Title:

Systematic discovery and classification of human cell line essential genes

Authors:

Traver Hart[1,†], Megha Chandrashekhar[1,2,†], Michael Aregger[1,†], Zachary Steinhart[3], Kevin R. Brown[1], Stephane Angers[3,4], and Jason Moffat[1,2,*]

Affiliations:

[1]Donnelly Centre, University of Toronto, Toronto, Canada.

[2]Department of Molecular Genetics, University of Toronto, Toronto, Canada.

[3]Department of Pharmaceutical Sciences, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, Canada.

[4]Department of Biochemistry, University of Toronto, Toronto, Canada.

*Correspondence should be addressed to: j.moffat@utoronto.ca (J.M.)

†These authors contributed equally to this work.

Abstract:

The study of gene essentiality in human cells is crucial for elucidating gene function and holds great potential for finding therapeutic targets for diseases such as cancer. Technological advances in genome editing using clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 systems have set the stage for identifying human cell line core and context-dependent essential genes. However, first generation negative selection screens using CRISPR technology demonstrate extreme variability across different cell lines. To advance the development of the catalogue of human core and context-dependent essential genes, we have developed an optimized, ultracomplex, genome-scale gRNA library of 176,500 guide RNAs targeting 17,661 genes and have applied it to negative and positive selection screens in a human cell line. Using an improved Bayesian analytical approach, we find CRISPR-based screens yield double to triple the number of essential genes than were previously observed using systematic RNA interference, including many genes at moderate expression levels that are largely refractory to RNAi methods. We further characterized four essential genes of unknown significance and found that they all likely exist in protein complexes with other essential genes. For example, RBM48 and ARMC7 are both essential nuclear proteins, strongly interact and are commonly amplified across major cancers. Our findings suggest the CRISPR-Cas9 system fundamentally alters the landscape for systematic reverse genetics in human cells for elucidating gene function, identifying disease genes, and uncovering therapeutic targets.

The Human Genome Project has yielded a fairly complete catalogue of cellular components, and a major goal moving forward will be to systematically discover and classify all genetic elements involved in normal biological processes and disease (*1*). With advances in gene editing technologies enabled by the CRISPR-Cas system (*2-5*), it is no longer quixotic to seek a comprehensive picture of cellular circuitry for human cells. Fortunately, yeast research has helped to establish a conceptual framework for systematic genetics, providing a glimpse of the basic modular organization of a cell and the corresponding genetic landscape (*6*). Much of the basis for advances in yeast systematic genetics was the discovery and classification of essential and non-essential genes (*7, 8*). However, this binary classification was later demonstrated to be an artifact of laboratory conditions: while some genes do indeed seem to be constitutively essential, many others revealed fitness defects only under environmental stress or in different genetic backgrounds (*9, 10*).

The distinction between core and context-dependent essentials is particularly relevant in humans, where a typical cell expresses perhaps two-thirds of the genome's complement of genes (*11*). Identifying a tissue's essential genes, and delineating them from the constitutive essentials shared across all tissues, may hold the key to unlocking tissue-specific disease. In tumors, this is the foundation for the concept of target identification by synthetic lethality (*12*): genes essential in tumor cells but not in adjacent normal tissues should make ideal therapeutic targets with high effectiveness and minimal side effects.

Identifying these context-specific essentials has been no easy feat. To date, this field has been constrained by technology; RNA interference (RNAi) has been the best available tool but, despite improvements in reagent design and analytical approaches, its utility is limited by imperfect mRNA knockdown as well as confounding off-target effects (*13-15*). We previously exploited the concept of core vs. context essentials – the Daisy model of gene essentiality, where each petal represents a tissue or context but all tissues share the core – in a systematic analysis of RNAi and first-generation CRISPR library screens (*16*). Our approach asserts that the degree to which a given screen recapitulates a known core essential set is a measure of its accuracy. One CRISPR screen showed marked improvement over comparable shRNA screens, hinting at the existence of a large pool of human essential genes that may be beyond the reach of RNAi. However, a screen in a second cell line by the same team yielded results only marginally better than random expectation, highlighting the wide variation in CRISPR efficiency and casting doubt on the generalizability of the system.

Generation and testing of an ultracomplex human CRISPR library

We sought to confirm the utility of CRISPR as an improved screening technology to extend the catalogue of both the core and context human essential genes. We designed a genome-scale library of 174,978 guide RNAs (gRNAs) targeting 92,741 coding exons or adjacent splice sites in 17,661 protein-coding genes (see Supplementary Information). Over 70% of the gRNAs (126,900) have no predicted off-target binding site in hg19, even allowing for up to 2 mismatches, with the remainder (48,078) having only a single predicted 2-mismatch off-target site. We also added control sequences targeting non-human genes including EGFP, LacZ, and luciferase, as well as 1,380 gRNAs targeting random loci on chromosome 10, more than half of which are “promiscuous gRNAs” with ≥20 genomic binding sites (range = 20-5x10^5). The total library contains 176,500 gRNAs.

We established clones of HCT116 colorectal cancer cells that constitutively express *Streptococcus pyogenes* Cas9, and confirmed the editing ability of two independent clones (C1 and C2) following lentiviral-based expression of gRNAs targeting PTK7, TP53 and CEP85 (Fig. 1A, S1, S2A). These gRNAs decreased protein expression of target genes as measured by western blot analysis (Fig.1B, S2B). RNA sequencing analysis of the parental HCT116 cell line, as well as the two Cas9-expressing clones, revealed no major gene expression changes (Fig. S3, R=0.99). We then tested whether gRNAs targeting essential genes caused anti-proliferative effects following expression in HCT116-C1 cells (hereafter referred to as “HCT116”). We cloned gRNAs targeting the essential genes PSMD1, EIF3D and PSMB2 into pLCKO from our own library designs, as well as from an independent design (Fig. 1C) (*17*). At 10 days post-puromycin selection, gRNAs targeting all three genes showed moderate to strong anti-proliferative effects relative to the control LacZ gRNA (Fig. 1D; $p<0.05$, Student’s T-test). gRNAs we designed caused anti-proliferative effects comparable to (or better than) other designs (Fig. 1D, see starred gRNAs). Together, these results suggested that our HCT116 cells and gRNA designs were capable of identifying cancer cell proliferation genes.

We then cloned the 177k library into the pLCKO vector at 650-fold representation and confirmed this by deep sequencing. The plasmid library was converted to a lentiviral library pool, which was used to infect HCT116 cells at an MOI~0.3, or an average of one clone per cell (*18*), where each gRNA was represented at 270-fold coverage in the cell population (Fig. 1E).

To evaluate our library, we conducted a straightforward positive selection screen to look for suppressors of a cell cycle block. We performed the positive selection screen in HCT116 cells by applying a thymidine block 6 days after puromycin selection for cells infected with our gRNA library. The thymidine block was maintained for >4 weeks until clonal proliferation and outgrowth was observed (Fig. 1E). Genomic DNA from the outgrown cells was harvested and subjected to high-throughput sequencing, resulting in a strong enrichment of cells expressing gRNAs targeting thymidine kinase 1 (TK1). In fact, 11 out of 12 TK1 gRNAs present in the library were significantly enriched, while no other gene was represented by more than one gRNA (Fig. 1F). In the presence of high concentrations of thymidine, dTTP production is elevated causing an imbalance of dNTPs, which leads to cell cycle arrest at the G1-S transition through the activity of TK1 (Fig. S4 and(*19*)). Although TK1 has been previously identified as a factor in eliciting a thymidine block in mammalian cells (*20-22*), it has not been reported as directly identified in an unbiased genetic suppressor screen.

To identify essential genes, we performed parallel negative selection screens and sampled evolving cell populations every 3 days from day 6 to day 18, or approximately 20 doublings (Fig. 1E). Genomic DNA from each time point was isolated, and gRNA abundance was measured by massively parallel sequencing of the integrated gRNA cassettes in order to monitor the change in abundance of each gRNA between the initial cell population and each of the subsequent 5 time points (see Supplementary Information). gRNAs targeting essential genes are expected to “dropout” from the initial population, while those targeting non-essential genes are expected to be maintained (*18, 23*).

The severity of phenotype that can be expected from off-target effects of a single gRNA is largely unknown. To address this question upfront, we analyzed the 827 gRNA sequences in our ultracomplex library that target random sites on chromosome 10, and that were detected with at least 50 reads at T0. Of these, 479 were designed using the same stringency filters as the gRNAs targeting coding exons; that is, none have predicted off-target binding sites within two mismatches. The remaining 348 gRNAs are “promiscuous gRNAs” with more than 20 or more predicted perfect match off-target binding sites. We measured the fold-change distribution of each set at each time point and observed that specific targeting of random sites on chr10 yields a fold-change distribution similar to that of gRNAs targeting nonessential genes, while promiscuous gRNAs induce a severe shift in fold-change distribution (Fig. S5). Whether this indicates that multiple genetic lesions are highly toxic or simply reflects the relationship between increased off-target sites and increased likelihood of cleaving a critical locus is unknown. Surprisingly, we do not observe a correlation between number of promiscuous target sites and population dropout rate (Fig. S6). Follow-up studies that measure the relationship between the number of predicted genomic target sites and the corresponding fitness defect could specifically assay the robustness of the genome and potentially improve the design of next-generation gRNA libraries by relaxing needlessly strict target specificity constraints.

Analyzing negative selection screens

We adapted the analytical pipeline described previously in Hart *et al* to evaluate the quality of our dropout screen and to classify essential genes (*16*). Using the gold-standard sets of 360 essential and 927 nonessential genes defined in that study, we observed that the fold-change distribution of gRNAs targeting essential genes is significantly shifted relative to those targeting nonessential genes, and that the shift increases with time (Fig. 2A). We then applied an updated version of the algorithm described in (*16*), which implements dynamic boundary selection and 5-fold cross-validation to facilitate general applicability to negative selection screens. The improved version of this algorithm, which can easily be adapted for custom screening purposes, is referred to as the Bayesian Analysis of Gene Essentiality ("BAGEL"). We used BAGEL to calculate a log Bayes Factor (BF) for each gene (see Supplementary Information), with more positive scores representing higher confidence that a given gene's knockout causes a decrease in fitness—for which we use the general term "essential". Unlike most other analytical methods for large-scale gene perturbation studies, BAGEL uses the data from all reagents and all samples in a screen, and therefore provides a ready framework for the integration of experiments with multiple time points. Moreover, the use of gold-standard reference sets enables the unbiased evaluation of screen performance using precision and recall (Fig. 2B).

We compare the performance of the HCT116 screen to a similar screen for essential genes in the A375 melanoma cell line (*24*), which we reanalyzed with the BAGEL pipeline, as well as a genome-scale pooled shRNA dropout screen we performed in HCT116 cells using a library of 78,432 shRNAs targeting 16,056 genes (*25*). Both CRISPR screens captured a large fraction of the reference essential gene set (~30% recall) without misclassifying any reference nonessentials; that is, a false discovery rate of 0%. (Fig. 2B). The shRNA screen has higher overall recall, presumably because the reference sets are derived from repeated observations in shRNA screens (*16, 23, 26*). However, the total number of genes classified as essential is markedly higher in CRISPR screens: 1,716 in HCT116 cells and 1,349 in A375 cells at a BF > 10, a threshold that identifies essential genes at high accuracy, compared to just 683 genes in parental HCT116 cells by shRNA (Fig. 2B).

We further evaluated screen quality by an independent method comparing essential genes to gene expression. Fig. 2C shows the distribution of $\log_2$(FPKM + 0.01) gene expression values (see Supplementary Information for details), with the main peak representing the bulk of active gene expression and the left peak representing unexpressed genes. We rank-ordered genes by expression level and divided them into bins of 500 genes; for each bin we then plotted the mean expression level versus the fraction of genes classified as essential in each screen (Fig. 2C). A background error rate for each screen was then calculated by taking all genes with trace/no expression (logFPKM < -2) and measuring the fraction classified as essential, assuming that these trace/no expression genes are false positives (Fig. 2C, dashed lines). At our threshold of BF>10, all three screens had a comparable background error rate of 1.1-1.6%, though the CRISPR screens clearly identified many more essential genes across all relevant gene expression levels (Fig. 2C). Moreover, our ultracomplex library detected 27% more essential genes than the 65K member GeCKO library used by Shalem *et al*. (*24*), across all expression levels.

The background error rate enables the analysis of how false discovery rates vary with gene expression. The binwise false discovery rates (the background error rate divided by the fraction of essential genes in each bin) for both the HCT116 and A375 CRISPR screens are virtually identical, with genes expressed at logFPKM > 3 having a binwise FDR < 20% (Fig. 2D). This compares very favorably against pooled-library shRNA screens, which achieve a comparable FDR above logFPKM 6.25. When projected onto a density plot of gene expression, the enormous advantage of CRISPR over RNAi is highlighted (Fig. 2D): over 8,100 genes are expressed in the effective CRISPR window, comprising some 70% of the ~11,600 genes with biologically relevant expression levels (logFPKM > 0), compared to just ~1,500 genes (13%) assayable by pooled-library shRNA with comparable accuracy. For CRISPR, the relationship between expression and error rate is likely driven by the fact that virtually all essential genes in a cell are expressed at these moderate to high levels, as there is no *a priori* reason why CRISPR, which targets DNA, should have an expression bias (contrary to RNAi, which targets transcripts).

The high accuracy of CRISPR screens analyzed by BAGEL across a wide range of expression levels is reflected in the concordance between the two CRISPR screens. Performed by different labs using different cell lines, libraries, and experimental designs, and using a fundamentally new technology, the two screens nevertheless detected 662 genes in common, or nearly 50% of the smaller Shalem *et al.* set. Of these genes, over 90% (n=601) are expressed at logFPKM > 3, where this intersection represents a 2.9-fold enrichment over random expectation (P = $1.4x10^{-11}$, hypergeometric test). Nevertheless, each screen captures a large number of unique essentials, and neither screen exceeds 60% recall against the gold-standard reference set (Fig. 2B) at the specified threshold, suggesting a substantial false negative rate and a large population of still-undetected essential genes. Intersection analysis estimates the total population of cell-line essentials at ~2,800 genes, which is comparable to the estimates from the precision-recall analysis for A375 (1,349 genes/50% recall = 2,698 genes) and HCT116 (1,716 genes/59% recall = 2,908 genes). These highly concordant estimates nearly triple the number of essential genes predicted by RNAi and represent a fundamental advance towards comprehensive reverse genetic screens in mammalian cell lines.

While the intersection between the two screens is a measure of their accuracy, the differences between them provide insight into real biological differences. Notably, knockout of mitochondrial genes causes fitness defects in HCT116 colorectal cancer cells but not in A375 melanoma cells (n=30 genes, 3.0-fold enrichment, P-value = 10^{-7}), consistent with the BRAF-driven melanoma cell line's dysfunction in oxidative phosphorylation and addiction to glycolysis (*27*). Conversely, all three members of the CDK-activating kinase (CAK) complex, including CCNH, CDK7, and MNAT1, are essential in A375 but not HCT116, suggesting a unique dependency of melanoma cells on CDK activity and a potential therapeutic window for existing CDK inhibitors..

Characteristics of CRISPR essential genes

Functional genomics in model organisms and in human cell lines has proven that subunits of stable protein complexes tend to share the same loss-of-function phenotype (*28*). We tested the 1,716 HCT116 essentials for enrichment in CORUM and GO protein complexes, identifying 30 largely non-overlapping protein complexes that are highly enriched for essential subunits

(Bonferroni-corrected P-value < 0.05; Fig. 3A). In addition, we compared the number of essential subunits in each complex to the number of subunits in the list of 823 global essentials gleaned from high-quality shRNA screens (*16*). In all cases, CRISPR increased the fraction of subunits classified as essential. Remarkably, over a third of these protein complexes are detected exclusively by CRISPR, where 0 or 1 subunits were detected by RNAi (Fig. 3A). Among the 121 essential subunits of these complexes, nearly 70% (n=83) have moderate mRNA expression (logFPKM from 3 - 6.25) and only 3 have lower expression (Fig. 3B). In most cases, these complexes perform basic biological processes such as DNA replication; for example, the origin recognition complex, the minichromosome maintenance complex, the DNA synthesome, kinetochore proteins, the centromere complex, the RNA polymerase I complex, the RNA polymerase III complex, and the transcription elongation complex (Fig. 3A). Of the 35 high-expression genes detected exclusively by CRISPR, 31 are mitochondrial, consisting of subunits of the mitochondrial ribosome and Complex I of the electron transport chain (Fig. 3A-B).

Beyond enrichment for protein complexes, the CRISPR essentials exhibit genomic and evolutionary characteristics consistent with previous observations of essential genes. More than half of CRISPR essentials are also essential in yeast (Fig. 3C), and more than 60 percent have mouse orthologs whose knockouts cause prenatal lethality (Fig. 3D). These ratios are strikingly similar to those observed when comparing a much smaller set of 231 essential genes in essential complexes derived from RNAi (*16*). In addition, CRISPR essentials also show the same depletion for paralogs as RNAi essentials (Fig. 3E). The similarity of these signatures, despite the fact that the CRISPR set is more than sevenfold larger than the set of subunits of RNAi essential complexes, argues strongly for the accuracy of the set and the much greater sensitivity of the CRISPR screening assay.

Characterizing novel essential genes.

There still remains a significant portion of human genes that are uncharacterized or unstudied, part of the “dark matter” of the genome. We selected a subset of essential but uncharacterized genes from the HCT116 screen for further study. We chose eleven hits which each have less than the median number of Pubmed citations and GO terms for essential genes (Fig. 4A-B and Supplementary Information). The list was winnowed further to genes that are available in the ORFeome v8.1 cDNA collection, that were sequence verified, and that could be expressed in HEK293T cells, resulting in a final list of four genes (ANKRD49, ZNF830, CCDC84, RBM48). Notably, all four of these genes are expressed at moderate levels refractory to RNAi perturbation (Figure 4C). We confirmed that all four genes are essential for cell proliferation in HCT116 cells by independent treatment with single gRNAs. At 10 days post-puromycin selection, gRNAs targeting the genes caused moderate to strong anti-proliferative effects relative to controls (Fig. 4D, p<0.05, Student's T-test).

We aimed to infer the function of each gene based on protein-protein interactions and subcellular localization. We expressed V5-tagged open reading frames (ORFs) of the four genes in HEK293T and HCT116 cells. ANKRD49 showed staining distributed throughout the cell, whereas ZNF830, CCDC84 and RBM48 exhibited exclusive nuclear staining (Fig. 4E and data not shown). These results indicate the subcellular locations where these proteins elicit their essential functions.

Next, we performed immunoprecipitations in HEK293T and HCT116 cells expressing V5- and 3xFLAG-tagged ORFs and analyzed the captured proteins by mass spectrometry (Fig. 4F and Supplementary Information). Interestingly, protein-protein interactions with other essential genes were detected for all ORFs tested, suggesting that these novel essential genes may also be assigned to essential complexes. For example, ANKRD49 was found to associate strongly with FKBPL, an essential gene in HCT116 cells with BF>20, and together likely function as an Hsp90 co-chaperone to promote aspects of protein folding (*29*). ZNF830 and CCDC84 reproducibly pulled down different sets of proteins that are predicted to participate in mRNA splicing (Fig. 4F). ZNF830 was previously found to associate with the XAB2 splicing complex, whereas CCDC84 has not been reported to interact with any splicing factors (*30*). Finally, the RNA-binding motif containing protein RBM48 captured ARMC7 as a prey, which itself was identified as a novel essential gene of unknown function. Interestingly, both RBM48 and ARMC7 are found to be amplified across several cancer tissues and cell lines (Fig. 4G, S7) (*31, 32*). Across all tumors with copy number data in cBioPortal (n=11,545), both genes show an enrichment for copy number amplification and a corresponding depletion for copy loss relative to a pool of nonessential genes. (Fig. 4H). Collectively, these observations point to a novel, stable protein complex with no known function that likely has an RNA-related essential role in proliferating of tumor cells.

Discussion

The adaptation of library RNA interference methods to mammalian cells a decade ago opened the door to large scale screens for synthetic lethal partners of known oncogenes. Though a great deal has been learned, the inherent limitations of RNAi and the research community's failure to rigorously evaluate and standardize both experimental and analytical methods have collectively fallen short of the payoff that systematic reverse genetic screens have provided in model organisms. Now, with the emergence of the CRISPR-Cas9 system as a tool for genetic manipulation, expectations are again justifiably elevated. CRISPR library screens strongly indicate the existence of a population of essential genes perhaps three times larger than that predicted by RNAi, with an as-yet completely unknown breakdown between constitutive and context essentials. This fundamentally alters the landscape of systematic cancer target identification. New context-specific essentials represent potential synthetic lethal partners of known oncogenes, and new constitutive essentials broaden the net for genes whose haploinsufficiency, resulting from copy loss induced by genomic instability, may offer a therapeutic window for inhibition.

Though the potential is enormous, many of the same pitfalls that beset RNAi now lay in the path of widespread CRISPR adoption. Despite rapid progress in characterizing the CRISPR system, the community is largely unaware of what systemic shortcomings may exist - and exist they surely do. We offer an analytical pipeline that can, in principle, be applied to any large-scale library screen. Perhaps more importantly, we describe an evaluation framework that can be used to compare and improve future analytical approaches, and ideally prevent the false starts that have limited RNAi. With a foundation of rigorous analytical methods, the CRISPR-Cas9 system may finally enable systematic genetic research in mammalian cell lines on a scale to rival that achieved in model organisms.

References:

1. E. S. Lander, Initial impact of the sequencing of the human genome. *Nature* **470**, 187-197 (2011); published online EpubFeb 10 (10.1038/nature09792).
2. S. W. Cho, S. Kim, J. M. Kim, J. S. Kim, Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31**, 230-232 (2013); published online EpubMar (10.1038/nbt.2507).
3. L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, F. Zhang, Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013); published online EpubFeb 15 (10.1126/science.1231143).
4. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012); published online EpubAug 17 (10.1126/science.1225829).
5. P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, G. M. Church, RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826 (2013); published online EpubFeb 15 (10.1126/science.1232033).
6. M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, J. Prinz, R. P. St Onge, B. VanderSluis, T. Makhnevych, F. J. Vizeacoumar, S. Alizadeh, S. Bahr, R. L. Brost, Y. Chen, M. Cokol, R. Deshpande, Z. Li, Z. Y. Lin, W. Liang, M. Marback, J. Paw, B. J. San Luis, E. Shuteriqi, A. H. Tong, N. van Dyk, I. M. Wallace, J. A. Whitney, M. T. Weirauch, G. Zhong, H. Zhu, W. A. Houry, M. Brudno, S. Ragibizadeh, B. Papp, C. Pal, F. P. Roth, G. Giaever, C. Nislow, O. G. Troyanskaya, H. Bussey, G. D. Bader, A. C. Gingras, Q. D. Morris, P. M. Kim, C. A. Kaiser, C. L. Myers, B. J. Andrews, C. Boone, The genetic landscape of a cell. *Science* **327**, 425-431 (2010); published online EpubJan 22 (10.1126/science.1180823).
7. G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, M. Johnston, Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-391 (2002); published online EpubJul 25 (10.1038/nature00935).
8. E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R.

Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, R. W. Davis, Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-906 (1999); published online EpubAug 6 (

9. M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, D. Koller, R. B. Altman, R. W. Davis, C. Nislow, G. Giaever, The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362-365 (2008); published online EpubApr 18 (10.1126/science.1150021).
10. A. K. Ramani, T. Chuluunbaatar, A. J. Verster, H. Na, V. Vu, N. Pelte, N. Wannissorn, A. Jiao, A. G. Fraser, The majority of animal genes are required for wild-type fitness. *Cell* **148**, 792-802 (2012); published online EpubFeb 17 (10.1016/j.cell.2012.01.019).
11. T. Hart, H. K. Komori, S. Lamere, K. Podshivalova, D. R. Salomon, Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14**, 778 (2013); published online EpubNov 11 (10.1186/1471-2164-14-778).
12. L. H. Hartwell, P. Szankasi, C. J. Roberts, A. W. Murray, S. H. Friend, Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278**, 1064-1068 (1997); published online EpubNov 7 (
13. C. J. Echeverri, P. A. Beachy, B. Baum, M. Boutros, F. Buchholz, S. K. Chanda, J. Downward, J. Ellenberg, A. G. Fraser, N. Hacohen, W. C. Hahn, A. L. Jackson, A. Kiger, P. S. Linsley, L. Lum, Y. Ma, B. Mathey-Prevot, D. E. Root, D. M. Sabatini, J. Taipale, N. Perrimon, R. Bernards, Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Methods* **3**, 777-779 (2006); published online EpubOct (nmeth1006-777 [pii]10.1038/nmeth1006-777).
14. W. G. Kaelin, Jr., Molecular biology. Use and abuse of RNAi to study mammalian gene function. *Science* **337**, 421-422 (2012); published online EpubJul 27 (10.1126/science.1225787).
15. J. Moffat, J. H. Reiling, D. M. Sabatini, Off-target effects associated with long dsRNAs in Drosophila RNAi screens. *Trends Pharmacol Sci* **28**, 149-151 (2007); published online EpubApr (10.1016/j.tips.2007.02.009).
16. T. Hart, K. R. Brown, F. Sircoulomb, R. Rottapel, J. Moffat, Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol Syst Biol* **10**, 733 (2014)10.15252/msb.20145216).
17. J. G. Doench, E. Hartenian, D. B. Graham, Z. Tothova, M. Hegde, I. Smith, M. Sullender, B. L. Ebert, R. J. Xavier, D. E. Root, Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol*, (2014); published online EpubSep 3 (10.1038/nbt.3026).
18. T. Ketela, L. E. Heisler, K. R. Brown, R. Ammar, D. Kasimer, A. Surendra, E. Ericson, K. Blakely, D. Karamboulas, A. M. Smith, T. Durbic, A. Arnoldo, K. Cheung-Ong, J. L. Koh, S. Gopal, G. S. Cowley, X. Yang, J. K. Grenier, G. Giaever, D. E. Root, J. Moffat, C. Nislow, A comprehensive platform for highly multiplexed mammalian functional genetic screens. *BMC Genomics* **12**, 213 (2011)10.1186/1471-2164-12-213).
19. M. Meuth, The molecular basis of mutations induced by deoxyribonucleoside triphosphate pool imbalances in mammalian cells. *Exp Cell Res* **181**, 305-316 (1989); published online EpubApr (
20. E. S. Arner, S. Eriksson, Mammalian deoxyribonucleoside kinases. *Pharmacology & therapeutics* **67**, 155-186 (1995).

21. P. Reichard, Interactions between deoxyribonucleotide and DNA synthesis. *Annu Rev Biochem* **57**, 349-374 (1988)10.1146/annurev.bi.57.070188.002025).
22. I. Young, G. J. Young, J. S. Wiley, M. B. van der Weyden, Nucleoside transport and cytosine arabinoside (araC) metabolism in human T lymphoblasts resistant to araC, thymidine and 6-methylmercaptopurine riboside. *Eur J Cancer Clin Oncol* **21**, 1077-1082 (1985); published online EpubSep (
23. R. Marcotte, K. R. Brown, F. Suarez, A. Sayad, K. Karamboulas, P. M. Krzyzanowski, F. Sircoulomb, M. Medrano, Y. Fedyshyn, J. L. Koh, D. van Dyk, B. Fedyshyn, M. Luhova, G. C. Brito, F. J. Vizeacoumar, F. S. Vizeacoumar, A. Datti, D. Kasimer, A. Buzina, P. Mero, C. Misquitta, J. Normand, M. Haider, T. Ketela, J. L. Wrana, R. Rottapel, B. G. Neel, J. Moffat, Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov* **2**, 172-189 (2012); published online EpubFeb (10.1158/2159-8290.CD-11-0224).
24. O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, F. Zhang, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014); published online EpubJan 3 (10.1126/science.1247005).
25. F. J. Vizeacoumar, R. Arnold, F. S. Vizeacoumar, M. Chandrashekhar, A. Buzina, J. T. Young, J. H. Kwan, A. Sayad, P. Mero, S. Lawo, H. Tanaka, K. R. Brown, A. Baryshnikova, A. B. Mak, Y. Fedyshyn, Y. Wang, G. C. Brito, D. Kasimer, T. Makhnevych, T. Ketela, A. Datti, M. Babu, A. Emili, L. Pelletier, J. Wrana, Z. Wainberg, P. M. Kim, R. Rottapel, C. A. O'Brien, B. Andrews, C. Boone, J. Moffat, A negative genetic interaction map in isogenic cancer cell lines reveals cancer cell vulnerabilities. *Mol Syst Biol* **9**, 696 (2013)10.1038/msb.2013.54).
26. H. W. Cheung, G. S. Cowley, B. A. Weir, J. S. Boehm, S. Rusin, J. A. Scott, A. East, L. D. Ali, P. H. Lizotte, T. C. Wong, G. Jiang, J. Hsiao, C. H. Mermel, G. Getz, J. Barretina, S. Gopal, P. Tamayo, J. Gould, A. Tsherniak, N. Stransky, B. Luo, Y. Ren, R. Drapkin, S. N. Bhatia, J. P. Mesirov, L. A. Garraway, M. Meyerson, E. S. Lander, D. E. Root, W. C. Hahn, Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A* **108**, 12372-12377 (2011); published online EpubJul 26 (10.1073/pnas.1109363108).
27. A. Hall, K. D. Meyle, M. K. Lange, M. Klima, M. Sanderhoff, C. Dahl, C. Abildgaard, K. Thorup, S. M. Moghimi, P. B. Jensen, J. Bartek, P. Guldberg, C. Christensen, Dysfunctional oxidative phosphorylation makes malignant melanoma cells addicted to glycolysis driven by the (V600E)BRAF oncogene. *Oncotarget* **4**, 584-599 (2013); published online EpubApr (
28. L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray, From molecular to modular cell biology. *Nature* **402**, C47-52 (1999); published online EpubDec 2 (10.1038/35011540).
29. M. Taipale, G. Tucker, J. Peng, I. Krykbaeva, Z. Y. Lin, B. Larsen, H. Choi, B. Berger, A. C. Gingras, S. Lindquist, A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. *Cell* **158**, 434-448 (2014); published online EpubJul 17 (10.1016/j.cell.2014.05.039).
30. I. Kuraoka, S. Ito, T. Wada, M. Hayashida, L. Lee, M. Saijo, Y. Nakatsu, M. Matsumoto, T. Matsunaga, H. Handa, J. Qin, Y. Nakatani, K. Tanaka, Isolation of XAB2 complex involved in pre-mRNA splicing, transcription, and transcription-coupled repair. *J Biol Chem* **283**, 940-950 (2008); published online EpubJan 11 (10.1074/jbc.M706647200).

31. E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, N. Schultz, The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-404 (2012); published online EpubMay (10.1158/2159-8290.CD-12-0095).

32. J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, N. Schultz, Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013); published online EpubApr 2 (10.1126/scisignal.2004088).

Acknowledgments:

We thank all members of the Moffat and Angers labs; Dax Torti and the Donnelly Sequencing Centre for assistance with sequencing. This work was supported by grants to J.M. from the Ontario Research Fund, Ontario Institute for Cancer Research, Canadian Foundation for Innovation, and Canada Research Chairs program. J.M. is a Tier 2 Canada Research Chair and a Senior Fellow at the Canadian Institute for Advanced Research.

Figure Legends:

Figure 1. gRNA library testing and experimental setup of a genome-scale screen for essential genes in HCT116 cells. (A) Validation of the editing efficiency of the CRISPR-Cas9 system. Parental HCT116 cells (WT) and two single-cell derived HCT116 clones stably expressing FLAG-Cas9 (C1 and C2) were infected at low MOI with lentivirus harboring gRNA-expressing cassettes targeting CEP85, PTK7 or TP53. Infected cells were selected with puromycin for 48 hours, and 8 days post-selection Cas9 editing efficiency was monitored by the T7 endonuclease cleavage assay. (B) As in (A) but Cas9 editing efficiency was monitored by western blotting with antibodies against PTK7. Cas9 was detected with anti-FLAG antibodies and β-Actin was used as a loading control. (C) Gene models of essential genes including PSMD1, EIF3D, and PSMB2 showing target locations of library-designed gRNAs as small boxes above each model. The numbers and asterisks indicate the individual gRNAs used in part (D). (D) Cell viability assays in HCT116 cells infected with lentiviral-based gRNA expression cassettes. Target genes and gRNAs are indicated on the horizontal axis, and cell viability as measured by AlamarBlue staining 10 days post-selection relative to the LacZ control is indicated by the vertical axis (n=3). (E) Construction of the lentiviral pLCKO gRNA library and experimental setup for negative and positive selection screens. (F) Positive selection screen for suppressors of thymidine block. Normalized read counts of all gRNAs at T0 are plotted against the mean normalized count for thymidine-treated samples. TK1 gRNAs are circled in red.

Figure 2. Evaluating screen performance. (A) Fold-change distributions from normalized gRNA read counts for gRNAs targeting gold-standard essential (solid) and nonessential (dashed)

genes for each of the indicated time points. (B) Precision-recall curves after applying BAGEL to the HCT116-gRNA, HCT116-shRNA and A375-gRNA screens. (C) Quantile normalized gene expression density estimate for HCT116 and A375 (gray curve; left Y-axis). Genes are rank ordered by expression and binned (n=500). For each bin, mean expression (+/- s.d.) is plotted against the fraction of essential genes (BF>10; right Y-axis; note log scale). Mean %essential of bins with FPKM<-2 (dashed line) estimates background error rate of each screen. (D) Background error rate divided by the %essential is a binwise FDR estimate. Shading highlights CRISPR and RNAi effective regions.

Figure 3. Characteristics of CRISPR essentials. (A) Selected essential protein complexes and detection rates by CRISPR (blue) and RNAi (green). (B) Genes detected by CRISPR but not RNAi, shaded by expression level (Fig. 2D). (C) Fraction of CRISPR essentials whose yeast orthologs are also essential (red, solid), compared to a strict set of “essentials in essential complexes” from RNAi (red border), CRISPR nonessentials (black, solid), and RNAi nonessentials (black border). (D) Fraction of CRISPR essentials whose mouse knockouts are prenatal lethal, as in (C). (E) Fraction of CRISPR essentials with one or more paralogs, as in (C).

Figure 4. Validation of uncharacterized essential genes. Distributions of the number of Pubmed citations (A) or GO terms (B) linked to each gene symbol in Entrez Gene for genes found to be essential by CRISPR (blue), or non-essential (red). Dotted lines indicate median value for each distribution. Position of selected uncharacterized essentials in the distribution is shown. CCDC84 had no associated GO terms. (C) RNAseq expression level of the uncharacterized genes in both HCT116 cells (red) and HEK293T cells (blue). Dotted black lines denote the effective regions for CRISPR (lower line) and RNAi (upper line). (D) Cell viability assays in HCT116 cells infected with lentiviral-based gRNA expression cassettes, as measured by AlamarBlue staining 10 days post-selection (n=3). (E) Immunofluorescence microscopy analysis of 293T cells stably expressing V5-tagged ORFs. Subcellular localization of ORFs was detected with an anti-V5 antibody (green), and nuclei were stained using DAPI (blue). (F) Summary of the protein-interaction network of ANKRD49, ZNF830, CCDC84 and RBM48 detected by AP-MS analysis. V5- or 3xFLAG-tagged ORFs of the indicated proteins were immunoprecipitated from HEK293T and HCT116 cells, and interactors were identified by mass

spectrometry analysis. Bait proteins are indicated by rectangles and prey proteins by circles. The color of the node corresponds to the Bayes Factor essentiality score, and the width of the edges corresponds to the number of times the interaction was detected. Results from 5 independent AP-MS experiments are shown for ANKRD49 and ZNF830, and 4 independent AP-MS experiments for CCDC84 and RBM48. (G) Cross-cancer alteration summary for RBM48. The histogram depicts the alteration frequency of RBM48 across tumor samples and cancer cell lines. The following alteration types are indicated: deletion (purple), amplification (green). This figure is modified from cBioPortal for Cancer Genomics {Cerami, 2012 #3498;Gao, 2013 #3499}. (H) Distribution of copy number across all tumors in cBioportal (n=11,545) for non-essential genes (black) and RBM48 (green).

Supplementary Information:

Materials and Methods, Figures S1-S11, and Tables S1-S6 are not included in this preprint.

FIGURE 1

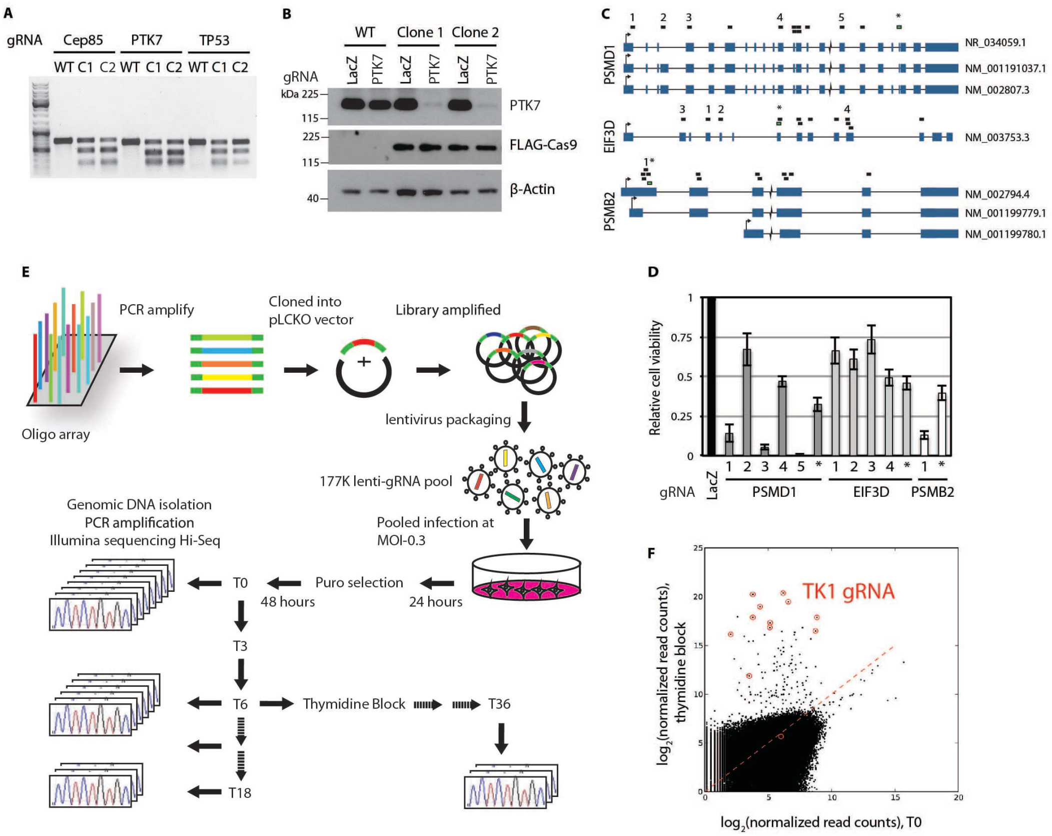

A
gRNA
Cep85
PTK7
TP53
WT C1 C2 WT C1 C2 WT C1 C2
B
WT
Clone 1
Clone 2
gRNA
LacZ
PTK7
LacZ
PTK7
LacZ
PTK7
kDa 225
115
225
115
40
PTK7
FLAG-Cas9
β-Actin
C
1
2
3
4
5
*
PSMD1
NR_034059.1
NM_001191037.1
NM_002807.3
3
1
2
*
4
EIF3D
NM_003753.3
1*
PSMB2
NM_002794.4
NM_001199779.1
NM_001199780.1
E
PCR amplify
Cloned into
pLCKO vector
Library amplified
Oligo array
lentivirus packaging
177K lenti-gRNA pool
Pooled infection at
MOI-0.3
Genomic DNA isolation
PCR amplification
Illumina sequencing Hi-Seq
T0
Puro selection
48 hours
24 hours
T3
T6
Thymidine Block
T36
T18
D
Relative cell viability
1
0.75
0.5
0.25
0
LacZ
1 2 3 4 5 *
1 2 3 4 *
1 *
gRNA
PSMD1
EIF3D
PSMB2
F
TK1 gRNA
log2(normalized read counts), thymidine block
log2(normalized read counts), T0
0
5
10
15
20
25
0
5
10
15
20

FIGURE 2

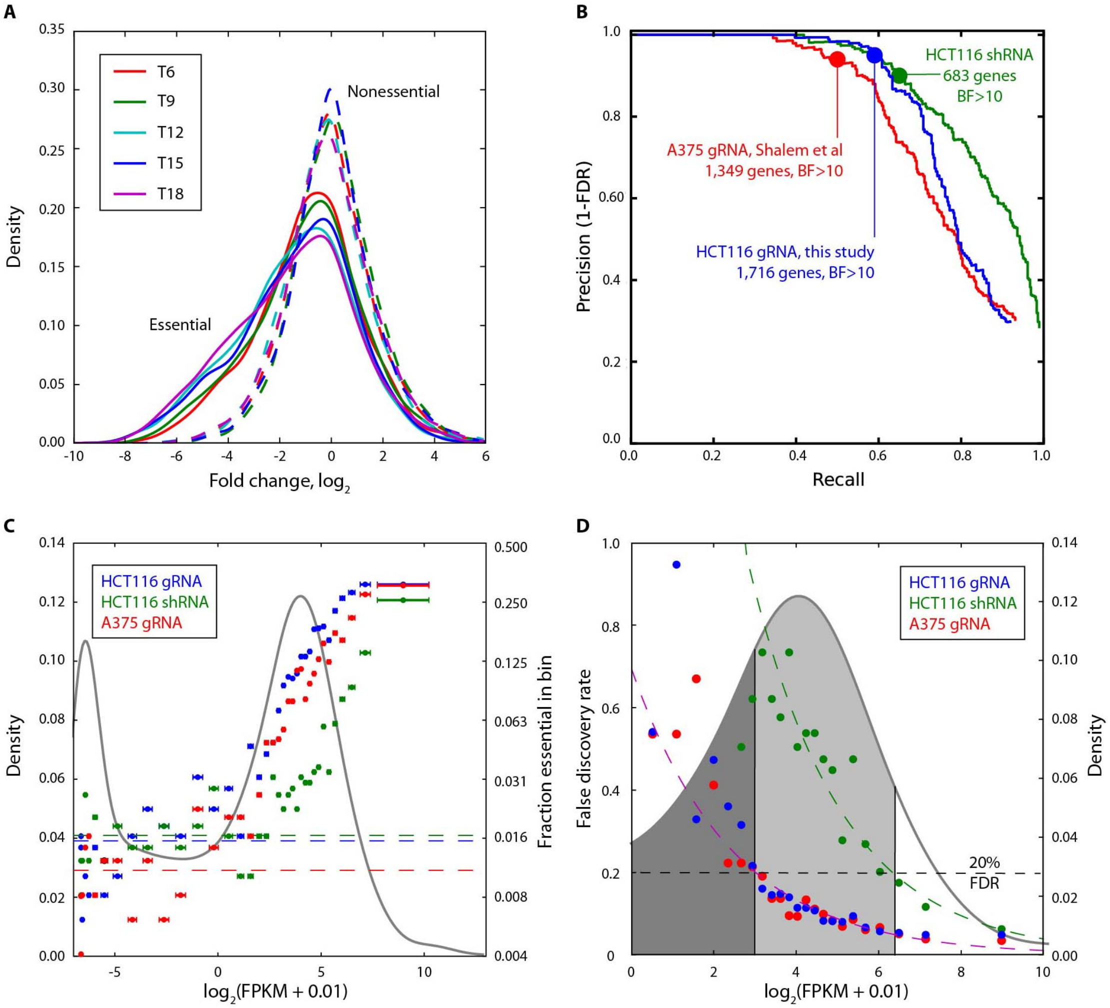

A
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-10
-8
-6
-4
-2
0
2
4
6
Fold change, $\log_2$
T6
T9
T12
T15
T18
Nonessential
Essential
B
Precision (1-FDR)
1.0
0.8
0.60
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall
HCT116 shRNA
683 genes
BF>10
A375 gRNA, Shalem et al
1,349 genes, BF>10
HCT116 gRNA, this study
1,716 genes, BF>10
C
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
-5
0
5
10
$\log_2$(FPKM + 0.01)
Fraction essential in bin
0.500
0.250
0.125
0.063
0.031
0.016
0.008
0.004
HCT116 gRNA
HCT116 shRNA
A375 gRNA
D
False discovery rate
1.0
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
$\log_2$(FPKM + 0.01)
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
HCT116 gRNA
HCT116 shRNA
A375 gRNA
20%
FDR

FIGURE 3

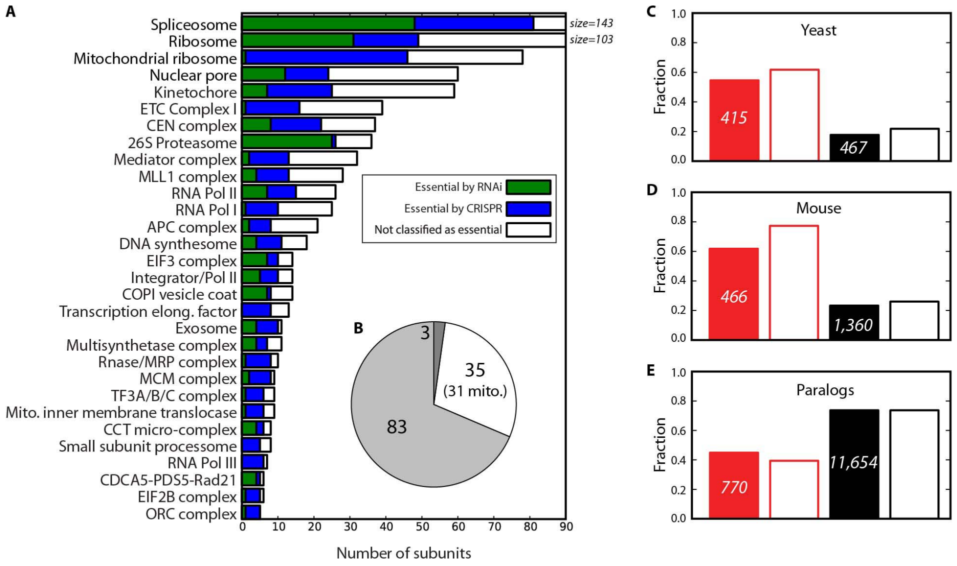
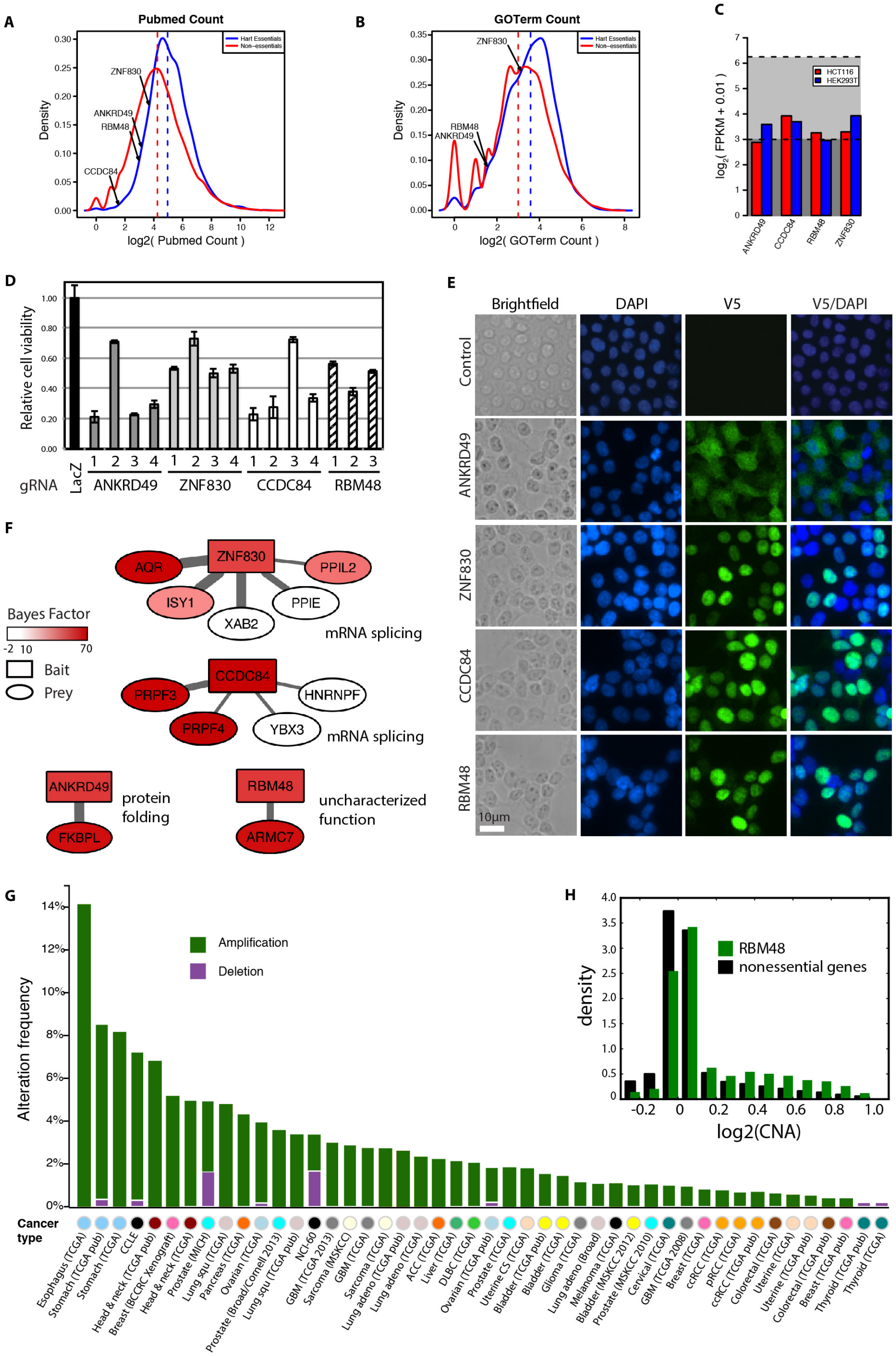

A
Spliceosome
Ribosome
Mitochondrial ribosome
Nuclear pore
Kinetochore
ETC Complex I
CEN complex
26S Proteasome
Mediator complex
MLL1 complex
RNA Pol II
RNA Pol I
APC complex
DNA synthesome
EIF3 complex
Integrator/Pol II
COPI vesicle coat
Transcription elong. factor
Exosome
Multisynthetase complex
Rnase/MRP complex
MCM complex
TF3A/B/C complex
Mito. inner membrane translocase
CCT micro-complex
Small subunit processome
RNA Pol III
CDCA5-PDS5-Rad21
EIF2B complex
ORC complex
size=143
size=103
Essential by RNAi 🟩
Essential by CRISPR 🟦
Not classified as essential □
0
10
20
30
40
50
60
70
80
90
Number of subunits
B
3
35
(31 mito.)
83
C
Yeast
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
415
467
D
Mouse
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
466
1,360
E
Paralogs
Fraction
1.0
0.8
0.6
0.4
0.2
0.0
770
11,654

FIGURE 4


A
Pubmed Count
Hart Essentials
Non−essentials
Density
ZNF830
ANKRD49
RBM48
CCDC84
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
12
log2(Pubmed Count)
B
GOTerm Count
Hart Essentials
Non−essentials
Density
ZNF830
RBM48
ANKRD49
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
2
4
6
8
log2(GOTerm Count)
C
log2(FPKM + 0.01)
HCT116
HEK293T
0
1
2
3
4
5
6
7
ANKRD49
CCDC84
RBM48
ZNF830
D
Relative cell viability
0.00
0.20
0.40
0.60
0.80
1.00
gRNA
LacZ
1 2 3 4
ANKRD49
1 2 3 4
ZNF830
1 2 3 4
CCDC84
1 2 3
RBM48
E
Brightfield
DAPI
V5
V5/DAPI
Control
ANKRD49
ZNF830
CCDC84
RBM48
10μm
F
Bayes Factor
-2 10
70
Bait
Prey
AQR
ZNF830
PPIL2
ISY1
PPIE
XAB2
mRNA splicing
CCDC84
PRPF3
HNRNPF
PRPF4
YBX3
mRNA splicing
ANKRD49
FKBPL
protein folding
RBM48
ARMC7
uncharacterized function
G
Alteration frequency
0%
2%
4%
6%
8%
10%
12%
14%
Amplification
Deletion
Cancer type
Esophagus (TCGA)
Stomach (TCGA pub)
Stomach (TCGA)
CCLE
Head & neck (TCGA pub)
Breast (BCCRC Xenograft)
Head & neck (TCGA)
Prostate (MICH)
Lung squ (TCGA)
Pancreas (TCGA)
Ovarian (TCGA)
Prostate (Broad/Cornell 2013)
Lung squ (TCGA pub)
NCI-60
GBM (TCGA 2013)
Sarcoma (MSKCC)
GBM (TCGA)
Sarcoma (TCGA)
Lung adeno (TCGA pub)
Lung adeno (TCGA)
ACC (TCGA)
Liver (TCGA)
DLBC (TCGA)
Ovarian (TCGA pub)
Prostate (TCGA)
Uterine CS (TCGA)
Bladder (TCGA pub)
Bladder (TCGA)
Glioma (TCGA)
Lung adeno (Broad)
Melanoma (TCGA)
Bladder (MSKCC 2012)
Prostate (MSKCC 2010)
Cervical (TCGA)
GBM (TCGA 2008)
Breast (TCGA)
ccRCC (TCGA)
pRCC (TCGA)
ccRCC (TCGA pub)
Colorectal (TCGA)
Uterine (TCGA)
Uterine (TCGA pub)
Colorectal (TCGA pub)
Breast (TCGA pub)
Thyroid (TCGA pub)
Thyroid (TCGA)
H
density
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
RBM48
nonessential genes
-0.2
0
0.2
0.4
0.6
0.8
1.0
log2(CNA)