Dede et al. Genome Biology (2020) 21:262
https://doi.org/10.1186/s13059-020-02173-2

Genome Biology

RESEARCH Open Access

Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens

Merve Dede[1,2†], Megan McLaughlin[1,2†], Eiru Kim[1] and Traver Hart[1,3*]

* Correspondence: traver@hart-lab.org
†Merve Dede and Megan McLaughlin contributed equally to this work.
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3]Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
Full list of author information is available at the end of the article

Abstract

Background: Pooled library CRISPR/Cas9 knockout screening across hundreds of cell lines has identified genes whose disruption leads to fitness defects, a critical step in identifying candidate cancer targets. However, the number of essential genes detected from these monogenic knockout screens is low compared to the number of constitutively expressed genes in a cell.

Results: Through a systematic analysis of screen data in cancer cell lines generated by the Cancer Dependency Map, we observe that half of all constitutively expressed genes are never detected in any CRISPR screen and that these never-essentials are highly enriched for paralogs. We investigated functional buffering among approximately 400 candidate paralog pairs using CRISPR/enCas12a dual-gene knockout screening in three cell lines. We observe 24 synthetic lethal paralog pairs that have escaped detection by monogenic knockout screens at stringent thresholds. Nineteen of 24 (79%) synthetic lethal interactions are present in at least two out of three cell lines and 14 of 24 (58%) are present in all three cell lines tested, including alternate subunits of stable protein complexes as well as functionally redundant enzymes.

Conclusions: Together, these observations strongly suggest that functionally redundant paralogs represent a targetable set of genetic dependencies that are systematically under-represented among cell-essential genes in monogenic CRISPR-based loss of function screens.

Background

The adaptation of CRISPR-Cas9 system to genome-wide knockout screens in mammalian cells has greatly transformed the search for cancer-specific genomic vulnerabilities that can be targeted therapeutically. Monogenic pooled library CRISPR-Cas9 knockout screens revealed that mammalian cells have as much as 3–4 times more essential genes than the previous RNAi technology was able to detect at the same false discovery rate [1]. Moreover, through immense monogenic screening efforts, multiple groups

revealed lists of ~ 2000 highly concordant human essential genes, and comparison of CRISPR technology to orthogonal techniques such as random insertion of gene traps also showed consistent results [2–4].

However, even with the CRISPR technology, the number of essential genes detected through these screens is still far less than the number of genes constitutively expressed in a given cell line. This phenomenon was previously observed in systematic gene knockout studies in *S. cerevisiae* [5, 6], where only 17% of yeast genes were essential for growth in rich medium [6]. A closer look at the biological characteristics that define essentiality revealed a modular nature of gene essentiality [7] in which essentiality is not a characteristic of the protein or gene itself, but is rather defined by the protein complex to which the protein belongs. While genes that encode for members of a protein complex were shown to be more likely to be essential, paralogous genes were less likely to be essential [8]. However, a later study showed that a binary classification of genes into essential and nonessential was misleading due to the context-dependent nature of gene essentiality and that 97% of yeast genes showed some growth phenotype under different environmental conditions [9]. A similar study in *C. elegans* [10] suggested that, at the organismal level, virtually every gene is required for optimal growth in some condition.

Paralogous genes arise from gene duplications, an evolutionary mechanism to create new genes. While gene duplication can result in two functionally distinct genes over time, more frequently, the genes preserve a proportion of functional overlap through the process of subfunctionalization [11, 12]. In yeast gene deletion studies, singletons (genes without paralogs) were more than twice as likely as paralogous genes to be essential [8], indicating the role of paralogs in genetic buffering and suggesting that paralogs can affect how yeast cells respond to genetic and environmental perturbation. The buffering ability of paralogs to each other's loss is explained by their functional redundancy. Double deletion studies of paralog gene pairs in yeast revealed that synthetic lethality occurred with depletion of both paralog pairs, resulting in a fitness defect that was more than the expected additive effect of individual gene depletions [13]. Further analyses determined sequence similarity of paralog pairs as a predictive characteristic for the level of functional redundancy [14]. A major open question remains whether, and to what extent, these findings hold true for human cells generally and cancer cells specifically.

Recent studies investigated paralog dependencies in monogenic genome-wide CRIS PR-Cas9 knockout screens in human cells, revealing differential effects of paralogs on cellular fitness. One study showed that paralogs are less likely to be essential in whole-genome CRISPR knockout fitness screens than singleton genes [15], while another study demonstrated that some paralogs that form heterodimers are more deleterious to the cell compared to non-heterodimer forming paralogs [16]. However, these studies did not take into account the effect of tissue-specific expression of the paralog pairs.

In this study, using publicly available genome-wide screen data of genetically heterogeneous cell lines from the Cancer Dependency Map initiative [17, 18], we investigate paralogs among constitutively expressed never-essential genes as a set of targetable genetic dependencies that are systematically excluded in monogenic CRISPR-Cas9 knockout screening. We further demonstrate experimentally, using CRISPR/enCas12a multiplex knockouts, that dual-gene screens reveal synthetic lethality among targeted paralogs.

Results and discussion

As part of our ongoing effort to understand differential gene essentiality, we looked at the relationship between gene expression and gene essentiality across hundreds of cancer cell lines. We looked at gene expression from all cell lines in the Cancer Cell Line Encyclopedia (CCLE) [19] and considered the role of tissue-specific vs. constitutive gene expression. We took the mean and standard deviation of gene expression across 684 cell lines with high-quality CRISPR screens from the Avana 19Q4 data release [17] and modeled the joint distribution with a linear combination of 2-d Gaussian mixture models. We find that three elements correspond to the three major populations in the data: constitutively expressed genes (high expression, low variance), never-expressed genes (low expression, low variance), and genes that show variable, sometimes tissue-specific gene expression ("sometimes expressed" genes, high variance) (Fig. 1a).

We evaluated the fraction of essential genes in each population. We defined essential genes as those with a BAGEL-derived BF > 10, a high-confidence threshold corresponding to a posterior probability of essentiality of ~ 99%. Common essential genes are largely constitutively expressed, as expected, while context-dependent essential genes are divided across the constitutive expression and tissue-specific expression. Interestingly, among constitutively expressed genes, many are never-essential in any CRISPR knockout fitness screen (3032 of 7282; 42%; Fig. 1b). These observations regarding the

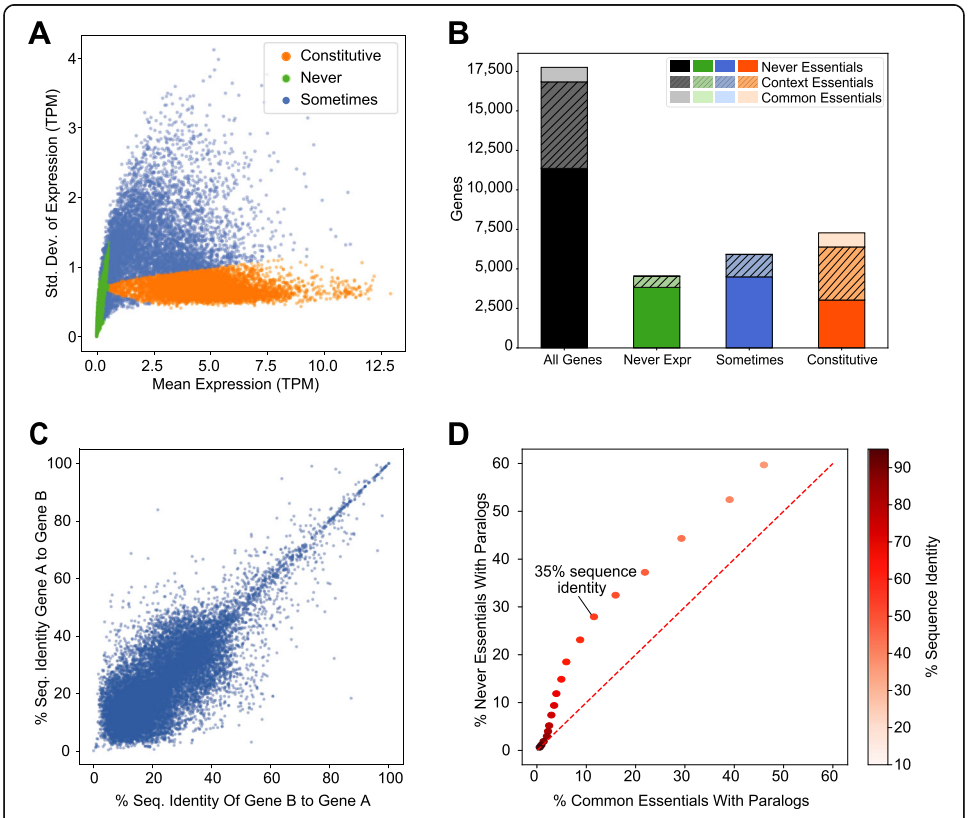

Fig. 1 Paralogs are under-represented in CRISPR-Cas9 screens. **a** Scatter plot of mean vs. standard deviation of log (TPM) gene expression in CCLE. Color coding by component of a 3-element Gaussian mixture model. **b** Fraction of context and common essentials as a function of gene expression. **c** Paralog pairwise sequence identity among constitutively expressed paralogs. **d** Fraction of common essentials with a paralog vs. fraction of never-essentials with a paralog, by paralog similarity

constitutively expressed genes raised the question about why we observe so few essential genes in these genetically heterogenous screens. Based on work in yeast and nematodes [10], we naively assumed that all constitutively expressed genes should be essential in some context and hypothesized that some combination of environmental or genetic buffering masks the fitness consequences of individual gene knockouts.

An important study by De Kegel and Ryan observed that paralogs are less likely to be essential in whole-genome CRISPR knockout fitness screens than singletons [15]. This work discovered more than 200 instances where higher essentiality in one paralog was accompanied by lower gene expression in the other, supporting the assertion that paralog buffering masks monogenic knockout fitness effects. We sought to extend this observation to include constitutively expressed genes. We obtained the list of the paralogs of human protein-coding genes from Ensembl Biomart [20] along with protein sequence similarity information (see the "Methods" section). After filtering for constitutively expressed genes, we observed that paralogs show a wide range of amino acid sequence similarity, with the majority showing relatively low identity (Fig. 1c). To evaluate whether paralogs are enriched in constitutively expressed never-essentials (hereafter "never-essentials"), we adopted a sliding scale of sequence identity and measured, at each threshold, the fraction of never-essentials and the fraction of common essentials captured. As shown in Fig. 1d, as sequence similarity stringency is relaxed, never-essentials are more likely to have a paralog than common essentials. At 35% or greater sequence similarity, nearly a third (27.9%) of constitutively expressed never-essentials have a paralog, compared with only 11.6% of common essentials ($P < 10^{-89}$, Z-test for difference in proportions).

To identify functionally redundant paralogs, we explored the Avana and Sanger data to find cases where loss of function of one member of a paralog pair resulted in increased dependency on the other (Fig. 2a). We limited the search for functional redundancy to genes classified as constitutively expressed according to our model, which excludes false associations arising from tissue-specific expression of paralog family members. The search is further constrained by requiring that one member of the pair show loss of function, either through predicted deleterious mutation or by severe decrease in gene expression (see the "Methods" section), in a sufficient number of cell lines to result in a statistically significant difference in gene essentiality of the other member. By applying this test to 628 gene pairs in the Avana data and 295 gene pairs in Project Score (Fig. 2a, Additional file 2: Table S1), we detected a total of 66 such cases of putative functional buffering at a P value < 0.01, of which 32 (48%) are common between the two sets (Fig. 2b, c). Two well-described cases in the BAF (mammalian SWI/SNF) complex were immediately apparent: mutations in *SMARCA4* are strongly associated with dependency on paralog *SMARCA2* ($P < 10^{-10}$; Fig. 2d), and mutations in *ARID1A* are associated with *ARID1B* dependency ($P < 10^{-9}$; Fig. 2e). Expanding loss-of-function to include significantly depleted gene expression also reveals an emergent dependency on *RPP25L* when *RPP25* is depleted ($P < 10^{-52}$; Fig. 2f). The two genes encode redundant subunits of RNAse P, a ribonuclease critical for maturation of tRNA, whose functional buffering was previously observed [4]. A fourth example is *FAM50A/FAM50B* putative functional redundancy (Fig. 2g). Interestingly, virtually nothing is known about the biological role of these genes.

Unfortunately, the cell lines screened by CRISPR knockout libraries only contain LOF alleles of a fraction of the candidate paralogs, limiting this discovery avenue to a

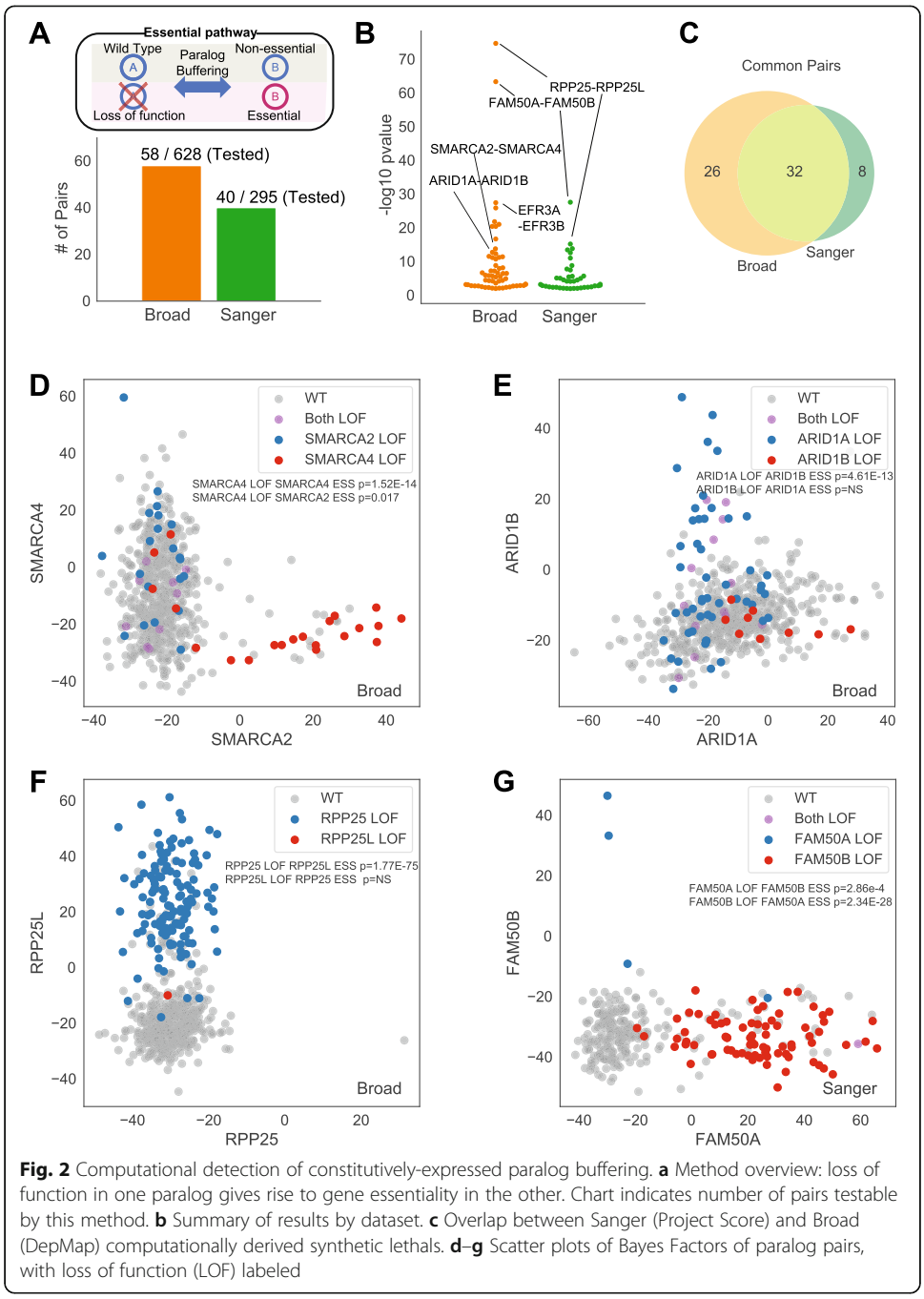

Fig. 2 Computational detection of constitutively-expressed paralog buffering. **a** Method overview: loss of function in one paralog gives rise to gene essentiality in the other. Chart indicates number of pairs testable by this method. **b** Summary of results by dataset. **c** Overlap between Sanger (Project Score) and Broad (DepMap) computationally derived synthetic lethals. **d**–**g** Scatter plots of Bayes Factors of paralog pairs, with loss of function (LOF) labeled

few dozen pairs. A comparison with De Kegel and Ryan [15] shows that more than half of our computationally derived hits (39 of 66, 59%; Additional file 1: Fig. S1) are present in their study, indicating strong concordance between the two approaches. Nevertheless, the large number of hits unique to each approach clearly indicates that neither approach is saturating, and additional approaches, both computational and experimental, are required to discover the complete catalog of paralog synthetic lethals.

Given the limitations of this computational approach, we sought to expand our knowledge of paralog buffering through systematic dual-gene CRISPR knockout screening. Cas12a, formerly Cpf1, offers an endogenous RNA endonuclease function that

enables processing and utilization of multiple gRNA from a single polycistronic transcript [21] and the modified enCas12a enzyme offers superior performance in genetic screens in mammalian cells [22, 23]. A key advantage of this system is that specific guide pairs can be synthesized in a single oligo, allowing one-step library design, a major advantage over multiplex Cas9 systems [24–27]. We therefore sought to apply the enCas12a multiplex knockout system to systematically identify paralog synthetic lethals. In our hands, cells with enCas12a effectively knocked out EGFP (Additional file 1: Fig. S2A) and achieved ~ 80% double knockout in a dual-guide construct targeting two cell surface markers (Additional file 1: Fig. S2B).

We chose 400 candidate paralog pairs to test experimentally. Gene pairs were selected based on several criteria, including amino acid sequence similarity, mRNA expression and co-expression, and whether either gene is frequently essential in DepMap. We manually added five additional candidate gene pairs from the literature: *SMARCA2-SMARCA4*, *CDH1-CDH3*, *ME2-ME3*, *BCL2L1-MCL1*, and *BRCA1-PARP1*, for a total of 405 targeted gene pairs. For each gene, up to three CRISPR RNA (crRNA) were selected using a library designed by DeWeirdt et al. [28]. Each gene pair was targeted with all 9 combinations of guides, in both A-B and B-A orientations, for a total of 18 clones targeting each pair. To evaluate single-knockout phenotype, we paired gene-targeted crRNA with three guides drawn from a pool of guides targeting 50 nonessential genes (Fig. 3a). We additionally targeted 50 essential genes, paired with random nonessential guides, as quality controls for the screens.

We transduced the library into enCas12a-expressing cells from three cancer cell lines of diverse origins: A549, a KRAS-driven lung cancer cell line; HT29, a BRAF-mutant colorectal cancer cell line; and OVCAR8 ovarian cancer cells. Cells were passaged in three replicates for 10 doublings and the relative abundance of each dual-guide construct was measured by 75-base single-end sequencing of the target amplicon, with fold changes measured relative to abundance in the plasmid pool. Quality control steps including abundance and distribution of read counts, clustering of raw read counts and fold changes, and separation of essential and nonessential control genes indicated effective screen performance (Fig. 3c and Additional file 1: Fig. S2). Additionally, high correlation of A-B and B-A guide pairs (Fig. 3d) indicates negligible positional bias in the enCas12a guide arrays. We therefore included both A-B and B-A pairs in all subsequent fitness calculations.

To calculate genetic interaction/synthetic lethality, we measured the single mutant fitness (SMF) for each gene as the mean fold change of the gene-control constructs. For control essential genes, SMF in our enCas12a screen correlates with BAGEL-derived Bayes Factor scores for the DepMap screens in the same cell lines (Fig. 3e). We then calculated the observed double mutant fitness (DMF) as the mean log fold change of the dual-gene knockout constructs (18 constructs per gene pair) and compared it to the expected DMF, the sum (in log space) of each gene's SMF (Fig. 3b). As has been widely observed in genetic interaction screens, most digenic knockouts do not result in an unexpected phenotype; here we observe that the distribution of "delta log fold change" (dLFC) values has most of its mass around zero (no synthetic effect), with a long tail of negative (synthetic sick/lethal) dLFC scores (Fig. 3f).

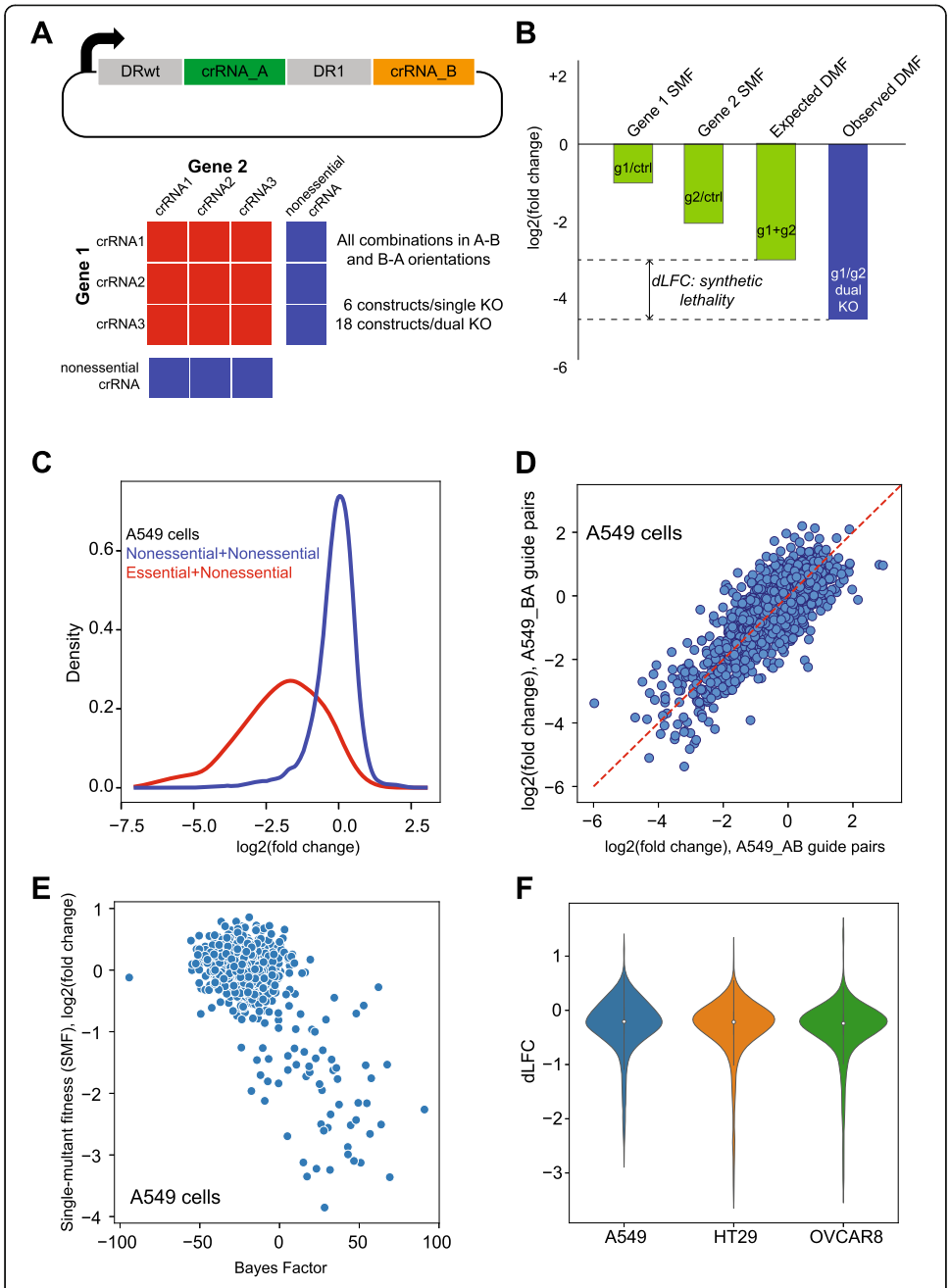

Fig. 3 Multiplex gene knockout with enCas12a. **a** Experimental design. EnCas12a crRNA dual-guide array design. Each construct targets two genes; each gene is targeted by 3 crRNA; each candidate paralog gene pair is targeted by 18 gene-gene constructs, with six gene-control constructs per gene (including both A-B and B-A orientations). **b** Evaluating synthetic lethality. Single mutant fitness (SMF) is the mean log fold change of control guides targeting a single gene. Expected double mutant fitness (DMF) is the sum of SMF. Observed DMF is the mean log fold change of dual-targeting constructs. Delta log fold change (dLFC) is the difference between observed and expected fold change. **c** QC plot showing separation of SMF of constructs targeting control essential and nonessential genes. **d** Scatter plot of all mirror constructs (same crRNA in A-B and B-A orientations) showing lack of positional effects. **e** SMF in this screen vs. gene BF in Avana data. **f** Distribution of dLFC scores for 403 gene pairs in each cell line

To compare across screens, we converted dLFC scores to a *Z* score, zdLFC, by truncating the top and bottom 2.5% of dLFC scores (Additional file 1: Fig. S2J, Additional file 3: Table S2, Additional file 4: Table S3). At a zdLFC score < – 3, all three screens showed high concordance, with 19 of 24 (79%) synthetic lethals present in at least two out of three cell lines and 14 of 24 (58%) present in all three (Fig. 4a, b). Fifteen of the 24 hits (62.5%) are ohnologs, gene copies resulting from whole-genome duplication [29], compared to 246 of the 405 gene pairs tested (60.7%), indicating neither enrichment nor depletion of synthetic lethals among ohnologs ($P = 0.86$, *Z*-test for difference of proportions). Despite prior work suggesting simple difference in log fold change is not an effective measure of genetic interaction [30], we find that zdLFC is highly correlated with more detailed approaches such as GEMINI [30], with R^2 values ranging from 0.59 (A549) to 0.74 (OVCAR8), and the two methods offer essentially no difference in hit calls (Additional file 1: Fig. S3).

Many top-scoring hits show strong concordance with other data corroborating a functional buffering/synthetic lethal relationship. RNA helicases *DDX19A* and *DDX19B* show characteristics of synthetic lethality as described by De Kegel and Ryan [15] across DepMap cell lines; DDX19A is strongly essential only when DDX19B is expressed at low levels (Fig. 4C). Similarly, TIAL1 low expression is associated with

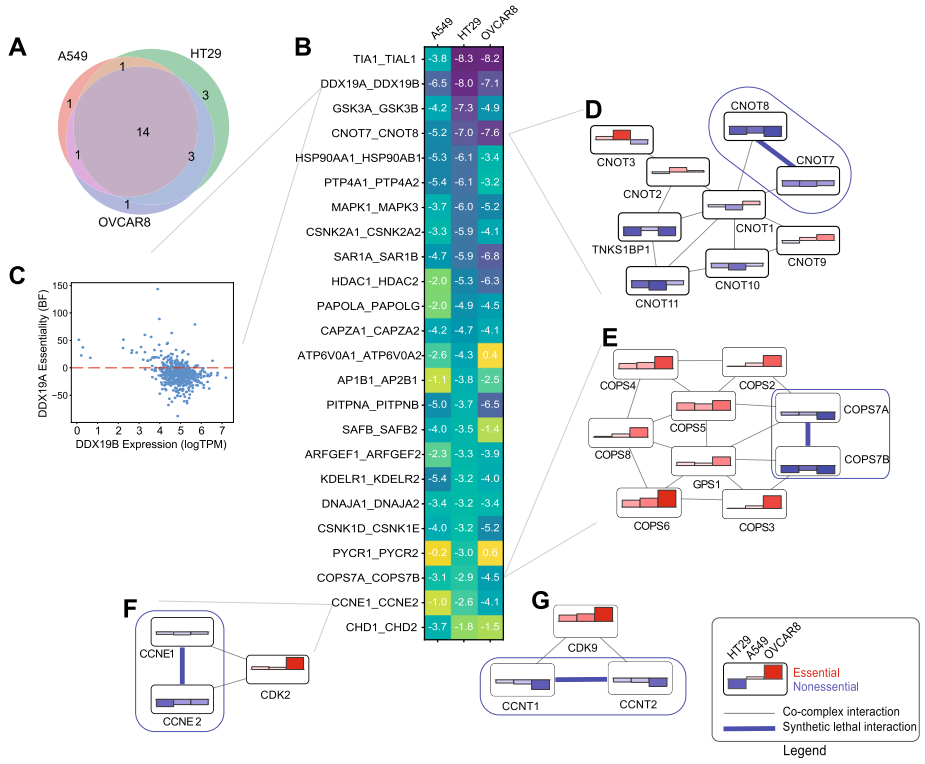

Fig. 4 Synthetic lethal paralogs. **a** Overlap of 24 hits with zdLFC < – 3 in any of the three cell lines. **b** Heatmap showing zdLFC score for the 24 hits. **c** *DDX19A* gene essentiality (BF) vs. *DDX19B* gene expression (logTPM) in Avana data. **d** Selected co-complex interactions in CCR4-NOT complex shows gene essentiality of several subunits and synthetic lethality of nonessential, alternate subunits *CNOT7/CNOT8*. **e** Co-complex interactions of COPS9 signalosome complex, showing strong gene essentiality of required components and synthetic lethality of alternate subunits *COPS7A/COPS7B*. **f** Synthetic lethality of *CCNE1/CCNE2* in OVCAR8 cells, where *CDK2* is highly essential. **g** Synthetic lethality of *CCNT1/CCNT2*, not shown in B (zdLFC = – 2.6), and essentiality of *CDK9*

TIA1 increased essentiality. Genes *CNOT7* and *CNOT8* encode alternate subunits of the CCR4-NOT complex, a critical regulator of eukaryotic gene expression [31]. Other subunits are sporadically essential in our three cell lines (Fig. 4d) but frequently essential across DepMap data [17, 18, 32], consistent with a constitutively essential protein complex. Moreover, CNOT7 essentiality is weakly but significantly anticorrelated with CNOT8 mRNA expression (Pearson correlation coefficient − 0.21, $P < 10^{-6}$). Likewise, *COPS7A* and *COPS7B* encode alternate, replaceable subunits of the COP9 signalosome complex; other subunits are irreplaceable and are uniformly essential in these cell lines (Fig. 4e).

Importantly, we note that synthetic lethality, even among paralogs, can also be context-dependent. Cyclin paralogs are often redundant interaction partners with their cognate cyclin-dependent kinases; here, CCNE1 and CCNE2 are synthetic lethal where CDK2 is highly essential, especially in OVCAR8 (Fig. 4f). Similarly, CCNT1-CCNT2 show weaker but significant synthetic lethality (zdLFC < − 2.5 in A549 and < − 1 in the other two cell lines) while their binding partner, CDK9, is highly essential in all three (Fig. 4g). Though the synthetic lethal relationships between SWI/SNF complex members *ARID1A/ARID1B* and *SMARCA2/SMARCA4* are well described in the literature and are detected in large scale screening data, their synthetic lethality only occurs where the SWI/SNF complex is itself essential. We test four paralog pairs in the BAF complex: *ARID1A/ARID1B*, *SMARCA2/SMARCA4*, *SMARCC1/SMARCC2*, and *SMARCD1/SMARCD2*, but we detect no synthetic lethal interactions, most likely because the complex itself is not essential in the cell lines we tested.

Synthetic lethality between our hits is corroborated by a dual-gene knockout screen using the CHyMErA hybrid Cas12/Cas9 system [33]. The 678 paralog pairs evaluated in the CHyMErA screens contain 110 pairs targeted in our library, including 12 of the 24 hits we defined. Our results are generally consistent, with *TIA1/TIAL1*, *SAR1A/SAR1B*, *PITNA/PITNB,* and *CNOT7/CNOT8* scoring strongly in both assays (Additional file 1: Fig. S4). In contrast, *MAPK1/MAPK3* and *CCNE1/CCNE2* are only hits in our cell lines. As with gene essentiality, synthetic lethality is often highly context-dependent.

Conclusions

CRISPR technology has revolutionized mammalian functional genomics and cancer targeting by leveraging endogenous DNA repair machinery to generate gene knockouts on a genomic scale. Extensive screening of cancer cell lines has been performed under the DepMap and Project Score initiatives to identify context-specific weaknesses and cancer biomarkers. Analyses of this data have revealed activation of oncogenic pathways and oncogene dependencies [18] as well as biomarker type dependencies such as Werner helicase, WRN, in colorectal and ovarian cell lines with MSI [34, 35]. However, despite these efforts, questions about what might be systematically missing from these data have, to our knowledge, not been rigorously explored.

We note that there are about 7000 genes that are constitutively expressed in each cell, but only about half of these are ever detected as essential. Studies in model organisms suggest that virtually every gene shows a growth phenotype under some environmental condition [9, 10]. It is unknown whether this holds true for individual mammalian cells, though tumors are often modeled as though they are colonies of single-celled organisms. It is also the case that most genetic screens of tumor cells are

carried out under permissive growth conditions, minimizing nutrient and oxidative stress to maximize growth rate and improve detection of dropouts. Thus, the degree of environmental buffering is largely unknown for these constitutively expressed never-essentials.

However, these never-essentials are highly enriched for paralogs. They are ~ 3 times more likely to have a paralog than always-essentials, suggesting that functional redundancy by related genes masks detection of a substantial population of genes in monogenic CRISPR knockout screens. This has profound implications for efforts to match targeted drugs with tumor genotypes, and to discover new candidate drug targets. Targeted small molecules often do not discriminate, or discriminate poorly, between closely related paralogs, and it is often their promiscuity rather than their specificity that renders them effective. For example, MEK inhibitor trametinib effectively targets the protein products of both *MAP2K1* and *MAP2K2*, redundant kinases downstream of RAS/RAF oncogenes, but the functional redundancy of these genes renders them both invisible to monogenic CRISPR screens, even in RAS/RAF backgrounds [36].

Recent developments in CRISPR screening technology enable effective genetic targeting of multiple genes simultaneously. Cas12a, previously known as Cpf1, is able to process a polycistronic mRNA to generate multiple CRISPR RNAs (crRNAs). This makes multiplexing much easier compared to inefficient Cas9 based multiplex systems which requires each guide RNA to be expressed by its own promoter. The improved version of this enzyme, enCas12a [22], coupled with an effective guide design algorithm [28] presents a powerful platform for multiplex genetic perturbation. Multiplex guide libraries can be synthesized directly, without requiring additional targeted or random mixing cloning steps, allowing direct assay of specific gene pairs as described here with roughly the same level of effort as a now-standard Cas9 monogenic screen. The robustness of predicted guide cutting efficiency remains untested relative to Cas9, given the relatively small amount of enCas12a data available, suggesting adopters of this technology should err toward caution when deciding on parameters for new experiments (e.g., number of guides per gene, number of gene-vs-control guide pairs). Nevertheless, as we demonstrate here, this platform holds enormous potential for exploring the stability and plasticity of genetic interactions in human cells.

Methods

DepMap essentiality data

A raw read count file of CRISPR pooled library screens for 690 cell lines using Avana library [17] (Broad DepMap project 19Q4) was downloaded from the data depository (https://depmap.org/portal/). Also, we downloaded Project Score (Sanger) screen [34] raw read counts for 323 cancer cells from the data depository (https://score.depmap.sanger.ac.uk/). We filtered the dataset to keep only the protein-coding genes for further analysis and updated their names using HGNC [13] and CCDS [37] database. We discarded sgRNAs targeting multiple genes in Avana library to avoid genetic interaction effects. The raw read counts were processed with the CRISPRcleanR [38] algorithm to correct for gene-independent fitness effects and calculate fold change. After that, the CRISPRcleanR processed fold changes of each cell line were analyzed through updated BAGEL2 build 114 (https://github.com/hart-lab/bagel). In comparison with published

BAGEL version v0.92 [39], the updated version employed a linear regression model to interpolate outliers and 10-fold cross-validation for data sampling. Essentiality of genes was measured as Bayes Factor (BF) based on gold standard reference sets of 681 core essential genes and 927 nonessential genes [1, 40]. Positive BF indicates essential genes and negative BF indicates nonessential genes. Lists of core essential genes and nonessential genes used in this study have been uploaded on the same repository with BAGEL2 software. To correct unexpected essentiality by sgRNAs targeting non-protein-coding regions in addition to desired target protein-coding gene, the multi-targeting effect of sgRNAs has been corrected using BAGEL2 -m option. The screen quality was evaluated by using "precision-recall" function in BAGEL2 software, and F-measure, the harmonic mean of precision and recall, was calculated for each screen at BF = 5. Finally, 581 cell lines for Broad screen and 320 cells for Sanger screen were selected for further study by F-measure threshold 0.8 to prevent noise from marginal quality of screens.

Defining constitutively expressed genes with GMM modeling

We utilized the log2 transformed RNA-seq TPM expression data from DepMap Data Portal expression data for Avana19Q4 release for 684 cell lines [17, 18]. The standard deviation of expression versus the mean expression values for all genes assayed in the Avana library (N = 17,755) across all cell lines, for which the expression data was available, were plotted. Python 3.6.9 package sklearn and its GaussianMixture function was used to classify genes by Gaussian mixture modeling based on mean and standard deviation of mRNA expression. A three-component model was selected as the best fit to the data (Additional file 1: Fig. S5) since the addition of a fourth component resulted in two highly overlapping component distributions. The group with the least expression and low standard deviation was labeled as never expressed, the second group with very high standard deviation and a range of mean expression values was labeled as sometimes expressed, and the constitutively expressed group with high mean expression and low standard deviation was classified as constitutively expressed genes. With this classification, we identified 7282 always expressed, 4544 never expressed, and 5929 sometimes expressed genes in the Avana dataset.

Paralogs

The human paralogous gene pairs for the protein-coding genes were utilized from Ensemble Release 95 Biomart with GRCh38.p12 genome assembly [20]. This release of Ensemble estimates paralogues from gene trees that are constructed with HMM as described in more detail at http://www.ensembl.org/info/genome/compara/homology_method.html. Other information such as chromosome location, paralogue percent sequence identity to human target gene, and percent sequence identity of target gene to the paralogous gene were also downloaded. After removing duplicate gene pairs and filtering for constitutively expressed genes, for each paralog pair (A-B pair), we obtained their percent sequence identities and we plotted the sequence similarities of A to B against those of B to A. We observed that the majority of the human paralogous gene pairs had low percentage sequence similarity. The paralog pairs which were both constitutively expressed gene lists were identified and were binned according to

different thresholds for percent sequence identity from a range of 10–95%. For each bin, the percentage of constitutively expressed never-essential genes with paralogs and the percentage of common essential genes (defined in [41]) with paralogs were calculated and their distributions were plotted. For downstream analysis, always expressed paralog pair lists were generated for each sequence identity threshold.

Discovering functional redundancy between paralogs in DepMap CRISPR/Cas9 screens

To investigate evidence for the functional redundancy between paralog genes in Broad and Sanger screens, we tested whether a gene is essential when the other paralog partner suffers loss of function. Firstly, we defined loss of function (LOF) call combining damaging mutations calls (frameshift or nonsense) adopted from CCLE mutation data [42] depletion of expression (mean log TPM < 1.0, CCLE RNA-seq) or deletion (copy number < 0.1, CCLE Copy number data). Then, we conducted statistical test of synthetic essentiality which is defined when a gene is observed as essential when its paralog partner loses its function. One-to-one paralog pairs with at least 30% sequence similarity were considered for this analysis to maximize the number of paralog pairs. We considered only pairs whose genes have at least two LOF calls and are essential in at least two cell lines. *P* value was calculated by the one-sided Fisher's exact test on the 2 × 2 contingency table of the number of cells classified by LOF and essential (BF > 10), and false discovery rate (FDR) was calculated by the method of Benjamini and Hochberg. We addressed pairs bidirectional ways, which test a significance of essentiality of gene A upon LOF of gene B and vice versa. A total of 57 pairs among 628 tested pairs in the Broad dataset and 40 pairs among 295 tested pairs in the Sanger dataset passed a threshold of *P* value < 0.01. Thirty-two pairs were common to both datasets.

Selecting paralogs for experimental testing

To identify and predict paralog pairs which we hypothesized to be enriched in synthetic lethal interactions, we applied multiple filters including percent sequence similarity, mean expression, standard deviation of expression, co-expression, and gene essentiality profiles. We built a network of paralogous gene families using Cytoscape [43] and filtered them initially for protein sequence similarity greater than or equal to 45%, mean expression (logTPM) > 1.5, standard deviation of expression < 1.25, and co-expression Pearson correlation coefficient > 0.1. Finally, we removed genes that were essential in more than 30 cell lines, resulting in a set of 400 pairs. In addition, we manually added several candidate synthetic lethals from the literature, including SMARCA2/SMARCA4, CDH1/CDH3, ME2/ME3, BCL2L1/MCL1, and BRCA1/PARP1.

Library design

We selected Cas12a CRISPR RNA sequences from a library from [28]. Guides were selected from an AsCas12a library design from July 2019, representing an intermediate phase of development of the DeWeirdt et al. work. Up to the top three guide sequences were selected from the library for each of the 793 candidate paralog genes (405 gene pairs), but given the restrictive TTTV PAM sequence for AsCas12a, three guides were not available for every gene. For two genes (ABHD16B, DGCR6), no crRNA were present in the library; pairs including these genes were removed in the downstream

analysis. As controls, a set of 50 nonessential and 50 pan-essential genes were chosen; genes were filtered for those with 1:1 orthologs in both rat and mouse to provide a useful multi-species reference set. These control genes are listed in Additional file 5: Table S4.

To design the library, we first pooled all crRNA targeting nonessential control genes (141 crRNA targeting 50 genes). Then, for each paralog gene pair, we collected all crRNA pairs in both orientations—for $n = 3$ crRNA per gene, there are $n^2 = 9$ crRNA pairs, or 18 total clones (A-B and B-A orientations for each). To generate single-knockout controls, we then took each crRNA targeting one of the paralogs and paired it with a crRNA randomly drawn from the nonessential pool, again designing clones in both A-B and B-A orientations, for a total of six control constructs per experimental gene (where $n = 3$ crRNA/gene). Finally, we took our set of control essential genes ($n = 149$ crRNA targeting 50 genes) and randomly paired each guide with a nonessential guide, in both orientations, as described above, for a total of 298 positive control guide constructs. The final library targets 841 genes (889 including nonessential genes) and 403 specified gene pairs with 12,328 constructs.

Vectors

The following vectors were a kind gift from John Doench:

pRDA_174 (enzyme expression): EF1a promoter drives EnCas12a enzyme expression; lentiviral vector; confers blasticidin resistance (Addgene #136476).
pRDA_052 (guide expression): U6 promoter drives gRNA expression; vector contains AsCas12a direct repeat upstream of dual BsmBI sites for insertion of guide arrays; lentiviral vector; confers puromycin resistance (Addgene #136474).
pRDA_221 (positive control): confers constitutive of short half-life EGFP and expression of two guides targeting EGFP; lentiviral vector; confers puromycin resistance [23, 28].
pDV204 (positive control): U6 promoter drives expression of guides targeting cell-surface markers CD47 and CD63; derived from pRDA_052 [23, 28].

Library production

An oligonucleotide pool comprising 12 k dual guide arrays was synthesized by CustomArray based on the following template:

5′*TATCTTGTGGAAAGGACGAAAC*ACCGG**TAATTTCTACTCTTGTAGAT**NNNNNNNNNNNNNNNNNNNN**TAATTTCTACTGTCGTAGAT**nnnnnnnnnnnnnnnnnnnnnnnnTTTTTTGAATT*CGCTAGCAAGCTTGGCGTAAC*-3′. The 145 nt fragment included the wildtype direct repeat for AsCas12a (bold) and an engineered variant direct repeat (bold underlined) [23, 28] for 23 nt guide sequences in the first (uppercase N) and second (lowercase n) positions, respectively. Flanking sequences (italic) enabled PCR amplification of the pool and cloning into BsmBI-linearized pRDA_052 by Gibson assembly.

The pool of guide arrays was amplified using Kapa HiFi 2X HotStart ReadyMix (Roche) using 10 ng of starting template per 50 μL reaction using primers DV202 (5′-TATCTTGTGGAAAGGACGAAAC) and DV203 (5′ GTTACGCCAAGCTTGCTAGC

G) at 0.3 μM final concentration and the following conditions: initial denaturation at 95 °C for 3 min, followed by twelve cycles of 20 s at 98 °C, 20 s 60 °C, 20 s at 72 °C using a ramp rate of 2.0 °C/s, and final extension at 72 °C for 5 min.

Full-length amplicon (145 bp) was purified by non-denaturing polyacrylamide gel electrophoresis using precast 10% acrylamide TBE gels (Bio-Rad). The guide expression vector pRDA_052 was digested with BsmBI (New England Biolabs), de-phosphorylated with Antarctic phosphatase (New England Biolabs), and concentrated using PCR cleanup columns (Life Technologies). Vector and insert were quantified using fluorimetry (Qubit dsDNA High Sensitivity kit, ThermoFisher). DNA assembly reactions using 0.4 pmol insert and 0.1 pmol vector per 20 μL HiFi Master Mix (New England Biolabs) were incubated at 50 °C for 1 h, re-digested with BsmBI, and desalted (Monarch low volume elution columns, New England Biolabs) for electroporation into Endura electrocompetent cells (Lucigen). After 1 h recovery at 37 °C, the bacteria were diluted 1:100 in 2xYT containing 200 μg mL-1 carbenicillin (AMS Bio) and grown at 30 °C for 16 h. Transfection grade plasmid was purified (PureLink HiPure Maxiprep, Invitrogen) and its guide arrays were sequenced to confirm complete and uniform library representation.

Cell culture

A549 and HT29 cells were a kind gift from Tim Heffernan. OVCAR8 cells were a kind gift from Phil Lorenzi. Cell line identities were confirmed by STR fingerprinting by MD Anderson's Cytogenetics and Cell Authentication Core (Powerplex 16 Locus High Sensitivity Assay, Promega).

Cells were grown at 37 °C in humidified 5.0% CO_2 atmosphere and passaged 2–3 times per week to maintain exponential growth. A549 and HT29 were grown in HEPES-modified DMEM (Sigma D7161); OVCAR8 was grown in HEPES-modified RPMI (Sigma R5886). Base media were supplemented with 10% FBS (Sigma), 1 mM sodium pyruvate (Gibco), 2 mM L-alanine-L-glutamine (Gibco), 1X penicillin-streptomycin (Gibco), and 100 μg mL^{-1} Normocin (Invivogen). Antibiotic-free cultures were routinely tested for mycoplasma contamination (PlasmoTest, Invivogen).

enCas12a screens

Lentivirus was produced by the University of Michigan Vector Core. Virus stocks were not titered in advance: all transductions were performed in multiple plates with a range of virus volumes and 8 μg mL^{-1} polybrene (EMD Millipore), but only the pool with the most optimal transduction efficiency was expanded and screened.

First, stable enCas12a expression was engineered by transduction with pRDA_174 at low MOI (10–20% transduction efficiency). Non-transduced cells were eliminated by selection with 10 μg mL^{-1} blasticidin (Invivogen). Selection was maintained until non-transduced controls reached 0% viability twice in succession (~ ten doublings). Editing efficiency was confirmed by transduction with control vectors targeting EGFP (pRDA_221) or cell surface markers CD47 and CD63 (pDV204) and flow cytometry. Cell lines lacking EnCas12a expression served as controls. Conjugated fluorescent antibodies and isotype controls were from BioLegend.

Second, enzyme-expressing pools were transduced with guide array virus. Multiple sub-confluent 15 cm plates were transduced to achieve a minimum of 12 M unique transductants without exceeding 50% transduction efficiency. Non-transduced cells were eliminated by 72 h treatment with 2 μg mL^{-1} puromycin (Gibco).

After puromycin selection was complete, three replicates were seeded, using 12 M viable cells per replicate, i.e., ~ 1000 cells per guide array. Screens were fed fresh medium every 2–3 days and passaged before reaching 80% confluency. Each replicate was re-seeded with 12 M viable cells to maintain coverage. Remaining cells were stored in 30 M aliquots at − 80 °C in cryopreservation medium (CellBanker 2, ZenoAq). Screens were terminated when replicates reached ten doublings.

Sequencing

Genomic DNA (gDNA) purification was automated in 24-well plates using a Kingfisher Flex instrument (ThermoFisher) and magnetic bead-compatible reagents (Mag-Bind Blood and Tissue DNA HDQ, Omega Biotek). Purified gDNA was eluted in 10 mM Tris-HCl pH 8.0, 1 mM EDTA and quantified by fluorimetry (Qubit dsDNA Broad Range kit, ThermoFisher).

Illumina-compatible guide array amplicons were amplified from gDNA in one step, as described [23, 28]. Indexed PCR primers were synthesized by Integrated DNA Technologies using Illumina's standard 8 nt indexes (D501-D508 and D701-D712). The forward primer design was 5'**AATGATACGGCGACCACCGAGATCTACAC**NNNNNNNN**ACACTCTTTCCCTACACGACGCTCTTCCGATCT***CTTGTGGAAAGGACGAAACACCG (5′-* **i5 flow cell adapter** – i5 index – **i5 read1 primer binding site** – *U6 annealing sequence)*. The reverse primer design was 5'**CAAGCAGAAGACGGCATACGAGAT**NNNNNNNN**GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT***GTTACGCCAAGCTTGCTAGCGAATTC* (*5′-* **i7 flow cell adapter** – i7 index – **i7 read2 primer binding site** – *pRDA_052 annealing sequence.)* Guide arrays were amplified from 80 μg of gDNA per replicate in multiple reactions, not exceeding 10 μg per 100 μl PCR volume. Eighty micrograms represents at least 500 cells per guide array for these hypotriploid cell lines (www.ATCC.org).

Each 100 μL reaction contained 0.5 μM of each primer, 200 μM dNTPs, and 1.25 μL of ExTaq polymerase (Takara). Guides were amplified using a slow ramp rate (2.0 °C/s) and minimum cycle number to limit bias, as follows: initial denaturation at 95 °C for 60 s, followed by 28 cycles of 30 s at 94 °C, 30 s at 52.5 °C, 30 s at 72 °C, and final extension at 72 °C for 10 m. Please note that Sanson et al. [23] now recommend using Titanium Taq plus DMSO (Takara), and we have observed slightly better mapping rates for Titanium Taq amplicons. The ~ 200 bp indexed amplicons were purified by size selection (2% agarose, E-Gel SureSelect II, ThermoFisher), quantified (QuBit, ThermoFisher), and pooled. Sequencing was performed using custom read primer oligo1210 (5′-CTTGTGGAAAGGACGAAACACCGGTAATTTCTACTCTTGTAGAT) (HPLC purified, Integrated DNA Technologies) using NextSeq 1 × 75 nt High Output reagents (Illumina).

Screen analysis

Construct sequences were combined into FASTA format ("paralog_2mer.fa") and indexed with bowtie-build, and sequencing reads were mapped to this database

with bowtie [44] with the following command line parameters (trim ten 3′ bases, allow 3 mismatches, discard sequences which map to more than one reference sequence): `bowtie --trim3 10 -v 3 -m 1 -S --sam-nohead paralog_2mer [fastq_files] > [output.sam]`

Sequence mapping rates ranged from 37 to 58%, averaging ~ 47%. Using this strict single-read mapping approach guarantees that only high-quality guide constructs were evaluated. Read counts were combined into a single matrix for further analysis (Additional file 4:Table S3).

Subsequent analysis was executed in Python notebooks, all of which are available at https://figshare.com/articles/software/enCas12a_screen_analysis_pipeline/12275642 [45]. Mean read depth for all samples exceeded 500 reads/guide, and all samples showed read distributions with minimal skew (Additional file 1: Fig. S2). A pseudocount of 5 reads was added to each construct in each sample, then read counts per sample were normalized to an average of 500 reads/guide (6.2 M reads/sample), and log fold change for each guide was calculated relative to the plasmid sequence counts (notebook cas12a-step01-screen_QC). Screen replicate quality was verified by plotting the kernel density estimate of the fold changes of all control essential constructs vs. all other constructs (see notebook cas12a-step04_calc_SMF; summarized in Fig. 3c). Screen-level fold change for each construct was then calculated as the mean of replicate fold changes.

Single mutant/knockout fitness, SMF, for each gene was calculated as the mean construct fold change of gene-control constructs, for both A and B position. Construct-level consistency is shown in Fig. 3d but gene-level SMF is even more consistent (see notebook cas12a-step04_calc_SMF), with Pearson correlation coefficients ranging from 0.87 to 0.94. A-B and B-A constructs were subsequently averaged to calculate sample-level SMF for each gene. The distribution of each shows a left skew consistent with the dropout (negative SMF) of a proportion of the genes in the sample (Table 1).

Difference in log fold change for a gene pair (dLFC) was calculated as observation, the mean LFC of all constructs targeting the gene pair, minus expectation, the sum of the SMF for the two genes. Given the skew of the SMF distributions, gene pairs with small positive SMF values sum yield an expectation of a positive LFC and, therefore, negative dLFC scores when the observed LFC is near zero. This explains the slight negative offset of dLFC distributions in Fig. 3f and necessitates normalization before calling hits. We normalized by Z-transformation after removing the top and bottom 2.5% of scores (see notebook cas12a-step07_robustZ_of_dLFC). The resulting zdLFC table was used for all subsequent analysis of synthetic lethality.

Table 1 Mean and median SMF

Cell line	Mean SMF	Median SMF
A549	− 0.10	0.014
HT29	− 0.10	0.053
OVCAR8	− 0.10	0.082

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02173-2.

Additional file 1: Fig. S1. Comparison of computationally derived hits from our analysis with hits in De Kegel and Ryan et al. [15]. **Fig. S2.** (A) Knockout of GPF with a crRNA targeting GFP in enCas12a knock-in cells. (B) Targeting two cell surface markers with a dual-guide crRNA in enCas12a-expressing OVCAR8 cells. (C) Total amplicon reads for the paralog screen. Dashed line indicates 500x sequencing depth (6 m reads for 12 k library). (D) Distribution of reads (boxplot) indicates good library representation in each sample. Dashed line = 30 reads/construct. (E) Clustering of normalized read counts. Clustering of replicates is consistent with high-quality screen data. (F, G) Lack of positional bias in mirror constructs containing the same two crRNA in A-B and B-A orientations. (H, I) SMF in this screen vs. BF from Avana data.(J) Z-transformation of distribution of dLFC (zdLFC) after truncating top/bottom 2.5% of values approximates a normal distribution. **Fig. S3.** Comparison of zdLFC scores to scores generated by GEMINI. (A) zdLFC vs GEMINI scores for 24 synthetic lethal pairs with their respective correlation coefficients. (B) zdLFC vs GEMINI scores for all tested paralog pairs with their respective correlation coefficients. **Fig. S4.** Comparison of common paralog pairs tested in our enCas12a screen with the CHyMErA screens. (A) Comparison of the 12 enCas12a hits in this study that were screened in HAP1 in the CHymErA study. (B) Comparison of all 110 paralog pairs tested in both enCas12a screen and the HAP1 CHymErA screen. **Fig. S5.** Gaussian mixture modeling (GMM) of gene expression of Avana 19Q4 cell lines. (A) Scatter plot of standard deviation of expression versus mean expression of gene assayed in Avana library in Avana19Q4 cell lines. (B) Contour plots of the two Gaussians from a two-component mixture model of data shown in A. (C) Contour plots of three-component GMM. (D) Contour plots of fourcomponent GMM.

Additional file 2: Table S1. Table of computational scores for predicted paralog pairs. (XLS 122 kb)

Additional file 3: Table S2. Table of zdLFC scores from the paralog screen.

Additional file 4: Table S3. Table of read counts of the paralog screen.

Additional file 5: Table S4. Control essential and nonessential genes.

Additional file 6. Review history.

Acknowledgements

A549 and HT29 cells were a kind gift from Tim Heffernan. OVCAR8 cells were a kind gift from Phil Lorenzi. Vectors pRDA_174 (Addgene #136476) and pRDA_052 (Addgene #136474) were a kind gift from John Doench.

Peer review information

Yixin Yao was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 6.

Authors' contributions

MD, MM, EK, and TH designed and executed the project. MD and EK conducted all bioinformatics analyses involving paralogs and DepMap/CCLE data. MM performed all experimental work constructing the enCas12a library, screening cell lines, and preparing libraries for quantitative sequencing. MD and TH wrote the manuscript, and MD, MM, EK, and TH edited it. The authors read and approved the final manuscript.

Authors' information

Twitter handles: @drmervedede (Merve Dede); @megsmclaugh (Megan McLaughlin); @paran_mir (Eiru Kim); @TraverHart (Traver Hart).

Funding

MD, MM, and TH were supported by NIGMS grant R35GM130119. TH is a CPRIT Scholar in Cancer Research (grant RR160032) and is additionally supported by MD Anderson Cancer Center Support Grant P30 CA016672.

Availability of data and materials

Python notebooks for the analysis of enCas12a screen can be found in the figshare repository at https://figshare.com/articles/software/enCas12a_screen_analysis_pipeline/12275642 [45].
Sequencing data are available on the NCBI BioProject Archive under accession no. PRJNA664967 [46].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

MM is a shareholder in Pionyr Immunotherapeutics. TH is a consultant for Repare Therapeutics. The remaining authors declare no competing interests.

Author details
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Graduate School of Biological Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [3]Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Received: 13 May 2020 Accepted: 30 September 2020
Published online: 15 October 2020

References

1. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014;10(7):733.
2. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350(6264):1092–6.
3. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell. 2015;163(6):1515–26.
4. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096–101.
5. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002;418(6896):387–91.
6. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science. 1999;285(5429):901–6.
7. Hart GT, Lee I, Marcotte ER. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics. 2007;8(1):236.
8. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H. Role of duplicate genes in genetic robustness against null mutations. Nature. 2003;421(6918):63–6.
9. Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. Science. 2008;320(5874):362–5.
10. Ramani AK, Chuluunbaatar T, Verster AJ, Na H, Vu V, Pelte N, et al. The majority of animal genes are required for wild-type fitness. Cell. 2012;148(4):792–802.
11. Brookfield JFY. Genetic redundancy: screening for selection in yeast. Curr Biol. 1997;7(6):R366–8.
12. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 2008;9(12):938–50.
13. DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, et al. Exposing the fitness contribution of duplicated genes. Nat Genet. 2008;40(5):676–81.
14. Li J, Yuan Z, Zhang Z. The cellular robustness by genetic redundancy in budding yeast. PLoS Genet. 2010;6(11):e1001187.
15. De Kegel B, Ryan CJ. Paralog buffering contributes to the variable essentiality of genes in cancer cell lines. PLOS Genet. 2019;15(10):e1008466.
16. Dandage R, Landry CR. Paralog dependency indirectly affects the robustness of human cells. Mol Syst Biol. 2019;15(9). [cited 2020 May 2] Available from: https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.20198871.
17. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49(12):1779–84.
18. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a Cancer Dependency Map. Cell. 2017;170(3):564–76 e16.
19. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7.
20. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2017;46(D1):D754–61.
21. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163(3):759–71.
22. Kleinstiver BP, Sousa AA, Walton RT, Tak YE, Hsu JY, Clement K, et al. Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. Nat Biotechnol. 2019;37(3):276–82.
23. Sanson KR, DeWeirdt PC, Sangree AK, Hanna RE, Hegde M, Teng T, et al. Optimization of AsCas12a for combinatorial genetic screens in human cells. Genetics; 2019 [cited 2020 May 3]. Available from: http://biorxiv.org/lookup/doi/10.1101/747170.
24. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121):819–23.
25. Kabadi AM, Ousterout DG, Hilton IB, Gersbach CA. Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. Nucleic Acids Res. 2014;42(19):e147.
26. Chen X, Xu F, Zhu C, Ji J, Zhou X, Feng X, et al. Dual sgRNA-directed gene knockout using CRISPR/Cas9 technology in Caenorhabditis elegans. Sci Rep. 2015;4(1):7581.
27. Shen JP, Zhao D, Sasik R, Luebeck J, Birmingham A, Bojorquez-Gomez A, et al. Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. Nat Methods. 2017;14(6):573–6.
28. DeWeirdt PC, Sanson KR, Sangree AK, Hegde M, Hanna RE, Feeley MN, et al. Optimization of AsCas12a for combinatorial genetic screens in human cells. Nat Biotechnol. 2020; [cited 2020 Aug 16] Available from: http://www.nature.com/articles/s41587-020-0600-6.
29. Singh PP, Arora J, Isambert H. Identification of ohnolog genes originating from whole genome duplication in early vertebrates, based on synteny comparison across multiple genomes. PLOS Comput Biol. 2015;11(7):e1004394.
30. Zamanighomi M, Jain SS, Ito T, Pal D, Daley TP, Sellers WR. GEMINI: a variational Bayesian approach to identify genetic interactions from combinatorial CRISPR screens. Genome Biol. 2019;20(1):137.

31. Lau N-C, Kolkman A, van Schaik FMA, Mulder KW, Pijnappel WWMP, Heck AJR, et al. Human Ccr4–Not complexes contain variable deadenylase subunits. Biochem J. 2009;422(3):443–53.
32. Lenoir WF, Lim TL, Hart T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. Nucleic Acids Res. 2018;46(D1):D776–80.
33. Gonatopoulos-Pournatzis T, Aregger M, Brown KR, Farhangmehr S, Braunschweig U, Ward HN, et al. Genetic interaction mapping and exon-resolution functional genomics with a hybrid Cas9–Cas12a platform. Nat Biotechnol. 2020 [cited 2020 May 3]; Available from: http://www.nature.com/articles/s41587-020-0437-z.
34. Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. Nature. 2019;568(7753):511–6.
35. Chan EM, Shibue T, McFarland JM, Gaeta B, Ghandi M, Dumont N, et al. WRN helicase is a synthetic lethal target in microsatellite unstable cancers. Nature. 2019;568(7753):551–6.
36. Kim E, Dede M, Lenoir WF, Wang G, Srinivasan S, Colic M, et al. A network of human functional gene interactions from knockout fitness screens in cancer cells. Life Sci Alliance. 2019;2(2):e201800278.
37. Farrell CM, O'Leary NA, Harte RA, Loveland JE, Wilming LG, Wallin C, et al. Current status and new features of the Consensus Coding Sequence database. Nucleic Acids Res. 2014;42(D1):D865–72.
38. Iorio F, Behan FM, Gonçalves E, Bhosle SG, Chen E, Shepherd R, et al. Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. BMC Genomics. 2018;19(1):604.
39. Hart T, Moffat J. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics. 2016;17(1):164.
40. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3amp58 GenesGenomesGenetics. 2017;7(8):2719–27.
41. Dede M, Kim E, Hart T. Biases and blind-spots in genome-wide CRISPR knockout screens. Systems Biology; 2020 [cited 2020 Apr 26]. Available from: http://biorxiv.org/lookup/doi/10.1101/2020.01.16.909606.
42. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569(7757):503–8.
43. Shannon P. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
44. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
45. Dede M, McLaughlin M, Kim E, Hart T. Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens. Python notebooks. figshare; 2020. Available from: https://figshare.com/articles/software/enCas12a_screen_analysis_pipeline/12275642.
46. Dede M, McLaughlin M, Kim E, Hart T. Multiplex enCas12a screens detect functional buffering among paralogs otherwise masked in monogenic Cas9 knockout screens. Sequencing data. Sequence Read Archive; 2020. Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA664967.

Publisher's Note