Methods and Materials

Cell lines and vectors

Materials were obtained from the following sources: Jiyoye and Raji cells from Dr. Robert Weinberg (Whitehead Institute for Biomedical Research), K562 cells from ATCC, KBM7 cells from Dr. Thijn Brummelkamp (Netherlands Cancer Institute), HEL cells from the Cancer Cell Line Encyclopedia (Broad Institute), pEGFP-C1-Fibrillarin and lentiCRISPR-v1 from Addgene, pGT-GFP, pGT+1-GFP, and pGT+2-GFP were kindly provided by Dr. Thijn Brummelkamp (Netherlands Cancer Institute) and pMXs-IRES-Bsd vector was purchased from Cell Biolabs, Inc. The identities of all cell lines used in this study were authenticated by STR profiling.

Cell culture

All cells were cultured in IMDM (Life Technologies) and supplemented with 20% Inactivated Fetal Calf Serum (Sigma), 5 mM glutamine, and penicillin/streptomycin.

Vector construction

The GFP and FLAG-tagged C16orf80, C3orf17, and C9orf114 expression vectors were constructed by cloning, via Gibson assembly, cDNA inserts generated by PCR from a KBM7 cDNA library into versions of the pMXs-IRES-Bsd vector containing a C-terminal GFP or FLAG tag. For sgDDX3Y rescue experiments, GFP and DDX3X expression vectors were constructed by cloning, via Gibson assembly, cDNA inserts generated by PCR from a GFP-tagged pMXs-IRES-Bsd vector (for GFP) and a KBM7 cDNA library into the pMXs-IRES-Bsd vector (for DDX3X). For live-cell imaging, the EGFP cassette in pEGFP-C1-Fibrillarin was replaced with turboRFP.

Genome-wide lentiviral sgRNA library construction

Oligonucleotides were synthesized on the CustomArray 90K arrays (CustomArray Inc.) as two separate sub-pools. PCR was performed to incorporate overhangs compatible for Gibson Assembly (NEB) into lentiCRISPR-v1 linearized with *Bsm*BI (primer sequences provided below). Gibson Assembly reaction products were transformed into E. cloni 10G SUPREME electrocompetent cells (Lucigen). To preserve the diversity of the library, at least 20-fold coverage in library representation was recovered in each transformation and grown in liquid culture for 16-18 hours.

PCR primers for library amplification
F-GGCTTTATATATCTTGTGGAAAGGACGAAACACCG
R-CTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC

Virus production and transduction

Virus was produced by co-transfecting the transfer vector of interest with VSV-G envelope plasmid and Delta-Vpr (lentiviral) or Gag-Pol (retroviral) packaging plasmids into HEK-293T cells using XTremeGene 9 transfection reagent (Roche). Media was changed 24 hours after transfection and the virus-containing supernatant was collected 72 hours after transfection and passed through a 0.45 μm filter to eliminate cells. Target cells in 6-well tissue culture plates were infected in media containing 8 μg/mL of polybrene by centrifugation at 2,220 RPM for 45 minutes. 24 hours after infection, virus was removed and cells were selected with the appropriate antibiotics.

RNA sequencing

Transcriptomic analysis was performed using the strand-specific RNA sequencing protocol described previously (*37*). Briefly, total RNA was extracted from KBM7, K562, Jiyoye and Raji cells using the RNeasy Mini kit (Qiagen). 5 µg of polyA-selected RNA was fragmented and dephosphorylated after which an ssRNA adapter was then ligated. Reverse transcription (RT) was performed using a primer complementary to the RNA adapter after which a DNA adapter was ligated onto the 3' end of the resulting cDNA product. The library was then PCR amplified, cleaned, quantified using a TapeStation (Agilent) and sequenced on a HiSeq 2500 (Illumina). All primer sequences for this protocol can be found in (*37*).

Reads were aligned against the human genome version hg19 using Tophat version 2.0.13 (*38*). Parameters used were '--transcriptome-only --no-novel-juncs' and '–transcriptome-index' for which we used RefSeq transcript models. Transcript abundances were quantified using Cufflinks version 2.2.1 (*39*).

Comparative essentiality testing

To compare human gene essentiality with yeast gene essentiality, as assessed by knockout viability (*1, 10*), 1-to-1 human-yeast homologs mappings were obtained from the Ensembl Gene release 79 database (*40*). To obtain gene scores for the RNAi dataset, we averaged hairpin scores across 216 cell lines and used the mean hairpin score as the gene score (*11*). For FPKM (Fragments Per Kilobase of exon per Million fragments mapped) values from the RNA sequencing experiment in KBM7 cells were used. Human genes common to the gene-trap, CRISPR, RNAi and KBM7 RNA-seq datasets were used in the essentiality analysis to control for any biases in mapping. Each dataset was ranked by their respective scores and used to predict the essentiality of yeast homologs. The sensitivity and specificity of these predictions were analyzed using receiver operator characteristic curves. Similar analyses were performed to assess the coverage of the library through down-sampling the number of sgRNAs per gene and the performance of the optimized library compared to previous studies using unoptimized libraries.

Features of essential genes

To understand the relationship between gene essentiality and various gene properties, CS was intersected with several datasets described in detail below. To visualize the data, genes were ranked by CS and median-binned into approximately 200 bins. To assess phyletic retention across species, the number of homologs was determined for each gene using the Homologene database release 68 (*41*). To assess sequence divergence between closely related species, the dN/dS ratio was calculated for all 1-1 mouse-human homologs identified in Ensembl Gene relase 79 (*42*). To assess the evolutionary constraint on human genes, essentiality scores were compared between genes with and without stop-gain alleles with an observed frequency above 0.05% and an average sample depth > 30x using data from the Exome Variant Server 6500 (*43*). To assess protein network connectivity the number of protein-protein interactions identified in the Biogrid database release 3.3.124 was used (*44*). To assess gene expression, RNA-sequencing was performed on the KBM7 cell lines and the FPKM values were calculated and used for the analysis. To assess genetic redundancy, essentiality scores were compared between genes with and without recognizable paralogs in the TreeFam database release 9 (*45*). Finally, genes of unknown function were defined as those that were not annotated in the Reactome database and contained no unique (i.e. gene-specific) Entrez Gene GeneRIF.

Gene set enrichment analysis (GSEA)

To identify pathways over- or underrepresented for essential genes, GSEA was performed using the KBM7 CS as a ranked gene list. To identify pathways that were highly expressed in Raji cells, GSEA was performed comparing the FPKM values from RNA-sequencing experiments performed in Raji cells with those from the other three lines. Both analyses were performed using 1,000 permutations and the C2 curated gene sets and the C5 GO gene sets.

Western blotting

Cells were lysed directly in Laemmeli sample buffer, sonicated, separated on a NuPAGE Novex 16% Tris-Glycine gel, and transferred to a polyvinylidene difluoride membrane (Millipore). Immunoblots were processed according to standard procedures, using primary antibodies directed to S6K1 (CST) and gamma-H2AX (Ser139), clone JBW301 (Millipore) and analyzed using enhanced chemiluminescence with HRP-conjugated anti-mouse and anti-rabbit secondary antibodies (Santa Cruz Biotechnology).

Short-term proliferation assay

Individual sgRNA constructs targeting *C16orf80*, *C3orf17*, *C9orf114*, *DDX3Y*, the non-genic region of the *BCR-ABL* amplicon in K562 cells, and the non-genic region of the *JAK2* amplicon in HEL cells were cloned into lentiCRISPR-v1 (sequences provided below). Lentivirus was produced and target cells were transduced and selected as described above for the screens. For the sgDDX3Y experiments, target Raji cells stably expressing GFP and DDX3X were first generated via retroviral transduction.

ATP-based measurements of cellular proliferation were performed by plating 2,000 cells per well, biologically replicated six times, in 96-well plates. After 1-5 hours (for the initial time point), 50 µl of Cell Titer-Glo reagent (Promega) were added to each well, mixed for 5 minutes, after which the luminescence was measured on the SpectraMax M5 Luminometer (Molecular Devices). At the indicated times, the same procedure was performed. For all samples, after removal of the highest and lowest outliers of the six measurements, the fold change in luminescence relative to the initial sample was computed.

sgC16orf80-1: CGATGCTGTAGAGGATGGAG
sgC16orf80-2: TGTCTGAGAAGTAAACCCGT

sgC3orf17-1: GTGTGAGAATCCCTAAGGCG
sgC3orf17-2: GGGCCAAATGGGGTTTGTGG

sgC9orf114-1: GCGGCAGAGAAGGAGGACCG
sgC9orf114-2: CAGGCGGGCTCACCTCCGTG

sgDDX3Y-1: GCAGTTTAGCGATATTGACA
sgDDX3Y-2: TCTTGTTGGGGCTAAAACCA

sgAmplicon-1 (BCR-ABL): GGATGACAGATGATGGATGG
sgAmplicon-2 (BCR-ABL): GGCTCCCTTCAAGTGGGATG

sgAmplicon-1 (JAK2): GGTTTAATGGAAGAGAAGGG
sgAmplicon-2 (JAK2): GAGGCATATTCTTCTCCTGG

sgAAVS1: GGGGCCACTAGGGACAGGAT (AAVS1-targeting control)

sgCTRL: GGATACTTCTTCGAACGTTT (non-targeting control)

Co-expression analysis

C16orf80, *C3orf17*, and *C9orf114* expression levels across all CCLE cell lines, as assessed by microarray analysis, was compared with gene expression levels across the CCLE of all other genes in a pairwise fashion to obtain a Pearson correlation coefficient for each gene as a measure of the degree of co-expression (*46*). The top 200 ranked genes that were not on the same chromosome as the analyzed gene (to filter out highly co-expressing neighboring genes) were then analyzed via DAVID to identify functional categories that were associated with genes concordantly expressed with *C16orf80*, *C3orf17*, and *C9orf114* (*47*).

Live-cell imaging

100,000 HEK-293T cells stably expressing C-terminal GFP-tagged C16orf80, C3orf17, and C9orf114, generated via retroviral transduction, were seeded on fibronectin-coated glass-bottom 35 mm dishes (MatTek Corporation). The next day, cells were transfected using XtremeGene9 transfection reagent (Roche) with 1.2μg of pTURBORFP-C1-Fibrillarin. 30 minutes before imaging, cells were stained with Hoechst 33342 (Invitrogen) and imaged on a spinning disk confocal microscope (Perkin Elmer) with a 488 nm and a 568 nm laser through a 63X objective.

Immunoprecipitation and mass spectrometry data analysis

KBM7 lines stably expressing C-terminal FLAG-tagged C16orf80, C3orf17, and C9orf114 were generated by viral transduction and FLAG-immunoprecipitated for mass spectrometric analysis. Briefly, 1 billion C16orf80-FLAG, C3orf17-FLAG, C9orf114-FLAG, and wild-type KBM7 cells were pelleted and rinsed twice with ice-cold PBS and lysed in 1mL lysis buffer (1% NP40, 50mM Tris-HCl pH 7.5 150mM NaCl, 1 tablet of EDTA-free protease inhibitor (Roche; per 25 ml buffer)). The soluble fractions of the cell lysates were isolated by centrifugation at 13,000 rpm in a microcentrifuge for 10 minutes. The Anti-FLAG-M2 magnetic beads were washed with lysis buffer three times. Subsequently, 50 µl was added to cleared cell lysates and incubated with rotation overnight at 4°C. Next, the beads were washed twice with lysis buffer and twice in wash buffer (50mM Tris-HCl pH 7.5 150mM NaCl, 1 tablet of EDTA-free protease inhibitor (Roche; per 25 ml buffer)). The resulting samples were processed and analyzed by mass spectrometry using iTRAQ as previously described (*48*).

We assessed enrichment as previously described (*49*). Briefly, for each peptide, a $\log_2$ fold change was calculated between its intensity in the ORF-expressing sample compared to the wild-type control pull-down. The peptides were mapped to the Uniprot database and, for each gene, the median fold-change value was used. Enrichment scores were then calculated by subtracting the median of the entire distribution of the $\log_2$ transformed values across all genes to center the fold change distribution.

Characterization of DDX3X

Genomic DNA was extracted from Jiyoye and Raji cells and amplified using primers directed against the exon-intron 8 boundary (primer sequences provided below). PCR amplicons were purified and analyzed by Sanger sequencing (Eton Bioscience Inc.).

cDNA libraries were prepared from RNA from Raji, Jiyoye, KBM7 and K562 cells and amplified with primers targeting exons 8 and 9 of DDX3X (primer sequences provided below). PCR amplicons were then analyzed by gel electrophoresis.

Genomic DNA primers for DDX3X
F-CTTGCGTGGTTTATGGTGGTG
R-GCCCATCCTAGTTGACTGTCC

cDNA primers for DDX3X
F-GGTATTAGCACCAACGAGAGAGT
R-GCCAACTCTTCCTACAGCCAA

Supplementary Text

Note S1. CRISPR/Cas9-based screen.

sgRNA Library Design

The optimized sgRNA library was designed with the same specificity requirements as previously described (*4*). Briefly, candidate sgRNAs were first filtered for potential off-target matches (for duplicated or highly homologous genes and gene families with less than 5 members, this requirement was relaxed to allow for sgRNAs that targeted multiple homologs).

In contrast to previous collections, the sgRNAs in this library were designed for high cleavage activity. Efficient sgRNAs are critical for essentiality screens because an sgRNA can only reliably assess essentiality if it cleaves and inactivates its gene target in a large proportion of the cells expressing it. To identify rules governing target cleavage efficiency, a support-vector-machine classifier was constructed using the target sequences (encoded by 80 binary features for positions 1-20 and nucleotides A, C, G, and T) to predict depletion scores of ribosomal protein-targeting sgRNAs from a previous pooled proliferation screen described in (*4*). As these sgRNAs are all expected to be essential, differences in their depletion levels reflected differences in cleavage efficiency. Using the SVM classifier, new sgRNA candidate sequences were ranked in order of their predicted cleavage efficacy and the best 4-10 (mode=10) candidates for each gene were selected for synthesis. In total, we constructed a novel library, containing 178,896 sgRNAs targeting 18,166 protein-coding genes in the human consensus CDS (CCDS) and 1,004 non-targeting control sgRNAs

Screen procedure

The Cas9-expressing sgRNA library lentivirus was produced in HEK-293T cells as described above. In all screens, 240 million target cells were transduced with the viral pool to achieve an average 1000-fold coverage of the library after selection. After 72 hours, cells were selected with puromycin and an initial pool of 80 million cells was harvested for genomic DNA extraction. The remaining cells were passaged every 3 days, and after 14 doublings, 80 million cells were harvested for genomic DNA extraction.

Screen deconvolution

sgRNA inserts were PCR amplified from 50-75 million genome equivalents of DNA from each initial and final sample, achieving an average coverage of ~275-400x of the sgRNA library. The resultant PCR products were purified and sequenced on a HiSeq 2500 (Illumina) (primer sequences provided below) to monitor the change in the abundance of each sgRNA between the initial and final cell populations. sgRNAs targeting essential genes are expected to be depleted from the population, while those targeting dispensable genes should be maintained. The results for two neighboring genes, *RPL14*, an essential ribosomal protein gene, and *ZNF619*, a dispensable gene encoding a zinc finger protein, illustrate the expected pattern (Fig. 1).

Primer sequences for sgRNA quantification
F-AATGATACGGCGACCACCGAGATCTAGAATACTGCCATTTGTCTCAAG
R-CAAGCAGAAGACGGCATACGAGATCnnnnnnTTTCTTGGGTAGTTTGCAGTTTT
(nnnnn denotes the sample barcode)

Illumina sequencing primer
CGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC

Illumina indexing primer
TTTCAAGTTACGGTAAGCATATGATAGTCCATTTTAAAACATAATTTTAAAACTGCAAACTACCCAAGAAA

Data analysis

Sequencing reads were aligned to the sgRNA library and the abundance of each sgRNA was calculated. The sgRNA counts from the initial populations of all four cells lines were combined to generate an initial reference set. sgRNAs with less than 400 counts in this set were removed from downstream analyses. The $\log_2$ fold change in abundance of each sgRNA was calculated for final population samples for each of the cell lines after adding a count of one as a pseudocount. Gene-based CRISPR scores (CS) were defined as the average $\log_2$ fold change of all sgRNAs targeting a given gene and calculated for all screens. The CS reported for the KBM7 cell line was the average of two independent replicate experiments.

To identify genes essential for optimal proliferation under standard media conditions, the $\log_2$ fold change distribution for all sgRNAs targeting a given gene was compared with the entire distribution using a Kolmogorov-Smirnov test using the ks_2samp function from the scipy.stats Python library. The resulting p-values were corrected using the Benjamini-Hochberg procedure. Genes with a CS < -0.1 and corrected $p< 0.05$ in the KBM7 cell line were defined as cell-essential in downstream analyses.

To identify cell line-specific essential genes, the CS distribution of each line was mean-normalized to zero. For each gene in each line, the CS in the given line was subtracted by the minimum CS in the other three lines to define a cell line-specific essentiality score (negative values indicate cell line specificity). For each line, genes with a differential score less than -1.5 (~4 standard deviations from the mean score) whose minimum CS in the other three lines was greater than -1 were defined as cell line-specific genes.

To identify cancer type-specific essential genes, the $\log_2$ sgRNA fold change distribution of each line was mean-normalized to zero. For each sgRNA, a differential essentiality score was then defined as the average $\log_2$ fold change in the two CML lines subtracted by the average $\log_2$ fold change in the two Burkitt's lymphoma lines (with positive and negative value representing Burkitt's lymphoma- and CML-specific essentiality, respectively). The differential essentiality score distribution for all sgRNAs targeting a given gene was compared with the entire distribution using a Kolmogorov-Smirnov test using the ks_2samp function from the scipy.stats Python library. The resulting p-values were corrected using the Benjamini-Hochberg procedure. Candidate genes with corrected $p < 0.05$ were then filtered by additional CS-based criteria after mean-normalization of the CS: (i) the CS must perfectly segregate between cell lines of the two cancer types (ii) CS must be less than -1 in both lines of the given cancer type (iii) the gene must not be essential (CS less than -1) in either cell line of the other cancer type (iv) the average difference in CS between cell lines of the two cancer types must be greater than 1. Genes fulfilling all of these criteria were designated as cancer-type specific. Identical criteria were applied to identify "set-selective" genes for the permuted sample groupings.

Note S2. Gene-trap-based screen.

Screen procedure

Gene-trap retrovirus was produced in HEK-293T cells as previously described (*7*). 100 million KBM7 cells were infected and, after 3 days, an initial pool of 100 million cells was harvested for genomic DNA extraction. The remaining cells were passaged every 3 days, and after 14 population doublings, 100 million cells were harvested for genomic DNA extraction.

Screen deconvolution

To date, haploid genetic screens have been limited to probing phenotypes amenable to positive selection. In positive selection screens, candidate genes are enriched for disruptive insertions and can be readily detected in the surviving mutant population by using inverse PCR. However, this protocol is unsuited for negative selection screening, in which inserts in the genes of interest are underrepresented. Identification of these genes requires a highly accurate and efficient method for measuring the presence of insertion sites in all genes. Toward this end, we developed such a protocol using 'splinkerette'-based PCR, which enables efficient amplification of DNA fragments with only one known end, coupled with massively parallel sequencing to map genomic regions proximal to each gene-trap insertion sites (*50*).

Briefly, 100 million genome equivalents of DNA were digested with *Nla*III and/or *Mse*I to produce sticky ends to which double stranded 'splinkerette' adapters were ligated (adapter sequences provided below). Two rounds of nested PCR were performed to generate an insert junction library. We sequenced the resultant library on a HiSeq 2000 (Illumina) (primer sequences provided below) and, after aligning the reads to the reference genome, tallied the number and orientation of unique integration events in each gene. Essential genes are expected to contain fewer inactivating insertions (i.e., in the 'sense' orientation) relative to the number of 'harmless' inserts (i.e., in the 'anti-sense' orientation), whereas non-essential genes are expected to show no such bias. The results for the neighboring genes *RPL14* and *ZNF619* show the expected pattern (Fig. 1).

Splinkerette adapters
*Mse*I adapters
F-CGCGAACAACGCTAACGACGCGAACGACAGC
R-TAGCTGTCGTTCGCGTCGTAAAAAAACTTTTTTT

*Nla*III adapters
F-CGCGAACAACGCTAACGACGCGAACGACAGCCATG
R-GCTGTCGTTCGCGTCGTAAAAAAACTTTTTTT

Primer sequences for insertion quantification
Outer primers
F-CGAGTCCACGATTCGGATGCAA
R-CGCGAACAACGCTAACGACG
Inner primers
F-AATGATACGGCGACCACCGAGATCTACACGAAACATCTGATGGTTCTCTAGCTTGCC
R-CAAGCAGAAGACGGCATACGAGATCnnnnnnCACTTACCGCTAACGACGCGAACGACAG
(nnnnn denotes the sample barcode)

Illumina sequencing primer
CTAGCTTGCCAAACCTACAGGTGGGGTCTTTCA

Illumina indexing primer
GCTGTCGTTCGCGTCGTTAGCGGTAAGTG

Data analysis

Reads were de-duplicated (as only unique insertions were counted) then mapped to the reference human genome and intersected with RefSeq transcript models. For each transcript, the number of unique sense and anti-sense insertions in all intronic regions was tallied. For genes with multiple transcript models, the fraction of inactivating inserts was calculated for each transcript and the transcript with the minimum value was defined as the gene-trap score (GTS).

The accuracy of the GTS for a given gene depends on the total number of insertions observed. Therefore, we determined a minimum number (n) of anti-sense inserts in a gene required for inclusion in downstream analyses by assessing the concordance between replicate experiments. For this test, genes shared between replicates 1 and 2 were compared. For increasing n, the correlation between gene-trap replicates increased, reaching a plateau at ~0.89 at n=65, which we used as a cutoff for subsequent GTS analyses (Fig. S1C-D). In the final dataset, we combined insertional data from 3 independent replicates of the final population at the transcript level, calculated the GTS for each gene, and removed all genes with less than n=65 anti-sense insertions.

Note S3. Analysis of cell-essential genes in the KBM7 gene-trap screen.

The gene-trap method allows for a principled estimate of the number of genes essential for optimal proliferation by comparing the GTS distributions between the initial and final populations. An excess of low GTS genes on the haploid chromosomes (all except 8) in the final population indicated that approximately 1,142 of the 6,694 genes with adequate insertional data (*n*=65 anti-sense inserts) are cell-essential (Fig. 2B). Notably, the proportion of genes that scored as cell-essential by the gene-trap method (17%) is an overestimate of the proportion of essential genes in the genome as a whole. The retrovirus utilized in these screens exhibits a substantial preference for integrating into regions of active transcription (*51*). As a result, silent or lowly-expressed genes (that are unlikely to be essential) are not targeted and therefore excluded from analysis.

We also compared our results with those from a co-published study by Blomen et al. in which a similar proliferation-based screen was performed in KBM7 cells using gene-trap mutagenesis. Despite two major methodological differences between our experiments (namely, the use of different splice acceptor sequences in the gene-trap vectors and an additional purification for haploid cells prior to mutagenesis implemented by Blomen et al.), the results were highly concordant.

For the overlap analysis, the set of genes on the haploid chromosomes where *n*=65 in both datasets was considered (6,285 genes). For this common set of genes, the GTS between the two studies was well-correlated (r=0.81) (Fig. S1E). Furthermore, 1,039 of the top 1,250 genes, as ranked by the GTS ratio score, from both data sets (83%) were overlapping (Fig. S1F). The high level of agreement reached between these two experiments performed in different labs with different subclones of KBM7 cells using different reagents (in addition to the aforementioned differences in experimental protocols) provides a strong demonstration of the robustness of this method and the validity of our results.

Note S4. Paralogous gene expression may underlie Jiyoye-specific essential genes.

We identified several additional instances of cell line-specific essentiality that could be attributed to paralogous gene pairs. For example, the cyclin-dependent kinase *CDK6* was specifically essential in Jiyoye cells, in which its paralog, *CDK4*, is specifically not expressed (Fig. S4C). Similar patterns were observed in this cell line for two other pairs of paralogous genes, *HK1*/*HK2* and *SLC2A1*/*SLC2A3* (also known as GLUT1/GLUT3), which are both involved in glucose metabolism (Fig. S4C). To assess the essentiality of functions supplied by a set of paralogous genes, it may be useful to design libraries containing multiple sgRNAs to simultaneously inactivate all members of the set.

Note S5. Cell line-specific essential functions in Raji and KBM7 cells.

In Raji cells, two of the three subunits comprising the heterotrimeric IκB kinase complex, a positive regulator of the NF-κB pathway, scored strongly (with the third subunit nearing our selectivity threshold) indicating the importance of this pathway in this cell line (Fig. S4E). Consistent with this hypothesis, we found that Raji cells show a distinctive gene-expression signature of NF-κB pathway activation that may underlie their unique dependence (Fig. S4F).

In KBM7 cells, we found several sets of cell line-specific essential genes that encode physically interacting or functionally related gene products, including *RAD51B*/*RAD51D*/*FANCM*/*RTEL1*, *MCL1*/*BCL2*, *RUNX1*/*CBFB*, *LIPT2*/*LIAS*, and *SEPHS2*/*SEPSECS*; the biological bases for the selective essentiality of these gene sets remain to be defined, although it is tempting to speculate whether the importance of the various DNA-repair components is related to the unique haploid karyotype of this line (Table S5).

Figures

Fig. S1

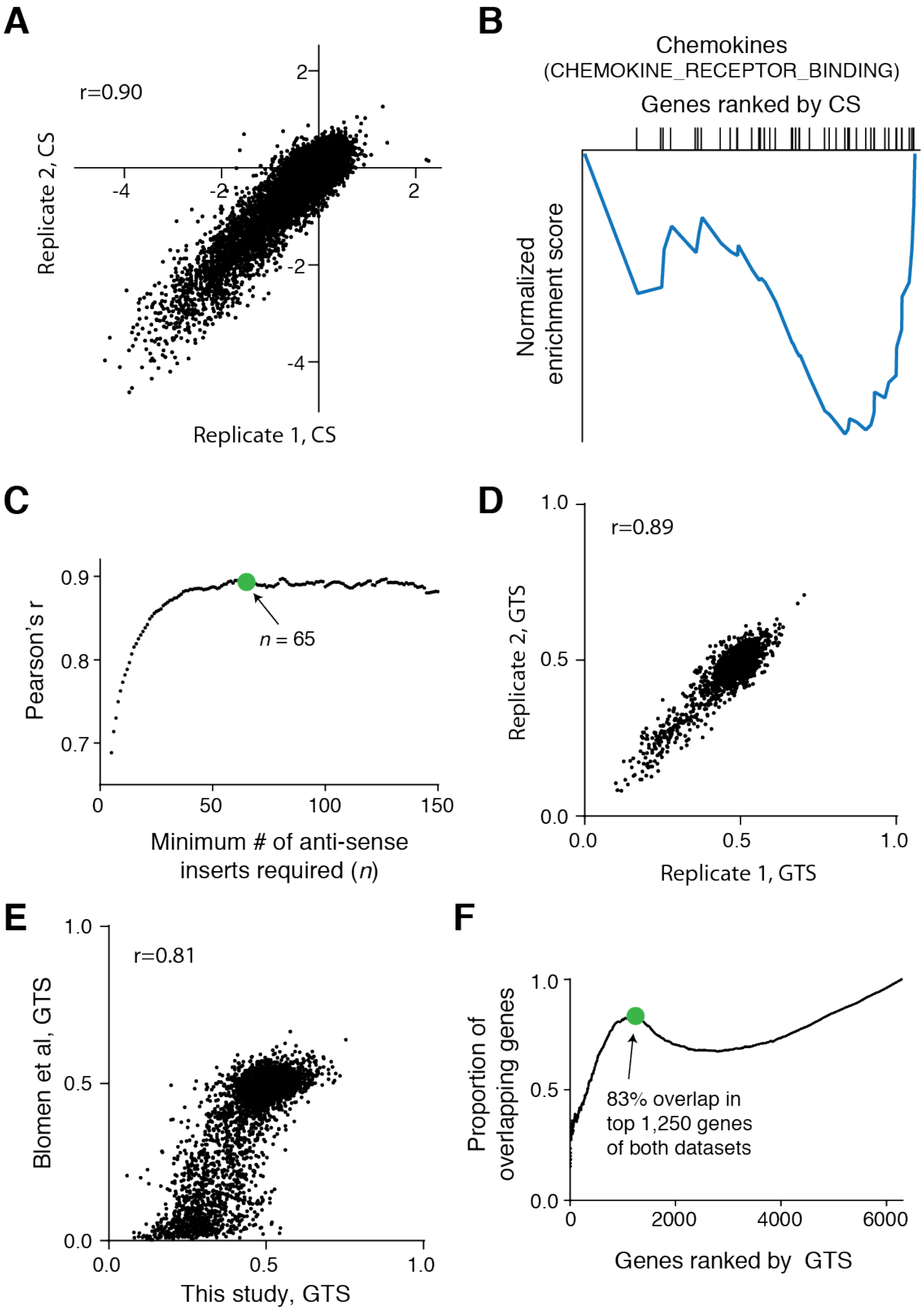


Fig. S1. Replicate screening experiments are well correlated. (**A**) CS from replicate CRISPR/Cas9 screens in KBM7 cells. (**B**) Gene-set enrichment analysis (GSEA). Chemokines are underrepresented among essential genes in the pooled screens. (**C**) Correlation between gene-trap replicates at various cutoffs for the required minimum number of anti-sense inserts observed in a gene, *n*. (**D**) GTS from replicate gene-trap screens for *n*=65. **(E)** Correlation between full gene-trap dataset from our study and co-published study for all genes on the haploid chromosomes where *n*=65 in both datasets (6,285 genes). **(F)** Overlap analysis. The set of genes in (E) was ranked by the GTS in both datasets and the proportion of overlapping genes between the top X genes of both datasets, for all X between 1 and 6,285, was determined. Of the top 1,250 genes in both datasets, 1,039 genes (83%) were overlapping, indicating the high concordance between these two studies.

Fig. S2

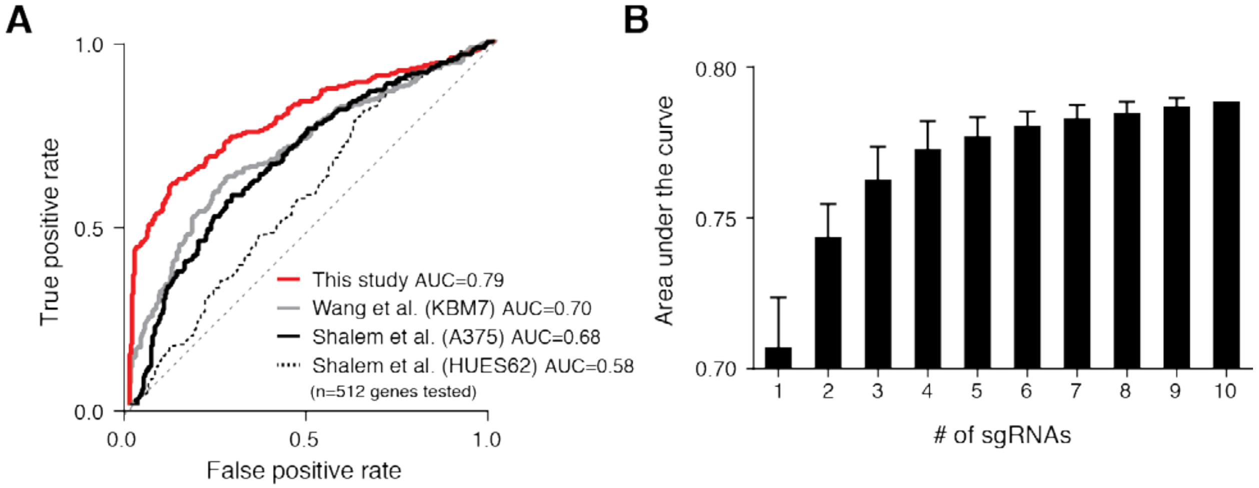


Fig. S2. sgRNA library characterization. (**A**) Gene scores from genes common to all three studies were used to predict the essentiality of yeast homologs. Receiver operator characteristic (ROC) curve analysis revealed a substantial improvement with the optimized library. (**B**) Area under the curve (AUC) values from ROC analysis using gene scores from down-sampling experiments. Error bars denote SD. (n=25 samplings).

Fig. S3

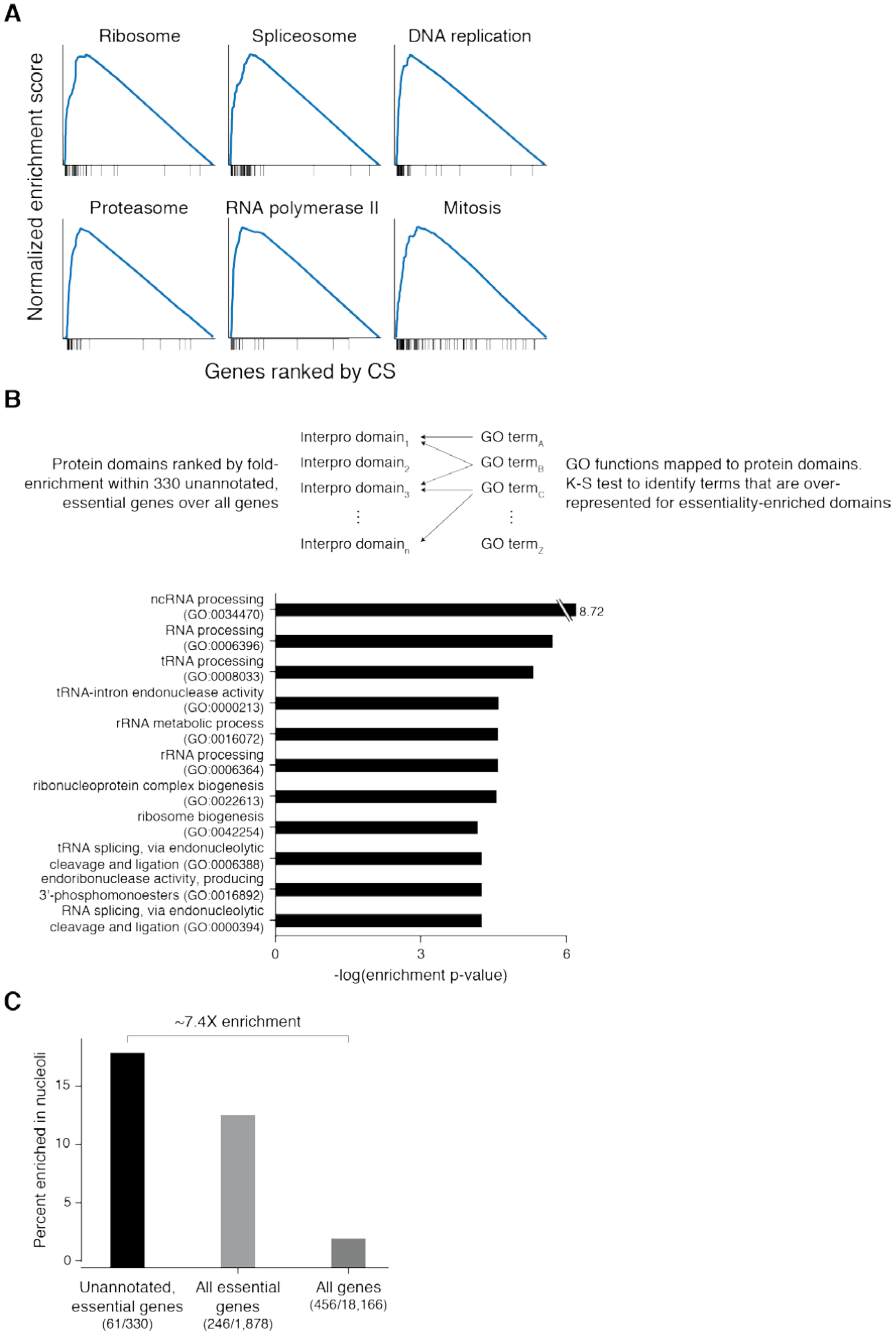
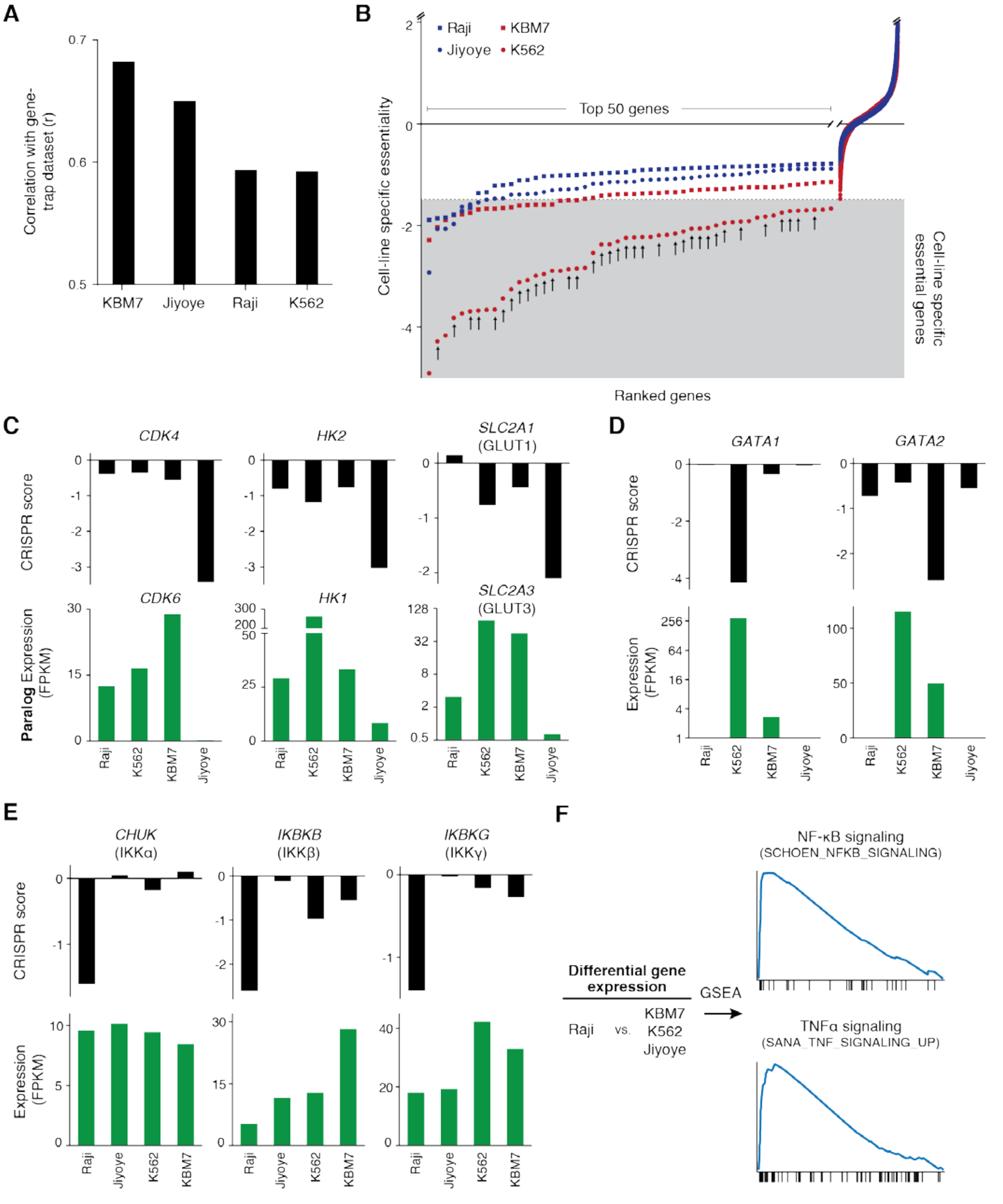

A
Ribosome
Spliceosome
DNA replication
Proteasome
RNA polymerase II
Mitosis
Normalized enrichment score
Genes ranked by CS
B
Interpro domain$_1$
Interpro domain$_2$
Interpro domain$_3$
⋮
Interpro domain$_n$
GO term$_A$
GO term$_B$
GO term$_C$
⋮
GO term$_Z$
Protein domains ranked by fold-enrichment within 330 unannotated, essential genes over all genes
GO functions mapped to protein domains. K-S test to identify terms that are over-represented for essentiality-enriched domains
ncRNA processing (GO:0034470)
8.72
RNA processing (GO:0006396)
tRNA processing (GO:0008033)
tRNA-intron endonuclease activity (GO:0000213)
rRNA metabolic process (GO:0016072)
rRNA processing (GO:0006364)
ribonucleoprotein complex biogenesis (GO:0022613)
ribosome biogenesis (GO:0042254)
tRNA splicing, via endonucleolytic cleavage and ligation (GO:0006388)
endoribonuclease activity, producing 3'-phosphomonoesters (GO:0016892)
RNA splicing, via endonucleolytic cleavage and ligation (GO:0000394)
0
3
6
-log(enrichment p-value)
C
~7.4X enrichment
Percent enriched in nucleoli
0
5
10
15
Unannotated, essential genes (61/330)
All essential genes (246/1,878)
All genes (456/18,166)

Fig. S3. Characteristics of unannotated essential genes. (A) Cell-essential genes are involved in fundamental biological processes. GSEA was performed on genes ranked by CS. (**B**) GO term analysis. Interpro domains were ranked by their fold-enrichment within the unannotated essential genes compared to all genes and mapped to GO terms. A K-S test was performed for each GO term to identify terms over-represented for high-ranking (i.e. unannotated essential gene-enriched) domains. (**C**) Comparisons with the nucleolar proteomic dataset from (*21*) revealed a substantial enrichment of nucleolar gene products encoded by the uncharacterized, cell-essential genes as compared to the rest of the genome. Parentheses denote the fraction of nucleolar genes.

Fig. S4



A
Correlation with gene-trap dataset (r)
0.7
0.6
0.5
KBM7
Jiyoye
Raji
K562
B
Raji
KBM7
Jiyoye
K562
Top 50 genes
Cell-line specific essentiality
2
0
-2
-4
Cell-line specific essential genes
Ranked genes
C
CDK4
HK2
SLC2A1 (GLUT1)
CRISPR score
0
-1
-2
-3
CDK6
HK1
SLC2A3 (GLUT3)
Paralog Expression (FPKM)
30
15
0
300
200
50
25
128
32
8
2
0.5
Raji
K562
KBM7
Jiyoye
D
GATA1
GATA2
CRISPR score
0
-2
-4
-1
-2
Expression (FPKM)
256
64
16
4
1
100
50
0
Raji
K562
KBM7
Jiyoye
E
CHUK (IKKα)
IKBKB (IKKβ)
IKBKG (IKKγ)
CRISPR score
0
-1
-2
Expression (FPKM)
10
5
0
30
15
40
20
Raji
Jiyoye
K562
KBM7
F
Differential gene expression
Raji vs. KBM7 K562 Jiyoye
GSEA
NF-κB signaling (SCHOEN_NFKB_SIGNALING)
TNFα signaling (SANA_TNF_SIGNALING_UP)

Fig. S4. Cell line-specific essentiality. (A) Correlation of GTS from screens conducted in KBM7 cells with CS from screens conducted in all four cells lines. **(B)** For each line, genes are ranked by the difference between mean-normalized CS in the line and the minimum, mean-normalized CS of the other three lines. Arrows along the distribution for K562 indicate genes residing the high-copy tandem amplification. **(C)** Specific gene essentiality of *CDK4*, *HK2*, and *SLC2A1* in Jiyoye cells due to absent/low expression of paralogs, *CDK6*, *HK1*, and *SLC2A3*. **(D)** Differential requirements for *GATA1* and *GATA* in K562 and KBM7 cells, respectively. **(E)** Raji-specific essentiality of *CHUK* and *IKBKB* (*IKBKG* approaches selectively threshold) which form the heterotrimeric IκB kinase complex. **(F)** GSEA analysis reveals NF-κB pathway activation specifically in Raji cells.

Fig. S5

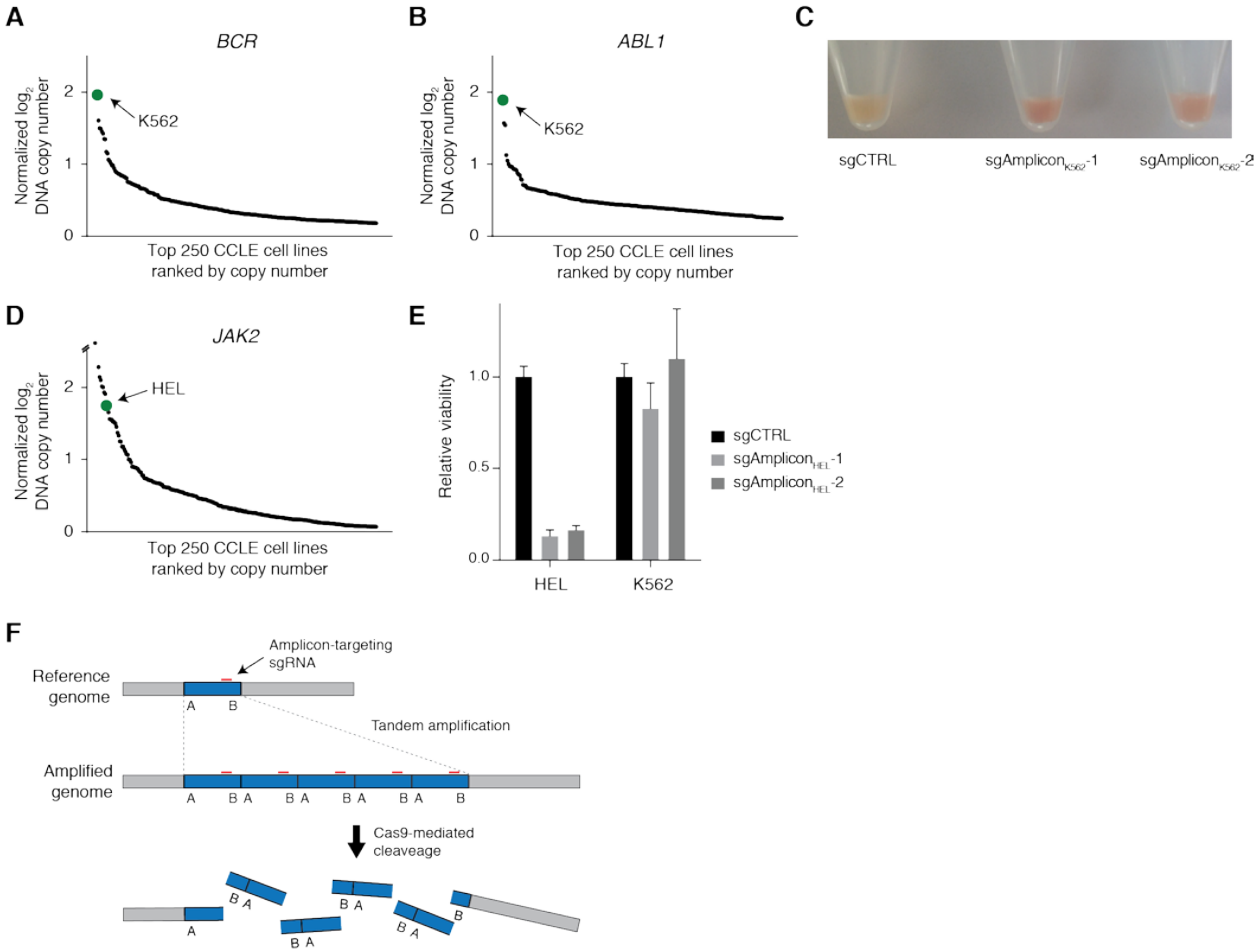
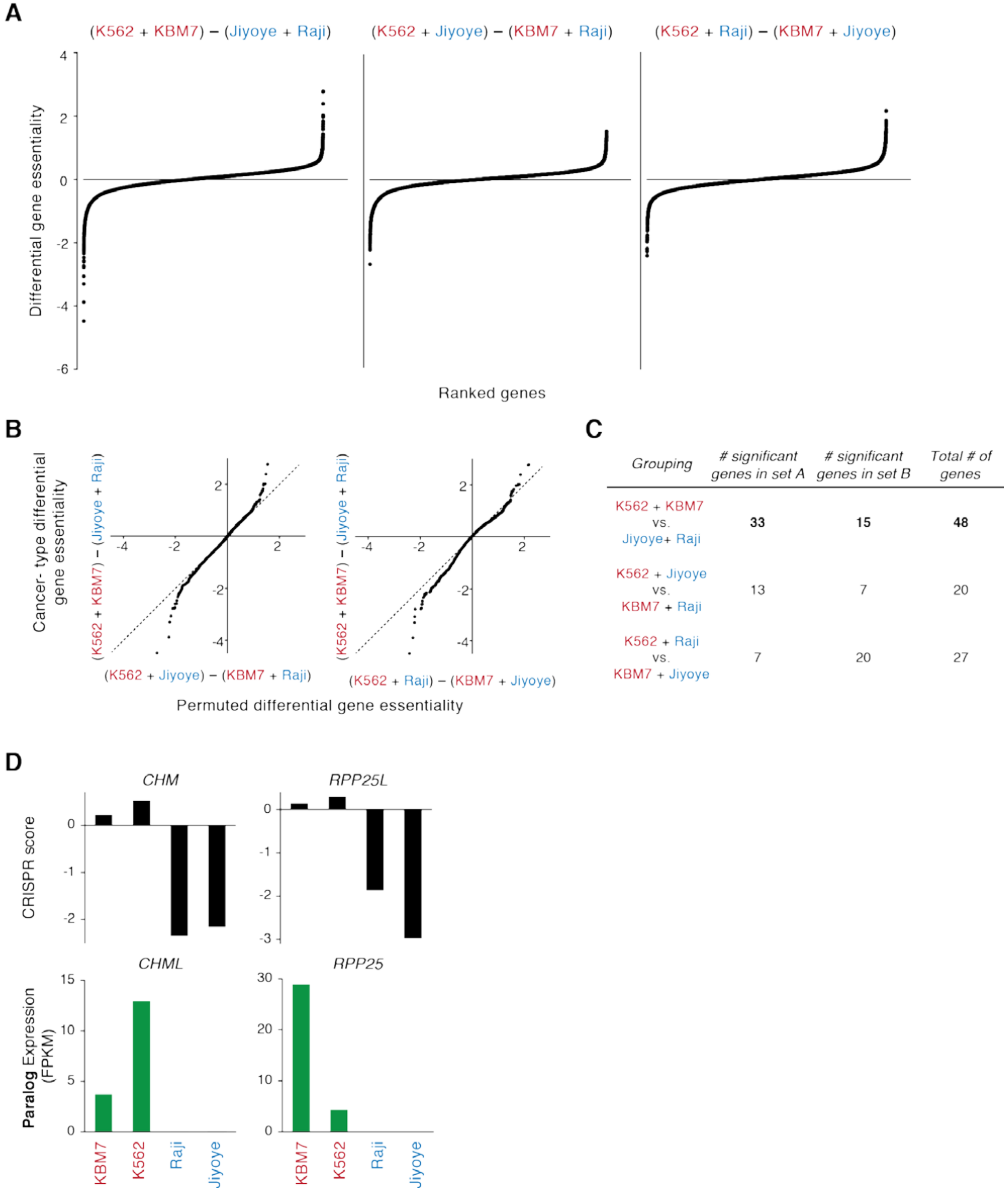


Fig. S5. Cytotoxicity induced by Cas9-mediated cleavage within highly amplified regions. (**A**) Out of all cell lines in the CCLE, K562 cells are the most amplified at the *BCR* locus, as assessed by DNA microarray copy number analysis. Only the top 250 cell lines are displayed for clarity. (**B**) Similar analysis as in (A) for the *ABL1* locus. (**C**) Two sgRNAs targeting non-genic sites within the BCR-ABL amplicon induce erythroid differentiation, as assessed by hemoglobin production. (**D**) Similar analysis as in (A) for the *JAK2* locus in HEL cells. (**E**) Two sgRNAs targeting non-genic sites within the JAK2 amplicon exhibit toxicity in HEL but not K562 cells. Error bars denote SD (n=4). (**F**) Model of Cas9-mediated cleavage in a prototypical region of high-copy tandem amplification.

Fig. S6



Grouping	# significant genes in set A	# significant genes in set B	Total # of genes
K562 + KBM7 vs. Jiyoye+ Raji	**33**	**15**	**48**
K562 + Jiyoye vs. KBM7 + Raji	13	7	20
K562 + Raji vs. KBM7 + Jiyoye	7	20	27

Fig. S6. Cancer-type specific essentiality. (**A**) Differential gene essentiality of cell lines paired by cancer type or paired randomly. (**B**) Quantile-quantile plot of cancer type-specific essentiality versus permuted set-specific essentiality. (**C**) Greater number of cancer type-specific essential genes versus permuted set-specific essential genes. (**D**) *CHM* and *RPP25L* essentiality in Raji and Jiyoye cells is likely due to the lack of expression of paralogs, CHML and RPP25, respectively.

Table Captions

Table S1 (separate file)
Annotations for the genome-wide sgRNA library containing spacer sequences and target gene information.

Table S2 (separate file)
Raw sgRNA counts from initial and final KBM7, K562, Raji, and Jiyoye cell populations.

Table S3 (separate file)
CRISPR scores (CS) and K-S test p-values adjusted for multiple hypotheses testing from screens in KBM7, K562, Raji, and Jiyoye cells. Values for KBM7 replicates are averaged.

Table S4 (separate file)
Gene-trap scores (GTS) from the gene-trap screen in KBM7 cells. Only genes with at least 65 or more anti-sense insertions were analyzed.

Tables

Rank	Gene	KBM7	K562	Jiyoye	Raji	Scoring cell line	Comment
1	*KIF18A*	-2.44	0.36	0.13	-0.15	KBM7	
2	*MCL1*	-2.56	-0.52	-0.11	-0.44	KBM7	BCL2/MCL1 pair
3	*ERG*	-1.89	0.41	0.01	0.21	KBM7	
4	*CBFB*	-2.21	-0.42	0.21	0.64	KBM7	CBFB/RUNX1 pair
5	*GATA2*	-2.32	0.00	-0.34	-0.53	KBM7	HSC/CMP master regulator
6	*RBM10*	-1.41	0.72	0.34	0.60	KBM7	
7	*SEPSECS*	-2.08	0.31	-0.39	0.19	KBM7	SEPSECS/SEPHS2 pair
8	*RUNX1*	-2.30	-0.63	-0.61	0.22	KBM7	CBFB/RUNX1 pair
9	*RAD51B*	-1.54	0.36	0.29	0.13	KBM7	RAD51B/RAD51D/RTEL1/FANCM set
10	*USP17L21*	-2.11	-0.05	-0.31	-0.45	KBM7	
11	*RAD51D*	-2.55	-0.86	-0.86	-0.90	KBM7	RAD51B/RAD51D/RTEL1/FANCM set
12	*TRAF2*	-1.35	0.28	0.53	0.25	KBM7	
13	*LIAS*	-2.17	-0.57	-0.27	-0.44	KBM7	LIAS/LIPT2 pair
14	*STAT5B*	-2.25	-0.66	0.27	0.19	KBM7	BCR-ABL pathway
15	*RTEL1*	-2.43	-0.83	-0.29	-0.69	KBM7	RAD51B/RAD51D/RTEL1/FANCM set
16	*BCL2*	-1.51	0.86	0.48	0.08	KBM7	BCL2/MCL1 pair
17	*FANCM*	-1.43	0.08	0.19	0.19	KBM7	RAD51B/RAD51D/RTEL1/FANCM set
18	*SEPHS2*	-1.90	-0.40	-0.24	-0.20	KBM7	SEPSECS/SEPHS2 pair
19	*LIPT2*	-1.59	-0.09	0.09	0.09	KBM7	LIAS/LIPT2 pair
1	*SMCHD1*	0.09	-4.82	0.15	0.65	K562	
2	*MAPK1*	-0.46	-4.74	0.25	-0.29	K562	In amplicon
3	*PELO*	-0.62	-4.79	-0.30	-0.57	K562	
4	*FIBCD1*	0.10	-3.72	0.28	0.27	K562	In amplicon
5	*AHCYL1*	-0.51	-4.25	-0.29	-0.50	K562	
6	*QRFP*	0.01	-3.68	0.61	0.40	K562	In amplicon
7	*SDF2L1*	0.39	-3.28	0.43	0.46	K562	In amplicon
8	*GATA1*	-0.06	-3.73	0.17	0.16	K562	MEP master regulator
9	*RAB36*	-0.11	-3.77	0.37	0.16	K562	In amplicon
10	*GNAZ*	0.29	-3.24	0.20	0.55	K562	In amplicon
11	*HIC2*	0.20	-3.79	-0.14	-0.53	K562	In amplicon
12	*IGLL5*	0.47	-2.76	0.35	0.55	K562	In amplicon
13	*GGTLC2*	-0.33	-3.73	-0.65	-0.51	K562	In amplicon
14	*ARVCF*	0.02	-3.18	0.17	-0.19	K562	In amplicon
15	*PPM1F*	-0.30	-3.52	-0.13	-0.57	K562	In amplicon
16	*YPEL1*	-0.02	-2.93	-0.04	0.00	K562	In amplicon
17	*REXO2*	-0.88	-3.78	-0.69	0.07	K562	
18	*LAMC3*	-0.04	-2.90	0.13	0.14	K562	In amplicon

19	*ZDHHC8*	0.03	-3.04	-0.08	-0.19	K562	In amplicon
20	*FLVCR1*	0.12	-3.19	0.10	-0.35	K562	
21	*KLHL22*	-0.06	-2.85	0.04	-0.31	K562	In amplicon
22	*PRAME*	0.13	-2.31	0.18	0.07	K562	In amplicon
23	*TOP3B*	0.15	-2.45	0.13	-0.07	K562	In amplicon
24	*TXNRD2*	0.15	-2.44	-0.12	-0.02	K562	In amplicon
25	*ZNF280B*	0.50	-2.07	0.18	0.17	K562	In amplicon
26	*AIF1L*	0.19	-2.05	0.26	0.29	K562	In amplicon
27	*TRMT2A*	-0.02	-2.50	-0.27	-0.16	K562	In amplicon
28	*FAM155A*	0.54	-1.91	0.31	0.65	K562	
29	*ZNF280A*	0.31	-1.91	0.31	0.29	K562	In amplicon
30	*CBFA2T3*	0.27	-2.17	0.02	0.12	K562	
31	*HIRA*	-0.23	-2.78	-0.62	-0.22	K562	In amplicon
32	*DGCR2*	0.45	-1.80	0.32	0.35	K562	In amplicon
33	*RTDR1*	0.23	-2.03	0.04	0.07	K562	In amplicon
34	*P2RX6*	0.25	-1.81	0.27	0.36	K562	In amplicon
35	*CRKL*	0.21	-1.93	0.12	0.15	K562	In amplicon
36	*SLC7A4*	0.41	-1.61	0.43	0.39	K562	In amplicon
37	*FAM78A*	0.17	-1.78	0.25	0.27	K562	In amplicon
38	*TRIP12*	0.22	-1.71	0.43	0.64	K562	
39	*TSSK2*	0.31	-2.02	-0.04	-0.10	K562	In amplicon
40	*EPS8L3*	-0.43	-2.34	-0.41	-0.32	K562	
41	*SIPA1*	-0.45	-2.30	-0.42	-0.29	K562	
42	*C22orf29*	-0.51	-2.40	-0.49	-0.58	K562	In amplicon
43	*L3MBTL2*	0.37	-1.49	0.32	0.65	K562	
44	*AIFM3*	-0.21	-2.30	-0.55	-0.19	K562	In amplicon
45	*VPREB1*	0.04	-1.92	-0.04	-0.20	K562	In amplicon
46	*YDJC*	0.46	-1.88	0.24	-0.17	K562	In amplicon
47	*EIF2AK4*	0.09	-1.73	0.13	-0.03	K562	
48	*RTN4R*	0.30	-1.49	0.21	0.22	K562	In amplicon
49	*CNNM4*	-0.31	-2.01	-0.30	-0.20	K562	
50	*SSBP3*	-0.46	-2.13	-0.19	-0.25	K562	
51	*DGCR6L*	-0.07	-2.26	0.07	-0.59	K562	In amplicon
52	*PTPN1*	0.64	-1.86	0.04	-0.21	K562	
53	*CLTCL1*	0.27	-1.36	0.36	0.40	K562	In amplicon
54	*ARL1*	0.34	-2.37	0.39	-0.74	K562	
55	*CCT8L2*	0.27	-1.48	0.45	0.14	K562	In amplicon
56	*MED16*	0.29	-2.07	-0.03	-0.45	K562	
57	*FRMPD2*	0.07	-1.54	0.61	0.50	K562	
58	*CRAMP1L*	0.47	-1.65	0.45	-0.04	K562	
59	*COMT*	0.22	-1.57	0.16	0.03	K562	In amplicon
60	*TMEM63A*	0.13	-1.60	0.64	0.00	K562	
61	*HSP90AB1*	-0.10	-1.68	0.27	-0.09	K562	
62	*TBC1D31*	0.14	-1.43	0.27	0.42	K562	
63	*SREBF1*	-0.65	-2.19	0.21	-0.22	K562	

1	*CDK4*	-0.27	0.08	-3.20	-0.19	Jiyoye	Paralog expression
2	*GOT1*	0.45	0.17	-1.90	0.24	Jiyoye	
3	*HK2*	-0.48	-0.75	-2.82	-0.61	Jiyoye	Paralog expression
4	*GPI*	-0.66	-0.65	-2.63	0.00	Jiyoye	
5	*DERL1*	0.02	-0.14	-1.91	-0.28	Jiyoye	
6	*SUCLG1*	-0.13	-0.44	-2.03	0.28	Jiyoye	
7	*SLC2A1*	-0.16	-0.34	-1.88	0.34	Jiyoye	Paralog expression
1	*DDX3Y*	0.55	1.11	0.52	-1.37	Raji	Paralog mutation
2	*IKBKB*	-0.27	-0.55	0.09	-2.41	Raji	NF-κB pathway
3	*CLCN3*	0.33	1.03	0.55	-1.52	Raji	
4	*ACSL1*	0.58	0.56	0.23	-1.56	Raji	
5	*SH3GL1*	0.05	0.58	-0.48	-2.21	Raji	
6	*CHUK*	0.38	0.24	0.26	-1.41	Raji	NF-κB pathway

Table S5. Cell-line specific hits. CRISPR scores for cell line-specific hits from all four cell lines. CRISPR scores are mean-normalized.

Rank	Gene	KBM7	K562	Jiyoye	Raji	Average difference	Adjusted p-value	Scoring cancer type	Comment
1	*ABL1*	-3.98	-4.80	0.29	-0.10	-4.48	1.2E-05	CML	BCR-ABL pathway
2	*BCR*	-3.05	-5.30	-0.29	-0.32	-3.88	6.3E-05	CML	BCR-ABL pathway
3	*SOS1*	-2.26	-3.47	0.50	0.36	-3.30	8.5E-05	CML	BCR-ABL pathway
4	*GRB2*	-1.71	-3.36	0.71	0.34	-3.06	1.1E-02	CML	BCR-ABL pathway
5	*LMO2*	-2.18	-2.83	0.29	0.25	-2.78	5.5E-05	CML	
6	*ATIC*	-2.96	-3.82	-0.43	-0.87	-2.74	5.5E-05	CML	One-carbon metabolism
7	*GAB2*	-1.54	-3.35	0.24	-0.20	-2.47	6.5E-05	CML	BCR-ABL pathway
8	*FKBPL*	-2.34	-2.37	-0.21	-0.10	-2.20	5.6E-05	CML	
9	*ATP1A1*	-2.70	-2.57	-0.29	-0.72	-2.13	4.2E-04	CML	
10	*MTHFD1*	-2.03	-3.25	-0.44	-0.63	-2.11	1.2E-04	CML	One-carbon metabolism
11	*PIK3C3*	-2.52	-1.71	-0.02	-0.15	-2.03	2.6E-04	CML	
12	*OBFC1*	-1.33	-1.86	0.35	0.18	-1.86	2.1E-03	CML	
13	*HSD17B12*	-1.50	-3.07	-0.15	-0.86	-1.78	4.9E-04	CML	
14	*PGM3*	-1.48	-1.39	0.20	0.35	-1.71	1.1E-02	CML	
15	*NSMCE1*	-2.34	-1.65	-0.20	-0.46	-1.67	1.0E-03	CML	
16	*PMVK*	-2.23	-1.68	-0.18	-0.44	-1.65	1.0E-02	CML	
17	*SHMT2*	-1.82	-1.03	0.28	0.14	-1.64	6.3E-05	CML	One-carbon metabolism
18	*NRF1*	-1.46	-3.13	-0.86	-0.74	-1.50	3.8E-02	CML	
19	*MYBL2*	-1.58	-1.71	-0.16	-0.14	-1.50	3.4E-02	CML	
20	*MINOS1*	-2.40	-1.61	-0.33	-0.71	-1.48	1.3E-02	CML	ETC assembly factor
21	*WRB*	-1.32	-1.29	-0.12	0.36	-1.42	8.6E-03	CML	
22	*CREBBP*	-1.85	-1.18	-0.30	-0.05	-1.34	4.9E-02	CML	
23	*BAG6*	-1.18	-1.75	-0.32	0.06	-1.34	1.1E-03	CML	
24	*ZNHIT3*	-1.01	-1.70	-0.12	0.03	-1.31	2.7E-02	CML	
25	*SCO2*	-2.04	-1.88	-0.64	-0.73	-1.28	1.6E-03	CML	ETC assembly factor
26	*CENPW*	-1.52	-1.87	-0.43	-0.41	-1.27	3.0E-02	CML	
27	*UBE2L3*	-1.38	-2.81	-0.96	-0.71	-1.26	4.3E-03	CML	
28	*THG1L*	-1.06	-2.59	-0.64	-0.51	-1.26	1.8E-02	CML	
29	*COASY*	-2.16	-2.18	-0.89	-0.95	-1.25	7.9E-03	CML	
30	*FASTKD5*	-1.73	-1.67	-0.62	-0.30	-1.24	5.0E-02	CML	

31	*NAE1*	-1.06	-1.80	-0.72	0.29	-1.21	1.0E-02	CML	
32	*PET117*	-2.09	-1.29	-0.86	-0.31	-1.10	2.0E-02	CML	ETC assembly factor
33	*BCS1L*	-1.61	-1.55	-0.52	-0.50	-1.07	4.1E-02	CML	ETC assembly factor
1	*RPP25L*	0.42	0.72	-2.75	-1.66	2.78	1.9E-04	BL	Paralog not expressed
2	*CHM*	0.50	0.94	-1.94	-2.14	2.76	1.2E-05	BL	Paralog not expressed
3	*EBF1*	0.39	0.22	-1.40	-2.76	2.38	6.7E-04	BL	B-cell transcription factor
4	*MEF2B*	-0.06	-0.62	-2.43	-2.27	2.01	5.5E-05	BL	Mutated in lymphoma
5	*MEF2BNB-MEF2B*	-0.06	-0.62	-2.43	-2.27	2.01	5.5E-05	BL	
6	*POU2AF1*	0.59	0.32	-1.01	-2.04	1.98	5.5E-05	BL	B-cell transcription factor
7	*CCND3*	0.08	0.05	-2.40	-1.41	1.97	1.7E-02	BL	Mutated in lymphoma
8	*PAX5*	0.18	0.55	-1.49	-1.28	1.75	1.1E-03	BL	B-cell transcription factor
9	*LOC100287177*	-0.35	0.87	-1.33	-1.50	1.68	3.5E-04	BL	
10	*PPIAL4D*	-0.57	-0.49	-1.91	-1.83	1.34	7.0E-05	BL	
11	*STAG2*	-0.07	-0.13	-1.48	-1.41	1.34	1.1E-03	BL	
12	*NBPF24*	-0.90	0.11	-1.74	-1.61	1.28	1.2E-03	BL	
13	*TTC7A*	-0.02	-0.48	-1.58	-1.40	1.24	6.7E-03	BL	
14	*LOC147646*	-0.14	0.19	-1.02	-1.27	1.17	2.7E-02	BL	
15	*GTF2E2*	-0.48	-0.94	-1.18	-2.25	1.00	1.3E-02	BL	

Table S6. Cancer type-specific hits. CRISPR scores for cancer type-specific hits from all four cell lines and average differences between cancer types. CRISPR scores are mean-normalized.

References

37. J. M. Engreitz *et al.*, The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science* **341**, (2013).
38. D. Kim *et al.*, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013).
39. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech* **28**, 511-515 (2010).
40. F. Cunningham *et al.*, Ensembl 2015. *Nucleic Acids Research* **43**, D662-D669 (2015).
41. L. Y. Geer *et al.*, The NCBI BioSystems database. *Nucleic Acids Research* **38**, D492-D496 (2010).
42. P. Flicek *et al.*, Ensembl 2014. *Nucleic Acids Research*, (2013).
43. J. A. Tennessen *et al.*, Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64-69 (2012).
44. C. Stark *et al.*, BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**, D535-D539 (2006).
45. H. Li *et al.*, TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research* **34**, D572-D580 (2006).
46. J. Barretina *et al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-307 (2012).
47. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44-57 (2008).
48. P. Mertins *et al.*, iTRAQ Labeling is Superior to mTRAQ for Quantitative Global Proteomics and Phosphoproteomics. *Molecular & Cellular Proteomics* **11**, (2012).
49. S. Schwartz *et al.*, Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites. *Cell Reports* **8**, 284-296 (2014).
50. A. G. Uren *et al.*, A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nature Protocols* **4**, 789-798 (2009).
51. T. Brady *et al.*, Integration target site selection by a resurrected human endogenous retrovirus. *Genes & Development* **23**, 633-642 (2009).