Supplemental Material
for
A high resolution genome-wide CRISPR/Cas9 viability screen reveals structural features and contextual diversity of the human cell-essential proteome

Thierry Bertomeu[1]*, Jasmin Coulombe-Huntington[1]*, Andrew Chatr-aryamontri[1]*, Karine Bourdages[1], Etienne Coyaud[2], Brian Raught[2], Yu Xia[3] and Mike Tyers[1]

[1]Institute for Research in Immunology and Cancer, Department of Medicine, University of Montreal, Montreal, Quebec, Canada
[2]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada
[3]Department of Bioengineering, McGill University Montreal, Quebec, Canada

Supplemental Methods

Cell culture

The 293T (CRL-3216) cell line was obtained from ATCC and the NALM-6 cell line was provided by Dr. Steven Elledge (Harvard Medical School). 293T cells were grown in 10% FBS (Wisent) DMEM medium and NALM-6 cells were grown in 10% FBS RPMI medium at 5% CO_2 and 37^{o}C. Cas9-expressing clones of NALM-6 cells were generated by transduction and selection (1 µg/mL puromycin, 4 days) of a lentivirus vector bearing a Dox-inducible Cas9 construct (pCW-Cas9) (1), followed by expansion of several clones obtained from single cell sorting on a BD FACSAria II. A clone that showed exhibited negligible FLAG-Cas9 expression in the absence of doxycycline and strong induction of FLAG-Cas9 expression following 2 days of doxycycline treatment (2 µg/mL), based on Western blots using an HRP-conjugated anti-FLAG antibody (Sigma), was selected for library screens.

sgRNA library design

The EKO library is an extension of a previously described sgRNA library (2) designed to target a set of 18,166 RefSeq genes based on previously optimized design principles (2). This set was supplemented with sgRNAs designed to target an additional 4,790 known or hypothetical protein-coding genes from AceView (2010 release) (3) and GENCODE v19 (2012 release) (4)

databases, as retrieved from the UCSC Human Genome Browser (5). We used the 37th genome assembly (hg19) of the human genome to generate the sgRNA designs since AceView gene annotations were not available for the most recent assembly. RefSeq gene annotations were also retrieved from the UCSC Human Genome Browser (May 2011 release). We identified all coding regions from the AceView and GENCODE annotations that did not overlap with RefSeq coding regions from a previous library (1). Up to 10 sgRNAs were designed to target each candidate AceView and GENCODE gene. Additional sgRNAs were also designed for genes that were targeted by fewer than 10 sgRNAs in the previous library (2). We later remapped all sgRNAs to RefSeq gene annotations (December 2015 release) to create a core sgRNA annotation set. Gene names were converted to standard HGNC symbols when possible by mapping previous gene symbols to new approved symbols as of December 2015. RefSeq gene names that failed to map to an HGNC entry but whose coding regions overlapped a coding region from an AceView or GENCODE gene that did map to an HGNC entry were renamed accordingly. In order to define alternatively spliced exons for sgRNA design, we identified all AceView coding regions that were overlapped by an intron in another transcript of the same gene to a maximum span of 3 exons. Expression and exon skipping were confirmed using the Illumina Human BodyMap 2.0 datase (6). All 75 base pair reads were retrieved from the Sequence Read Archive [Accession ID: ERP000546] and Bowtie 0.12.7 was set to output all alignments (-a option) with default settings (7) to map reads to all exons and splice junctions from the AceView database. We preserved exons with 30 or more reads that mapped to the exon and 8 or more reads that mapped to exon-exon junctions that skipped over the exon or exonic region. This analysis yielded a set of 38,292 alternatively spliced exons or exonic regions for sgRNA design (see below).

Design of sgRNAs

The reference hg19 genome sequence was used to identify all protospacer adjacent motif (PAM) sequences (NGG) for which the Cas9 cleavage site would occur within a coding region and for which the potential sgRNA sequence would be 40-80% GC content. For each potential sgRNA sequence, we used Bowtie 0.12.7 set to output all alignments (-a option) with default settings to identify all perfect or near perfect matches to the genome. Near-perfect matches were defined as possessing a single mismatched position within the first 12 bases of the sgRNA, which are predicted to have a reasonable probability of inducing Cas9 cleavage (2). Only matches positioned immediately upstream of a PAM (NGG) sequence were considered. All sgRNAs with greater than 4 perfect matches to the genome were eliminated in order to avoid

multiple cleavage events. We assigned each sgRNA a score based on its predicted efficiency for Cas9 binding and/or on-target cleavage. Based on two position weight matrices from a previous study (1), the following features were assigned a +1 contribution to the overall score: (i) a G at positions 1, 17, 19 or 20, (ii) an A at position 11 or 12, (iii) a C at position 18, and (iv) annealing to the coding strand. Conversely, a T at positions 17, 19 or 20 was assigned a -1 contribution to the overall score. Up to 10 sgRNAs were designed per gene and up to 3 sgRNAs per alternative exon based on in order of priority: (i) the smallest number of perfect matches to the genome, (ii) the smallest number of genomic alignments with a single-base mismatch, (iii) the largest number of transcripts targeted for a given gene, (iv) the highest Cas9 efficiency sequence score, and (v) the most 5'-targeting sequence. Non-targeting control sgRNAs were designed by generating random 20mers and selecting a subset of 2,043 with 40-80% GC content that could not be matched to the human genome by Bowtie 0.12.7 and that had average similar Cas9 efficiency scores as the rest of the library. All sgRNAs used to generate the EKO library are listed in Table S1.

Library construction

pCW-Cas9 (#50661), pLX-sgRNA (#50662), psPAX2 (#12260), pCMV-VSV-G (#8454) and LentiCRISPR v2 (#52961) were obtained from Addgene. To reduce the number of empty ligation events, pLX-sgRNA was modified as follows. pLX-sgRNA was amplified using PCR primers pLX-sgRNA REV and pLX-sgRNA FWD (see Table S6 for primer sequences) and HiFi KAPA polymerase (KAPA Biosystems). A 1032 bp fragment from the AAVS1 locus was amplified using primers F3 AAVS1 and R3 AAVS1 using NALM-6 genomic DNA as template, followed by re-amplification using primers pLX-sgRNA5' BfuA1 F3 AAVS1 and pLX-sgRNA3' BfuA1 R3 AAVS1 to add BfuA1 restriction sites and Gibson assembly (GA) homology sequences on each end. This fragment was cloned into amplified pLX-sgRNA plasmid by GA to create the pLX-sgRNA 2XBfuA1 plasmid, which became a BfuA1 Golden Gate-compatible vector for efficient insertion of sgRNA cassettes. Three 92K custom oligonucleotide pools (Custom Array) were independently amplified using PAGE-purified oligos FWD GA sgRNA -50 to 0 and REV GA sgRNA +50 to 0. The resulting 120 bp amplicons were cloned by GA into BfuA1-digested and gel-purified pLX-sgRNA 2XBfua1, followed by transformation of ultra-competent DH5α bacteria at a representation of at least 100 colonies per sgRNA. Following overnight growth on LB ampicillin plates at a density of 50,000-100,000 colonies per 15 cm plate, bacteria were eluted and plasmid DNA recovered by maxiprep DNA purification. Each plasmid sub-library was maintained separately to permit lower-complexity whole-genome

screens if desired. Single sgRNAs were cloned into the LentiCRISPR v2 plasmid according to the Zhang lab protocol available at Addgene.

Lentivirus production

For pooled library lentivirus production, plasmids psPAX2 (151 µg), pCMV-VSV-G (82 µg) and equal amounts of each three pooled libraries of sgRNAs (78 µg each) were combined in 16 mL of water and vortexed with 1.4 mL of 1 mg/mL polyethyleneimine. After 15 min, this mixture was added to 550 mL of DMEM with 10% FBS and overlaid on 90% confluent 293T cells grown in a Corning HYPERFlask. For single sgRNA lentivirus production, 6 µg psPAX2, 3 µg pCMV-VSV-G and 9 µg LentiCRISPR v2 were mixed in 666 µL water and 53 µL 1 mg/mL polyethyleneimine for 15 min and the transfection mixture overlaid on 90% confluent 293T cells in 10 cm dishes. After 16 h, media was changed to 2% FBS DMEM and after 32 h, lentivirus-bearing supernatant was collected, passed through a 0.45 µM filter and adjusted to a final concentration of 5% sucrose, 2 mM $MgCl_2$ and 10 mM HEPES pH7.2.

sgRNA library transduction

$8x10^6$ cells of the doxycycline-inducible Cas9 NALM-6 clone were grown in a spinner flask in 1L media supplemented with 10 µg/mL protamine sulfate. After 15 min, a fraction of the EKO 278K whole-genome sgRNA lentivirus pooled library was added. After 48 h, multiplicity of infection (MOI) was evaluated by Q-PCR. Genomic DNA was extracted from 100,000 cells from the library and from cells bearing pLX-sgRNA insertions of known copy numbers as standards using the prepGEM DNA extraction kit (ZyGEM). This DNA was used as a template for Q-PCR reactions with primers FWD 1 BLAST and REV 1 BLAST matched with Roche's Universal Primer Library #89. A MOI of 0.5 was used for the EKO whole-genome sgRNA library screen. Cells were selected and expanded for 6 days in 3L media with 10 µg/mL blasticidin. 140 million cells (the EKO library primary titer, corresponding to 500 cells per sgRNA for the 278,754 different sgRNAs) were collected to establish sgRNA frequencies in the library pool prior to Cas9 induction. Most of the primary infection was preserved in cryovials (10 million cells per mL in 50% v/v FBS, 10% DMSO v/v and 40% v/v RPMI) as the uninduced version of the EKO library for future use. A further 140 million cells were expanded for 7 days in 2 µg/mL doxycycline to induce Cas9 expression and indel formation, and then cultured for an additional 14 days without doxycycline. Every 2-3 days, cells were counted and a fraction of cells (a minimum of 140 million cells for each time point) were washed with PBS and stored at -20^{o}C for later genomic DNA extraction. The remaining cells were split back to 400,000 cells per mL at a

minimum total of 140 million viable cells to maintain pool representation. Over the course of this entire growth regime, cells divided approximately once every 24 hours.

Genomic DNA extraction

Frozen cell pellets were resuspended in 1.4 mL TE (10 mM Tris-HCl pH8.0, 1 mM EDTA) and lysis buffer (10 mM Tris-HCl pH8.0, 10 mM EDTA, 0.5% w/v SDS, 0.2 mg/mL proteinase K) was added for a final volume of 14 mL. Tubes were incubated for 3 h in a 55^{o}C water bath with occasional vortexing. 5M NaCl was added to extracts to a final concentration of 0.2M, lysates were extracted twice with equal volumes of pH7.5 phenol/chloroform/isoamyl alcohol (25:24:1) followed by chloroform extraction using phase-lock tubes. RNAse A (final concentration 50 µg/mL) was added and an overnight digestion carried out at 37^{o}C. After phenol/chloroform and chloroform extraction, DNA was recovered by precipitation with 2.5 volumes ethanol and 1/30 volume 3M sodium acetate (pH 5.2). The pellet was washed with 70% ethanol, briefly dried and resuspended in 1 mL TE in a 55^{o}C dry block, and finally sheared by several passes through a 27G needle.

Next generation sequencing of sgRNA cassettes

924 µg genomic DNA (corresponding to 140 million cells, given that a human diploid cell contains 6.6 pg DNA) were used as template in a first round of PCR (1150 µL 10X reaction buffer, 230 µl 10 mM dNTPs, 9.2 µl 500 µM primer Outer 1, 9.2 µl 500 µM primer Outer 2, 230 µl DMSO and 290 units of GenScript Green Taq DNA polymerase in a total volume of 11.5 mL). Multiple 100 µl reactions were setup in 96-well format on a BioRad T100 thermal cycler (95^{o}C 5 min, 25 cycles of 35 sec at 94^{o}C, 35 sec at 52^{o}C and 36 sec at 72^{o}C, final step of 10 min at 72^{o}C after the last cycle). Completed reaction mixes were combined into one 15 mL tube and vortexed and 1.5 mL aliquots were concentrated to 100 µl by ethanol-precipitation, followed by separation on a 1% agarose gel and extraction of the 475 bp amplicon. A second PCR reaction was performed to add Illumina sequencing adapters and 6 bp indexing primers (250 ng PCR1 template, 5 µl 10X buffer, 5 µl 2,5 mM dNTPs, 1 µl of PAGE-purified equimolar premix 100 µM TruSeq Universal Adapters -2/0/+2/+5/+7 to shuffle the constant DNA region to be sequenced, 1 µl of 100 µM PAGE-purified TruSeq Adapter with appropriate index, 1 µl DMSO and 5 units Takara Taq to 50 µl total volume (5 min at 95^{o}C, 5 cycles of 15 sec at 95^{o}C, 30 sec at 50^{o}C and 30 sec at 72^{o}C, 5 cycles of 15 sec at 95^{o}C, 30 sec at 56^{o}C and 30 sec at 72^{o}C, followed by 5 min final step at 72^{o}C after the last cycle). A 236-245 bp amplicon was gel extracted from a 2% agarose gel and sequenced on a Illumina HiSeq 2000 in a 50 bp single read configuration with a

target average coverage of 100 reads per sgRNA. Next generation sequencing was performed at the McGill University and Génome Québec Innovation Centre (Montréal, Canada).

Assessment of single sgRNA growth inhibition

200 µL of single sgRNA lentiviral stock solution was mixed with 800 µL 10% FBS RPMI and 2 µL 10 mg/mL protamine sulfate and incubated at room temperature for 15 min, then added to 1 million NALM-6 cells suspended in 1 mL medium and incubated at 37^{o}C in 5% CO_2 for 2 days. Cells were resuspended, 1 mL discarded and 2 mL fresh medium added with puromycin to 1 µg/mL final concentration. After 2 further days, 1 mL of cultures was resuspended in 4 mL fresh media with puromycin and selected for 4 additional days. Cell concentration was then measured in triplicate on a Beckman-Coulter Z2 Counter using a standard threshold to exclude debris counts.

Identification of protein interactions by BioID analysis

Generation and culture of cell lines. Detection of protein interactions by the BioID proximity labeling method (8) was carried out as described previously (9). Briefly, tetracycline-inducible N-terminal FlagBirA(R118G)-tagged proteins were expressed in human Flp-In T-REx 293 cells (Invitrogen, Carlsbad, CA). Protein expression was induced by adding 1 µg/mL tetracycline to the culture medium (DMEM, 10% fetal calf serum, 10 mM HEPES (pH 8.0), 1% penicillin-streptomycin) for 24h. Cells were treated concurrently with 50 µM biotin (BioShop, Burlington, ON, Canada) to effect proximity labeling. Each biological replicate, consisting of 5 x 150 cm^2 pooled plates of subconfluent (80%) cells, was then scraped into PBS, washed twice in 10 mL PBS, collected by centrifugation at 1000 x g for 5 min at 4°C and stored at -80°C until lysis.

Biotin-streptavidin affinity purification. The cell pellet was resuspended in 10 mL of lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton X-100, 0.1% SDS, 1:500 protease inhibitor cocktail (Sigma-Aldrich), 1:1000 benzonase nuclease (Novagen)), incubated on an end-over-end rotator at 4°C for 1 h, briefly sonicated to disrupt any visible aggregates, then centrifuged at 16,000*g* for 30 min at 4°C. The supernatant was transferred to a fresh 15 mL conical tube, 30 µL of packed, pre-equilibrated streptavidin-sepharose beads (GE) were added, and the mixture incubated for 3 h at 4°C with end-over-end rotation. Beads were pelleted by centrifugation at 2000 rpm for 2 min and transferred with 1mL of lysis buffer to a fresh Eppendorf tube. Beads were washed once with 1 mL lysis buffer and twice with 1 mL of 50 mM ammonium bicarbonate (pH 8.3). Beads were transferred in ammonium bicarbonate to a

fresh centrifuge tube, and washed two more times with 1 mL ammonium bicarbonate buffer. Tryptic digestion was performed by incubating the beads with 1 µg MS grade TPCK trypsin (Promega, Madison, WI) dissolved in 200 µL of 50 mM ammonium bicarbonate (pH 8.3) overnight at 37°C. The following morning, an additional 0.5 µg trypsin was added, and the beads incubated 2 h at 37°C. Beads were pelleted by centrifugation at 2000 x *g* for 2 min, and the supernatant was transferred to a fresh Eppendorf tube. Beads were washed twice with 150 µL of 50 mM ammonium bicarbonate, and these washes were pooled with the first eluate. The sample was lyophilized, and resuspended in buffer A (0.1% formic acid) and 1/5th of the sample analyzed per MS run.

Mass spectrometry. High performance liquid chromatography was conducted using a 2cm pre-column (Acclaim PepMap 50 mm x 100 µm inner diameter (ID)), and 50 cm analytical column (Acclaim PepMap, 500 mm x 75 um diameter; C18; 2 um; 100 Å, Thermo Fisher Scientific, Waltham, MA), running a 120 min reversed-phase buffer gradient at 225 nL/min on a Proxeon EASY-nLC 1000 pump in-line with a Thermo Q-Exactive HF quadrupole-Orbitrap mass spectrometer. A parent ion scan was performed using a resolving power of 60,000, then up to the twenty most intense peaks were selected for MS/MS (minimum ion count of 1,000 for activation), using higher energy collision induced dissociation (HCD) fragmentation. Dynamic exclusion was activated such that MS/MS of the same *m*/*z* (within a range of 10 ppm; exclusion list size = 500) detected twice within 5 sec were excluded from analysis for 15 sec. For protein identification, Thermo .RAW files were converted to .mzXML format using Proteowizard (10), then searched using X!Tandem (11) and Comet (12) against the human Human RefSeq Version 45 database (containing 36113 entries). Search parameters specified a parent ion mass tolerance of 10 ppm, and an MS/MS fragment ion tolerance of 0.4 Da, with up to 2 missed cleavages allowed for trypsin. Variable modifications of +16@M and W, +32@M and W, +42@N-terminus, and +1@N and Q were allowed. Proteins identified with an iProphet cut-off of 0.9 (corresponding to ≤1% FDR) and at least two unique peptides were analyzed with SAINT Express v.3.3. Twelve control runs from cells expressing the FlagBirA* epitope tag alone were collapsed to the two highest spectral counts for each prey, and compared to the two technical replicates of each bait analysis. High confidence interactors were defined as those with BFDR≤0.02. Raw mass spectrometry data has been submitted to the MassIVE database under accession number MSV000081460.

Comparison of RANKS versus other scoring methods

Several methods have been proposed for scoring changes in sgRNA and/or gene representation in CRISPR/Cas9 screens. The most straightforward approach determines the average $\log_2$ read frequency fold-change of all sgRNAs for each gene as the gene score (1). The MAGeCK method applies median-based normalization and models the read count variance of sgRNAs as a function of read count (13). MAGeCK integrates individual sgRNA p-values by comparing the combined sgRNA p-values for a gene to a reference distribution generated by combining random sgRNAs together, similarly to RANKS. The BAGEL algorithm is a probabilistic approach developed to identify essential genes from CRISPR screens (14, 15). BAGEL is a naïve Bayes classifier that is trained on a gold-standard set of essential and non-essential genes. To evaluate how the RANKS scoring system compares to these existing approaches, we applied all four methods to our dataset to identify likely essential genes. We calculated the average $\log_2$ read frequency fold-changes from the sgRNA scores we had already calculated for RANKS. We used the publicly available package for MAGeCK set with default parameters, selecting as control sgRNAs the 2,043 non-targeting sgRNAs, as we did for RANKS. BAGEL was implemented using the publically available package together with the gold-standard training sets provided and with minimal read counts set to 20. For all four methods, we restricted the list of sgRNAs to those in the main RefSeq annotation set of 19,084 genes used in the current study, and restricted analyzed genes to those with ≥4 sgRNAs covered by ≥20 reads in at least one sample. We assessed the top 2,000 genes from each method, where we expect most essential genes to be scored, and found that 88% of these genes were shared across all 4 methods. This concordance illustrates that the assignment of gene essentiality is quite robust to the actual scoring scheme. To identify differences in the performance of each algorithm as compared to RANKS, we further assessed the genes within the top 2,000 that were not shared with RANKS. In the case of BAGEL, we discarded all genes that appeared in the training set, since the selection of such genes is expected to be favored by the training procedure. Within the top 2,000 genes, those ranked by $\log_2$ fold-change, MAGeCK and BAGEL differed from RANKS by 108, 160 and 70 genes, respectively. We compared these sets of non-overlapping genes considering four features: (i) the number of different cell lines in which the genes were found to be essential out of three whole-genome studies based on CRISPR screens(1, 14) or gene trap score in HAP1 cells(16) (Fig. S3A); (ii) the proportion of mutated residues across 46 vertebrate species (Fig. S3B); (iii) gene trap scores in HAP1 cells (Fig. S3C) and (iv) mRNA expression level in NALM-6 cells, as derived from public RNA-seq data (Fig. S3D) (17). In pair-wise comparisons, we found that RANKS significantly outperformed the other three methods in 10 out of the 12 possible comparisons (p-value<0.05). Performance

in terms of mutation rate for MAGeCK and gene-trap score for BAGEL were not significantly different from that of RANKS (p-value>0.05).

Correlates of gene and protein features

Sequence conservation. MultiZ RefSeq protein sequence alignments of 45 vertebrate genomes (18) aligned to the hg19 human genome assembly were retrieved from the UCSC Genome Browser (5). For each RefSeq gene, we defined the mutation rate as the total number of residues differing from the human residue over the total number of residues aligned, excluding gaps. We classified more conserved sgRNAs as those for which the 30 residues centered on the Cas9-targeted codon were more conserved than the average for the gene. Human-yeast gene ortholog mappings were taken from NCBI (19).

mRNA expression. The Sequence Read Archive for all reads of the first RNA-seq sequencing run for the untreated parental NALM-6 cell line (17) (accession #:GSM1872078) was aligned with Bowtie 2.2.5 (20) with default settings to map RNA-seq reads to the union of all RefSeq, GENCODE and AceView transcripts. We defined gene expression as the $\log_2$ of the number of reads per million reads mapped including a pseudocount of 1 read.

DNAse 1 hypersensitivity. Peak data for naïve B-cells based on DNAse-seq analysis from the ENCODE consortium (21) (accession #:GSM1008557) was retrieved from the UCSC Genome Browser for the hg19 genome assembly (5). Peak density was defined as the number of peaks per 1 kb overlapping a RefSeq (May 2011) gene locus. The correlation with mRNA expression was performed using the integrated sum of DNAse 1 hypersensitivity reads per locus for the ~15,000 core sgRNA set genes that scored as least essential.

Protein domains. Using HMMer 3.0 (22) and an e-value threshold of 1e-4, we mapped all Pfam-A domains (23) (version dated May 2011) to AceView and GENCODE predicted protein sequences. We then classified each sgRNA as either targeting a predicted domain-coding region or not. To correlate effects with known or candidate secondary structures, we mapped the above 30 amino acid blocks to all protein sequences from the PDB database (24) using BLATv34 (25) . For PDB models containing two or more proteins, we defined interfacial residues as those within 6 Å from a non-hydrogen atom of the other protein (24).

Intrinsically disordered regions. IUPred v1.0 was used to scan all RefSeq protein sequences for

long disordered stretches (26). The average probability per residue of belonging to a disordered stretch over the length of the exon was used for the alternatively-spliced exon comparison and a cut-off of 40% was used to classify residues as potentially disordered for the residue-level essentiality analysis.

Identification of paralogs. Blast v2.2.17 with default parameters was used to search for matches between any two RefSeq protein sequences. For each gene, the highest bit score of any alignment with a different gene was compared to the highest bit score of the target protein alignment to itself in order to estimate the percent sequence identity. Protein pairs with a sequence identity of 30% or greater were considered paralogs.

Protein interaction network analysis. Protein network features (degree, clustering coefficient) were computed using the iGraph python library (27). The 486 universal essential genes were mapped to a generic GO slim term set (see http://www.geneontology.org/ontology/subsets/goslim_generic.obo) (28).

Non-specific effects of Cas9 cleavage on sgRNA representation

The phenotypic effect of DSBs on sgRNA depletion was evident from both the relative enrichment of non-targeting control sgRNAs compared to sgRNAs targeting non-essential genes in the library (Fig. 1G) and the stronger depletion of sgRNAs with multiple predicted matches to the genome (Fig. 2B). To mitigate this effect, we recalculated RANKS-based p-values using all targeting sgRNAs in the library to model the control distribution except for sgRNAs that targeted known essential genes (1, 14, 16). Note that the non-targeting sgRNA set should still serve as an optimal control for any comparative screen in which the untreated pool undergoes the same doxycycline induction and outgrowth as the treated sample. Based on the results presented in Fig. 2B, we also opted to exclude all sgRNAs with ≥2 potential off-target cleavage sites. As removing all sgRNAs with potential off-target cleavage sites would have significantly reduced the coverage of the EKO library, we chose to preserve those sgRNAs with 1 or 2 potential off-target sites. To prevent overweighting of depletion scores for these sgRNAs, we added a fixed correction factor to the $\log_2$ read frequency fold-change, based on the results of linear regression analysis (+0.20 per perfect match and +0.13 per near-perfect match). As compared to removal of all multi-matching sgRNAs, the retention of sgRNAs with 1 or 2 potential off-target cleavage sites allowed inclusion of an additional 548 RefSeq genes with ≥4 sgRNAs, 62 of which scored as essential (FDR<0.05). As compared to other essential genes,

these 62 genes possessed on average lower gene trap scores in HAP1 cells (p=7.78e-5, Wilcoxon test, Fig. S5A) and are more conserved across 46 vertebrate genomes (p=3.86e-7, Wilcoxon test, Fig. S5B). Based on these comparisons, the inclusion of sgRNAs with 2 or 3 genomic matches with the applied score correction, should not represent a disproportionate source of false positive essential gene identifications. As above, we note that for comparative screens in which the untreated sample undergoes similar outgrowth to a treated sample, score correction for sgRNAs with multiple potential cleavage sites should not be necessary. Using this new set of p-values, the ranking of genes in the NALM-6 screen remained virtually identical but with only 2,236 genes below the 0.05 FDR threshold (Table S1), which closely matched expectation based on previous genome-wide screens in human cell lines (1, 14, 16).

Extraction of protein and peptide evidence

Raw mass spectral data files for a recent draft map of the human proteome covering 73 different tissues and body fluids (29) were retrieved from ProteomicsDB (www.proteomicsdb.org). All files were converted to mgf format using Proteowizard (10) and XTandem! was used with default settings to identify peptides (11) that corresponded to RefSeq AceView or GenCode genes included in the EKO library. For each peptide spectrum, only the most significant peptide identification with an FDR<0.1 was tabulated. Protein-level expression was defined as the number of peptide identifications divided by the length of the longest encoded isoform. Alternative exon inclusion/expression was defined as the number of identified peptides in an exon divided by the length of the exon and then divided by the protein-level expression. Evidence of protein expression for hypothetical proteins was restricted to spectra matched with an FDR<0.001 and not matched to any RefSeq protein. For protein level expression of subunits of CORUM complexes across 30 different tissues, expressed proteins were designated by an average spectral count of ≥ 5 across replicates (30). For proximity of subunits with in a common PDB structure (24), two subunits were considered to interact directly if any α-carbon atoms from each chain were within 10 Å of each other. Location of residues within an α-helix, an extended β-strand or neither was drawn from PDB as annotated by DSSP (24, 31).

Gene identifiers

For EKO library design and sgRNA annotation, we mapped as many genes as possible to the official HGNC symbol. The same annotation procedure was applied to gene lists reported in previous genome-wide essentiality screens (1, 14, 16). This approach enabled the annotation of a shared set of 16,996 gene symbols shared across the NALM-6 screen and 9 previously

reported genome-wide CRISPR screens. The determination of universal and cell line-specific essential gene sets was based on this list of common genes. Note that KBM7 gene trap data (16) was not considered for this analysis since the identical cell line was also screened by the CRISPR/Cas9 method (1).

Modeling gene essentiality

We modeled the distribution of essential genes recovered across 10 different screens. A simple random model with a single parameter for the rate of spontaneous cell type-specific essentiality (α) was effectively a nested model (i.e., a version with some parameters fixed) within a binary model. The binary model possessed two additional parameters, the fraction of core essential genes (μ) and the false negative rate (FNR), which were fixed to zero in the random model. Similarly, the binary model was nested within the continuous model. In the continuous model, core essentiality follows a sigmoidal curve between 0 and 1 with two fitted parameters corresponding to the curve's midpoint (μ) and inverse steepness (σ). In this context, the binary model corresponds to a version of the continuous model where the steepness is infinite (fixed at μ=0) and the midpoint is the fraction of core essential genes (fitted). In each case, the distributions were created by generating a base probability (ρ) of gene essentiality in any given cell line (i.e., the core essentiality of the gene) for each of the 16,996 genes tested across all 10 cell lines. For the random model, this probability was equal to the parameter α. For the binary model, this value was set to 1-FNR with probability μ, or α otherwise. The random model produces a binomial distribution and the binary model produces a mixture of two binomials. We could thus apply the binomial probability function to compute the probabilities of a gene being essential in a given number of cell lines. We manually adjusted the parameters in order to maximize the log likelihood of the model generating the observed distribution. For the continuous model, the probability (ρ) was set to

$$\rho = \frac{1 - FNR}{1 + e^{\frac{\mu - x}{\sigma}}}$$

where σ is the inverse steepness of the curve, μ the midpoint and x is a random number between -1 and +1. Each non-essential gene is then reassigned as essential with probability α. The probability of a gene being essential in a given number of cell lines was estimated by randomLy generating distributions from the model. Parameters were manually adjusted order to maximize the log likelihood of the observed distribution, averaged over 20 runs. The final optimized parameters for the random model were: α=0.105; for the binary model: α =0.027,

FNR=0.27, μ =0.11; for the continuous model: α=0.01, FNR=0, μ=0.075, σ=0.77. The obtained maximum likelihoods of each optimized model were used to calculate (via the likelihood ratio test) the probability that the more complex model would produce a better fit by chance given fewer degrees of freedom (i.e., additional fitted parameters). This test indicated that the binary model reproduced the distribution significantly better than the random model ($p<1e-10$), and the continuous model significantly better than the binary model ($p<1e-10$).

Statistical analysis

The R suite was used to calculate all statistical significance values. Beanplot figures were generated using the Beanplot v1.2 package for R, with default settings except for the masking of striplines and customization of the y-axis range. Boxplots were generated with the R Boxplot function using the "outline=FALSE" and "range=1" custom options. Independence of the different residue-level features in predicting phenotypic effects of each sgRNA was assessed by multi-variate linear regression. Depletion scores of the individual sgRNAs within each gene were compared to the average depletion score for each gene. In addition to each residue-level variables, the number of predicted potential off-target cleavage sites with perfect matches or near perfect matches, as well as the sgRNA sequence scores used during the sgRNA design step were included to ensure that the observed correlations were not due to biases in sgRNA sequences that mapped to specific protein regions.

Supplemental References

1. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. Science 350:1096-1101.
2. Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. Science 343:80-4.
3. Thierry-Mieg D, Thierry-Mieg J. 2006. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7 Suppl 1:S12 1-14.
4. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. 2012. GENCODE: the reference human

genome annotation for The ENCODE Project. Genome Res 22:1760-74.

5. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hickey G, Hinrichs AS, Hubley R, Karolchik D, Learned K, Lee BT, Li CH, Miga KH, Nguyen N, Paten B, Raney BJ, Smit AF, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43:D670-81.
6. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez Banet J, Billis K, Garcia Giron C, Hourlier T, Howe K, Kahari A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel JH, White S, Zadissa A, Flicek P, Searle SM. 2016. The Ensembl gene annotation system. Database (Oxford) 2016.
7. Langmead B. 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics Chapter 11:Unit 11 7.
8. Roux KJ, Kim DI, Raida M, Burke B. 2012. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. J Cell Biol 196:801-10.
9. Gupta GD, Coyaud E, Goncalves J, Mojarad BA, Liu Y, Wu Q, Gheiratmand L, Comartin D, Tkach JM, Cheung SW, Bashkurov M, Hasegan M, Knight JD, Lin ZY, Schueler M, Hildebrandt F, Moffat J, Gingras AC, Raught B, Pelletier L. 2015. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. Cell 163:1484-99.
10. Kessner D, Chambers M, Burke R, Agus D, Mallick P. 2008. ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24:2534-6.
11. Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466-7.
12. Eng JK, Jahan TA, Hoopmann MR. 2013. Comet: an open-source MS/MS sequence database search tool. Proteomics 13:22-4.
13. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, Irizarry RA, Liu JS, Brown M, Liu XS. 2014. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol 15:554.
14. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J. 2015. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell 163:1515-1526.
15. Hart T, Moffat J. 2016. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics 17:164.
16. Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van

Diemen FR, Olk N, Stukalov A, Marceau C, Janssen H, Carette JE, Bennett KL, Colinge J, Superti-Furga G, Brummelkamp TR. 2015. Gene essentiality and synthetic lethality in haploid human cells. Science 350:1092-6.

17. Darman RB, Seiler M, Agrawal AA, Lim KH, Peng S, Aird D, Bailey SL, Bhavsar EB, Chan B, Colla S, Corson L, Feala J, Fekkes P, Ichikawa K, Keaney GF, Lee L, Kumar P, Kunii K, MacKenzie C, Matijevic M, Mizui Y, Myint K, Park ES, Puyang X, Selvaraj A, Thomas MP, Tsai J, Wang JY, Warmuth M, Yang H, Zhu P, Garcia-Manero G, Furman RR, Yu L, Smith PG, Buonamici S. 2015. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point. Cell Rep 13:1033-45.
18. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14:708-15.
19. NCBIResourceCoordinators. 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7-D19.
20. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-9.
21. EncodeProjectConsortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.
22. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.
23. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222-30.
24. Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK. 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43:D345-56.
25. Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12:656-64.
26. Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433-4.
27. Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems:1695.
28. GeneOntologyConsortium. 2015. Gene Ontology Consortium: going forward. Nucleic

Acids Res 43:D1049-56.
29. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. 2014. Mass-spectrometry-based draft of the human proteome. Nature 509:582-7.
30. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, et al. 2014. A draft map of the human proteome. Nature 509:575-81.
31. Touw WG, Baakman C, Black J, te Beek TA, Krieger E, Joosten RP, Vriend G. 2015. A series of PDB-related databanks for everyday needs. Nucleic Acids Res 43:D364-8.
32. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M. 2015. The BioGRID interaction database: 2015 update. Nucleic Acids Res 43:D470-8.
33. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. 2010. CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 38:D497-501.
34. Pu S, Wong J, Turner B, Cho E, Wodak SJ. 2009. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 37:825-31.

Supplemental Figure Legends

Supplemental Figure 1. sgRNA read count distributions at each time point in the NALM-6 screen. Read count distribution for day 0 (after 6 days blasticidin selection and before Cas9 induction) is shown in each panel. All other read count distributions were normalized by a constant factor to the day 0 average to allow overlay comparisons. Doxycycline induction of Cas9 was initiated at day 0 and terminated after day 7, with subsequent outgrowth in the absence of doxycycline. Cells were diluted to $4x10^5$ cells per mL every other day. **(A)** Input

plasmid library. **(B)** Cas9 induction for 3 days. **(C)** Cas9 induction for 7 days. **(D)** Outgrowth for 4 days after completion of Cas9 induction (day 11). **(E)** Outgrowth for 8 days after completion of Cas9 induction (day 15). **(F)** Outgrowth for 14 days after completion of Cas9 induction (day 21). **(G)** Comparison of gene-trap scores from a HAP1 cell line screen (16) for non-overlapping genes from the top 2,000 ranked genes in the NALM-6 screen at day 15 and at day 21. **(H)** Comparison of gene-trap scores from a HAP1 cell line screen (16) for non-overlapping genes from the top 2,000 ranked genes in the NALM-6 screen at day 7 and at day 15. **(I)** Gene ontology enrichment (red)/depletion (blue) analysis in the top scoring 1,000 essential genes detected at day 7 as compared to the top 1,000 essential genes detected at day 15.

Supplemental Figure 2. Validation of NALM-6 screen and specific screen hits. **(A)** For each of the 10 independent CRISPR/Cas9 screens analyzed in this study, the fraction of non-essential genes essential in at least one other cell line (gene dropout rate) was plotted as a function of essentiality across the other 9 lines (number of screens). The NALM-6 screen had the highest inclusion rate of highly cross-validated essential genes. **(B)** Network graph representation of interactions for UBALD1 and C19orf53 as determined by the BioID proximity labeling method. UE genes are indicated in green, CE genes in blue, LE genes in red and NE genes in black. Small nodes represent connected proteins from other studies. Genes implicated in ribosome biogenesis are diamond shaped. Genes in the indicated main enriched functional category for each protein are circled. **(C)** Four NALM-6 specific LE genes identified in the EKO library screen were each targeted by two independent sgRNAs in NALM-6 cells. Effects on proliferation were assessed after 6 days of puromycin selection and compared to two independent non-targeting sgRNAs (AAVS1 and Azami-Green). **(D)** Three essential genes that contained an alternatively spliced exon that scored as non-essential in the NALM-6 screen were each targeted by two sgRNAs against a constitutive and an alternatively spliced exons and assessed as in panel C. **(E)** Three hypothetical ORFs that scored as essential in the NALM-6 screen were assessed as in panel C.

Supplemental Figure 3. Comparison of RANKS to other CRISPR/Cas9 screen scoring systems. Non-overlapping genes from the top 2,000 most essential genes were compared for each ranking method at the gene level. **(A)** Number of different cell lines in which a gene was scored as essential based on nine cell lines screened by CRISPR/Cas9 methodology (1, 14) and one cell line by gene-trap (16). **(B)** Mutation rate determined as fraction of aligned residues differing from the human sequence across 45 vertebrate species in a 46-way Multi-Z whole-

genome alignment. **(C)** Gene trap scores in HAP1 cells expressed as the ratio of gene-disrupting to non-disrupting intronic insertions per gene (16). **(D)** mRNA expression levels as measured by $\log_2$(reads/kb) from RNA-seq in NALM-6 cells (17). In all instances, p-values were calculated with the Wilcoxon rank sum test.

Supplemental Figure 4. Gene features by sgRNA depletion rank. Targeted RefSeq genes ranked from most depleted to the least were binned into groups of 2,000 genes and assessed for the indicated features. **(A)** Mutation rate as determined by fraction of aligned residues differing from the human sequence across 45 vertebrate species in 46-way Multi-Z whole-genome alignment. **(B)** Number of protein-protein interaction partners reported in the BioGRID database (3.4.133 release) (32). **(C)** Log_2 RNA-seq reads per million (RPM) in the NALM-6 cell line. **(D)** Gene-trap score in HAP1 cell line screen (16) calculated as the ratio of sense to anti-sense intronic insertions of the splice-disrupting gene trap insert. **(E)** Gene-level sgRNA depletion score for KBM7 haploid cell line (1) based on average $\log_2$ read frequency change of sgRNAs targeting each gene. **(F)** Number of DNAse I hypersensitivity read peaks per kbp in naïve B-cells from ENCODE (21). The histograms shown are the basis for the summary panel shown in Figure 2.

Supplemental Figure 5. Properties of essential genes targeted by sgRNAs with 1 or 2 potential off-target cleavage sites as compared to other essential genes. **(A)** Gene trap scores in HAP1 cell line screen (16). **(B)** Mutation rate as determined by fraction of aligned residues differing from the human sequence across 45 vertebrate species in the 46-way MultiZ whole-genome alignment.

Supplemental Figure 6. Distribution of essential genes across cell lines. **(A)** Distribution of gene rank by significance scores in individual screens of essential and non-essential genes as a function of the number of lines in which each gene was essential. Blue bins show non-essential gene rank (FDR>0.05) and red bins essential gene rank (FDR>0.05). White circles represent the median rank of each bin and error bars show the 5% and 95% percentile gene rank. **(B)** Number of genes scored as essential in a given number of screens as a function of different gene rank-based essentiality thresholds. **(C)** Number of genes with or without a paralog of >30% sequence identity scored as essential in a given number of screens. **(D)** Evaluation of different models to account for essential gene distribution across different cell lines. Fitted values for each of the three models, corresponding to the probability of being essential in any

given cell line for the top 5,000 genes, ranked in decreasing order.

Supplemental Figure 7. Restriction of functional diversity as essentiality becomes distributed over more cell lines. The percentage of genes essential in 1 to 3, 4 to 7 or 8 to 10 cell lines as annotated with the indicated GO Slim biological process (28).

Supplemental Figure 8. Clustering of cell lines by essential genes versus essential protein complexes. **(A)** Fraction of protein complexes that contain essential subunits (blue) and the fraction of essential proteins that are present in protein complexes (red). Human complexes were from the CORUM database (33) and yeast complexes from the CYC2008 dataset (34). Human values were an average across the 10 cell lines analyzed in this study for either all essential genes or only CE and UE genes. **(B)** Fraction of protein complexes that contain essential subunits (blue) and the fraction of essential proteins that are present in protein complexes (red) for each cell line analyzed in this study. **(C)** Cell lines clustered by the number of shared essential complexes, defined as containing at least one essential subunit in at least one cell line. **(D)** Cell lines clustered by the number of shared essential genes identified in CRISPR/Cas9 screens.

Supplemental Figure 9. Properties of proteins encoded by genes that were uniquely essential to the NALM-6 cell line. **(A)** Interaction degree of NALM-6 specific essential proteins (red) compared to universal essential proteins (blue) as a function of mRNA expression level. **(B)** Interaction degree of NALM-6 specific essential proteins compared to universal essential proteins and non-essential proteins. **(C)** Number of interaction of NALM-6 specific essential proteins with universal essential proteins as a function of mRNA expression level. **(D)** Interactions of NALM-6 specific essential proteins with universal essential proteins, as compared to interactions of non-essential proteins and self-interactions of universal proteins. All p-values were calculated with the Wilcoxon rank sum test.

Supplemental Tables

Supplemental Table S1. List of sgRNA sequences in the EKO library, the targeted genomic region, the targeted features and the number of potential cleavage sites, and read counts for each time point in the NALM-6 screen.

Supplemental Table S2. Genes targeted by the EKO library, and when available, RANKS

score, corrected RANKS score, FDR values, MAGeCK score, BAGEL score, average sgRNA $\log_2$ fold-change, gene targeting in other libraries, and the number of cell lines in which the gene scored as essential.

Supplemental Table S3. Universal essential, NALM-6 specific lone essential and non-RefSeq essential genes identified in NALM-6 screen with the EKO library.

Supplemental Table S4. Universal essential, contextual essential, lone essential and non-essential subunit composition of human protein complexes with at least one essential subunit.

Supplemental Table S5. Scores for all alternatively spliced exons in NALM-6 screen with the EKO library with FDR values for exons in essential genes.

Supplemental Table S6. Oligonucleotides used in this study.

Supplemental Table S7. Significant protein interactions detected by BioID method.

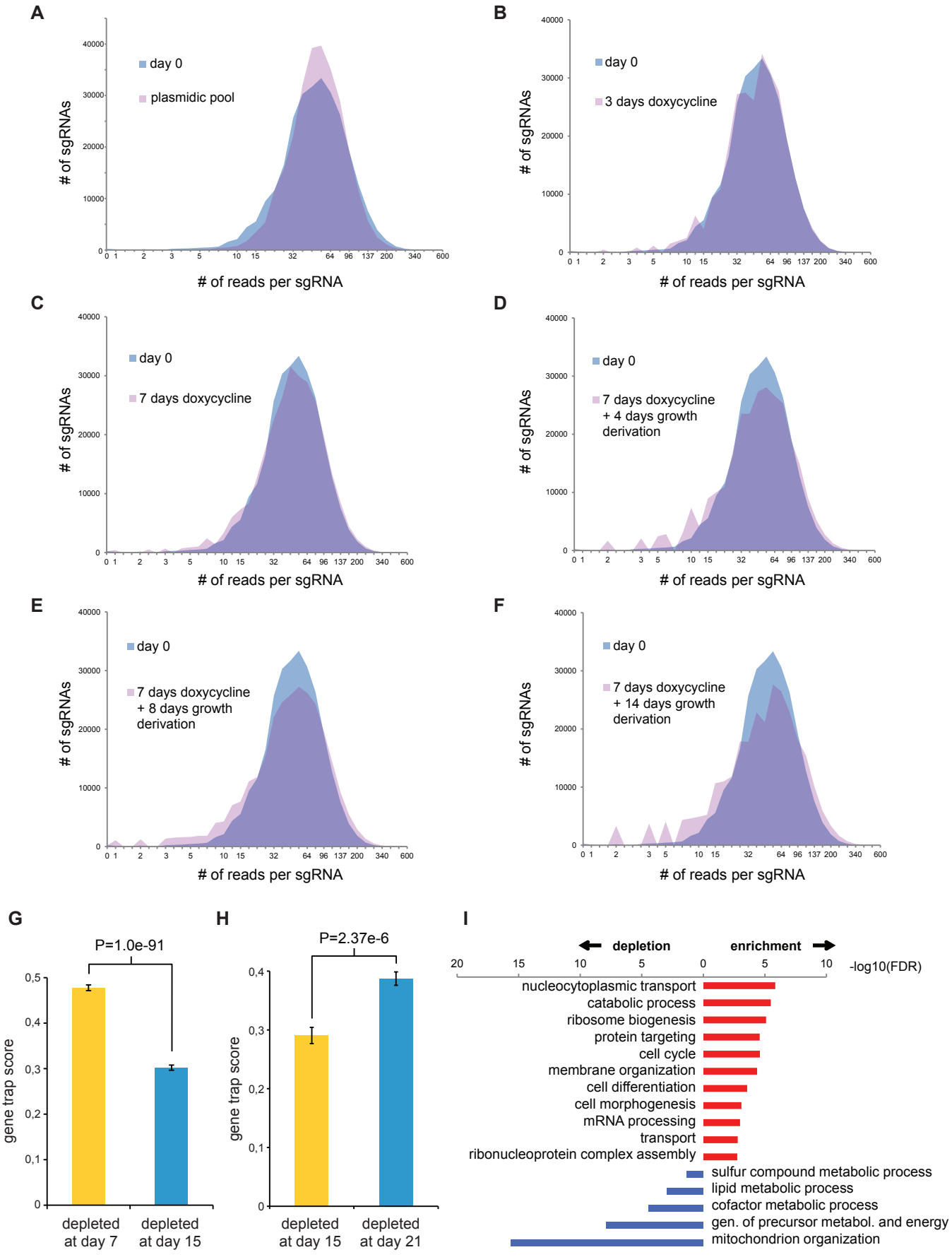
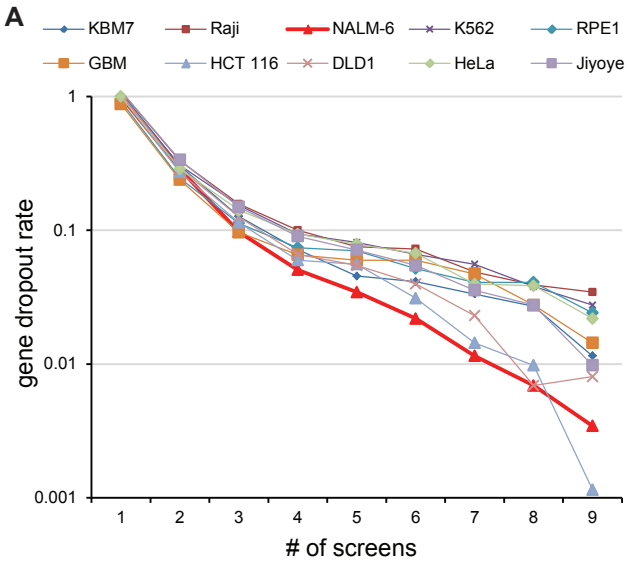
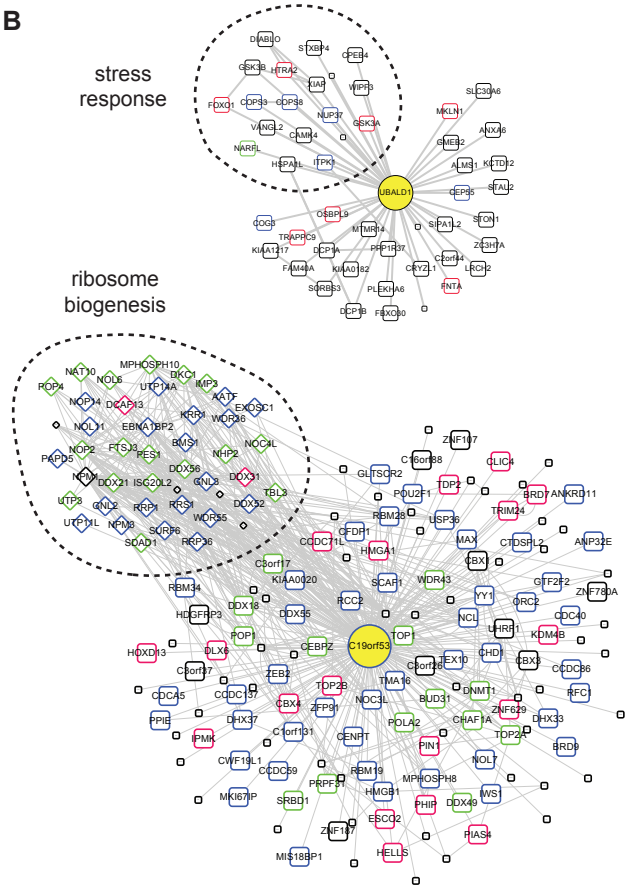
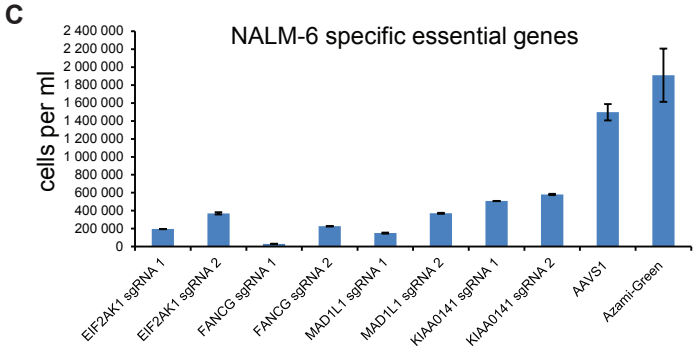
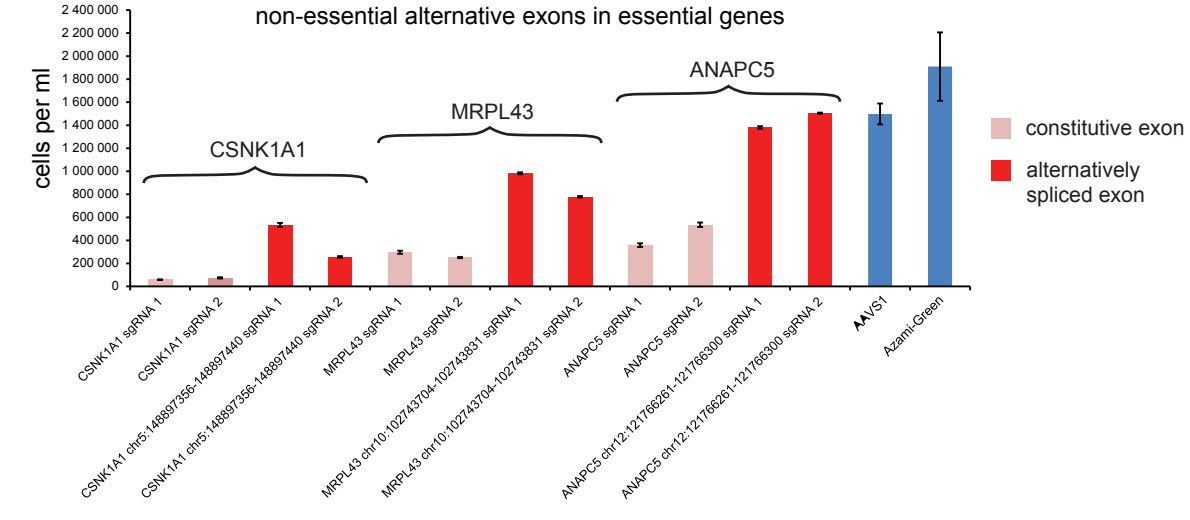

A
day 0
plasmidic pool
# of sgRNAs
# of reads per sgRNA
B
day 0
3 days doxycycline
C
day 0
7 days doxycycline
D
day 0
7 days doxycycline
+ 4 days growth
derivation
E
day 0
7 days doxycycline
+ 8 days growth
derivation
F
day 0
7 days doxycycline
+ 14 days growth
derivation
G
P=1.0e-91
gene trap score
depleted
at day 7
depleted
at day 15
H
P=2.37e-6
depleted
at day 15
depleted
at day 21
I
depletion
enrichment
-log10(FDR)
nucleocytoplasmic transport
catabolic process
ribosome biogenesis
protein targeting
cell cycle
membrane organization
cell differentiation
cell morphogenesis
mRNA processing
transport
ribonucleoprotein complex assembly
sulfur compound metabolic process
lipid metabolic process
cofactor metabolic process
gen. of precursor metabol. and energy
mitochondrion organization

Figure S1


A
KBM7
Raji
NALM-6
K562
RPE1
GBM
HCT 116
DLD1
HeLa
Jiyoye
gene dropout rate
1
0.1
0.01
0.001
1
2
3
4
5
6
7
8
9
# of screens


B
stress response
ribosome biogenesis


C
NALM-6 specific essential genes
cells per ml
2 400 000
2 200 000
2 000 000
1 800 000
1 600 000
1 400 000
1 200 000
1 000 000
800 000
600 000
400 000
200 000
0
EIF2AK1 sgRNA 1
EIF2AK1 sgRNA 2
FANCG sgRNA 1
FANCG sgRNA 2
MAD1L1 sgRNA 1
MAD1L1 sgRNA 2
KIAA0141 sgRNA 1
KIAA0141 sgRNA 2
AAVS1
Azami-Green


D
non-essential alternative exons in essential genes
cells per ml
CSNK1A1
MRPL43
ANAPC5
constitutive exon
alternatively spliced exon
CSNK1A1 sgRNA 1
CSNK1A1 sgRNA 2
CSNK1A1 chr5:148897356-148897440 sgRNA 1
CSNK1A1 chr5:148897356-148897440 sgRNA 2
MRPL43 sgRNA 1
MRPL43 sgRNA 2
MRPL43 chr10:102743704-102743831 sgRNA 1
MRPL43 chr10:102743704-102743831 sgRNA 2
ANAPC5 sgRNA 1
ANAPC5 sgRNA 2
ANAPC5 chr12:121766261-121766300 sgRNA 1
ANAPC5 chr12:121766261-121766300 sgRNA 2
AAVS1
Azami-Green

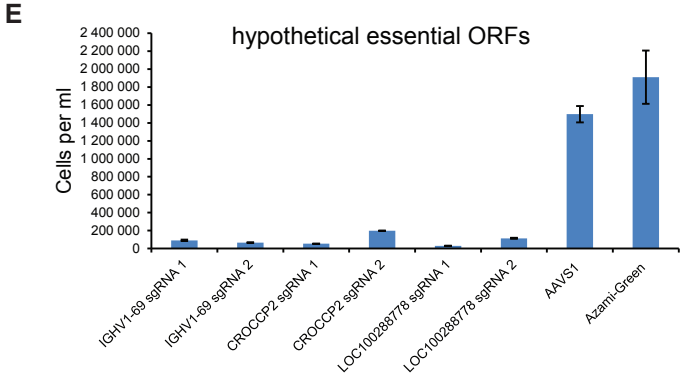

E
hypothetical essential ORFs
Cells per ml
IGHV1-69 sgRNA 1
IGHV1-69 sgRNA 2
CROCCP2 sgRNA 1
CROCCP2 sgRNA 2
LOC100288778 sgRNA 1
LOC100288778 sgRNA 2
AAVS1
Azami-Green

Figure S2

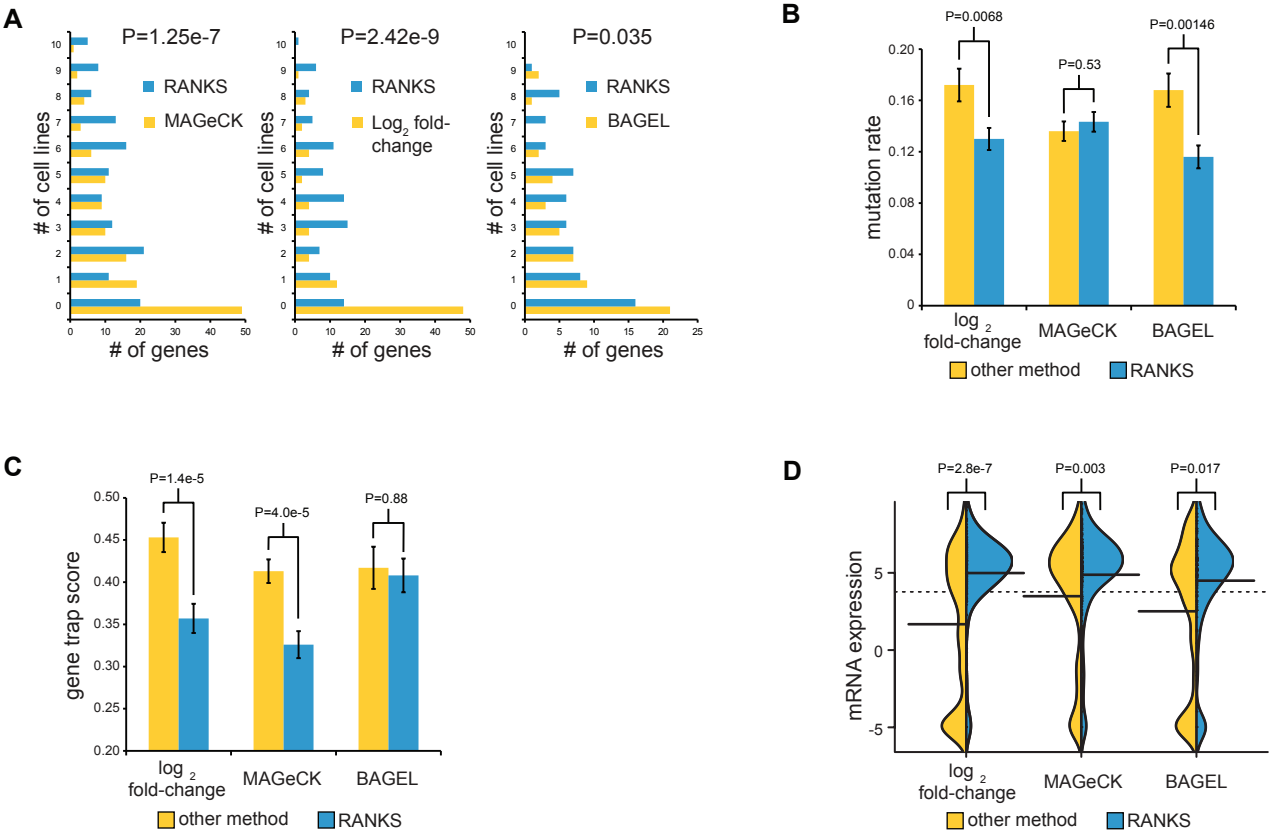

A
P=1.25e-7
RANKS
MAGeCK
# of cell lines
# of genes
P=2.42e-9
RANKS
Log2 fold-change
# of cell lines
# of genes
P=0.035
RANKS
BAGEL
# of cell lines
# of genes
B
P=0.0068
P=0.53
P=0.00146
mutation rate
log 2 fold-change
MAGeCK
BAGEL
other method
RANKS
C
P=1.4e-5
P=4.0e-5
P=0.88
gene trap score
log 2 fold-change
MAGeCK
BAGEL
other method
RANKS
D
P=2.8e-7
P=0.003
P=0.017
mRNA expression
log 2 fold-change
MAGeCK
BAGEL
other method
RANKS

Figure S3

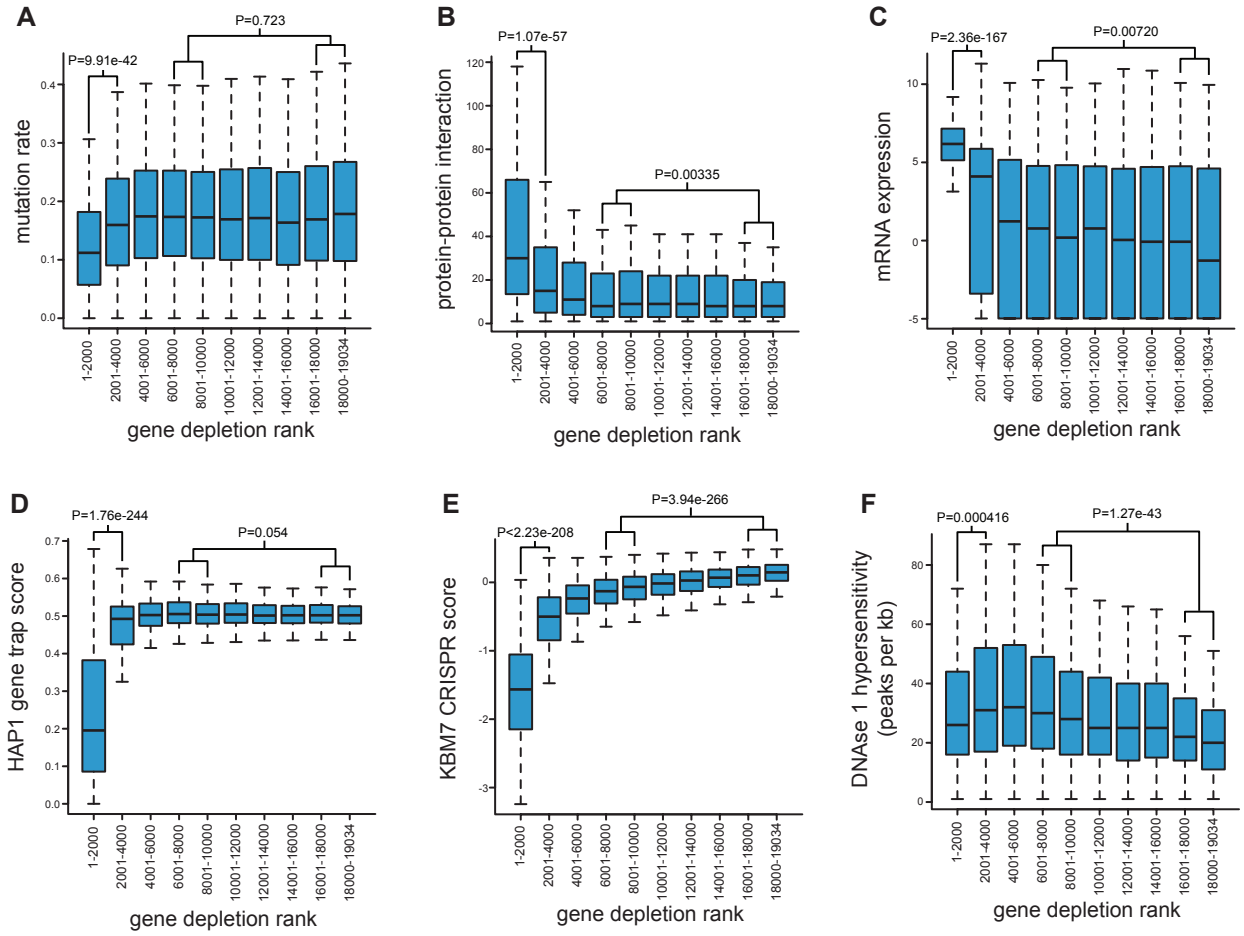

A
P=9.91e-42
P=0.723
mutation rate
gene depletion rank
B
P=1.07e-57
P=0.00335
protein-protein interaction
gene depletion rank
C
P=2.36e-167
P=0.00720
mRNA expression
gene depletion rank
D
P=1.76e-244
P=0.054
HAP1 gene trap score
gene depletion rank
E
P<2.23e-208
P=3.94e-266
KBM7 CRISPR score
gene depletion rank
F
P=0.000416
P=1.27e-43
DNAse 1 hypersensitivity (peaks per kb)
gene depletion rank
1-2000
2001-4000
4001-6000
6001-8000
8001-10000
10001-12000
12001-14000
14001-16000
16001-18000
18000-19034

Figure S4

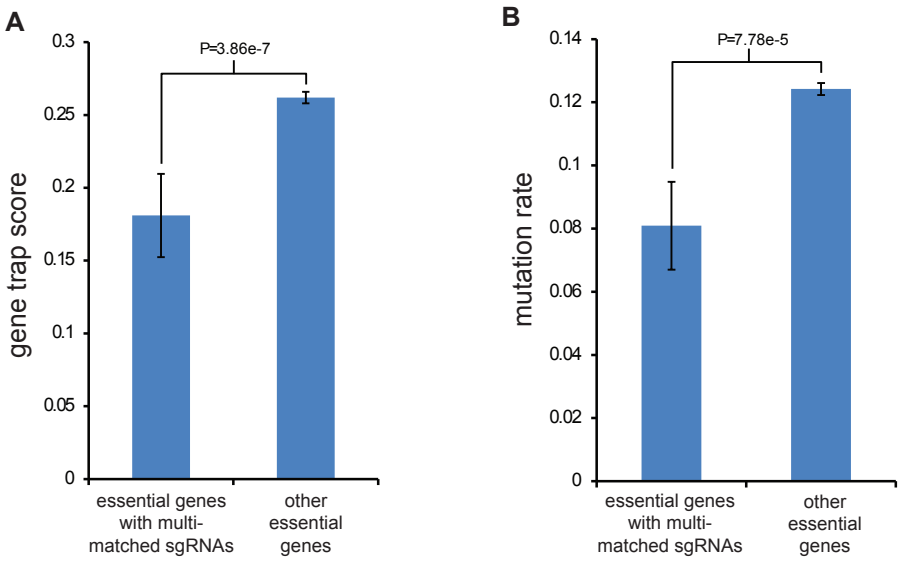

A
P=3.86e-7
gene trap score
0.3
0.25
0.2
0.15
0.1
0.05
0
essential genes with multi-matched sgRNAs
other essential genes
B
P=7.78e-5
mutation rate
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
essential genes with multi-matched sgRNAs
other essential genes

Figure S5

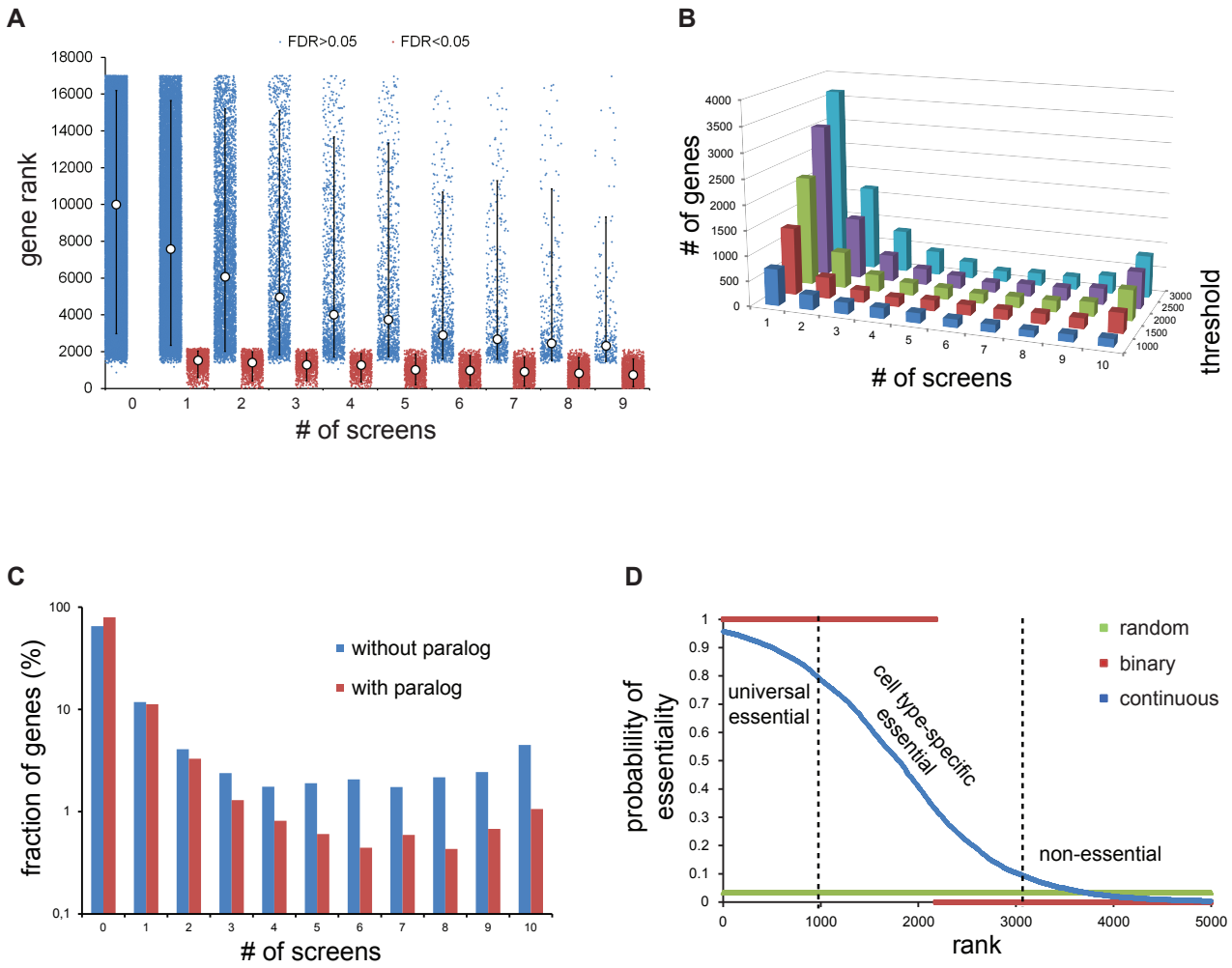

A
FDR>0.05
FDR<0.05
gene rank
# of screens
B
# of genes
# of screens
threshold
C
without paralog
with paralog
fraction of genes (%)
# of screens
D
random
binary
continuous
probability of essentiality
universal essential
cell type-specific essential
non-essential
rank

Figure S6

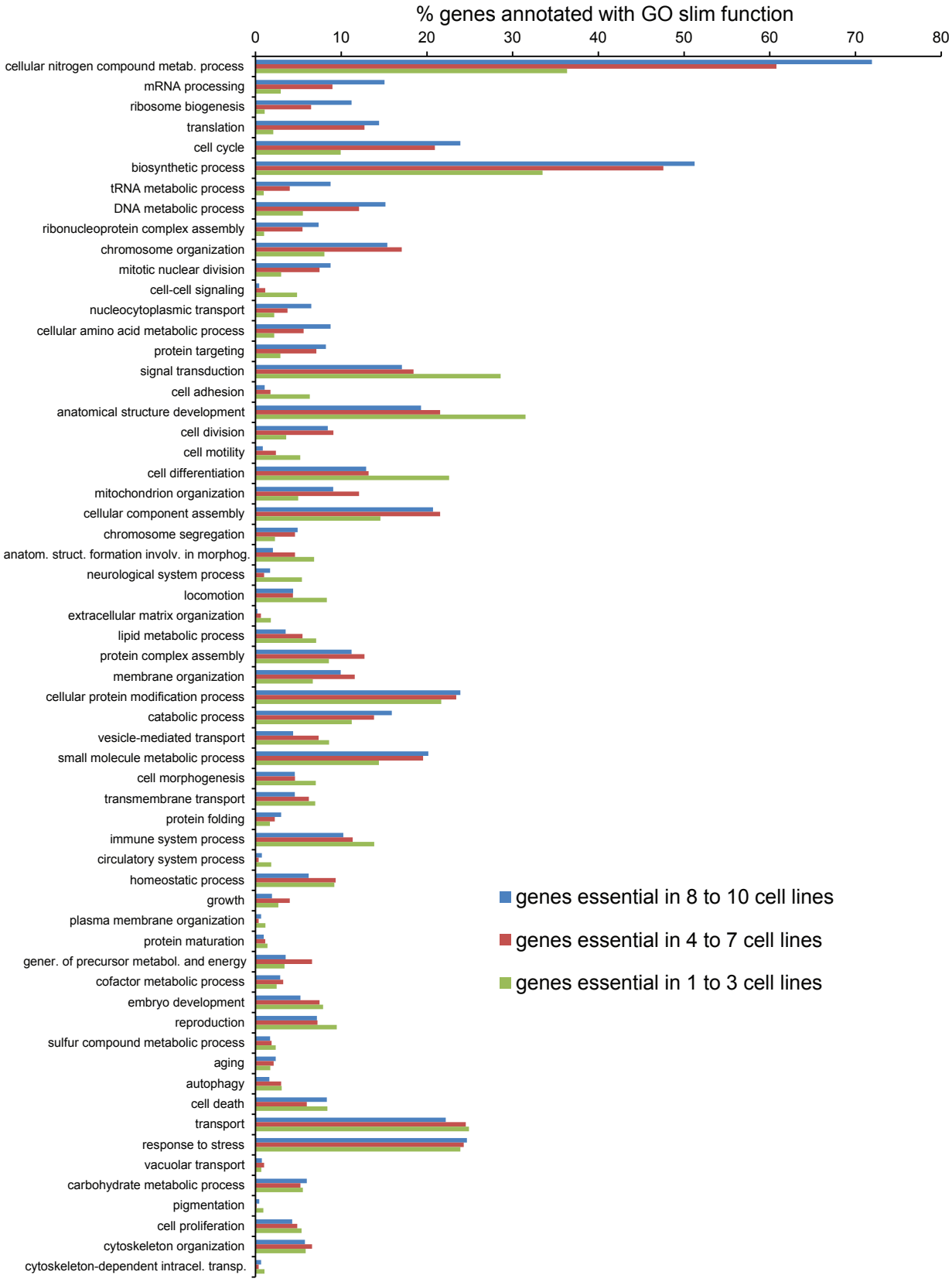

% genes annotated with GO slim function
0
10
20
30
40
50
60
70
80
cellular nitrogen compound metab. process
mRNA processing
ribosome biogenesis
translation
cell cycle
biosynthetic process
tRNA metabolic process
DNA metabolic process
ribonucleoprotein complex assembly
chromosome organization
mitotic nuclear division
cell-cell signaling
nucleocytoplasmic transport
cellular amino acid metabolic process
protein targeting
signal transduction
cell adhesion
anatomical structure development
cell division
cell motility
cell differentiation
mitochondrion organization
cellular component assembly
chromosome segregation
anatom. struct. formation involv. in morphog.
neurological system process
locomotion
extracellular matrix organization
lipid metabolic process
protein complex assembly
membrane organization
cellular protein modification process
catabolic process
vesicle-mediated transport
small molecule metabolic process
cell morphogenesis
transmembrane transport
protein folding
immune system process
circulatory system process
homeostatic process
growth
plasma membrane organization
protein maturation
gener. of precursor metabol. and energy
cofactor metabolic process
embryo development
reproduction
sulfur compound metabolic process
aging
autophagy
cell death
transport
response to stress
vacuolar transport
carbohydrate metabolic process
pigmentation
cell proliferation
cytoskeleton organization
cytoskeleton-dependent intracel. transp.
genes essential in 8 to 10 cell lines
genes essential in 4 to 7 cell lines
genes essential in 1 to 3 cell lines

Figure S7

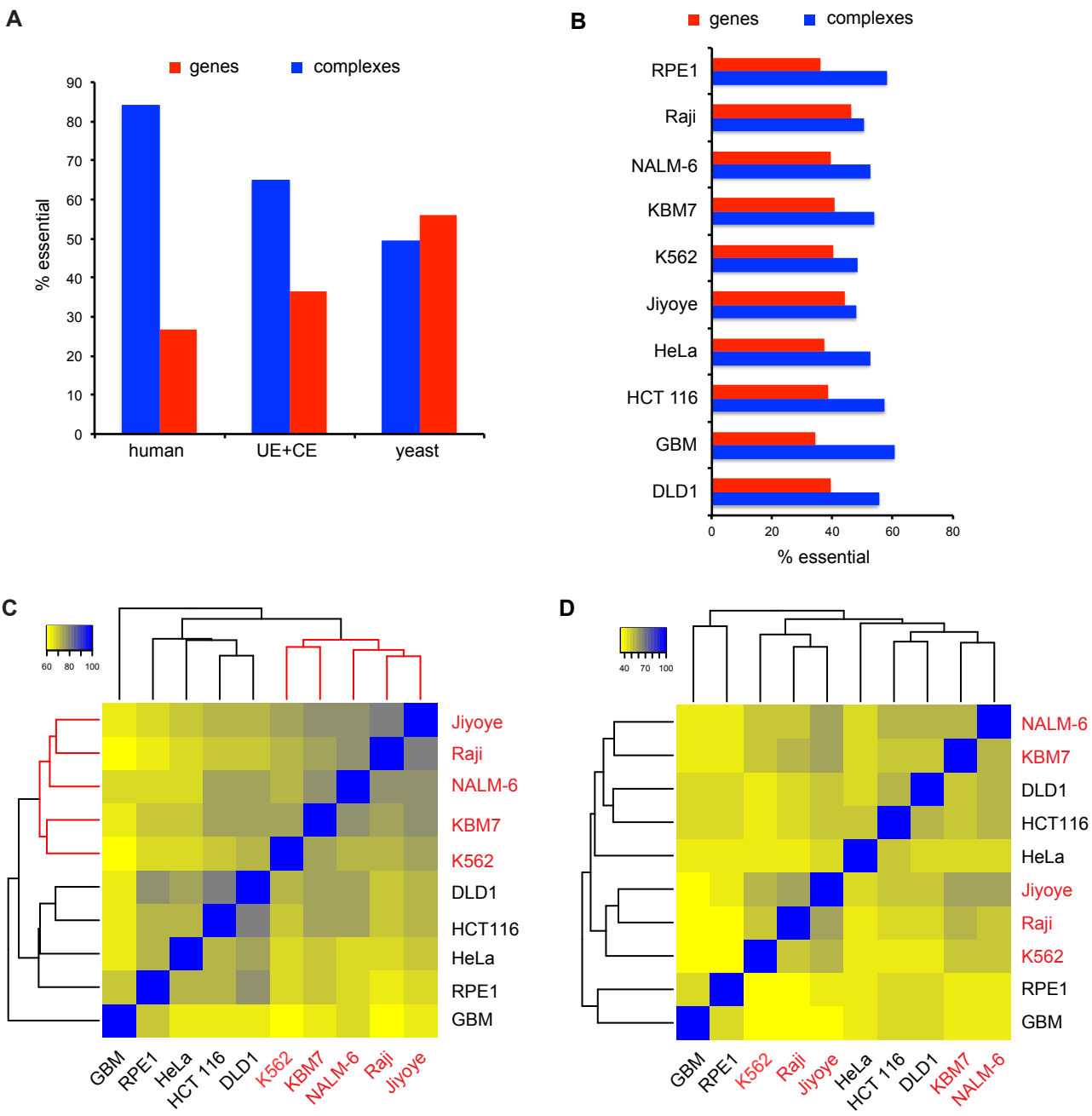

A
genes
complexes
% essential
human
UE+CE
yeast
B
genes
complexes
RPE1
Raji
NALM-6
KBM7
K562
Jiyoye
HeLa
HCT 116
GBM
DLD1
% essential
C
Jiyoye
Raji
NALM-6
KBM7
K562
DLD1
HCT116
HeLa
RPE1
GBM
D
NALM-6
KBM7
DLD1
HCT116
HeLa
Jiyoye
Raji
K562
RPE1
GBM

Figure S8

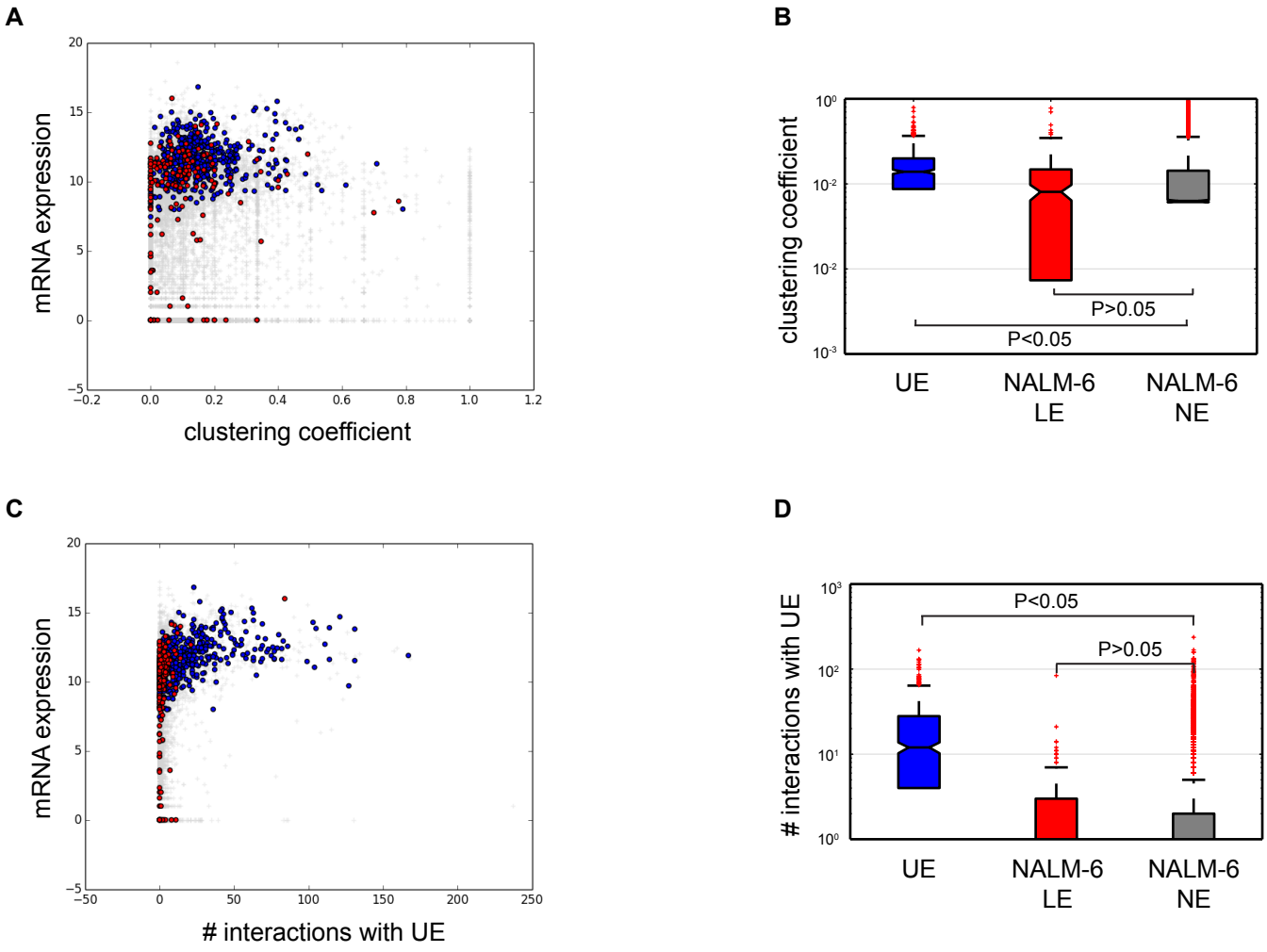

A
mRNA expression
clustering coefficient
B
clustering coefficient
P>0.05
P<0.05
UE
NALM-6 LE
NALM-6 NE
C
mRNA expression
interactions with UE
D
interactions with UE
P<0.05
P>0.05
UE
NALM-6 LE
NALM-6 NE

Figure S9