralogs' protein sequences are and shared protein-protein interactions measuring how similar the two paralogs are in terms of their role in the protein-protein interaction network. Other features that measure similarity (fraction of shared domains, co-localisation, and co-expression) are predictive of synthetic lethality but to a much lesser extent (Fig. 2A).

Features that capture evolutionary conservation appear to be highly predictive of synthetic lethality, in particular a measure of the number of species in which a paralog pair has an identifiable ortholog ('Conservation Score', see Methods). Genes that have been conserved over a long evolutionary period are more likely to encode essential functionality (Chen *et al*, 2012), which could explain why highly conserved paralog pairs are more likely to be SL. Similarly, we found that gene age (Capra *et al*, 2012), a measure of the phylogenetic age of a protein, is very predictive of SL.

We included a number of features relating to gene conservation in the model eukaryotes *S. cerevisiae* and *Schizosaccharomyces pombe* (Fig. 2A, Table S6 and Methods for feature

details). Although both are yeasts, they are very evolutionary distant (over 400 million years (Sipiczki, 2000)) and differ significantly in terms of their basic biology (e.g. *S. pombe* has RNAi machinery and splicing much more similar to metazoans), and in terms of their genome composition (*S. cerevisiae* has undergone a whole genome duplication while *S. pombe* has not (Wood, 2006)). For both species we have a comprehensive overview of which genes in their genome are essential, thanks to large scale gene deletion projects (Giaever *et al*, 2002; Kim *et al*, 2010). We anticipated that having an essential *S. cerevisiae* or *S. pombe* ortholog would be very predictive of an SL relationship between a human paralog pair, but surprisingly we found that these two features were only weakly predictive. In fact, having any identifiable ortholog in *S. pombe* was more predictive of SL than having an essential *S. pombe* ortholog (average precision ~6.2% vs. ~5.3%), and the same trend was evident for *S. cerevisiae* (average precision ~4.9% vs. ~4.5%) (Fig. 2A). One possible explanation for this is that the orthologs in the two yeast species may themselves have been subject to a gene duplication event and hence may appear non-essential due to paralog buffering. However, even when restricting our analysis to cases in which there was a single ortholog in either species (i.e. a many-to-one mapping) the addition of essentiality information did not improve predictive power (Fig. S1). This suggests that in general having any identifiable ortholog in budding or fission yeast is more informative than having an essential ortholog.

Species conservation and protein age are both indicators of gene importance – as noted, a gene that is widely conserved is likely to perform some essential function. We also included in our feature set a number of more direct measures of gene importance derived from gene essentiality profiles calculated from CRISPR screens in cancer cell lines. We anticipated that paralog pairs that function in largely essential protein complexes (i.e. complexes with subunits that are frequently essential across the DepMap cell lines) or that have protein-protein interactions with essential genes are more likely to be SL and found that this is the case (the average precision associated with protein complex and shared PPI essentiality is ~7.8% and ~5.2% respectively) (Fig 2A).

Previous work has suggested that genes that encode proteins with a high number of protein-protein interactions are more likely to be essential, although the reasons for this association are much disputed (Hahn & Kern, 2005; He & Zhang, 2006; Zotenko *et al*, 2008; Wang *et al*, 2009). We tested whether paralog pairs with more protein-protein interactions might be more likely to be SL and found that this is indeed the case (Fig 2A, ROC AUC ~0.64, average precision ~6%). However, this feature was nowhere near as predictive as shared protein-protein interactions (ROC AUC ~0.73, average precision ~10%), suggesting that the total number of interactions that the genes in a paralog pair have is less important than the fraction they have in common.

The final set of features we found to be predictive relate to the possibility of compensation by other paralogs in a family. In many cases a pair of paralogs belongs to a much larger family of paralogs. We found that SL paralogs tend to come from smaller families (the median family size for SL paralogs is 5, compared to 8 for non-SL paralogs) and are more likely to be each other's closest paralog (Fig. 2A).

Overall, the features we have assembled for paralog pairs quantify 1) the extent of their functional similarity, e.g. sequence identity or co-expression; 2) whether their shared

functionality is likely to be essential, e.g. membership of an essential protein complex; and 3) whether there are other potential sources of genetic redundancy that would mask a synthetic lethal relationship, e.g. paralog family size. We observe that paralog pairs are likely to be SL if they are functionally similar, if their shared function is important, and if they uniquely compensate for each other's loss.

An ensemble classifier to predict synthetic lethality among paralog pairs

Having identified several features that are individually predictive of synthetic lethality between paralog pairs in our bronze standard dataset, we next asked whether combining these features into an ensemble classifier could provide greater predictive power. Though several of the features are expected to be correlated with each other, we retain all features as they can still contribute unique information to the classifier. For example, we previously found that whole genome duplicates are more likely to be in protein complexes, and while this could partially explain why whole genome duplicates are more likely to be synthetic lethal, we also found that duplication mode was still predictive when we controlled for protein complex membership (De Kegel & Ryan, 2019).

We trained a random forest classifier (Breiman, 2001), i.e. an ensemble of decision trees, using our bronze standard set of paralog pairs annotated with all 22 features (see Methods). A tree-based ensemble classifier makes it possible to capture interactions between features as well as non-linear relationships between a given feature and paralog synthetic lethality. We find that the predictions made by the random forest classifier confidently outperform sequence identity, one the best individual features, which we use as the baseline for comparison (Fig. 2B). Specifically, the classifier achieves a ROC AUC of ~0.80 and average precision of ~23% over the training data, while the metrics for sequence identity were ~0.73 and ~10% respectively.

An ensemble classifier can accurately predict synthetic lethality in combinatorial CRISPR screens

We found that, for our bronze standard set of SL and non-SL paralog pairs, our classifier significantly outperformed sequence identity alone. To evaluate the classifier using an orthogonal approach, we next evaluated its ability to predict the results of dual-gene knockout screens of paralog pairs. We obtained data from two independent combinatorial CRISPR screening efforts - one using a Cas9 based approach (Thompson *et al*, in press) and one using a Cas12a approach (Dede *et al*, 2020) (see Methods). For these screens individual and paired guide RNAs are used to determine the expected and observed fitness consequences, respectively, of perturbing each gene pair. Synthetic lethality can then be identified where the observed fitness defect is significantly greater than that expected from the fitness defects induced by the individual guide RNAs.

Thompson et al. screened 1,191 gene pairs in three cell lines (A375, Mewo and RPE); we filtered this down to 541 paralog pairs that overlap with our list from Ensembl and that are not broadly essential (see Methods). Of these pairs, 71 were identified as SL in at least one cell line and 23 were identified as SL in at least two cell lines. Dede et al. screened 393 paralog pairs that overlap with our list from Ensembl (see Methods), also in three different cell lines (A549, HT29 and OVCAR8), and identified 24 pairs as SL in at least one cell line,

and 19 as SL in at least two cell lines. We initially tested how well our classifier could distinguish those paralog pairs that are SL in multiple cell lines from those that are SL in zero cell lines. We found that our classifier can make accurate predictions for both screens - up to a recall of ~29-34% we observe precision of at least 50% (Fig. 3A, Fig. 3B). In addition, we note that our classifier clearly outperforms sequence identity - the average precision associated with the two screens is ~33% and ~38% for our predictions, compared to ~12% and ~18% for sequence identity alone. We repeated this analysis using SLs observed in one or more cell lines (Fig. S2A, Fig. S2B) and found that our classifier again outperforms sequence identity.

Some of the paralog pairs screened by Dede et al. and Thompson et al. were present in our bronze standard dataset. To ensure that our classifier wasn't overfitting to our training (bronze standard) dataset we repeated the above evaluation for just the unseen pairs - i.e. the pairs in each screen that were not included in our bronze standard dataset. We found that our classifier still performs well when only considering the unseen pairs (Fig. S2C, Fig. S2D), indicating that the classifier's performance on these independent datasets is not based on overfitting to the common pairs.

There are no previous methods for specifically predicting synthetic lethality between paralog pairs, but there are several approaches for predicting SL between all gene pairs (i.e. both paralog pairs and unrelated gene pairs). To further evaluate our classifier, we compared its performance to a recently developed high-performing SL classifier called SLant (Synthetic Lethal analysis via Network topology) (Benstead-Hume *et al*, 2019). Due to the input requirements of SLant, not all paralog pairs in our analysis datasets were assigned a predicted score and consequently we focussed our analysis on those pairs for which predictions were available. We found that our classifier outperformed the predictions from SLant on both the Dede et al. and the Thompson et al. datasets (Fig. S3). This was true when considering pairs identified as SL in two or more cell lines (Fig. S3A and S3B) or in at least one single cell line (Fig. S3C and S3D). For instance - analysing the hits from Dede et al. observed in at least one cell line our classifier achieved an average precision of 54% vs 21% for SLant (Fig. S3D). Overall this suggests that in comparison to a generic SL classifier, our paralog-specific SL classifier is better able to capture the factors of a potential synthetic lethal relationship between paralog pairs.

Distinguishing robust synthetic lethal interactions from cell-line specific synthetic lethals

It has become clear that many synthetic lethal interactions operate only within specific genetic backgrounds or contexts (Ryan *et al*, 2018; Henkel *et al*, 2019; Lord *et al*, 2020; Ku *et al*, 2020). Interactions that are observed in multiple cell lines are more likely to be robust in the face of the molecular heterogeneity that exists in tumours, and consequently may be of significantly greater therapeutic importance (Ryan *et al*, 2018). To assess whether our classifier could distinguish these robust synthetic lethal interactions, observed in multiple cell lines, from those seen in only a single cell line we made predictions for the full set of 541 paralog pairs screened by Thompson et al. We split these according to the number cell lines in which the pair was identified as SL and found that our classifier captures some information as to which synthetic lethal relationships might be more robust - on average, pairs that were hits in three cell lines have a higher predicted score than those that were hits in only two,

which in turn have higher scores than those that were hits in only one cell line (Fig. 3C). We find that the difference in the prediction scores is significant for hits in zero vs. one cell line (Mann-Whitney test $p<1x10^{-5}$), one vs. three cell lines (Mann-Whitney test $p\simeq0.005$), and one vs. two cell lines (Mann-Whitney test $p\simeq0.035$), but not for hits in two vs. three cell lines. We observed a similar trend in the predictions for the full set of paralog pairs screened by Dede et al. but the differences between groups are not significant, likely because their screens showed strong agreement across cell lines, i.e. there were few cell-line specific hits (Fig. S4).

Understanding why specific paralog pairs are likely or unlikely to be synthetic lethal

Having validated our classifier using data from two combinatorial CRISPR screens, we next made predictions for all ~36.6k paralog pairs from Ensembl that have at least 20% sequence identity (in at least 1 direction of gene-to-gene comparison) (Table S7). These scores provide a means of ranking all paralog pairs according to their likelihood of being synthetic lethal. We found that, with the exception of one pair, the paralog SLs with previous experimental validation (Table S1) were in the top 3% most highly ranked gene pairs (Fig. 4A).

Although we found a number of features that are individually predictive of synthetic lethality (Fig. 2A) we note that not all of the pairs with high predicted scores have what might be expected values for these features. In general, pairs with high-sequence identity are more likely to be synthetic lethal, but a number of highly ranked pairs have only moderate sequence identity (e.g. *UBB*/*UBC*, a pair which is not present in our bronze standard dataset but has previously been validated as SL (Table S1) (Tsherniak *et al*, 2017), is ranked in the top 3% of gene pairs but has sequence identity close to the average of all paralog pairs). Similarly, in general pairs from small paralog families are more likely to be synthetic lethal but some paralog pairs from larger families are highly ranked (e.g. *RAB1A/RAB1B*, a pair not present in our bronze standard but recently identified as synthetic lethal in a high-throughput screen (Blomen *et al*, 2015), is ranked in the top 0.6% of pairs but belongs to a paralog family of size 15). To understand which factors influence the predictions for a given paralog pair, we made use of TreeExplainer, a recently developed algorithm to compute local explanations for ensemble tree-based predictions such as those from our random forest classifier (Lundberg *et al*, 2020). Briefly, TreeExplainer computes how much each feature pushes a specific prediction higher or lower. This numeric measure of feature credit is termed a SHAP (SHapley Additive exPlanations) value. More technically speaking, the SHAP values for a given paralog pair sum to the difference between the prediction for that pair and the average prediction for the classifier (i.e. the average prediction over all pairs in the training data). To complement the SHAP values (Table S8) we also compute z-scores for all the feature values (see Methods) - this provides us with a SHAP profile that links feature values and feature impacts (on prediction) for each pair. An example SHAP profile, for *AKT1/AKT2*, is shown in Fig. 4B.

These SHAP profiles provide a visual overview of what features influence the prediction for a specific paralog pair. For example, the paralog pairs *ARID1A/ARID1B* and *VPS4A/VPS4B* are similarly ranked and have both been previously validated as synthetic lethal (Helming *et al*, 2014; Szymańska *et al*, 2020; Lord *et al*, 2020). A comparison of their SHAP profiles reveals that the two pairs are highly ranked for very different reasons (Fig. 4C). The profile

for *ARID1A/ARID1B* shows that its prediction is mainly influenced by its membership of essential protein complex(es) and a significant overlap in the protein-protein interactors of *ARID1A* and *ARID1B*. The sequence identity of the two proteins, which is moderate, does not significantly influence the prediction. In contrast, the prediction for *VPS4A/VPS4B* is mainly driven by their high sequence similarity and by their old evolutionary age.

In cases where a gene has multiple paralogs, the SHAP profiles can be used to explain why one paralog pair is ranked more highly than others. For example, the tumour suppressor *STAG2* has previously been shown to be synthetic lethal with its paralog *STAG1* (van der Lelij *et al*, 2020). Consistent with this, one of the highest ranked paralog pairs from our classifier is *STAG1/STAG2* (Fig. 4A). *STAG2* has an additional paralog, *STAG3*, but our classifier ranks this pair significantly lower than *STAG1/STAG2* (Fig. 4A). Consistent with these rankings *STAG1*, but not *STAG3*, was recently identified as a hit in a genome-wide CRISPR screen for *STAG2* synthetic lethal interactions (van der Lelij *et al*, 2020). Looking at the SHAP profile for *STAG1/STAG2,* we observe that its prediction was pushed up substantially by high sequence identity and membership in essential complex(es), and to a lesser extent by some other features, such as shared protein-protein interactors and *STAG1* and *STAG2* being each other's closest paralogs (Fig. 4D, left). While still being pushed up by its membership of essential complex(es), the prediction for *STAG2/STAG3* is simultaneously pushed down by its moderate sequence identity, lack of shared interactors, and the genes not being each other's closest paralogs (Fig. 4D, right). This example indicates that SHAP profiles can reveal why one paralog pair from a paralog family is ranked more highly than another despite the two pairs having some common characteristics. Fig. S5 provides an additional example, showing why *AKT1* is predicted to be synthetic lethal with *AKT2* and *AKT3* but not with its 11 other paralogs - *AKT1/AKT2 and AKT1/AKT3* were also previously identified as SL in a combinatorial screen (Najm *et al*, 2018).

***ASF1A/ASF1B* and *COPS7A/COPS7B* are synthetic lethal pairs**

We selected two of our highly ranked predicted synthetic lethal interactions for further validation - *ASF1A/ASF1B* and *COPS7A/COPS7B* (Fig. 5A). Neither pair is present in our 'bronze standard' set of synthetic lethals nor in our manually curated set of validated synthetic lethals (Table S1). We selected *ASF1A/ASF1B* for validation because it is among the highest ranked predictions (top 0.1%) and because *ASF1A* is subject to frequent homozygous deletion in multiple cancer types including prostate adenocarcinoma, pancreatic adenocarcinoma and diffuse large B cell lymphoma (Lee *et al*, 2017).The SHAP profile for *ASF1A/ASF1B* reveals that it is highly ranked among our predictions because the two paralogs share high sequence identity (70%), have a high proportion of shared interactions, and function as part of multiple largely essential complexes (Histone H3.1 and H3.3 complexes (Tagami *et al*, 2004)) (Fig. 5B). To test this synthetic lethality we used the same approach as in (Lord *et al*, 2020) – we obtained knockouts of *ASF1A* and *ASF1B* in the near haploid HAP1 cell line and tested whether these knockout cells were more sensitive to the inhibition, via siRNA, of the other paralog. In both knockouts we found this to be the case – *ASF1A* knockout cell lines were more sensitive than wild-type parental cells to the inhibition of *ASF1B* while *ASF1B* knockouts were more sensitive to the inhibition of *ASF1A* (Fig. 5C, Table S9). The fitness impact of the siRNA targeting *ASF1A/ASF1B* in the wild-type cells was similar to that observed for a scrambled siRNA control, suggesting that inhibition of either gene in wild-type cells is well tolerated.

We next sought to validate the synthetic lethality between *COPS7A* and *COPS7B.* This pair is further down the list of predictions (top 1.5%) than *ASF1A*/*ASF1B* and has significantly lower sequence identity (47%). The SHAP profile reveals that the prediction for *COPS7A/COPS7B* is pushed down slightly by their moderate sequence identity, but pushed up by their high proportion of shared protein-protein interactions and their membership of an essential complex (the COP9 Signalosome (Seeger *et al*, 1998)) (Fig. 5D). We used the same approach to test the synthetic lethality between this pair and again found that *COPS7A* knockout cells were more sensitive to the inhibition of *COPS7B* and vice-versa (Fig. 5E, Table S9).

Discussion

In this work we have developed an approach for predicting synthetic lethality among human paralog pairs. We established a bronze standard set of (non-)SL paralog pairs that allowed us to evaluate the predictive power of 22 diverse features of paralog pairs, as well as a random forest classifier that incorporates these features. We find that sequence identity and shared protein-protein interactors are the best individual predictors of synthetic lethality between paralogs. However, the predictions made by our classifier significantly outperform these features alone. Furthermore, we show that our classifier, trained with data from single gene perturbation screens, can accurately predict the results of two independent combinatorial CRISPR screens of paralog pairs (Thompson *et al*, in press) and (Dede *et al*, 2020). Finally, we generate synthetic lethality predictions, with local explanations based on feature credit, for ~36.6k paralog pairs. These predictions and their explanations can be viewed at cancergd.org/static/paralogs.

A bronze standard set of synthetic lethal and non-synthetic lethal pairs

We have developed a ‘bronze standard’ set of 126 synthetic lethal and 3,508 non-synthetic lethal pairs. We deliberately refer to this as a bronze standard, rather than gold standard, because we anticipate that it contains both false positives (pairs that are incorrectly identified as synthetic lethal) and false negatives (pairs that are incorrectly identified as not synthetic lethal). We anticipate that there are fewer false positives as we have selected a relatively stringent false discovery rate (10%) and controlled for factors that might contribute to false positives. In particular we exclude from our analysis gene pairs with tissue specific expression patterns and incorporate tissue as a covariate into our analysis to identify synthetic lethality. This should prevent us calling a synthetic lethal relationship in cases where paralogs might simply be essential in a specific lineage because they are only expressed in that lineage. We anticipate that there are a significantly higher number of false negatives. The principal reason for this is that our ability to identify genes with loss-of-function alterations is imperfect. We may have called gene loss where a gene only has a reduction in expression or copy number, or a heterozygous loss-of-function mutation. Such limitations could be overcome by, e.g. incorporating protein expression datasets for the same cell lines (Nusinow *et al*, 2020), but currently this would severely limit the number of cell lines and paralog pairs we could analyse. Despite its limitations we show that our bronze standard dataset can be used to build a classifier that can accurately predict synthetic

lethality in combinatorial screens, suggesting that it is more than sufficient to learn general trends that distinguish synthetic lethal paralog pairs from non-synthetic lethal paralog pairs. We anticipate that as more combinatorial screens of paralog pairs are performed it may be possible to improve our bronze standard set by integrating these with our analysis of the DepMap data - e.g. only calling positives/negatives that are supported by both types of screen.

Frequency of synthetic lethality between paralogs in human cell lines

Our bronze standard dataset of human SLs has a synthetic lethality rate of ~3.5%, which is significantly lower than that reported for budding yeast, where 25-35% of paralog pairs have been identified as SL. However, this lower frequency of synthetic lethality is reasonably consistent with the results of the two combinatorial screens of paralog pairs analysed in this study. Dede et al. (Dede *et al*, 2020) and Thompson et al. (Thompson *et al*, in press) both performed combinatorial CRISPR screens in three cell lines to identify synthetic lethality between paralog pairs. Both screens had some selection bias in terms of the paralogs screened (high sequence identity for Dede et al., species conservation for Thompson et al.) which might be expected to increase the synthetic lethality rate. However, both screens still had a relatively low synthetic lethality rate – only ~4.8% (Dede *et al*, 2020) and ~4.2% (Thompson *et al*, in press) of screened paralog pairs were identified as hits in two or more cell lines. One explanation for this reduced synthetic lethality rate in human cells is that the median paralog family size is larger in humans (4 genes) than in *S. cerevisiae* (2 genes), in line with the ancestor of humans having undergone an extra round of whole genome duplication (Dehal & Boore, 2005). The consequences of perturbing any individual paralog pair in human cells may thus be masked by the presence of additional paralogs. Such higher-order relationships could be revealed by perturbing three or more genes simultaneously, as has been done in budding yeast (Haber *et al*, 2013; Kuzmin *et al*, 2020).

What makes a paralog pair likely to be synthetic lethal?

In this work we tried to answer this question generally, through analysis of the predictive power of individual features of paralog pairs (Fig. 2A) and on a pair-specific level, through analysis of the SHAP profiles of our predictions (Fig. 4-5).

Overall, we found that synthetic lethal paralog pairs are typically older and more widely conserved than non-synthetic lethal pairs. They are on average more highly expressed and are more likely to interact with essential genes. These findings suggest that the general importance of the genes' function is a key factor in making a paralog pair SL. Another key factor seems to be the degree of functional similarity between the genes - we found that SL paralog pairs tend to have higher sequence identity and greater overlap of protein-protein interactors compared to non-SL pairs. Finally, we found that, although many paralogs belong to large families, synthetic lethal pairs generally come from smaller families, suggesting that a paralog pair is more likely to be SL if the genes uniquely compensate for each other's loss.

Interestingly we found that a specific paralog pair can be highly ranked by our classifier even if it doesn't follow the overall trends that we observed. For example, although SL paralog pairs generally have higher sequence identity compared to non-SL paralog pairs, we identified pairs (e.g. *ARID1A/ARID1B* and *COPS7A/COPS7B*) that receive high predictions

despite average or below average sequence identity. This suggests that the causes of synthetic lethality among paralog pairs can differ significantly.

Robust vs. cell-line specific synthetic lethal interactions

The SL pairs in our bronze standard dataset were identified from combined analysis of heterogeneous cancer cell lines so they are expected to be largely context-agnostic, i.e. robust to cancer type and genetic background. There are likely additional context-specific synthetic lethal relationships (Ryan *et al*, 2018; Henkel *et al*, 2019) among our candidate paralog pairs but we cannot detect them here. In contrast, the paralog pairs screened by Thompson et al. and Dede et al., who both tested three different cell lines, are initially called SL in a cell-line specific fashion. We found that on average our classifier assigned higher prediction scores to pairs that were identified as SL in multiple cell lines, compared to pairs that were identified as SL in only one cell line. This suggests that our classifier, perhaps due to the training data, has some ability to distinguish robust synthetic lethalities from those that might be cell-line specific. Our classifier is based on context-agnostic features, e.g. average rather than cell-line specific complex essentiality, and can thus only make context-agnostic predictions of SL. However, as genetic interactions screens are carried out across a wider variety of cell lines it may become possible to develop a classifier that can make context-specific predictions.

***ASF1A/ASF1B* and *COPS7A/COPS7B* are validated synthetic lethal pairs**

We have demonstrated, using a low-throughput approach, that *ASF1A*/*ASF1B* and *COPS7A/COPS7B* are synthetic lethal in the HAP1 cell line. These paralog pairs represent a very high ranking prediction (*ASF1A/ASF1B*, top 0.1%) and a lower ranking prediction (*COPS7A/COPS7B,* top 1.5%) and a pair with high sequence identity (*ASF1A/ASF1B*, 70% id) and moderate sequence identity (*COPS7A/COPS7B*, 47% id). We note that *ASF1A*/*ASF1B* was also identified as a robust hit in the recent CRISPR-Cas9 screen of Thompson et al. (Thompson *et al*, in press), while *COPS7A/COPS7B* was screened by both Thompson et al. and Dede et al. (Dede *et al*, 2020), but only identified as a hit in Dede et al. Further work is required to establish if this is due to context-specific effects or if it is a false-negative in the Thompson et al. screen. Due to the relatively high frequency of homozygous deletion of *ASF1A* in different cancer types (Lee *et al*, 2017), the *ASF1A/ASF1B* synthetic lethal interaction may be of use as a therapeutic target.

Ranking paralog pairs most likely to have a synthetic lethal relationship

Paralogs are a rich source of synthetic lethality and exploiting this could be a widely applicable approach to targeting gene loss in cancer. However, there is a large number of paralog pairs in the human genome (>36k), and only a small fraction have been tested to date. The predictions made by our classifier provide a ranking of all paralog pairs in terms of how likely they are to exhibit a synthetic lethal relationship. These rankings can serve as a resource for selecting paralog pairs for inclusion in future combinatorial screens. To facilitate such analyses we have provided full details of our predictions, along with their SHAP profiles, in the supplementary materials.

Methods

Unless otherwise stated, all computational analysis was carried out using Python 3.7, Pandas 1.0.1 (McKinney, 2011), SciPy 1.4.1 (Virtanen *et al*, 2020), StatsModels 0.11.0 (Seabold & Perktold, 2010), numpy 1.18.5 (Harris *et al*, 2020) and scikit-learn 0.23.1 (Pedregosa *et al*, 2011). Notebooks containing the code for this project are available at github.com/cancergenetics/paralog_SL_prediction.

Paralog pairs

Protein-coding paralog gene pairs and their sequence identities were obtained from Ensembl 93 (GRCh38.p12 genome assembly) (Zerbino *et al*, 2018). These pairs were filtered down to those where both genes are annotated as 'gene with protein product' in the HGNC (Braschi *et al*, 2018), and where the sequence identity is 20% in at least 1 direction of comparison. HGNC approved gene symbols and Entrez identifiers were also assigned to all paralogs using the HGNC ID mappings.

Filtering paralog pairs

To derive our bronze standard dataset, the set of paralog pairs obtained from Ensembl (see above) was filtered according to several additional criteria (Fig. 1A). We required paralog pairs to have at least 20% sequence identity in both directions of comparison (rather than just one), and filtered out pairs belonging to paralog families with more than 20 genes - family size was calculated as the length of the union of both genes' (A1 and A2) paralogs. We further filtered out paralog pairs where both genes are not "broadly expressed". To determine this we obtained tissue distribution and specificity for mRNA expression from the Human Protein Atlas (version 19.3, file: proteinatlas.tsv) (Uhlén *et al*, 2015). We consider genes to be "broadly expressed" if their expression is detected in all tissues, or if their expression is detected in at least a third of tissues ('detected in many') and their tissue specificity is classified as low or not detected.

Gene dependency scores

We reprocessed the DepMap screen data for 769 cancer cell lines (release 20Q2, file: logfold_changes.csv) with CERES - which corrects for the known effects of copy number alterations (Meyers *et al*, 2017) - after dropping multi-targeting single-guide RNAs (Table S2). Multi-targeting guides were identified by the same methodology as in (De Kegel & Ryan, 2019) and include guides that align to multiple genes with either a perfect match, a single mismatch, or a PAM-distal double mismatch. Only scores for genes that were targeted by at least three unique guides are retained for further analysis - overall 1,681 of the 18,119 genes with CERES scores in DepMap were dropped.

Calling gene loss from expression, mutation and copy number data

We obtained RNA-seq (file: CCLE_expression.csv), gene mutation (file: CCLE_mutations.csv) and gene level copy number (file: CCLE_gene_cn.csv) data from the DepMap portal (release 20Q2) (Ghandi *et al*, 2019). Data from all three sources was available for 762 of the 769 cell lines included in the DepMap CRISPR screens, so gene loss was identified for these 762 cell lines (Table S3). A gene is marked as lost in a given cell line if at least one of the following conditions is met: gene expression (log TPM+1) < 1, gene expression z-score < -4, the gene has a LOF mutation (splice site, nonsense, or frameshift),

or the gene is homozygously deleted. We consider a gene to be homozygously deleted if its copy number ratio is < -1.28 and the gene's homozygous deletions are in general supported by a relative decrease in expression. Specifically, we test that the gene's expression is significantly lower in the cell lines where it is homozygously deleted, compared to the cell lines where it is not (t-test $p<0.05$).

Identifying putative synthetic lethal paralog pairs

To identify a significant association between A2 gene loss and A1 dependency we used a model with the form: A1_dependency ~ A2_status + C(cell_line_lineage). Cell line lineage information was obtained from the DepMap portal (release 20Q2, file: sample_info.csv). Pairs where A1 was not essential (CERES score < -0.6) in any of 762 cell lines for which we had data were directly marked as non-SL. For the remaining pairs we applied Benjamini-Hochberg multiple testing correction (Benjamini & Hochberg, 1995) on the p-value for the A2_status variable. Pairs with a negative A2_status coefficient were considered SL at a false discovery rate (FDR) of 10%.

Budding and fission yeast orthologs

S. pombe orthologs were obtained from PomBase (Lock *et al*, 2019) and their essentiality status (essential vs. non-essential) was obtained from (Kim *et al*, 2010). Orthologs without available essentiality data were marked as non-essential. *S. cerevisiae* orthologs were downloaded from Ensembl 93 (Zerbino *et al*, 2018) and their essentiality status was obtained from OGEE v2 (Chen *et al*, 2017).

Conservation score

The conservation score of a paralog pair was calculated as the number of species, out of nine, in which at least one member of the pair has an identifiable ortholog. The nine species considered are: *Arabidopsis thaliana, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Escherichia coli, Mus musculus, Rattus norvegicus, S. cerevisiae* and *Xenopus tropicalis*. Orthologs were obtained from Ensembl 93 (Zerbino *et al*, 2018) with the exception of *A. thaliana* and *E. coli* orthologs which were not available in Ensembl v93 and were instead obtained from InParanoid (Sonnhammer & Östlund, 2015).

Whole genome vs. small-scale duplicates

Paralog pairs are annotated as whole genome duplicates (WGDs) if they were included in the WGDs identified by (Makino & McLysaght, 2010) and/or the high-confidence (strict) list of WGDs in the OHNOLOGS v2 resource (Singh & Isambert, 2020). All other paralog pairs are classified as small-scale duplicates.

Gene age

Gene age for each individual gene in the paralog pairs was obtained from ProteinHistorian (Capra *et al*, 2012), with the default age estimation options (Family database = PPODv4_PTHR7-OrthoMCL, Reconstruction algorithm = wagner1.0). Although the genes in a paralog pair are theoretically expected to have the same age, discrepancies in the gene trees used to determine age and paralogy can lead to different age estimates. Consequently we used the average age of the two paralogs as the age for the pair.

Protein-protein interactions
Protein-protein interactions were obtained from BioGRID (version 3.5.187) (Oughtred *et al*, 2019). The full BioGRID data set was filtered for physical interactions between two human protein-coding genes. Genes in a paralog pair are annotated as interacting if there is any experimental evidence of an interaction between them.

Protein complex membership
Sub-units for known protein complexes were obtained from CORUM (Giurgiu *et al*, 2019) (complete complexes dataset). A paralog pair was labelled as a protein complex member if at least one of its member genes is annotated as coding for a protein complex sub-unit.

Gene expression features
Gene expression in 17,382 non-diseased tissue samples was obtained from the GTeX database (GTEx Consortium, 2020). For each gene in each paralog pair we computed the mean expression over all samples. This was used to populate two features for each pair: mean expression (lower), i.e. the mean of the lower expressed gene in the pair, and mean expression (higher), i.e. the mean of the higher expressed gene in the pair. We also computed Spearman's correlation between the expression of the two genes in each paralog pair across all samples.

Feature value imputation
For some features, such as gene expression and gene age, data was not available for all paralog genes and/or pairs. Missing feature values were imputed with the mean value across all paralog pairs, with the exception of "mean_complex_essentiality" for pairs that don't include a protein complex member, and "shared_ppi_mean_essentiality" for pairs that do not have any shared interactors - these were filled in with zero.

Metrics for individual features
For each feature we computed the ROC AUC and average precision for the feature values as well as the inverse of the feature values (i.e. we considered both a positive and negative relationship between the feature and synthetic lethality). We then retained the maximum ROC AUC and average precision values. For all features except family size and expression correlation the maximum metric values were observed for a positive relationship between the feature and synthetic lethality.

The random forest classifier
We use the random forest classifier implementation in the scikit-learn package (class RandomForestClassifier). With the aim of finding a balance between under- and over-fitting, we performed a gridsearch based on stratified cross-validation of our bronze standard dataset to select the hyper-parameters of the random forest classifier. The parameters used in the final model are as follows: n_estimators=600 (i.e. number of trees), max_features=11, max_tree_depth=3, min_child_weight=8, and random_state=8 (for reproducibility).

Paralog pairs screened by Thompson et al.
Thompson et al. (Thompson *et al*, in press) screened a total of 1,191 gene pairs, including both paralog and non-paralog pairs. Of these gene pairs, 592 are paralog pairs with at least 20% sequence identity that are included on the list we obtained from Ensembl. In their identification of SL gene pairs, Thompson et al. filtered out candidate pairs that contained a

gene that had been identified as essential in the cell line(s) they screened. We applied a similar filtering step - specifically, we filtered out pairs that included a gene that was identified as essential in all 3 cell lines screened by Thompson et al. This left us with 541 paralog pairs for further analysis.

Paralog pairs screened by Dede et al.

Dede et al. (Dede *et al*, 2020) screened 400 paralog pairs, of which 393 overlap with our list of paralog pairs (with at least 20% sequence identity) from Ensembl. We obtained screen results for all of these paralog pairs (Additional file 3: Table S2) (Dede *et al*, 2020) and considered a pair as SL in a given cell line if it had zdLFC score < -3 (this is the cutoff used by the authors).

SL predictions from SLant

We obtained SLant predictions for an extensive list of gene pairs via personal communication with the authors (Benstead-Hume *et al*, 2019) - this is an extended version of what is available in the Slorth database, which only includes scores for those pairs that were predicted to be SL.

Feature value z-scores

For numeric features we compute z-scores using feature values for all ~36.6k paralog pairs with at least 20% sequence identity in at least one direction of gene-to-gene comparison. These z-scores were then capped at +/-2. For boolean features the 'z-scores' are simply set to 2 and -2 for True and False feature values respectively, i.e. we don't take the ratio of True/False into account.

siRNA experiments

HAP1 cell lines were obtained from Horizon Discovery: HAP1_WT (C631), HAP1_COPS7A_KO (HZGHC000914c001), HAP1_COPS7B_KO (HZGHC000636c011), HAP1_ASF1A_KO (HZGHC007215c005) and HAP1_ASF1B_KO (HZGHC55723). Cells were cultured in IMDM (10-016-CVR; Corning) supplemented with 10% FBS (10270-106; Thermo Fisher Scientific). ON-TARGETplus siRNA SMARTpools targeting COPS7A (L-020873-01-0005), COPS7B (L-014207-02-0005), PLK1 (L-003290-00-0005), ASF1A (LQ-020222-02-0002), ASF1B (LQ-020553-00-0002) and a non-targeting scramble control (D-001810-10-20) were obtained from Dharmacon. HAP1 cells were seeded to a density of 5000 cells per well of a 96-well plate and 5nM siRNA was transfected with Lipofectamine 3000 (L3000015; Thermo Fisher Scientific) in Opti-MEM I Reduced Serum Medium (31985070; Thermo Fisher Scientific). Cell viability was measured 72 hours after siRNA transfection using CellTiter-Glo Luminescent Cell Viability Assay (G7570; Promega). The 96 well plates were read using a SpectraMax® M3 Microplate Reader (Molecular devices). Viability effects for each siRNA targeting each gene X in each cell line y were normalised using the following formula:

$$\text{NormalisedViability}\ (\text{siRNA_X}_y) = 1 - (\text{siCTRL}_y - \text{siRNA_X}_y) / (\text{siCTRL}_y - \text{siPLK1}_y)$$

where siCTRL_y is the average of 3 measurements of non-targeting scramble control in cell line y and siPLK1_y is the average of 3 measurements of an siRNA smartpool targeting PLK1 in cell line y. Raw and normalised viability data are in Table S9.

Acknowledgements

B.D.K., N.Q. and C.J.R. were funded through an Irish Research Council 2017/2018 Laureate Award awarded to C.J.R.. D.J.A. and N.A. were funded from the Wellcome Trust, Cancer Research UK and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC synergy grant agreement n° 319661 COMBATCANCER. We thank Graeme Benstead-Hume and Frances Pearl for sharing the complete list of predictions generated by SLant and Traver Hart for sharing data ahead of publication. We thank Barry Scott for the introduction to SHAP values.

Conflict of Interests

The authors declare that they have no conflict of interest.

References

Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, Santos R, Rao Y, Sassi F, Pinnelli M, *et al* (2019) Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* 568: 511–516

Benedetti L, Cereda M, Monteverde L, Desai N & Ciccarelli FD (2017) Synthetic lethal interaction between the tumour suppressor STAG2 and its paralog STAG1. *Oncotarget* 8: 37619

Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc*

Benstead-Hume G, Chen X, Hopkins SR, Lane KA, Downs JA & Pearl FMG (2019) Predicting synthetic lethal interactions using conserved patterns in protein interaction networks. *PLoS Comput Biol* 15: e1006888

Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A, *et al* (2015) Gene essentiality and synthetic lethality in haploid human cells. *Science* 350: 1092–1096

Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, Yates B & Bruford E (2018) Genenames. org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res* 47: D786–D792

Breiman L (2001) Random Forests. *Mach Learn* 45: 5–32

Capra JA, Williams AG & Pollard KS (2012) ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput Biol* 8: e1002567

Chen W-H, Lu G, Chen X, Zhao X-M & Bork P (2017) OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res* 45: D940–D944

Chen W-H, Trachana K, Lercher MJ & Bork P (2012) Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age. *Mol Biol Evol* 29: 1703–1706

Conant GC & Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9: 938–950

Dandage R & Landry CR (2019) Paralog dependency indirectly affects the robustness of human cells. *Mol Syst Biol* 15: e8871

Dean EJ, Davis JC, Davis RW & Petrov DA (2008) Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet* 4: e1000113

Dede M, McLaughlin M, Kim E & Hart T (2020) Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens. *Genome Biol* 21: 262

Dehal P & Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3: e314

De Kegel B & Ryan CJ (2019) Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. *PLoS Genet* 15: e1008466

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S & Kishony R (2008) Exposing the fitness contribution of duplicated genes. *Nat Genet* 40: 676–681

Dey P, Baddour J, Muller F, Wu CC, Wang H, Liao W-T, Lan Z, Chen A, Gutschner T, Kang Y, *et al* (2017) Genomic deletion of malic enzyme 2 confers collateral lethality in pancreatic cancer. *Nature* 542: 119–123

Ehrenhöfer-Wölfer K, Puchner T, Schwarz C, Rippka J, Blaha-Ostermann S, Strobl U, Hörmann A, Bader G, Kornigg S, Zahn S, *et al* (2019) SMARCA2-deficiency confers sensitivity to

targeted inhibition of SMARCA4 in esophageal squamous cell carcinoma cell lines. *Sci Rep* 9: 11661

Fortin J-P, Tan J, Gascoigne KE, Haverty PM, Forrest WF, Costa MR & Martin SE (2019) Multiple-gene targeting and mismatch tolerance can confound analysis of genome-wide pooled CRISPR screens. *Genome Biol* 20: 21

Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, *et al* (2019) Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, *et al* (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418: 387–391

Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C & Ruepp A (2019) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res* 47: D559–D563

Gonatopoulos-Pournatzis T, Aregger M, Brown KR, Farhangmehr S, Braunschweig U, Ward HN, Ha KCH, Weiss A, Billmann M, Durbic T, *et al* (2020) Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9–Cas12a platform. *Nat Biotechnol*: 1–11

GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369: 1318–1330

Guan Y, Dunham MJ & Troyanskaya OG (2007) Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics* 175: 933–943

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW & Li W-H (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421: 63–66

Haber JE, Braberg H, Wu Q, Alexander R, Haase J, Ryan C, Lipkin-Moore Z, Franks-Skiba KE, Johnson T, Shales M, *et al* (2013) Systematic triple-mutant analysis uncovers functional connectivity between pathways involved in chromosome regulation. *Cell Rep* 3: 2168–2178

Hahn MW & Kern AD (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* 22: 803–806

Han K, Jeng EE, Hess GT, Morgens DW, Li A & Bassik MC (2017) Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat Biotechnol* 35: 463–474

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, *et al* (2020) Array programming with NumPy. *Nature* 585: 357–362

Hartwell LH, Szankasi P, Roberts CJ, Murray AW & Friend SH (1997) Integrating genetic approaches into the discovery of anticancer drugs. *Science* 278: 1064–1068

Helming KC, Wang X, Wilson BG, Vazquez F, Haswell JR, Manchester HE, Kim Y, Kryukov GV, Ghandi M, Aguirre AJ, *et al* (2014) ARID1B is a specific vulnerability in ARID1A-mutant cancers. *Nat Med* 20: 251–254

Henkel L, Rauscher B & Boutros M (2019) Context-dependent genetic interactions in cancer. *Curr Opin Genet Dev* 54: 73–82

He X & Zhang J (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet* 2: e88

Hoffman GR, Rahal R, Buxton F, Xiang K, McAllister G, Frias E, Bagdasarian L, Huber J, Lindeman A, Chen D, *et al* (2014) Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers. *Proc Natl Acad Sci U S A* 111: 3128–3133

Huang A, Garraway LA, Ashworth A & Weber B (2020) Synthetic lethality as an engine for cancer drug target discovery. *Nat Rev Drug Discov* 19: 23–38

Kaelin WG Jr (2005) The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer* 5: 689–698

Kim D-U, Hayles J, Kim D, Wood V, Park H-O, Won M, Yoo H-S, Duhig T, Nam M, Palmer G, *et al* (2010) Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotechnol* 28: 617–623

Ku AA, Hu H-M, Zhao X, Shah KN, Kongara S, Wu D, McCormick F, Balmain A & Bandyopadhyay S (2020) Integration of multiple biological contexts reveals principles of synthetic lethality that affect reproducibility. *Nat Commun* 11: 2375

Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, *et al* (2020) Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. *Science* 368

Lee JS, Das A, Jerby-Arnon L, Arafeh R, Auslander N, Davidson M, McGarry L, James D, Amzallag A, Park SG, *et al* (2018) Harnessing synthetic lethality to predict the response to cancer treatment. *Nat Commun* 9: 2546

Lee KY, Im J-S, Shibata E & Dutta A (2017) ASF1a Promotes Non-homologous End Joining Repair by Facilitating Phosphorylation of MDC1 by ATM at Double-Strand Breaks. *Mol Cell* 68: 61–75.e5

van der Lelij P, Newman JA, Lieb S, Jude J, Katis V, Hoffmann T, Hinterndorfer M, Bader G, Kraut N, Pearson MA, *et al* (2020) STAG1 vulnerabilities for exploiting cohesin synthetic lethality in STAG2-deficient cancers. *Life Sci Alliance* 3

Lever J, Krzywinski M & Altman N (2016) Classification evaluation. *Nat Methods* 13: 603–604

Lock A, Rutherford K, Harris MA, Hayles J, Oliver SG, Bähler J & Wood V (2019) PomBase 2018: user-driven reimplementation of the fission yeast database provides rapid and intuitive access to diverse, interconnected information. *Nucleic Acids Res* 47: D821–D827

Lord CJ & Ashworth A (2017) PARP inhibitors: Synthetic lethality in the clinic. *Science* 355: 1152–1158

Lord CJ, Quinn N & Ryan CJ (2020) Integrative analysis of large-scale loss-of-function screens identifies robust cancer-associated genetic interactions. *Elife* 9

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N & Lee S-I (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2: 56–67

Makino T & McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A* 107: 9270–9274

McKinney W (2011) pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 14

Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, *et al* (2017) Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat Genet* 49: 1779

Muller FL, Colla S, Aquilanti E, Manzo VE, Genovese G, Lee J, Eisenson D, Narurkar R, Deng P, Nezi L, *et al* (2012) Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* 488: 337–342

Najm FJ, Strand C, Donovan KF, Hegde M, Sanson KR, Vaimberg EW, Sullender ME, Hartenian E, Kalani Z, Fusi N, *et al* (2018) Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. *Nat Biotechnol* 36: 179–189

Nusinow DP, Szpyt J, Ghandi M, Rose CM, McDonald ER 3rd, Kalocsay M, Jané-Valbuena J, Gelfand E, Schweppe DK, Jedrychowski M, *et al* (2020) Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* 180: 387–402.e16

Ogiwara H, Sasaki M, Mitachi T, Oike T, Higuchi S, Tominaga Y & Kohno T (2016) Targeting p300 Addiction in CBP-Deficient Cancers Causes Synthetic Lethality by Apoptotic Cell Death due to Abrogation of MYC Expression. *Cancer Discovery* 6: 430–445 doi:10.1158/2159-8290.cd-15-0754 [PREPRINT]

Oike T, Ogiwara H, Tominaga Y, Ito K, Ando O, Tsuta K, Mizukami T, Shimada Y, Isomura H, Komachi M, *et al* (2013) A synthetic lethality-based strategy to treat cancers harboring a genetic deficiency in the chromatin remodeling factor BRG1. *Cancer Res* 73: 5508–5518

O'Neil NJ, Bailey ML & Hieter P (2017) Synthetic lethality and cancer. *Nat Rev Genet* 18: 613–623

Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, *et al* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 47: D529–D541

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12: 2825–2830

Ryan CJ, Bajrami I & Lord CJ (2018) Synthetic Lethality and Cancer--Penetrance as the Major Barrier. *Trends Cancer Res*

Schick S, Rendeiro AF, Runggatscher K, Ringler A, Boidol B, Hinkel M, Májek P, Vulliard L, Penz T, Parapatics K, *et al* (2019) Systematic characterization of BAF mutations provides insights into intracomplex synthetic lethalities in human cancers. *Nat Genet* 51: 1399–1410

Seabold S & Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference* p 61. Scipy

Seeger M, Kraft R, Ferrell K, Bech-Otschir D, Dumdey R, Schade R, Gordon C, Naumann M & Dubiel W (1998) A novel protein complex involved in signal transduction possessing similarities to 26S proteasome subunits. *FASEB J* 12: 469–478

Singh PP & Isambert H (2020) OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. *Nucleic Acids Res* 48: D724–D730

Sipiczki M (2000) Where does fission yeast sit on the tree of life? *Genome Biol* 1: REVIEWS1011

Sonnhammer ELL & Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43: D234–9

Szymańska E, Nowak P, Kolmus K, Cybulska M, Goryca K, Derezińska-Wołek E, Szumera-Ciećkiewicz A, Brewińska-Olchowik M, Grochowska A, Piwocka K, *et al* (2020) Synthetic lethality between VPS4A and VPS4B triggers an inflammatory response in colorectal cancer. *EMBO Mol Med* 12: e10812

Tagami H, Ray-Gallet D, Almouzni G & Nakatani Y (2004) Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell* 116: 51–61

Thompson NA, Ranzani M, van der Weyden L, Iyer V, Offord V, Droop A, Behan F, Gonçalves E, Speak A, Iorio F, Hewinson J, Harle V, Robertson H, Anderson E, Fu B, Wong K, Yang F, Zagnoli-Vieira G, Chapman P, Del Castillo Velasco-Herrera M, Garnett M, Jackson SP, Adams DJ (in press) Combinatorial-CRISPR screening defines FAM50A/FAM50B as a targetable synthetic lethal gene pair. *Nat Commun*

Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, *et al* (2017) Defining a Cancer Dependency Map. *Cell* 170: 564–576.e16

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, *et al* (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347: 1260419

VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C & Myers CL (2010) Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol* 6: 429

Vavouri T, Semple JI, Garcia-Verdugo R & Lehner B (2009) Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell* 138: 198–208

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al* (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 17: 261–272

Viswanathan SR, Nogueira MF, Buss CG, Krill-Burger JM, Wawer MJ, Malolepsza E, Berger AC, Choi PS, Shih J, Taylor AM, *et al* (2018) Genome-scale analysis identifies paralog lethality as a vulnerability of chromosome 1p loss in cancer. *Nat Genet* 50: 937–943

Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ & Koller D (2009) A complex-based reconstruction of the Saccharomyces cerevisiae interactome. *Mol Cell Proteomics* 8: 1361–1381

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES & Sabatini DM (2015) Identification and characterization of essential genes in the human genome. *Science* 350: 1096–1101

Wood V (2006) Schizosaccharomyces pombe comparative genomics; from sequence to systems. In *Comparative Genomics: Using Fungi as Models*, Sunnerhagen P & Piskur J (eds) pp 233–285. Berlin, Heidelberg: Springer Berlin Heidelberg

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, *et al* (2018) Ensembl 2018. *Nucleic Acids Res* 46: D754–D761

Zhang J (2003) Evolution by gene duplication: an update. *Trends Ecol Evol* 18: 292–298

Zotenko E, Mestre J, O'Leary DP & Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput Biol* 4: e1000140

Figure Legends

Figure 1. Identifying a bronze standard set of SL paralog pairs
(A) Workflow of our computational approach to identify a bronze standard set of synthetic lethal paralog pairs from the full list of Ensembl protein-coding paralog pairs. The two genes that make up each paralog pair are referred to as A1 and A2. (B) Boxplots showing A1 dependency (CERES score) in cell lines where A2 is lost (blue box) vs. where A2 is wild-type (white box), for six of the putative SL paralog pairs that we identified using the approach in (A). For each box plot, the black central line represents the median, the top and bottom represent the 1st and 3rd quartile, and the whiskers extend 1.5 times the interquartile range past the box. Each dot represents one cell line in which A1 (the gene on the y-axis) was screened.

Figure 2. Assessing the predictive power of 22 features of paralog pairs and a random forest classifier
(A) Bar chart showing the area under the ROC curve (ROC AUC) and point plot showing the average precision for each feature as an individual classifier over the bronze standard dataset. The error bars indicate the mean standard error based on computing the ROC AUC and average precision for 10 randomly selected, stratified fourths of the dataset. Bars and dots are colored according to the feature category. (B) ROC and PR curves for the random forest classifier (RF, black), based on stratified 4-fold cross-validation of the bronze standard data, and for sequence identity alone (orange). The values shown in the legend are the ROC AUC and average precision, along with standard deviation where applicable.

Figure 3. Assessing the performance of our random forest classifier on data from independent genetic interaction screens
(A) ROC and PR curves showing the performance of our predictions (black) and sequence identity (orange) in distinguishing paralog pairs that were found to be SL in 2-3 vs.0 cell lines in genetic interaction screens carried out by Thompson et al. (B) Same as (A) but for paralog pairs screened by (Dede *et al*, 2020). (C) Left: Venn diagram showing the number of paralog pairs that were identified as SL in the indicated cell lines by Thompson et al. Right: Boxplots showing the distribution of prediction scores for pairs essential in zero, one, two or three cell lines. Individual predictions are shown as grey dots. Hues match to the regions of the venn diagram on the left.

Figure 4. Predictions for all paralog pairs and some of their local explanations
(A) Waterfall plot showing the prediction score vs. the prediction rank for all paralog pairs. The predictions highlighted in medium blue correspond to previously validated SL pairs, while those in dark blue correspond to paralog pairs that feature in (B), (C) and (D). (B) Annotated SHAP profile for the pair *AKT1/AKT2.* SHAP values are shown on the x-axis and can be either positive (feature increases prediction score) or negative (feature decreases prediction score). The 10 features with the largest global impact are shown on the y-axis - this feature subset was determined by calculating the median absolute SHAP value associated with each feature across all ~36.6k paralog pairs. The hue indicates the feature z-score, i.e. whether the feature has a relatively high (dark orange) or relatively low (dark blue) value for the given pair. (C) SHAP profiles for the pairs *VPS4A/VPS4B* and *ARID1A/ARID1B*. (D) SHAP profiles for the pairs *STAG1/STAG2* and *STAG2/STAG3*.

Figure 5. Prediction and validation of two new paralog synthetic lethalities
(A) Waterfall plot showing the prediction score vs. rank for all paralog pairs (as in Fig. 4A), with the two pairs featured in this figure highlighted in dark blue. (B) SHAP profile for the paralog pair *ASF1A/ASF1B*. (C) Mean viability of HAP1 cells treated with siRNA smartpools targeting *ASF1A* or *ASF1B*. Individual data points are shown as black dots. Data are normalized within each cell line such that the mean viability of cells treated with a negative control (non-targeting scrambled siRNA) is equal to one and the mean viability treated with a positive control (siRNA smartpool targeting the broadly essential PLK1 gene) is equal to 0. The p-values come from two sided heteroscedastic t-tests. (D) SHAP profile for *COPS7A/COPS7B*. (E) Bar chart showing the mean viability of HAP1 cells treated with siRNA smartpools targeting *COPS7A* or *COPS7B*; normalization and statistics as in (B).

Supplemental Figure Legends

Figure S1. Bar chart showing the area under the ROC curve (ROC AUC) and point plot showing the average precision for six budding and fission yeast features treated as individual classifiers. The error bars show the mean standard error based on computing the ROC AUC and average precision for 10 randomly selected, stratified fourths of our bronze standard dataset.

Figure S2. (A) ROC and PR curves showing the performance of our predictions (black) and sequence identity (orange) in distinguishing paralog pairs that were found to be SL in 1+ vs. 0 cell lines in genetic interaction screens carried out by Thompson et al. Numbers shown in the legends are ROC AUC and average precision respectively. (B) Same as (A) but for the paralog pairs screened by Dede et al. (C) Same as Fig. 3A but for just the subset of paralog pairs screened by Thompson et al. that were not included in our training (bronze standard) data. (D). Same as Fig. 3B but for just the subset of paralog pairs screened by Dede et al. that were not included in our training data.

Figure S3. (A) ROC and PR curves showing the performance of SLant (purple), our random forest classifier (black) and sequence identity (orange) in distinguishing paralog pairs that were identified as SL in 2+ vs. 0 cell lines in genetic interaction screens carried out by Thompson et al. Numbers shown in the legends are ROC AUC and average precision respectively. (B) Same as (A) but for the paralog pairs screened by Dede et al. (C) Same as (A) but for distinguishing paralog pairs identified as SL in 1+ vs. 0 cell lines by Thompson et al. (D) Same as (B) but for distinguishing paralog pairs identified as SL in 1+ vs. 0 cell lines by Dede et al.

Figure S4. Similar to Fig. 3C but for the genetic interaction screens carried out by Dede et al. Left: Venn diagram showing the number of paralog pairs that were identified as SL in the indicated cell lines by Dede et al. Right: Strip plot showing individual prediction scores for pairs essential in 0, 1, 2 or 3 cell lines. The black lines indicate the median prediction scores for each category. Hues match to the regions of the venn diagram on the left.

Figure S5. Top left: Waterfall plot showing the prediction score vs. rank for all paralog pairs (as in Fig. 4A) with paralog pairs that include the gene *AKT1* highlighted in dark blue. Top right/bottom: SHAP profiles for all paralog pairs that include *AKT1*.

Supplemental Files

Table S1. Paralog synthetic lethalities with low throughput experimental validation curated from the literature.

Table S2. Gene dependency scores computed by CERES across 769 cell lines, after filtering out potential multi-targeting guides. Columns are gene Entrez IDs, rows are DepMap cell line IDs.

Table S3. Binary matrix that indicates, for each candidate gene (A2 in the candidate paralog pairs) in each cell line, whether the gene is lost (1) or not (0) in that cell line. Rows are Entrez gene IDs (for A2) and columns are DepMap cell lines IDs.

Table S4. Results of testing 3,810 candidate paralog pairs for an association between A1 dependency and A2 loss. Pairs are listed twice if their genes were considered as the A1 and A2 gene in turn. The A2_status_coef, A2_status_p and A2_status_p_adj columns provide, respectively, the coefficient of the A2_status variable, the raw p-value and the FDR corrected p-value for the association between A1 dependency and A2 loss for each pair.

Table S5. Bronze standard set of putative SL and non-SL paralog pairs annotated with 22 paralog pair features. These pairs are a subset of those in Table S4.

Table S6. Descriptions of all 22 paralog pair features, plus the ROC AUC and average precision obtained from treating each feature as an individual classifier for our bronze standard dataset (Table S5).

Table S7. Feature values and predictions (score and rank) for all ~36.6k paralog pairs.

Table S8. SHAP values associated with all features for all predictions from Table S7.

Table S9. *ASF1A/ASF1B* and *COPS7A/COPS7B* viability data.

Fig. 1

A

List of ~36.6k protein coding paralog pairs (A1, A2) from Ensembl

↓

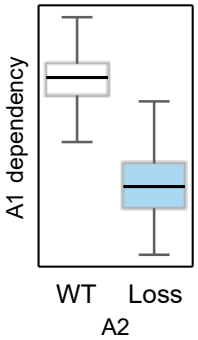
- ✓ Min. 20% sequence identity
- ✓ Paralog family size is max. 20
- ✓ Both genes are broadly expressed
- ✓ A1 has dependency scores
- ✓ A2 is lost in 10+ cell lines

↓

3,810 candidate paralog pairs

↓

Identify associations between A2 gene loss and A1 dependency using a linear model with cell line lineage as a covariate



↓

126 putative SL (~3.5%) +
3,508 putative non-SL paralog pairs

B

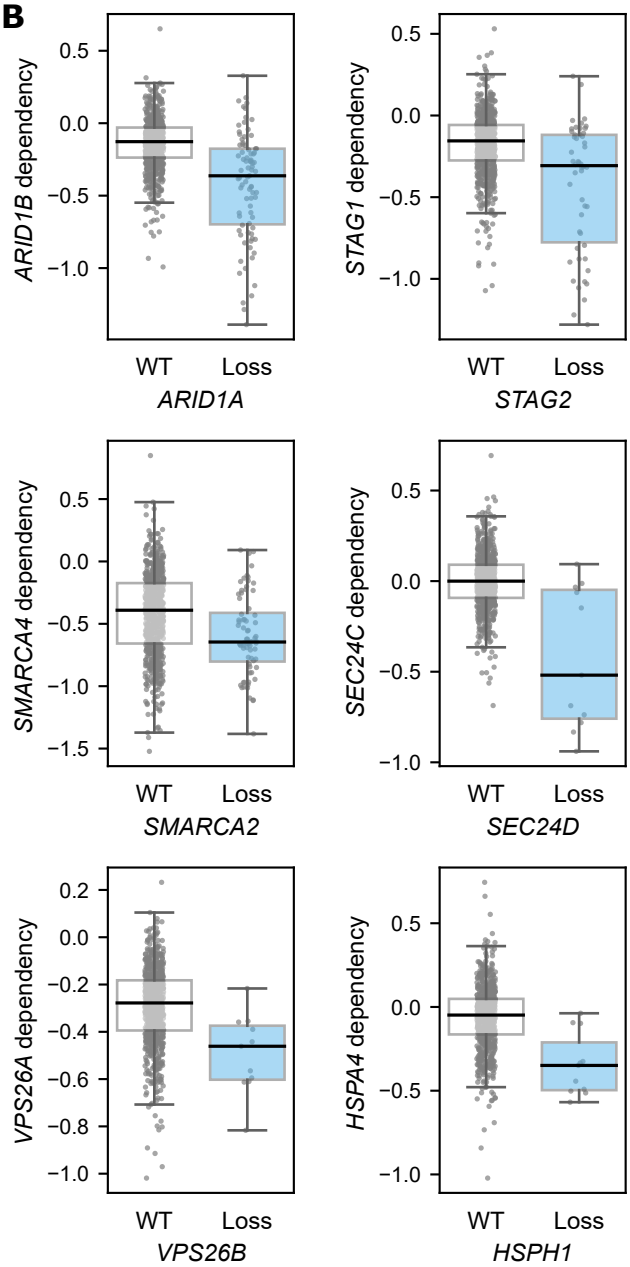
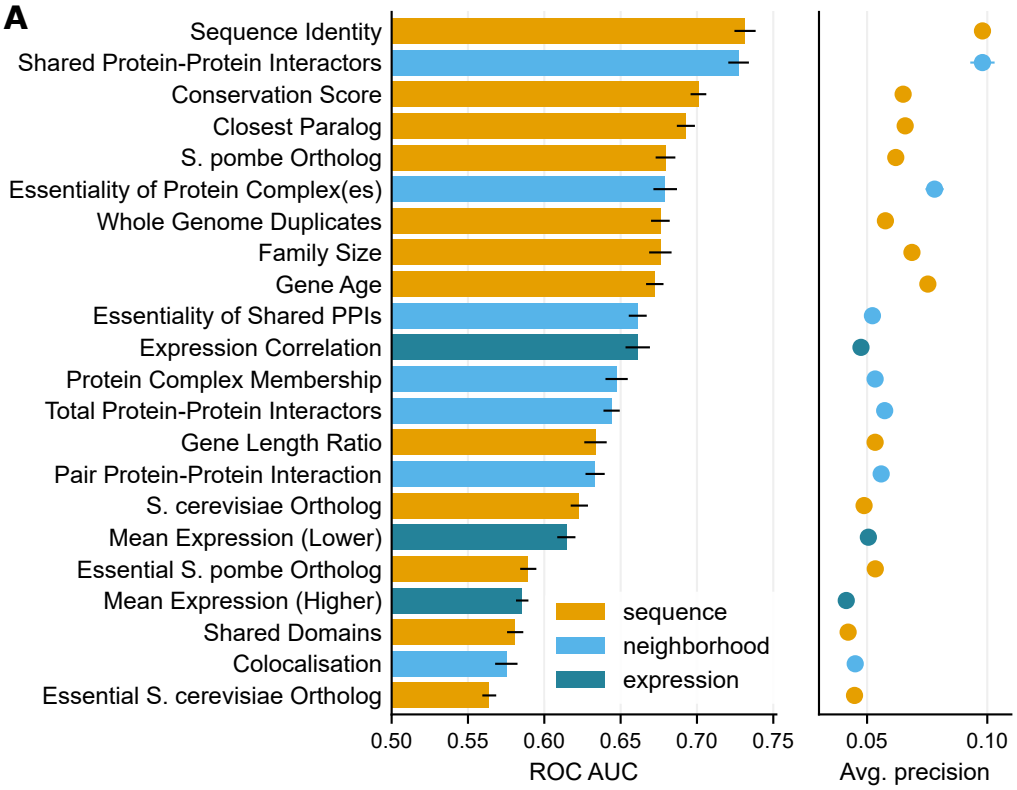


Fig. 2

A


Sequence Identity
Shared Protein-Protein Interactors
Conservation Score
Closest Paralog
S. pombe Ortholog
Essentiality of Protein Complex(es)
Whole Genome Duplicates
Family Size
Gene Age
Essentiality of Shared PPIs
Expression Correlation
Protein Complex Membership
Total Protein-Protein Interactors
Gene Length Ratio
Pair Protein-Protein Interaction
S. cerevisiae Ortholog
Mean Expression (Lower)
Essential S. pombe Ortholog
Mean Expression (Higher)
Shared Domains
Colocalisation
Essential S. cerevisiae Ortholog
sequence
neighborhood
expression
0.50
0.55
0.60
0.65
0.70
0.75
ROC AUC
0.05
0.10
Avg. precision

B

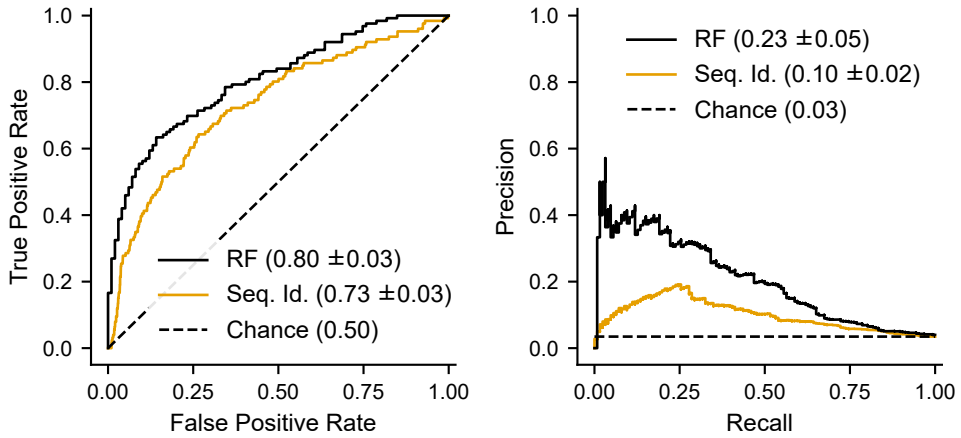
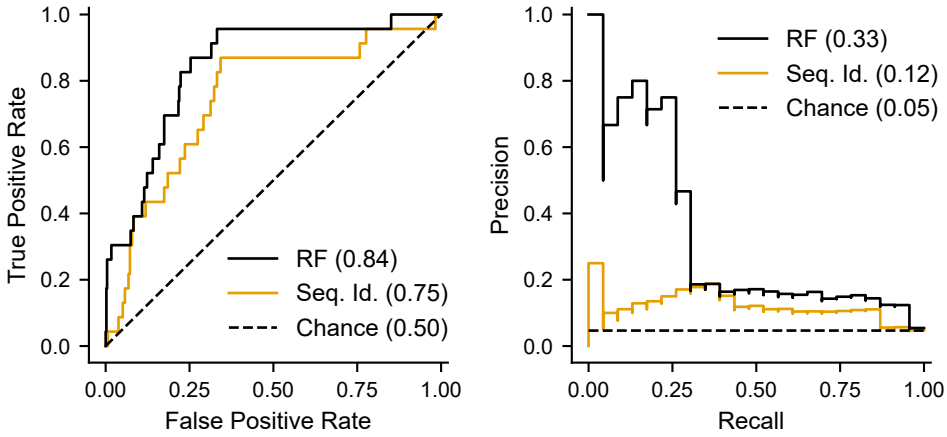
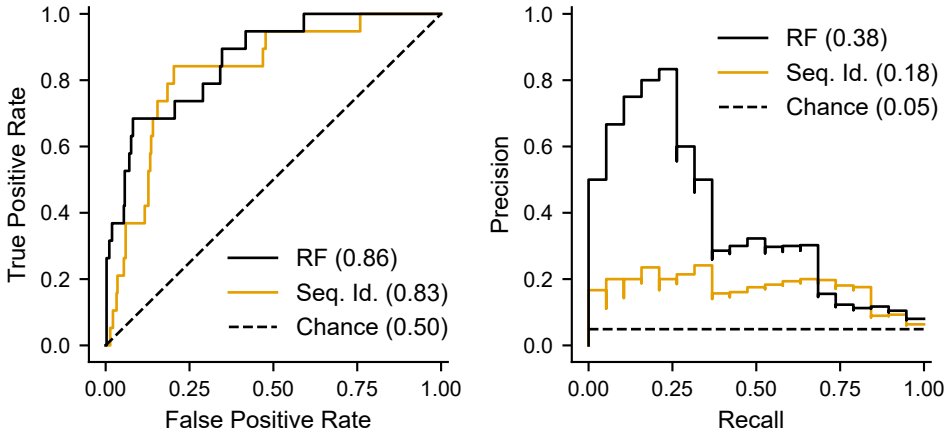

True Positive Rate
False Positive Rate
RF (0.80 ±0.03)
Seq. Id. (0.73 ±0.03)
Chance (0.50)
Precision
Recall
RF (0.23 ±0.05)
Seq. Id. (0.10 ±0.02)
Chance (0.03)

Fig. 3

A **Thompson et al. screens, SL in 2+ vs. 0 cell lines (SL=23, non-SL=470)**



B **Dede et al. screens, SL in 2+ vs. 0 cell lines (SL=19, non-SL=369)**



C

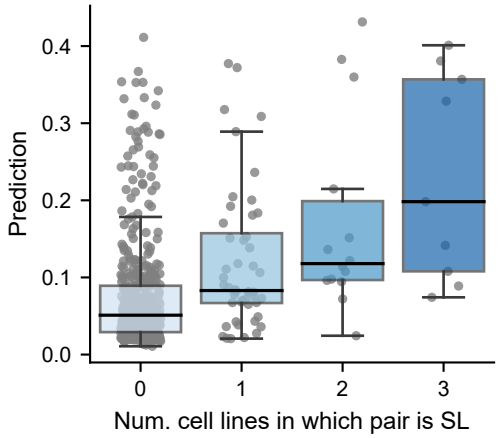


Fig. 4

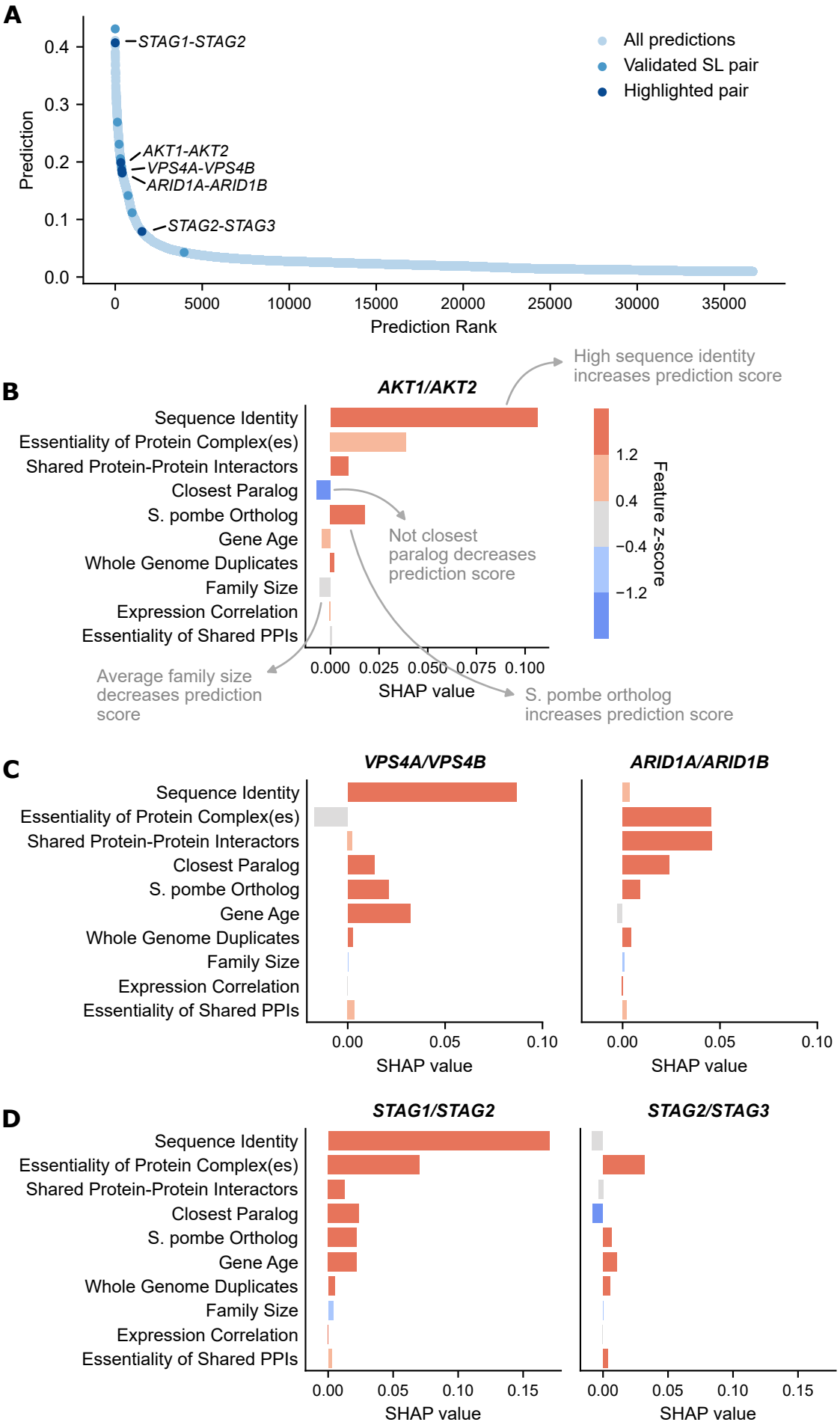

A
STAG1-STAG2
AKT1-AKT2
VPS4A-VPS4B
ARID1A-ARID1B
STAG2-STAG3
All predictions
Validated SL pair
Highlighted pair
Prediction
Prediction Rank
B
AKT1/AKT2
High sequence identity increases prediction score
Sequence Identity
Essentiality of Protein Complex(es)
Shared Protein-Protein Interactors
Closest Paralog
S. pombe Ortholog
Gene Age
Whole Genome Duplicates
Family Size
Expression Correlation
Essentiality of Shared PPIs
Not closest paralog decreases prediction score
Average family size decreases prediction score
S. pombe ortholog increases prediction score
SHAP value
Feature z-score
C
VPS4A/VPS4B
ARID1A/ARID1B
SHAP value
D
STAG1/STAG2
STAG2/STAG3
SHAP value

Fig. 5

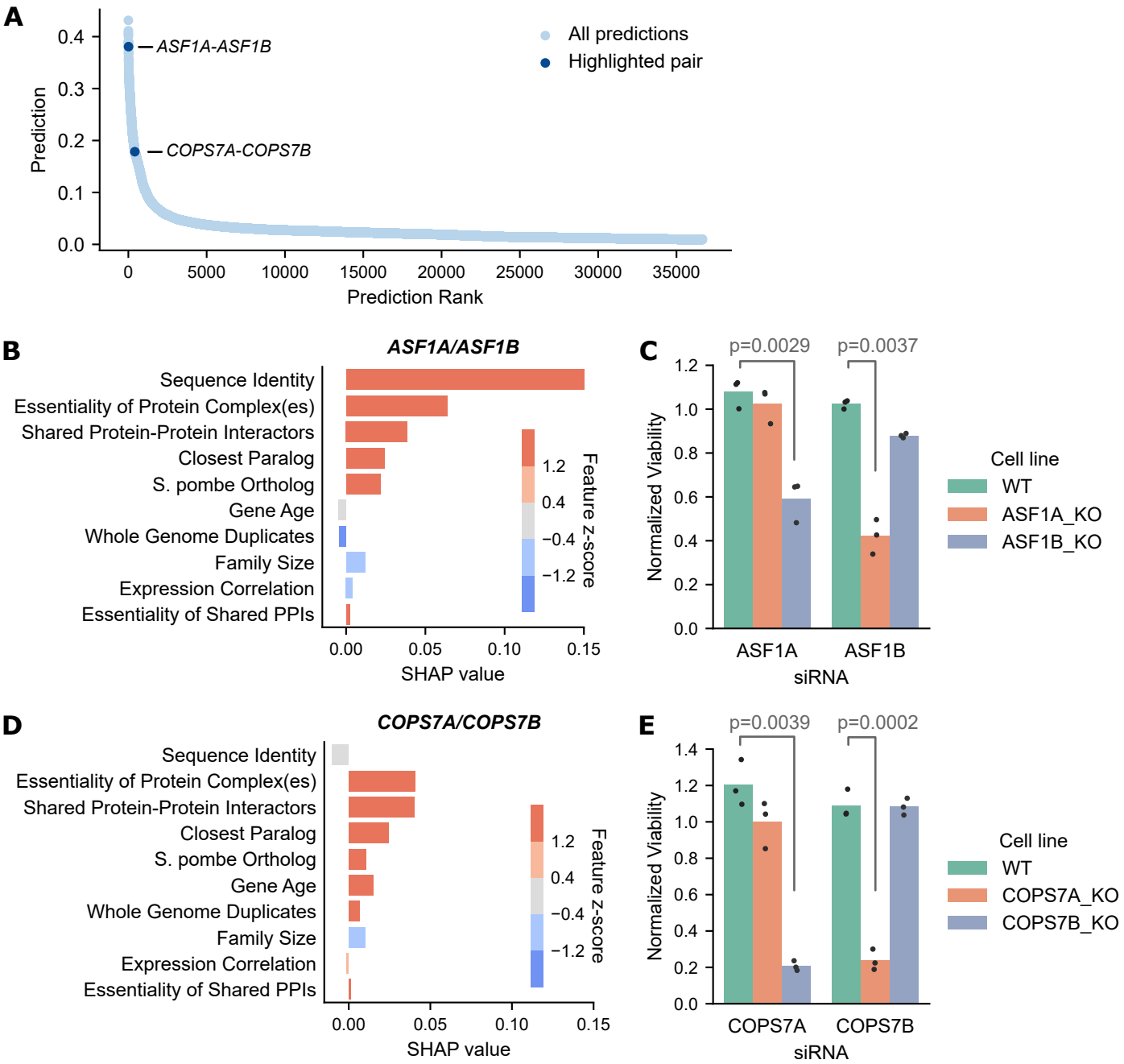

A
Prediction
Prediction Rank
ASF1A-ASF1B
COPS7A-COPS7B
All predictions
Highlighted pair
B
ASF1A/ASF1B
Sequence Identity
Essentiality of Protein Complex(es)
Shared Protein-Protein Interactors
Closest Paralog
S. pombe Ortholog
Gene Age
Whole Genome Duplicates
Family Size
Expression Correlation
Essentiality of Shared PPIs
SHAP value
Feature z-score
C
p=0.0029
p=0.0037
Normalized Viability
ASF1A
ASF1B
siRNA
Cell line
WT
ASF1A_KO
ASF1B_KO
D
COPS7A/COPS7B
E
p=0.0039
p=0.0002
COPS7A
COPS7B
COPS7A_KO
COPS7B_KO

Fig. S1

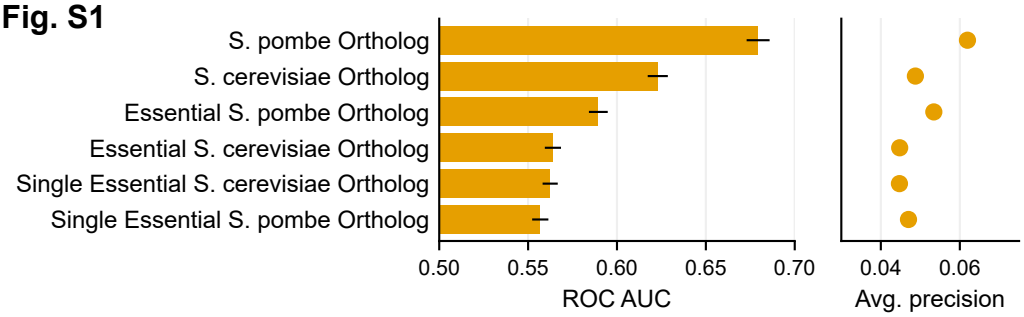
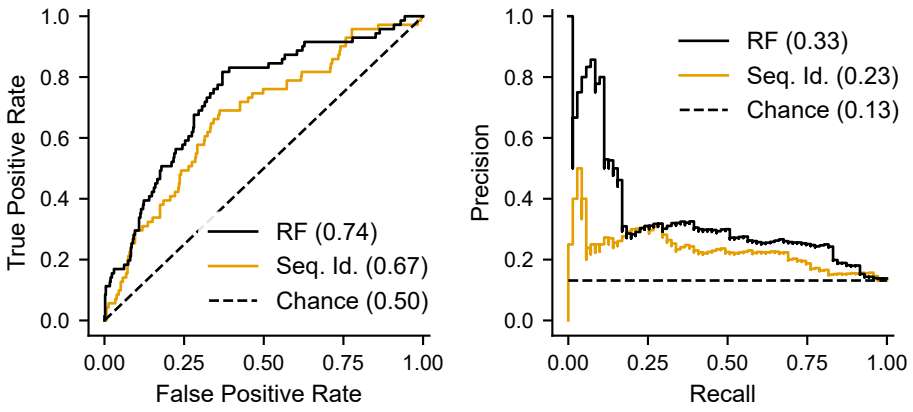
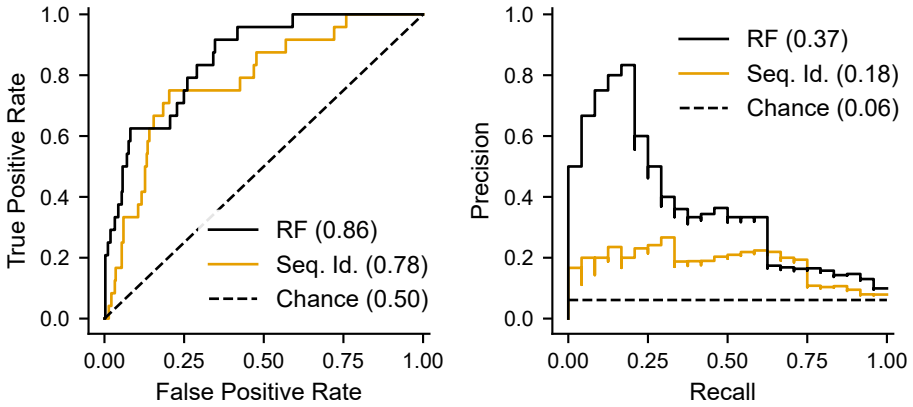

S. pombe Ortholog
S. cerevisiae Ortholog
Essential S. pombe Ortholog
Essential S. cerevisiae Ortholog
Single Essential S. cerevisiae Ortholog
Single Essential S. pombe Ortholog
0.50
0.55
0.60
0.65
0.70
ROC AUC
0.04
0.06
Avg. precision

Fig. S2

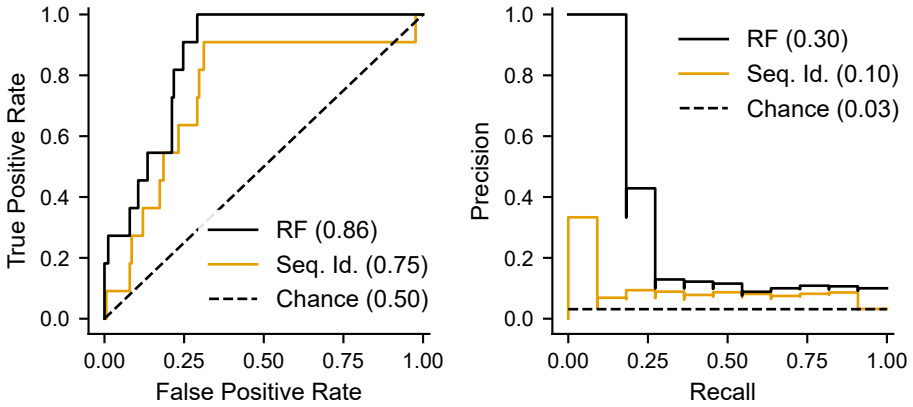
A **Thompson et al. screens, SL in 1+ vs. 0 cell lines (SL=71, non-SL=470)**



B **Dede et al. screens, SL in 1+ vs. 0 cell lines (SL=24, non-SL=369)**



C **Thompson et al. screens, SL in 2+ vs. 0 cell lines (SL=11, non-SL=340)**



D **Dede et al. screens, SL in 2+ vs. 0 cell lines (SL=14, non-SL=219)**

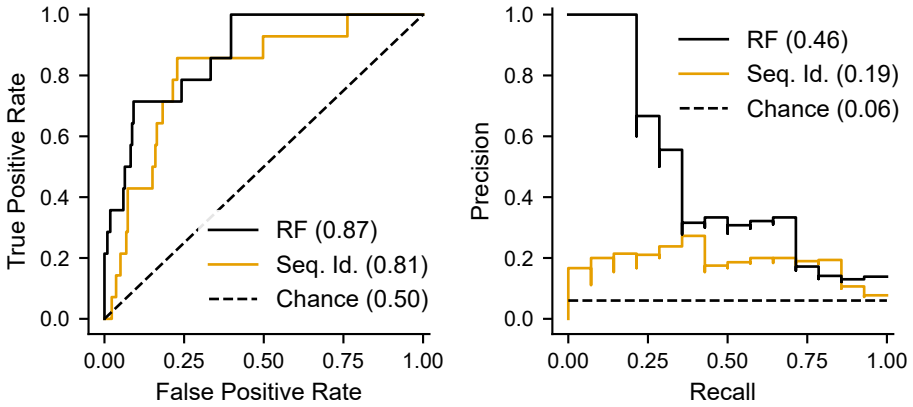
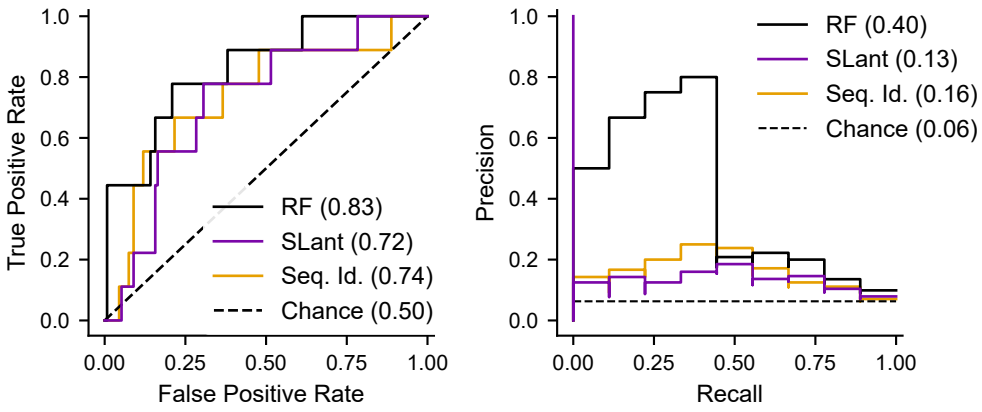
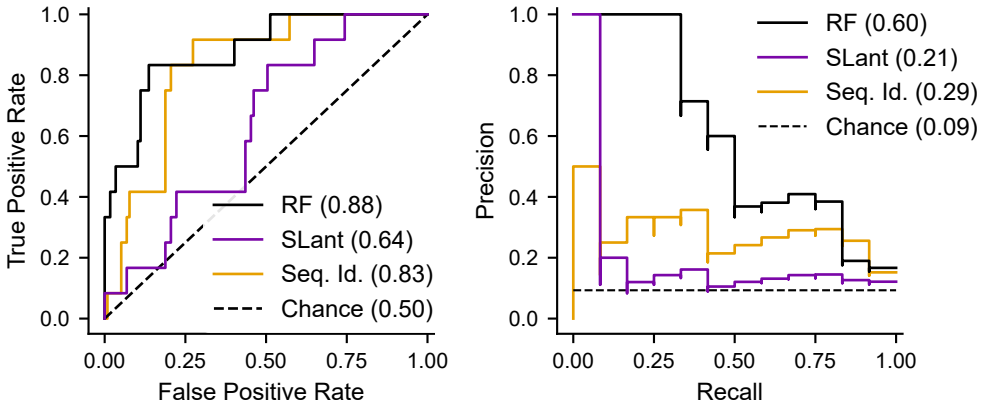


Fig. S3

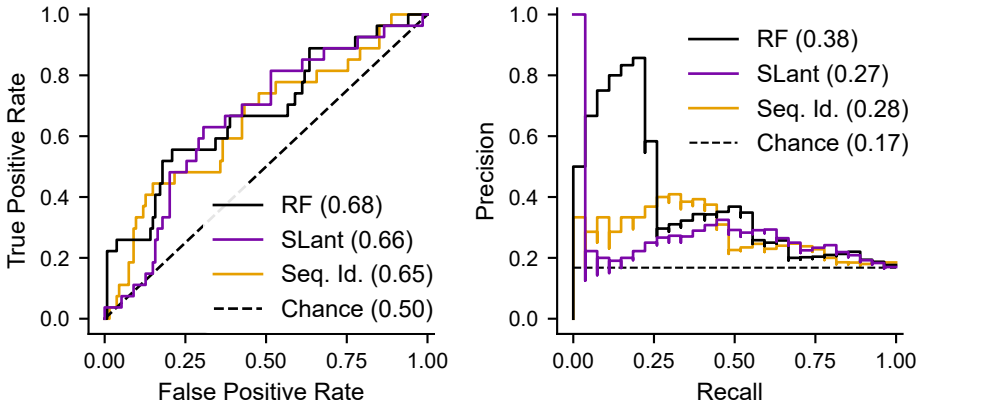
A **Thompson et al. screens, SL in 2+ vs. 0 cell lines (SL=9, non-SL=134)**



B **Dede et al. screens, SL in 2+ vs. 0 cell lines (SL=12, non-SL=117)**



C **Thompson et al. screens, SL in 1+ vs. 0 cell lines (SL=27, non-SL=134)**



D **Dede et al. screens, SL in 1+ vs. 0 cell lines (SL=14, non-SL=117)**

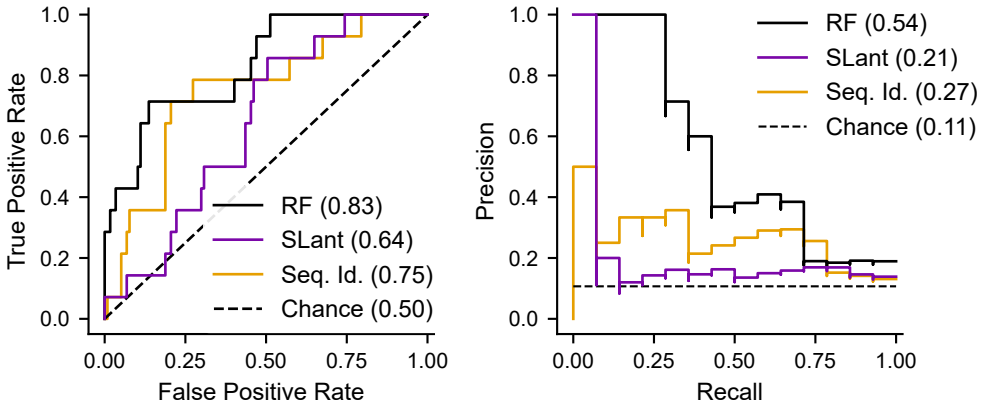
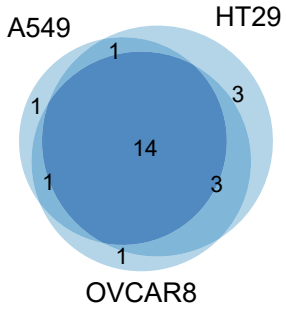


Fig. S4


A549
HT29
OVCAR8
1
1
3
14
1
3
1

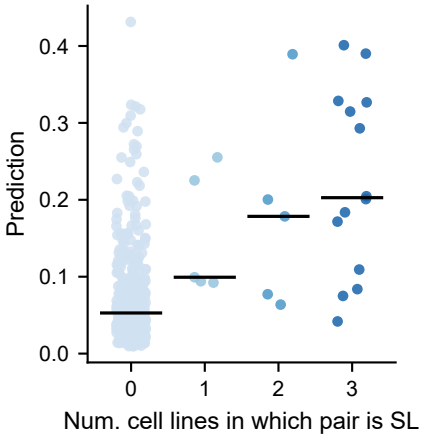
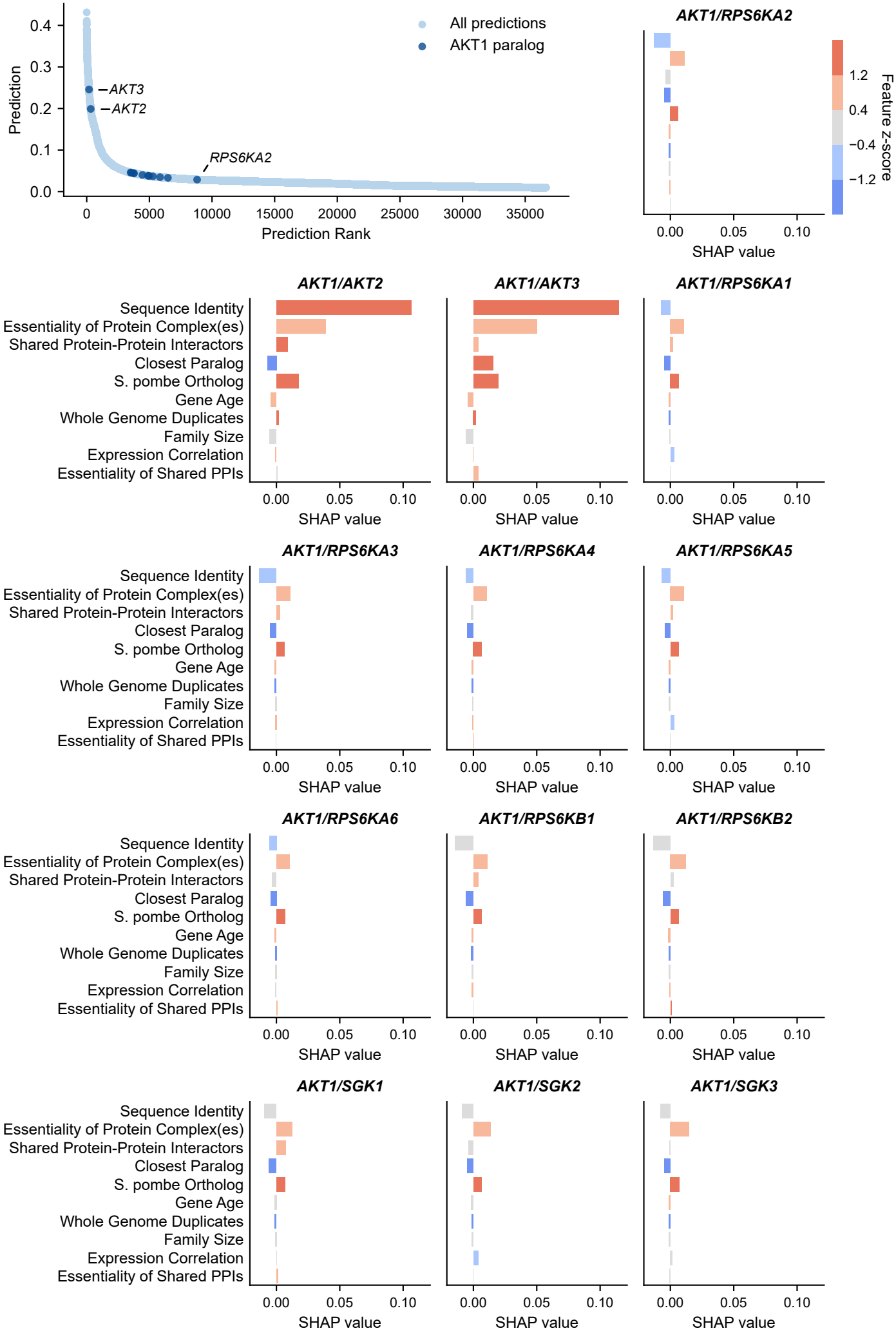

Prediction
0.0
0.1
0.2
0.3
0.4
0
1
2
3
Num. cell lines in which pair is SL

Fig. S5


All predictions
AKT1 paralog
Prediction
Prediction Rank
AKT3
AKT2
RPS6KA2
0.0
0.1
0.2
0.3
0.4
0
5000
10000
15000
20000
25000
30000
35000
AKT1/RPS6KA2
Feature z-score
1.2
0.4
−0.4
−1.2
AKT1/AKT2
AKT1/AKT3
AKT1/RPS6KA1
AKT1/RPS6KA3
AKT1/RPS6KA4
AKT1/RPS6KA5
AKT1/RPS6KA6
AKT1/RPS6KB1
AKT1/RPS6KB2
AKT1/SGK1
AKT1/SGK2
AKT1/SGK3
Sequence Identity
Essentiality of Protein Complex(es)
Shared Protein-Protein Interactors
Closest Paralog
S. pombe Ortholog
Gene Age
Whole Genome Duplicates
Family Size
Expression Correlation
Essentiality of Shared PPIs
0.00
0.05
0.10
SHAP value