CRISPR/Cas9 screening using unique molecular identifiers

Bernhard Schmierer[1,#], Sandeep K. Botla[1,#], Jilin Zhang[1], Mikko Turunen[2], Teemu Kivioja[2] and Jussi Taipale[1,2,*]

[1]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden.
[2]Genome-Scale Biology Research Program, Faculty of Medicine, University of Helsinki, PO Box 63 FI-00014 Helsinki, Finland.

* corresponding author
# equal contribution

Loss of function screening by CRISPR/Cas9 gene knockout with pooled, lentiviral guide libraries is a widely applicable method for systematic identification of genes contributing to diverse cellular phenotypes. Here, random sequence labels (RSLs) were incorporated into the guide-library. RSLs function as internal replicates for robust and reproducible hit calling, and act as unique molecular identifiers (UMIs) to allow massively parallel lineage tracing (MPLT) and true dropout screening.

Pooled CRISPR/Cas9 loss of function screening is a powerful approach to identify genes contributing to a wide range of phenotypes. Most commonly, a library of guide sequences is transduced lentivirally into a population of Cas9-expressing cells, which are then subjected to some form of selection pressure. Relative guide frequencies in the genome of the population before and after selection are quantified by next generation sequencing (NGS).

The approach has been applied successfully, but suffers from two major shortcomings: First, the presence of a guide in a cell does not necessarily lead to loss of function of the corresponding gene, as the total read count for a guide reflects cells with distinct genotypes, which are the result of mono- or bi-allelic frameshifts or in-frame deletions[1] (**Supplementary Fig. 1a**), as well as off target effects. These different genotypes also result in a range of phenotypes for the cells, and optimal identification of hit genes ideally requires a method that would individually track each clonal cell lineage derived from a single editing event. Secondly, identification of fitness genes whose guides are under negative selection can be statistically challenging, because of confounders such as random drift or undersampling.

Finally, creating a sufficient number of replicates in this type of experiment is labor intensive and costly. As a consequence, the read count variance in CRISPR/Cas9 pooled screening is commonly estimated globally from a single data set[2, 3], similar to statistical methods developed for RNA-Seq and ChIP-Seq[4, 5]. This is only valid if the vast majority of guides lack a detectable effect, which might not always be the case, for instance in smaller, targeted libraries. With an insufficient number of replicates, outliers tend to be called as hits and technical artefacts such as PCR bias or other random effects cannot be distinguished from real biological effects.

To address these issues, we have developed a method that allows tracing of individual virus-transduced cell lineages during a CRISPR/Cas9 screen. Depending on the kinetics of genome editing, we can either follow up single clones of identically edited cells, or small populations of sublineages with

different editing outcomes at the same locus (**Supplementary Fig. 1b**). To allow tracking individual viral integrants, we incorporated a random sequence label (RSL) in the guide-library plasmid. This creates "guide-sets" for each particular guide sequence consisting of thousands of distinguishable "RSL-guides" (**Fig. 1a**). These unique molecular identifiers (UMIs)[6] trace the fate of thousands of distinct cell lineages per guide sequence in a massively parallel fashion. The method enables "lineage dropout" screening and the creation of internal replicates, which facilitates data analysis and greatly improves significance and reproducibility of hit calling, while retaining the option of pooled, conventional analysis (**Fig. 1b**). To demonstrate the power and flexibility of the approach, we carried out a gene essentiality screen in the human colorectal carcinoma cell line RKO using a guide library targeting 2325 genes with 10 guides per gene[1], including all human transcription factors[7], essential ribosomal proteins[8, 9], other genes of interest and 101 non-targeting guides as negative controls.

For experimental details see **Methods**. Briefly, a PCR library product containing the guide sequences and the RSLs was cloned into a lentiviral vector, and the plasmid library was sequenced (Input). Libraries were packaged into lentiviruses and Cas9-expressing RKO cells were transduced. Cells were then propagated, genomic DNA was sampled at Day 4 (control) and Day 28 after transduction, and the relative frequencies of integrated guide sequences were determined by NGS.

The input plasmid library contained on average 3600 RSLs per guide. Sequencing analysis revealed that 93% of the RSL-guides present in the plasmid library were incorporated into the target cells (**Fig. 1c, right**). Based on the Poisson-distribution, this indicates that approximately half of the RSL-guides were incorporated to one or two original cell lineages. This also highlights the convenience of RSLs in tracking library representation. Because only a subset of the cells can be harvested at each time-point, undersampling is unavoidable, and some RSL-guides and thus cell lineages were present only in one of the time points (**Fig. 1c, right**). It should be emphasized that such undersampling and loss of cell lineages occurs whether or not RSLs are present. In the presence of RSLs, the effect becomes apparent and can be used in quality control of individual experiments. Where necessary, the RSL data can also be filtered to remove inconsistently sampled lineages prior to data analysis.

RSL-guide counts can be used to create internal replicates. Using the current state-of-the-art analysis package MAGeCK[3], the data was analyzed in the conventional way with a single read count per guide-set (as if the RSL had not been sequenced); and by binning the read counts of RSL-guides into 16 internal replicates (**Fig. 1d**). As a quality measure, we used replicate concordance between two experimental replicates, as well as the behavior of a set of guides targeting positive control genes and non-targeting control guides. Guide score correlation between experimental replicates was much higher in the 16 replicate analysis compared to the pooled analysis (**Fig. 1d**). More hit genes were found (262 in internal replicate- versus 251 in pooled analysis, false discovery rate (FDR) < 1%), and the hit gene overlap between experimental replicates was improved (172 versus 156 genes found in both replicates at FDR<1%, **Fig. 1e**). Thus, RSLs enable creation of internal replicates, which improves both statistical significance and reproducibility of hit calling by the current state-of-the-art analysis method. This result also suggests that even a minimal design with a low number of RSLs (10 – 20 per guide) might already be advantageous.

Internal replicates allow guide-set specific estimation of read count variance, and mean effect sizes can now be compared to non-targeting control guide-sets. Thus, in the presence of RSLs, classical

statistical tools such as t-statistics, z-factor, or strictly standardized mean difference (SSMD)[10, 11] become available to test for significant differences. To illustrate the approach, we binned RSL-guide counts for each guide-set to create 64 replicates and calculated a median effect size (MES) for each guide-set (median log2 ratio of all replicates in the set, **Fig 2a**). For statistical significance, a robust, median-based version of SSMD was used. Guide-sets were ranked according to a score defined as the product of MES and SSMD (**Supplementary Fig. 1c)**. For details, see **Methods**. Results obtained by this simple and intuitive method were comparable to those obtained using the state-of-the-art pipeline.

Finally, to investigate the lineage dropout approach, we calculated the average number of dropped out RSL-guides per gene between days 4 and 28, and called hits based on this data. Hit gene overlap was at least as good as with the other methods (**Fig. 2b, inset**). Concordance of the gene ranks between replicates in this analysis was higher than that observed in any other type of analysis (**Fig. 2b**, **c**). Furthermore, the mean rank of known positive control genes (20 ribosomal proteins; **Fig. 2d**) was lower than that found using the other analyses, suggesting higher precision as well as higher accuracy of true lineage dropout compared to methods based on read-count changes.

To summarize, including RSLs into the guide-vector allows analysis of multiple internal replicates, and more accurate hit calling based solely on loss of virus-transduced cell lineages. RSLs empower simple and robust hit calling with increased reproducibility, and allow efficient quality control of screens and the identification of outliers. The RSL strategy is not limited to CRISPR knockout screening, but can be applied in other screening methods such as CRISPR-dependent inhibition or activation screens. We expect that the RSL method will be particularly useful in the interrogation of genomic features that are small, e.g. exons, promoters, and even individual transcription factor binding sites. In many of these cases there is just one possible guide sequence, and the inclusion of RSLs is the only way to obtain the replicates that are required for hit calling. In the absence of precise knowledge of both on- and off-target activity, inclusion of multiple guide positions is however still important, and rescue experiments and/or analysis of the mutational spectrum of the cutsite are necessary to establish that the mutation induced by the guide results in the observed phenotype.

Methods

Oligo synthesis and library cloning. All sgRNA sequences used in this library were taken from a previously published, genome-wide library[1] (Supplementary file *RSL_guide_library.csv*). The 5' part of the library construct (blue + black, 122bp), containing the sgRNA was synthesized by oligo array (CustomArray). The 3' part containing the RSL and the Illumina i7 index primer sequence (green) was synthesized as a single 119 bp oligo (red + green + black). These two oligos were annealed to each other at the overlapping part (black) and double stranded by PCR (T_M = 64C) using outer primers (underlined).

GGCTTTATATATCTTGTGGAAAGGACGAAACACCGNNNNNNNNNNNNNNNNNNNNgtttAagagctagaaa
tagcaagttTaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTTgatcg
gaagagcacacgtctgaactccagtcacNNNNNNaagcttggcgtaactagatcttgagacaaa

The PCR product was cloned by Gibson assembly into the lentiviral vector pLenti-Puro-AU-flip-3xBsMBI, which was created by modifying lentiGuide-Puro (a gift from Feng Zhang, Addgene #52963) by replacing the sequence

gttttagagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTT

with

gtttAagagctagaaatagcaagttTaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTTcgtctct).

Gibson assembly, transformation and amplification of the library. 100 ng vector and 12 ng insert where assembled in a total reaction volume of 100 µl (NEBuilder® HiFi DNA Assembly Master Mix, NEB). The reaction was cleaned via a Minelute reaction cleanup column (Qiagen) and transformed into 6 x 50 µl electrocompetent *E. coli* (Endura™ ElectroCompetent Cells, Lucigen) using a 1.0 mm cuvette, 25µF, 400 Ohms, 1800 Volts. Bacteria were plated on several 24x24 cm agar plates and colonies were grown overnight.
Colonies were scraped into LB medium and the contained plasmids were isolated by Maxiprep.

Library packaging. The library was packaged in HEK 293T cells by cotransfecting the library plasmid and the two packaging plasmids psPAX2 (a gift from Didier Trono, Addgene #12260) and pCMV-VSV-G (a gift from Bob Weinberg, Addgene # 8454) in equimolar ratios. After 48 hours, the virus-containing supernatant was concentrated 40-fold using Lenti-X concentrator (Clontech), aliquoted for one time use and stored at -140C.

Cell lines and cell culture. All the cells used in this study were purchased directly from ATCC. Cells were regularly tested for mycoplasma using the Mycoalert detection kit (Lonza; cat# LT07-218)

Creating editing-proficient Cas9 cell lines. To rapidly generate editing-proficient cell lines, we synthesized a lentiviral construct (pLenti-Cas9-sgHPRT1) that encodes a codon optimized WT-SpCas9 that is flanked by two nuclear localization signals (derived from lenti-dCAS-VP64_Blast, a gift from Feng Zhang, Addgene #61425). In addition, the construct codes for blasticidin resistance, and carries an sgRNA against HPRT1 (GATGTGATGAAGGAGATGGG). HPRT1 loss confers resistance to the antimetabolite 6-thioguanine (6-TG). Lentivirally transduced cells were selected in 5 µg/ml Blasticidin and after one week to 10 days additionally with 5µg/ml 6-TG until control cells had died. Only cells that both express Cas9 and are editing proficient, as indicated by loss of HPRT1 function, will survive. The method allows rapid establishment of a pool of editing proficient cells. Compared to single cell clones, this method retains the genetic heterogeneity of the original cell line, avoids potential clonal effects of the particular integration site of Cas9, and greatly accelerates cell line generation. These benefits need to be weighed carefully against possible disadvantages, such as synthetic lethality with HPRT1 loss, or potential effects of the presence of a second guide in the cell.

Library transduction. A minimum of 100 million RKO Cas9 cells were transduced with the library virus. Cells were then selected for guide integration and expression by 1 µg/ml puromycin selection for 4 days. During this time, approximately 40 % of cells died, indicating a transduction efficiency of 60 % and an

estimated MOI of 1, corresponding to about 60% of the surviving cells containing a single guide and 40% containing more than one guide. Because of the vast number of RSL-guides, such passenger guides will associate with effective guides randomly and will not be significantly enriched or depleted in the population.

Cell propagation and sample preparation. Cells were kept in culture for a total of 28 days after transduction by sub-culturing them every three to four days. 100 million cells were reseeded at each split, and genomic DNA was prepared from 50 – 80 million cells at Days 4 and 28 after transduction. Day 4 after transduction was considered the control time point.

Preparation of the sequencing library from genomic DNA. The sequencing library preparation consists of 3 PCR steps, PCR1 amplifies the genomic region containing the guide sequence using the primers 1F and 1R. PCR2 and PCR3 then incorporate the Illumina adaptors with primers 2F/2R and 3F/3R, respectively. 3F contains the Illumina index for multiplexing, indicated by NNNNNN in the sequence given.

200 µg of genomic DNA (theoretically corresponding to 30 million diploid cells) were used as PCR template in 40 parallel PCR reactions (5 µg template DNA each) using KAPA HiFi HotStart polymerase (KAPA Biosystems)14 cycles were run and the reactions were pooled together. PCR2 used 5 µl of pooled PCR1 as template and was run for 19 cycles, PCR3 used 2µl of PCR2 as template and was run for another 14 cycles. The resulting product of 288 bp was gel purified and sequenced on an Illumina HiSeq 4000 instrument using a single read 20 cycles plus two 6bp index reads, where index read 1 reads the RSL and index read 2 reads the sample index.

1F GGACTATCATATGCTTACCGTAACTTGAAAGTATTTCG (59.6C)

1R CTTTAGTTTGTATGTCTGTTGCTATTATGTCTACTATTCTTTCC (59.4C)

2F TCTTTCCCTACACGACGCTCTTCCGATCtcttgtggaaaggacgaaacac (55.4C)

2R AGAAGACGGCATACGAGATctgccatttgtctcaagatctagttac (55.4)

3F (62.0C) AATGATACGGCGACCACCGAGATCTACAC NNNNNN TCTTTCCCTACACGACGCTCTTCCG

3R (61.8C) CAAGCagaagacggcatacgagatctgccatttg

The final library product was sequenced with a custom primer and the i5 and i7 index primers (underlined) by running 20+6+6 cycles on the Illumina HiSeq4000.

AATGATACGGCGACCACCGAGATCTACAC[i5]**NNNNNN**TCTTTCCCTACACGACGCTCTTCCGATCtcttgtggaaaggacgaaacacCG**NNNNNNNNNNNNNNNNNNNN**gtttAagagctagaaatagcaagttTaaataaGgctagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTTgatcggaagagcacacgtctgaactccagtcac[i7]**BBBBBB**aagcttggcgtaactagatcttgagacaaatggcagATCTCGTATGCCGTCTTCTGCTTG

Scripts used for counting RSL-guides and for binning. RSL-guides were counted in the original fastq files with the perl scripts *BatchRun-pub2.pl*, which requires the script *GuideUMI_pub2.pl*. Binning of RSL-guide counts was done using the script *countTruncatedRSLs.pl*.

SSMD analysis of read count data

Normalization. In RNASeq, methods such as median normalization are commonly preferred to total read-count normalization, mainly to compensate for the effect of a few very highly expressed genes that can take up a significant proportion of the total read count. CRISPR/Cas9 screening data are comparably well balanced and we thus chose the most basic normalisation method, total read count normalisation, to compensate for different sequencing depths. c_{ij} and t_{ij} represent the raw read counts for RSL-guide j in guide-set i for control (Day 4 after lentiviral transduction) and treatment (Day 28 after lentiviral transduction), respectively. The normalised read counts c'_{ij} and t'_{ij} are then

$$c'_{ij} = c_{ij} \frac{\sum_{ij}(c_{ij} + t_{ij})}{2\sum_{ij} c_{ij}}$$

$$t'_{ij} = t_{ij} \frac{\sum_{ij}(c_{ij} + t_{ij})}{2\sum_{ij} t_{ij}}$$

Median effect size and variability of the guide-sets. We define the effect size ES_{ij} for each RSL-guide j in guide-set i as the log2 of the fold change between treatment count and control count. To handle total loss of an RSL-guide in the treatment sample, we add a pseudo-count of 1 to all counts:

$$ES_{ij} = log_2 \frac{t'_{ij} + 1}{c'_{ij} + 1}$$

Next, we calculate the median effect size for guide set *i*, MES_i, and the median of the absolute deviations (MAD) of all RSL-guides j in guide-set i from MES_i

$$MES_i = \underset{j}{median}\, ES_{ij}$$

$$MAD_i = 1.4826\, \underset{j}{median} |ES_{ij} - MES_i|$$

The factor 1.4826 is chosen such that the MAD is approximately equal to the standard deviation under the assumption of normal distribution.

Median effect size and variability of the control guide-sets. The RSL library contains 101 non-targeting guide-sets. We calculate a single median effect size and MAD for this control set in the following way:

Median effect size of all non-targeting RSL-guides

$$MES_{CON} = \underset{ij}{median}\, ES_{ij}^{NONT}$$

Median absolute deviation of all non-targeting RSL-guides:

$$MAD_{CON} = 1.4826\, \underset{ij}{median}\, |ES_{ij}^{NONT} - MES_{CON}|$$

The strictly standardized mean difference (SSMD). SSMD is a measure for the significance of the difference in behaviour of sample *i* and the non-targeting controls. It takes into account both the effect size and the variability of the data.

$$SSMD_i = \frac{MES_i - MES_{CON}}{\sqrt{MAD_i^2 + MAD_{CON}^2}}$$

For samples with relatively small effect size, the SSMD can still become large if the spread is small. We thus introduce a score in which the effect size weighs more strongly, and which is used as a ranking parameter:

$$Score_i = MES_i\, |SSMD_i|$$

For hit calling, the average score and standard deviation were calculated for all non-targeting guide sets. The script used in these calculations is *SSMD.sh*, which calls the script R-script *SSMD.R*. Guide-sets were then ranked according to their score and the resulting ranked list was analysed with α-RRA, a robust rank aggregation algorithm as implemented in the "pathway" function of MAGeCK[3,4] using Supplementary file *RSL_guide_library.gmt.*

Figure Legends.

Figure 1.

a. **Library design. Top.** Guide library structure. The plasmid library contains an index read primer sequence (i7, green) and a random sequence label (RSL, red). These are placed downstream of the poly-T termination signal for guide-transcription, and are thus not part of the guide-RNA**. Bottom.** Structure of the sequencing library. Integrated guide sequences were amplified from genomic DNA by PCR, and adaptor sequences and sample index for multiplexing (Index) were attached in subsequent rounds of PCR. Sequencing was done with a custom primer (Seq) directly adjacent to the guide, and sample index and RSL are read in two index reads (standard illumina primers i5 and i7, respectively). The protocol requires only 32 read cycles (20 cycles followed by two 6 cycle index reads on an Illumina HiSeq4000 instrument).

b. **The presence of RSL guides allow two additional analysis methods.** In the presence of RSLs, the data can be analyzed at three different levels. First, the presence of RSLs can be ignored, this corresponds to the conventional analysis without RSLs present (Total count, left). Secondly, the readcounts of RSL-guides for a given guide-set can be binned into any number of bins, thus creating any number of internal replicates (middle). Finally, in massively parallel lineage tracing (MPLT, right), the behavior of each RSL-guide is monitored separately, which allows lineage drop out screening by comparing the number of distinct RSL-guides per guide-set at the control- and treatment time points. The bottom schematic illustrates the binning of individual sequence read counts.

c. **Number of distinct sequences carried through the experiment. Left.** Boxplot. The input plasmid library contained an average of about 3600 RSLs per guide, this number reduced to about 2800 RSLs per guide in the samples taken from the cell populations at Day 4 and Day 28. **Right.** Venn diagram. Very few (0.01%) RSL-guides are found in the samples that were not present in the input, and 93% of RSL-guides found in the input passed through virus packaging and transduction and were found in at least one of the samples. With 23,280 guides in the library, this corresponds to 78 million unique sequences in the cell population. Due to undersampling, the overlap between day 4 and day 28 was two thirds, with about one sixth of sequences found either only in day 4 or only in day 28.

d. **Conventional, total count analysis is improved by creating internal replicates.** The data was analyzed using the pipeline MAGeCK, either at the total count level ignoring RSLs (conventional method, **left**) or by splitting the total readcount into 16 replicates by binning readcounts for RSL-guides prior to analysis (**right**). The guide scores for two experimental replicates are plotted for all guides. Positive controls (guides targeting ribosomal proteins, red) as well as non-targeting control guides are indicated (blue). Guide score correlation between experimental replicates is much improved when the data is split into 16 internal replicates (Pearson correlation PC=0.53 versus

PC=0.26, respectively). The outliers in experimental replicate 2 (orange diamonds) were found to result from overrepresentation of unique RSL-guides in the outlier guide-set and are most likely due to PCR bias. Such guides are invariably called as hits if analysis is carried out at the guide-level only, but to a much lesser extent when internal replicates are used.

e. **Internal replicates increase the number of hit genes and reproducibility of hit calling.** At the gene level, usage of internal replicates increased both the number of significant hit genes, and the hit-list overlap between the two experimental replicates (FDR < 1%).

Figure 2.

a. **Internal replicates allow the application of standard statistics.** For standard statistical analysis of the dataset, RSL-guides in each guide set were binned into 64 bins. Effect sizes for each bin (log2 fold-change in normalized read counts from Day 4 to Day 28 after virus transduction) are plotted in ascending order for the different guide-sets, 10 guide-sets for MYCN (**left**) and MYC (**middle**), and 50 representative non-targeting guide-sets (**right**). The median effect size of the bins (MES) is indicated in red. The black line indicates the median effect size of all guides in the library. Note that non-targeting guide-sets tend to be slightly above this line, which might reflect a small fitness advantage of cells harboring guides that lack a target sequence (non-cutters). Guide hits were called by assessing effect size and strictly standardized mean difference (SSMD). The robust rank aggregation algorithm α-RRA[12] as implemented in the MAGeCK pipeline[3] was used to call hit genes from the obtained guide ranking. See **Methods** for details.

b. **Massively parallel lineage tracing allows hit calling by lineage dropout.** Scatter plot shows fraction of RSL-guides lost in each experimental replicate, red and blue dots indicate positive and negative controls, respectively. Inset shows overlap between hits called in each replicate at FDR < 1%. The simplest way of assessing guide effects on cell viability is by counting the number of RSL-guides in the treatment time point and compare to the number that were present in control. The black line indicates linear regression. The number of virus-transduced cell lineages that were lost is a direct indicator of the guide effect on cell viability. Note that, as in Fig. 2a, there seems to be a general fitness disadvantage in harboring a targeting guide as compared to a non-targeting guide. Targeting guides against irrelevant genomic regions might thus be a superior control in these types of screens.

c. **Comparison of the precision of the four analysis methods used.** Comparison of replicate concordance using rank correlation reveals that lineage dropout screening has the highest replicate concordance, followed by SSMD analysis.

d. **Assessment of accuracy of the different analysis methods.** The average ranks in the experimental replicates of known positive controls (20 ribosomal proteins) found by the different analysis

methods are shown as boxplots. Out of a total of 2335 genes in the library, the majority of the ribosomal protein controls rank in the top 5%. The mean rank (red line) found was lowest when the data were analyzed by lineage dropout, suggesting that this method most accurately measures the effect of loss of the respective gene functions.

Accession codes

Read data will be loaded to European Nucleotide Archive under accession PRJEB18436. Scripts will be released under public license and made available on Github.

Acknowledgements

The authors would like to thank Drs. Inderpreet Kaur Sur, Jenna Persson and Minna Taipale for comments and suggestions on the manuscript. Part of this work was carried out at Karolinska High Throughput Center (KHTC) and the High Throughput Genome Engineering Facility (HTGE) funded by SciLife Lab.

Authorship contributions.

B.S. and J.T. developed the approach, B.S., S.K.B. and M.T. performed the experiments, B.S., S.K.B, J.Z. and T.K. analyzed the data, B.S. and J.T. wrote the manuscript.

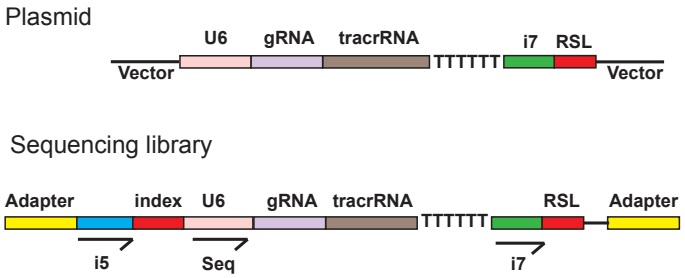
References

1. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014).
2. Li, W. et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol* **16**, 281 (2015).
3. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).
4. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
5. Anders, S. et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**, 1765-1786 (2013).
6. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72-74 (2012).
7. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**, 252-263 (2009).
8. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101 (2015).
9. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 164 (2016).

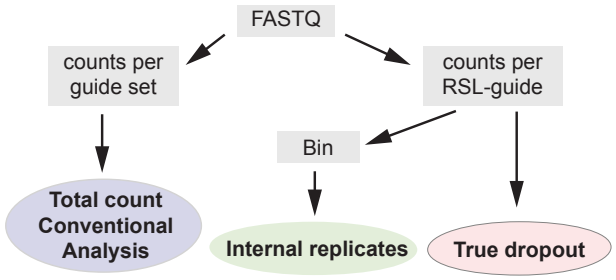
10. Zhang, X.D. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics* **89**, 552-561 (2007).
11. Zhang, X.D. et al. The use of strictly standardized mean difference for hit selection in primary RNA interference high-throughput screening experiments. *J Biomol Screen* **12**, 497-509 (2007).
12. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573-580 (2012).

Figure 1

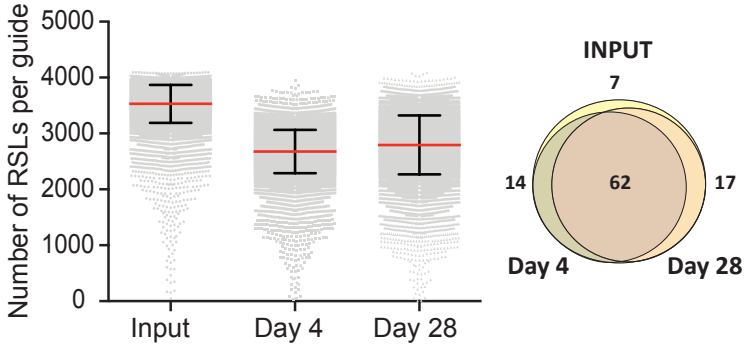
a. Library design.



b. Levels of analysis

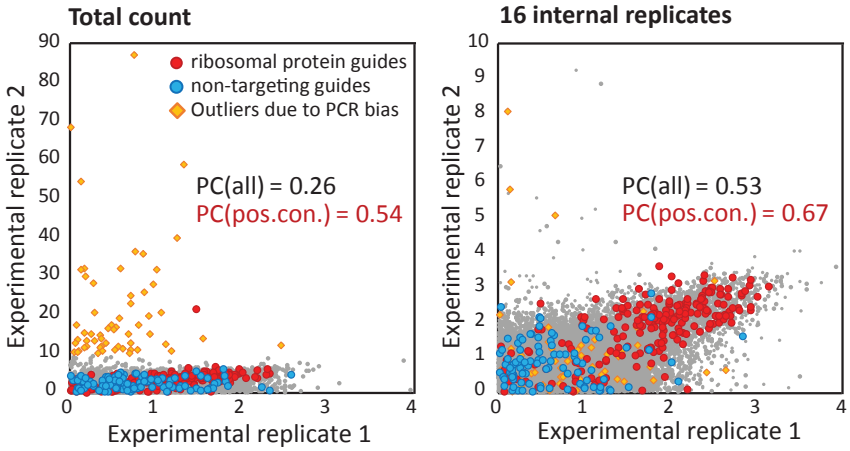


c. Library complexity and carry-through



Total count (Conventional analysis, counts per guide)		Internal Replicates (Counts per bin)		Massively parallel lineage tracing (MPLT) (Counts per RSL-guide)	
guide-$(N)_6$	12780	guide-A$(N)_5$	2645	guide-ACTGTC	51
				guide-ATACGA	10
		guide-C$(N)_5$	3984	guide-CTGGAA	22
				guide-CATCGT	9
		guide-G$(N)_5$	2895	guide-GTGACA	2
				guide-GTTGCC	4
		guide-T$(N)_5$	3256	guide-TGTATT	25
				guide-TTGGTG	65

d. Guide score correlation between experimental replicates.



e. Internal replicates increase number and experimental replicate overlap of hit genes

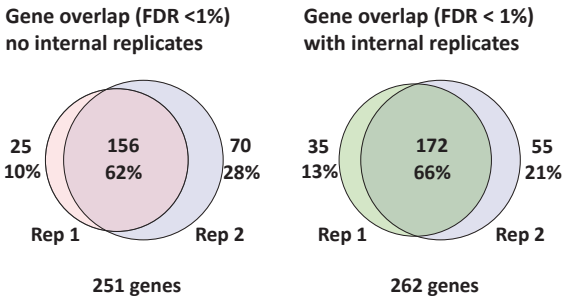
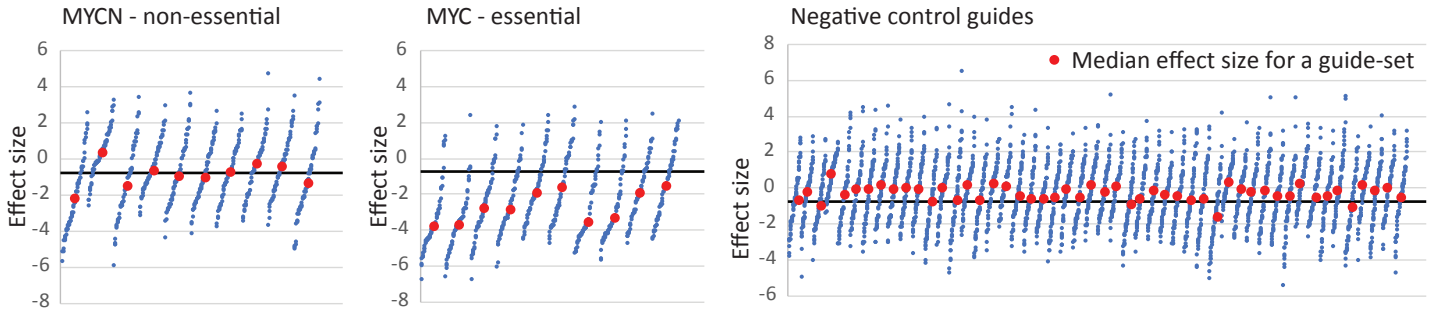
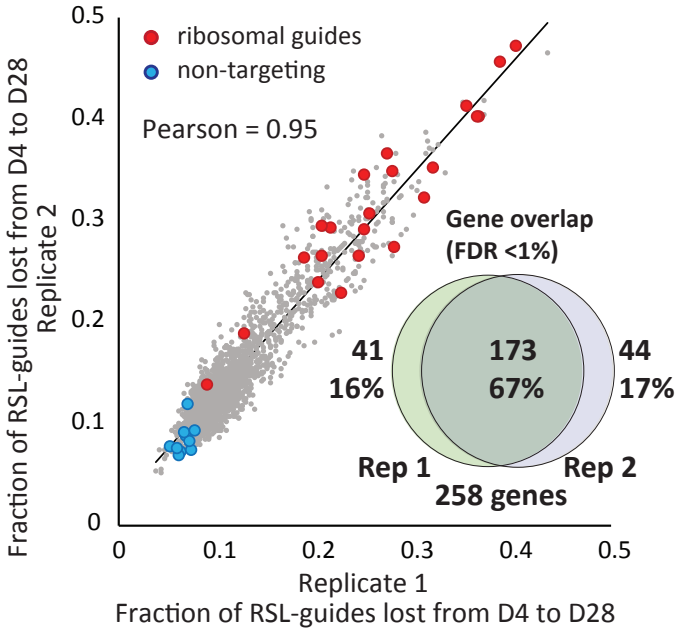


Figure 2

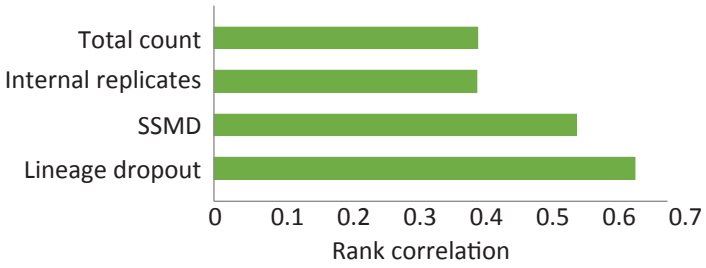
a. Internal replicates allow the application of standard statistics.



b. MPLT allows hit calling by real dropout



c. Rank correlation between replicates, top 200 hit genes



d. Average rank of ribosomal proteins