Minimal genome-wide human CRISPR-Cas9 library

Emanuel Gonçalves[1], Mark Thomas[1], Fiona M Behan[1], Gabriele Picco[1], Clare Pacini[1], Felicity Allen[1], David Parry-Smith[1], Francesco Iorio[1], Leopold Parts[1], Kosuke Yusa[2], Mathew J Garnett[1,*]

1. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
2. Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto 606-8507, Japan

* Corresponding author: mg12@sanger.ac.uk

Keywords: CRISPR-Cas9; Genome-wide; Minimal Library, Organoid

Abstract

CRISPR guide-RNA libraries have been iteratively optimised to provide increasingly efficient reagents although their large size is a barrier for some applications. We designed a minimal genome-wide human CRISPR-Cas9 library (MinLibCas9), optimised by mining existing large-scale gene loss-of-function datasets, resulting in a greater than 46% reduction in size compared to other libraries while preserving assay sensitivity and specificity. MinLibCas9 improves the efficiency of CRISPR-Cas9 loss-of-function screens and expands their application to complex models and assays.

Main

CRISPR-Cas9 loss-of-function screens have been used in a variety of model organisms, including human cells [1,2]. Genome-wide CRISPR-Cas9 single-guide RNA (sgRNA) libraries have been optimised to reduce off-target activity and increase on-target efficiency [3–5]. All current genome-wide libraries have more than 4 sgRNAs per gene and contain over 69,000 sgRNAs [2,4–14] (Figure 1a) (Supplementary Table 1). *In silico* downsampling analyses have shown that 2 sgRNAs per gene can recover previously defined essential genes [5,15] (Supplementary Figure 1). Using recent genome-wide CRISPR-Cas9 loss-of-function screens performed in hundreds of cancer cell lines [14,16], it is now possible to prioritise sgRNA selection and thereby improve design and reduce library sizes. Smaller genome-wide CRISPR libraries are more efficient and cost effective, and increase feasibility when using complex models (e.g. primary cultures, organoids, co-cultures, *in vivo* screens) or measuring complex phenotypic endpoints (e.g. scRNAseq or perturbations). Thus, to increase the utility of CRISPR screens, we designed an optimised minimal genome-wide human library (MinLibCas9) by mining data from Project Score[14] using a >100,000 sgRNA CRISPR-Cas9 library screened across 245 unique cancer cell lines.

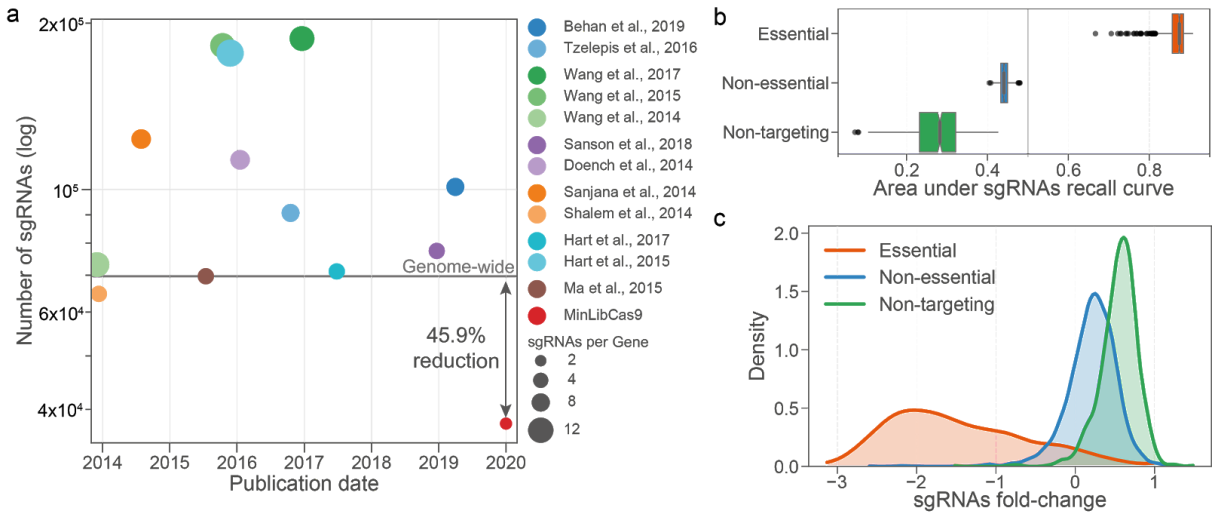


Figure 1. Human CRISPR-Cas9 libraries. ***a****, number of sgRNAs in each CRISPR-Cas9 library since the first reported genome-wide screens.* ***b****, area under the recall curve of sgRNAs targeting known essential (n=1,469) and non-essential (n=3,251) genes, and non-targeting guides non-targeting (n=1,004). Recall curves were drawn for each replicate of Project Score (n=663) and represent the cumulative distribution of each sgRNA group across all sgRNAs sorted ascendingly by fold-changes. Box-and-whisker plots show 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians.* ***c****, fold-change distribution on Project Score data-set of the different sgRNAs groups.*

We first minimised potential sgRNA off-target activities. The Project Score library [14] was optimised according to multiple factors [3,11]. Here we updated sgRNAs alignments to the GRCh38 build and computed new off-target summaries, which were then used to deprioritize non-selective guides[17]. In addition, JACKS scores [15] were used to identify sgRNAs with fitness profiles similar to the mean of all other sgRNAs targeting the same gene, and thereby exclude sgRNAs with outlier profiles suggestive of off-target or lack of activity.

Next, we selected sgRNAs with maximal on target activity. Approximately one third of all protein-coding genes induce a cellular loss-of-fitness effect upon knockout [14,16], thus for the remaining two thirds of genes it is difficult to distinguish efficient from non-efficient targeting sgRNAs. The introduction of CRISPR-Cas9 mediated DNA double-strand breaks induces a weak loss-of-fitness effect in cells regardless of the targeted site or gene [18–20]. The Project Score library included 997 non-targeting sgRNAs that do not align to any region in the human genome. These non-targeting sgRNAs were enriched towards positive fold-changes across all samples, which represents the relative growth advantage provided by not introducing a double-stranded break (Figure 1b and 1c). Thus, to prioritise the selection of working sgRNAs, we performed a non-parametric Kolmogorov-Smirnov test (KS scores) comparing the distribution of the fitness fold-changes of all sgRNAs to that of the

non-targeting guides (Figure 1c). Guides with high KS scores (values closer to 1) have a strong positive or negative median fold-changes, whereas those with low KS scores are more likely to have weak or no activity (Supplementary Figure 2). No strong association between different sgRNA design metrics (i.e. JACKS [15], Rule Set 2 [4] and FORECasT percentage of in-frame deletions [21]) was observed, suggesting that these provide complementary information (Supplementary Figure 2). To select guides, a combined strategy of using JACKS to exclude guides with outlier effects and ranking the remaining guides using the KS scores improved recall rates of gene dependencies identified with the original library (Supplementary Figure 3). Notably, the top 2 selected sgRNAs performed very similarly compared to using the full library, and limited improvement was observed when considering more than 2 guides.

For 2,165 genes, either none or only 1 sgRNA in the Project Score library was available or passed the off-target filters defined before, so alternative sgRNAs were collected from the most recent and broadly adopted CRISPR-Cas9 libraries, namely Avana, Brunello and TKOv3 [4,5,13]. To facilitate cross-library guide selection, we assembled a reference master library with standardized annotation for 300,168 unique sgRNAs (median 19 sgRNA per gene) including updated off-target summaries using the GRCh38 build, gene symbols, and guide efficacy metrics where available (Supplementary Table 2). Guide selection was performed iteratively across the four different libraries using increasingly less stringent efficacy filters, and therefore sgRNAs were grouped into one of three colour-coded groups (green, amber and red). Based on these criteria, the minimal genome-wide library targets 18,761 genes, using 2 optimal sgRNAs per gene, and has a total of 37,522 gene-targeting and 200 non-targeting sgRNAs (Supplementary Table 3). To preserve library consistency [22], 90.6% of the sgRNAs belong to the original Project Score library. MinLibCas9 library includes an additional 964 protein-coding genes while using 62.7% fewer sgRNAs compared to the original Project Score library [14], and it is over 46% smaller in size compared to any publicly available genome-wide human library (Figure 1a).

We benchmarked MinLibCas9 against the original library by *in silico* subsetting the guides from the Project Score library (Yusa V1.1) and tested it across the previously screened cell lines. The selected optimal two guides provided a systematic improvement over a random selection of 2 guides per gene (Supplementary Figure 4a) and gene-level fold-changes were largely concordant with the original library (mean Spearman's R=0.77). The cumulative number of significantly dependent cell lines identified per gene was well

correlated (Spearman's R=0.88, p-value < 0.001) (Supplementary Figure 4c), and dependencies not identified with MinLibCas9 had on average lower fold-changes (two-sided Welch's t-test p-value < 0.001) (Supplementary Figure 4d, 4e). A total of 107 genes had discordant significant dependencies, called as a significant dependency with one library but not the other, in more than 100 cell lines. This disagreement was primarily due to sub-groups of sgRNAs targeting the same gene with very distinct fold-change profiles (Supplementary Figure 4c, 4f and 4g). We therefore iteratively improved the minimal library by removing these guides and selecting new guides from the reference master library. Lastly, different levels of sgRNA coverage (25x, 50x, 75x, 100x and 500x) during screening revealed that for both the minimal and full library lower coverage has no impact on identifying essential genes, although replicate correlation at the gene-level is lower with the minimal library (Supplementary Figure 5a and b). Overall, MinLibCas9 preserved the capacity to identify known essential genes (Figure 2a) and recovered well significant dependencies found with the original library with an average precision (AP) greater than 89.8% in at least 80% of the 245 cancer cell lines (Figure 2b).

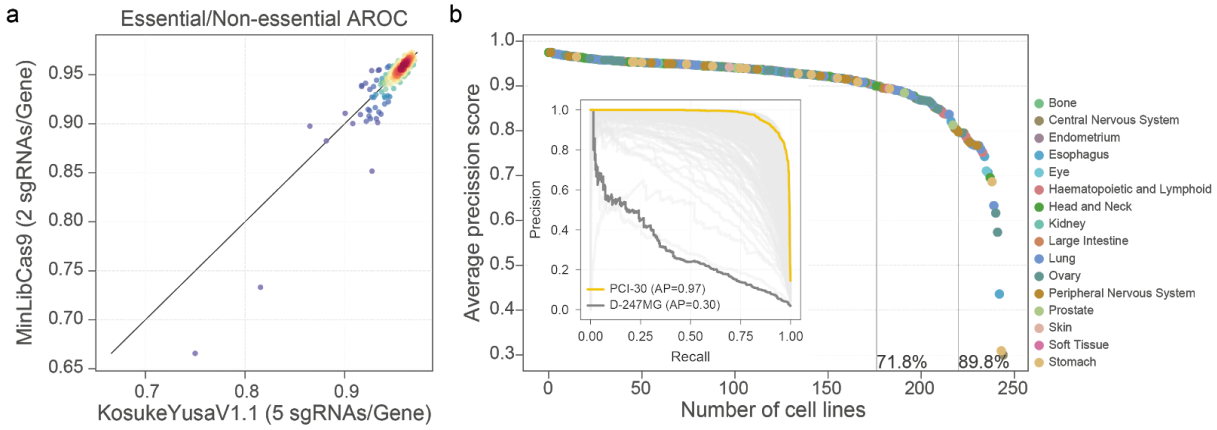


***Figure 2. Benchmark of minimal CRISPR-Cas9 library. a**, Standardized area under the receiver operating characteristic curve for 245 cell lines at 20% FDR for the essential genes calculated using the minimal library (2 optimal sgRNAs per gene) and the original full library (median of 5 sgRNAs per gene). **b**, Average Precision (AP) scores to classify significant gene dependencies identified at 1% FDR in the original library using gene fold-changes from MinLibCas9. Recall-Precision curves for all cell lines are represented in the inset and cell lines with the lowest and highest AP score are highlighted.*

To assess if the minimal library could recapitulate dependencies in more complex models and assays, we began by analysing a CRISPR-Cas9 screens to identify genes that enhance or suppress sensitivity to BRAF inhibitor (dabrafenib) in a partially-sensitive

BRAF-mutant colorectal cancer cell line (HT-29) (Supplementary Figure 6a). Using *in silico* downsampling analysis, gene fold-changes with both libraries were strongly correlated (Spearman's R=0.72, p-value < 0.001) and top dependencies time-series profiles were consistently identified (Supplementary Figure 6b and c). We also performed genome-wide screens in three 3D organoid cultures and confirmed that gene fold-changes between the original and *in silico* downsampled minimal library were strongly correlated (average Spearman's R=0.70), and the minimal library provides similar replicates correlation and capacity to identify known essential genes (Supplementary Figure 7).

In summary, we designed a minimal genome-wide human CRISPR-Cas9 library by using experimental data to select and rank sgRNAs. This reduced library size by over 46% compared to most currently used libraries and preserves sensitivity and specificity to identify gene dependencies. Further work will be required to confirm the *in silico* down-sampling analysis by independently synthesising and delivering MinLibCas9. Our reference guide library also provides a highly annotated and comprehensive resource of independently optimised sgRNAs that can be exploited to prioritise reagents for smaller focused libraries. MinLibCas9 unlocks the application of genome-wide screens to complex models which are currently limited to libraries focused on predefined sets of genes, moreover it provides a data-driven approach to prioritise the selection of the most effective sgRNAs for assays using more complex read-outs (e.g. Perturb-seq [23,24]).

Methods

CRISPR-Cas9 screens analysis

Screen analysis started with the sgRNA read count matrices. Guides with less than 30 counts in the control condition, i.e. plasmid DNA (pDNA), were excluded. Read counts were normalised to reads per million (G') within each sample using the following formula:

$$G'_i = (G_i / \sum_j^n G_j) \times 10^6$$

where G_i represents the raw counts of sgRNA i. A pseudo count of 1 was added to the whole matrix and $\log_2$ fold-changes were then calculated comparing to pDNA. sgRNAs recall curves are drawn by sorting the guides by fold-change, from the most negative to the most positive, and then the cumulative distribution is calculated for the different guide groups (i.e.

targeting essential genes, targeting non-essential genes and non-targeting sgRNAs). Then the area under the recall curve is calculated which represents the enrichment of each group towards negative or positive fold-changes, being an area of 0.5 the random expectation. sgRNA downsampling analyses were performed by randomly sampling *n* sgRNAs without replacement.

Gene-level fold-changes were calculated by grouping all sgRNAs by their targeting gene and taking the mean of the fold-changes. Similarly, replicates of the same cell line were mean-averaged. Gene dependencies were defined as significant, on a per sample basis, if the gene $\log_2$ fold-change was lower than the fold-change threshold at which essential genes were found at 1% False Discovery Rate (FDR) from non-essential genes in the Receiver Operating Characteristic (ROC) curve [9,22]. Similarly to Allen et al. [15], the same ROC curve was used to estimate the performance of the sample to recapitulate previously defined essential genes by taking the Area Under the ROC (AROC) curve at 20% FDR, i.e. standardized partial AUC at maximum 20% false positive rate.

Recall-Precision curves of gene dependencies were drawn for each cell line by taking the significant gene dependencies (1% FDR) identified with the Project Score library (Yusa V1.1) and using the gene fold-changes obtained with MinLibCas9. Curves were summarized using average precision (AP) scores, defined as follows:

$$\text{AP} = \sum_{j}^{n} (R_j - R_{j-1}) P_j$$

where P_n and R_n are the precision and recall at the nth threshold. AP score is a similar metric to the area under the Precision-Recall curve.

Guide efficacy KS-score

CRISPR-Cas9 sgRNAs Kolmogorov–Smirnov scores (KS scores) is a two-sided test assessing if the sgRNA fold-changes and the median fold-changes of all non-targeting sgRNAs are drawn from the same distribution (function ks_2samp from scipy [25] Python package was used). KS scores range between 0 and 1, and values closer to 0 represent sgRNAs with a distribution similar to non-targeting sgRNAs, in contrast values closer to 1 represent the most dissimilar sgRNAs. KS scores were estimated for 100,262 sgRNAs across 663 samples (245 unique cancer cell lines) of Project Score data-set [14] and for

73,911 sgRNAs across 1,257 samples (562 unique cancer cell lines) of the Broad DepMap19Q2 data-set [16,26] (Supplementary Table 4).

Reference Master CRISPR-Cas9 library

All the sgRNAs described in the Yusa [14], Avana [4], Brunello [4,5] and TKOv3 [13] libraries were assembled in a single reference master library. The master library contains a total of 355,011 sgRNAs with a median of 19 guides per gene. Where available, guides were annotated with efficiency scores from Rule Set 2 [4], JACKS [15], FORECasT in-frame indels [21] and KS. All sgRNAs were annotated with updated GRCh38 genomic coordinates and off-target summaries using the WGE database [17], as well as with updated gene symbol IDs from HGNC [27].

Minimal genome-wide CRISPR-Cas9 library

A minimal genome-wide library was assembled from the master library by ranking sgRNAs that minimise off-target and maximise on-target effects. A small number, 302, of gene-targeting sgRNAs that did not match any position in the GRCh38 were removed. Additionally, 339 sgRNAs with conflicting gene-targeting annotation across different libraries were also discarded.

Three different groups of sgRNAs corresponding to increasingly relaxed selection stringency levels were defined, termed as green, amber and red. Green represents guides with a single perfect match to the GRCh38 build and no other alignment with one sequence mismatch. Additionally, green sgRNAs have either a JACKS scores within a range between 0 and 2 (Yusa V1.1 or Avana guides) or a Rule Set 2 score higher than 0.4 (Brunello guides), with the exception to TKOv3 guides where no filter was applied. Amber represents sgRNAs with more relaxed off-target constraints, only requiring a single perfect alignment to the genome, and no filter based on JACKS or Rule Set 2 metrics was used. Lastly, red level sgRNAs can have up to 3 perfect alignments, similar to Koike-Yusa et al. [3], and, similar to amber sgRNAs, no filter based on guide efficacy metric was used.

For all protein-coding genes defined in HGNC [27] we tried to identify 2 optimal sgRNAs within these three different stringency levels. For each gene, guides were ranked using either KS or Rule Set 2 scores, and selection was performed until 2 sgRNAs successfully

passed the defined thresholds: (i) the Yusa library was queried and the top 2 sgRNAs ranked by KS scores were picked; (ii) the Avana library was ranked by KS scores and searched to pick the outstanding number of sgRNAs; (iii) the Brunello library was used to pick the outstanding number of sgRNAs and Rule Set 2 scores were used to rank the guides; and lastly (iv), sgRNAs from the TKOv3 library were considered. To minimise library-specific biases we prioritised the use of sgRNAs originating from Yusa library.

The assembled minimal library covers 18,761 protein-coding genes with 37,522 sgRNAs (33,986 Yusa V1.1; 1,732 Brunello; 1,493 Avana and 311 TKOv3) with 36,337 green, 740 amber and 445 red confidence level sgRNAs. An additional set of 200 non-targeting sgRNAs, chosen by their similarity to the median fold-changes of all non-targeting guides and with no perfect alignment, no 1nt-mismatch alignment and at most three 2nt-mismatch alignments to the GRCh38 build were added to allow future benchmarks and design improvement. For 107 genes the sgRNA selection was forced to exclude Yusa V1.1 library as these generated conflicting gene-level fold-changes (i.e. significant gene essentiality profiles discordant in more than 100 cell lines between the original and minimal library) (Supplementary Figure 4c, 4f and 4g).

CRISPR-Cas9 sgRNA coverage

KM-12 colorectal carcinoma cancer cell lines were CRISPR-Cas9 screened with Yusa V1.1 library similarly to Behan, et al. [14] using 25x, 50x, 75x, 100x and 500x library coverage, i.e. (Supplementary Table 5). Transduction efficiency of KM-12 was maintained at ~30% while cell numbers were adjusted to achieve different levels of library coverage. The different library coverage levels were performed in two independent experiments in technical triplicate; experiment A tested 100x and 500x coverage and experiment B tested 25x, 50x, 75x and 100x.

Drug perturbed CRISPR-Cas9 screens

We conducted time-series CRISPR-Cas9 screens, performed similarly to Behan et al. [14] in technical triplicate, with dabrafenib treatment in HT-29 cancer cell lines (Supplementary Table 6). HT-29 cells were transduced at 30% efficiency on day 1. Following puromycin selection, DNA was extracted on day 8 from a subset of cells representing the baseline undrugged condition. The remaining cells were treated with either dabrafenib (0.1μM) or

DMSO on day 8. Subsequently, DNA extraction, sgRNA amplification and sequencing was performed at day 10, 14, 18 and 21. Read count matrices were processed as described before and statistical analysis to identify the most significantly differential essential genes over-time was performed using R package limma [28] using the F-statistic and respective aggregated p-value. P-values were adjusted for false discovery rates (FDR) using Benjamini-Hochberg false discovery rate methods. Identical analysis was performed for the original Yusa V1.1 library and for the *in silico* down-sample minimal library and then compared.

Organoid genome-wide CRISPR-Cas9 screens

Genome-wide CRISPR-Cas9 screens were performed in 3 organoids, 1 derived from colorectal carcinoma patient sample (COLO021), and 2 organoids derived from esophageal cancer (CAM277 and CAM338) (Supplementary Table 7). CAM338 was screened in technical duplicate. Organoids were derived and maintained as previously described [29]. To express Cas9, tumoral organoids were dissociated into single cells and incubated overnight in suspension and complete media supplemented with pKLV2-EF1a-BsdCas9-W lentiviral particles and polybrene (8 µg ml−1). The day after, cells were seeded in matrigel and grown as organoids. Blasticidin selection (20 mg/ml) commenced 48h after transduction and maintained until the end of the experiment. All the organoid lines displayed Cas9 activity over 75%. The genome-wide sgRNA library transduction was adapted from a previous protocol recently reported to screen cancer cell lines [14]. Briefly, tumor organoids were dissociated into single cells and a total of $3.3x10^7$ cells were transduced overnight, in suspension, with an appropriate volume of the lentiviral-packaged whole-genome sgRNA library to achieve 30% transduction efficiency (100x library coverage) and polybrene (8 µg ml−1). The following day, cells were seeded in matrigel and grown as organoids. After 48h organoids were selected with Puromycin (2 mg/ml). After 14 days, approximately $2x10^7$ cells were collected, pelleted and stored at −80 °C for DNA extraction. Genomic DNA was extracted using the Qiagen, Blood & Cell Culture DNA Maxi Kit, 13362 as per the manufacturer's instructions. PCR amplification, Illumina sequencing (19-bp single-end sequencing with custom primers on the HiSeq2000 v.4 platform) and sgRNA counting were performed as described previously.

References

1. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
2. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
3. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).
4. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
5. Sanson, K. R. *et al.* Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun.* **9**, 5416 (2018).
6. 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).
7. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
8. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
9. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).
10. Ma, H. *et al.* A CRISPR-Based Screen Identifies Genes Essential for West-Nile-Virus-Induced Cell Death. *Cell Rep.* **12**, 673–683 (2015).
11. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17**, 1193–1205 (2016).
12. Park, R. J. *et al.* A genome-wide CRISPR screen identifies a restricted set of HIV host

dependency factors. *Nat. Genet.* **49**, 193–203 (2017).

13. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3* **7**, 2719–2727 (2017).
14. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).
15. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Res.* **29**, 464–471 (2019).
16. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
17. Hodgkins, A. *et al.* WGE: a CRISPR database for genome engineering. *Bioinformatics* **31**, 3078–3080 (2015).
18. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).
19. Munoz, D. M. *et al.* CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discov.* **6**, 900–913 (2016).
20. Gonçalves, E. *et al.* Structural rearrangements generate cell-specific, gene-independent CRISPR-Cas9 loss of fitness effects. *Genome Biol.* **20**, 27 (2019).
21. Allen, F. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4317
22. Dempster, J. M. *et al.* Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines. *bioRxiv* 720243 (2019). doi:10.1101/720243
23. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867–1882.e21

(2016).

24. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).
25. Jones, E., Oliphant, T., Peterson, P. & Others. SciPy: Open source scientific tools for Python. (2016).
26. DepMap, B. DepMap 19Q2 Public. (2019). doi:10.6084/m9.figshare.8061398.v1
27. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017).
28. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
29. Li, X. *et al.* Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. *Nat. Commun.* **9**, 2983 (2018).

Data availability

All data is contained in Supplementary Information.

Code availability

All code and results are publically available at github.com/EmanuelGoncalves/crispy/tree/master/notebooks/minlib .

Acknowledgements

We acknowledge Joel Rein for helpful comments on the integration of WGE annotations. Work in M.J.G was funded by the Wellcome Trust (206194) and Open Targets.

Author information

Conceptualization: E.G., M.G.

Software: E.G., M.T., C.P., F.A.

Validation: F.B., G.P.

Formal analysis: E.G.

Data curation: E.G., M.T.

Writing - original draft preparation: E.G., M.G.

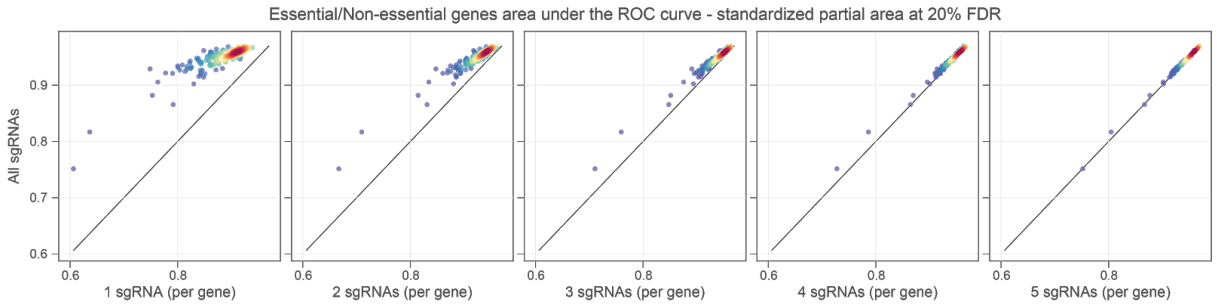
Writing - reviewing and editing: All authors

Visualisation: E.G.

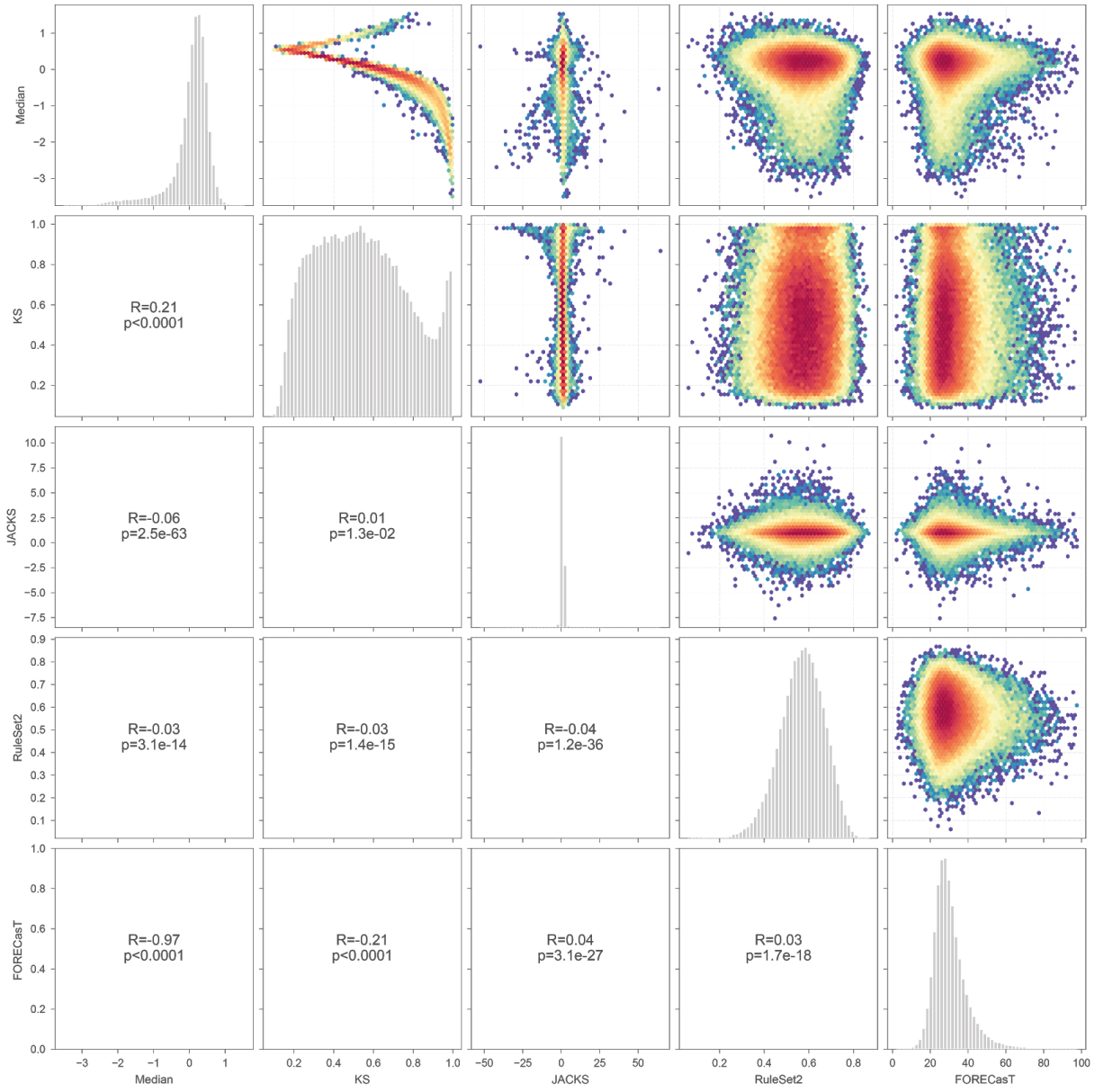
Supervision: D.P.S., F.I., L.P., K.Y., M.G.

Funding acquisition: M.G.

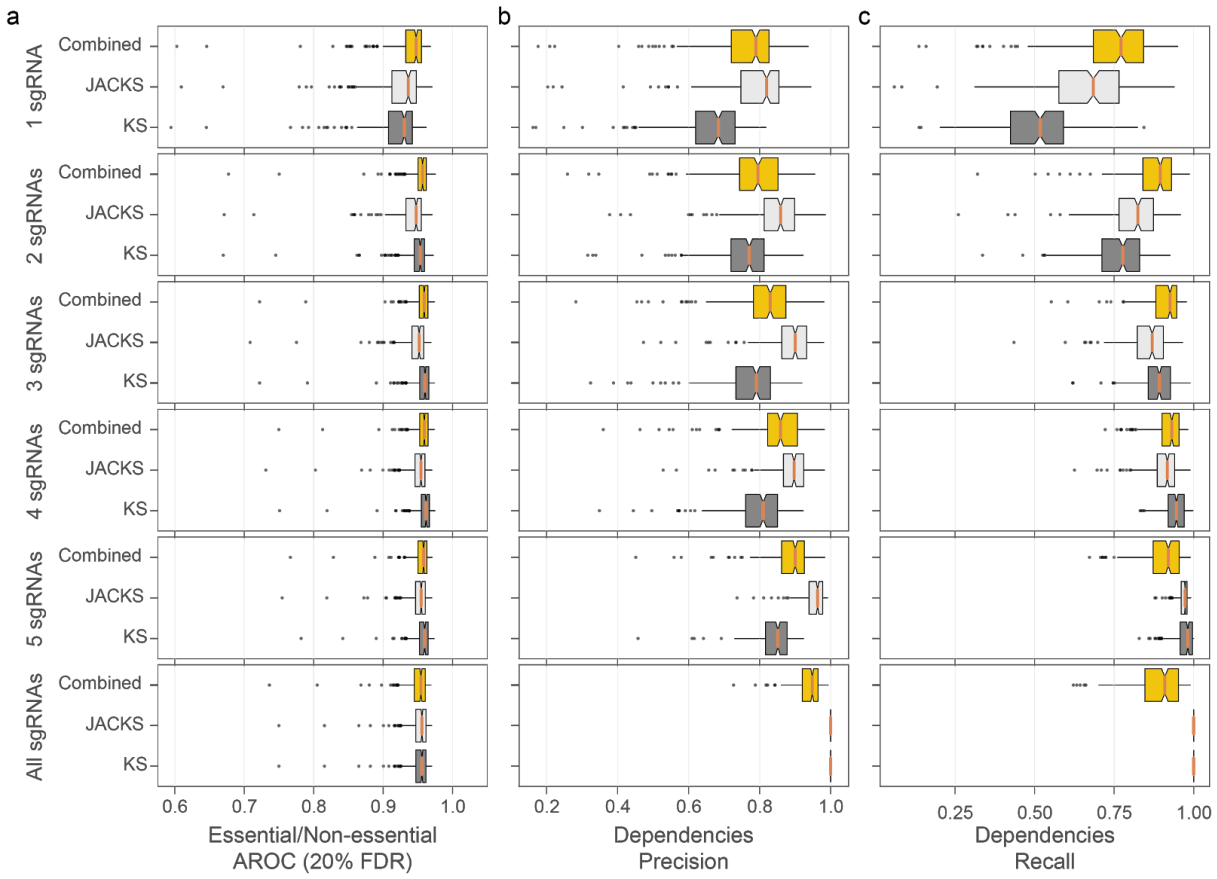
Supplementary figures



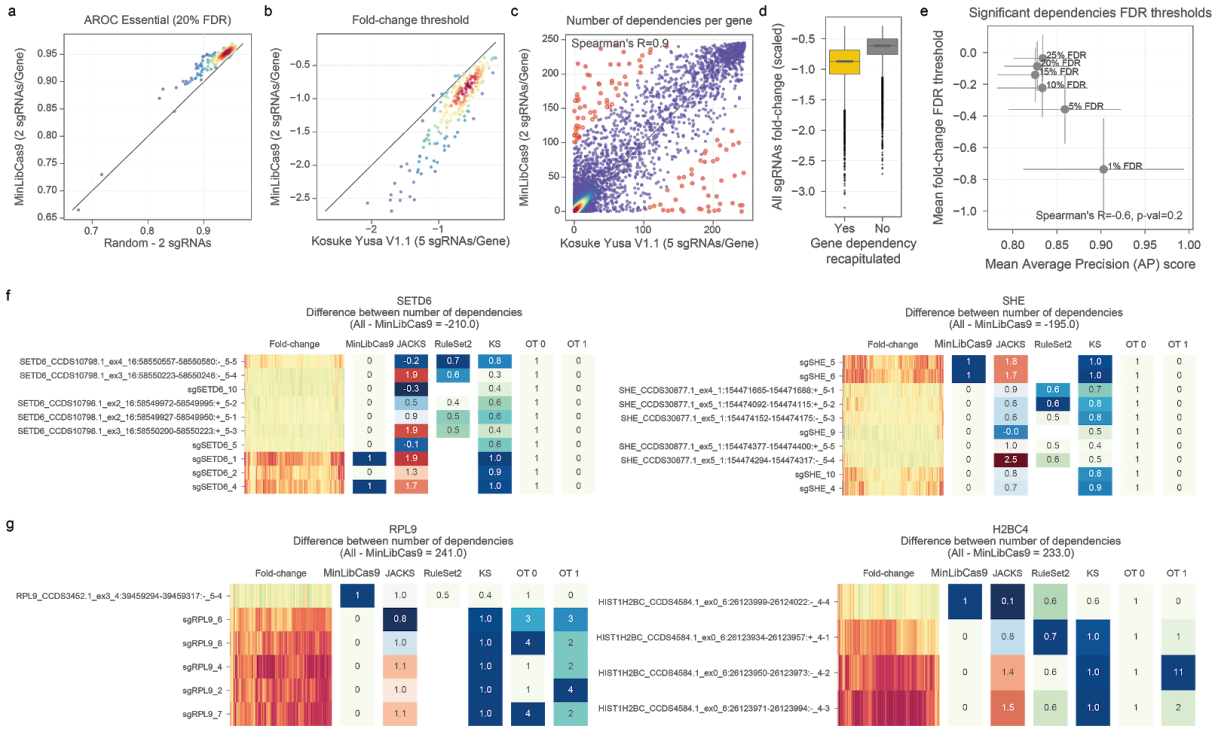
Supplementary Figure 1. Randomised selection of n sgRNAs per gene. *ROC curve derived from previously defined sets of essential and non-essential genes [9]. Standardized partial area under the ROC curve (AROC) calculated per cell line over the range of maximum false discovery rate of 20%. AROCs compared between downsampled sgRNAs and all sgRNAs available for all the covered genes. 10 random sgRNAs permutations without replacement per cell line and per n guides were performed and AROCs mean values are plotted.*



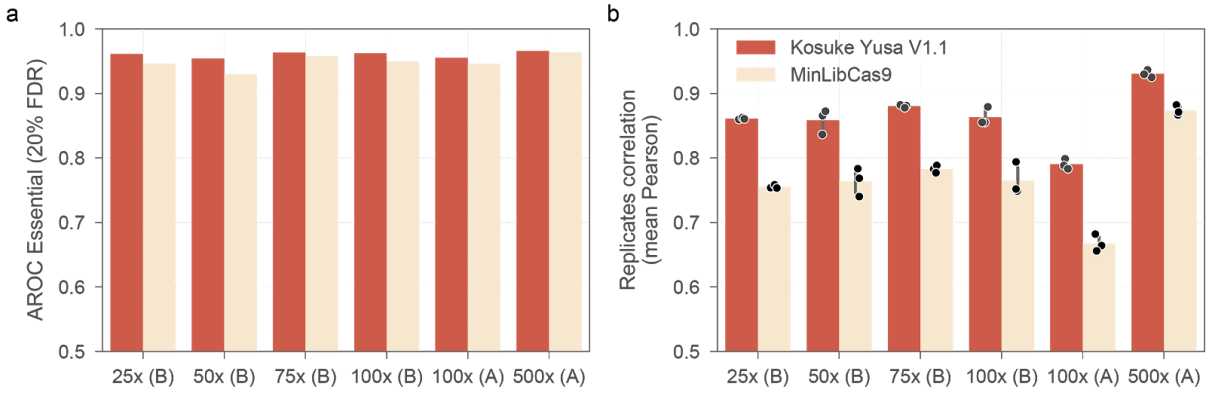
Supplementary Figure 2. Comparison of sgRNA metrics. *Efficiency metrics of Project Score library (Yusa V1.1) sgRNAs - KS metric (Kolmogorov Smirnov test) comparison to non-targeting guides, JACKS scores [15], Rule Set 2 [4] scores, and FORECasT [21] predicted percentage of in frame deletions produced - plotted together with guides median fold-changes calculated across 663 samples. Spearman correlation coefficients are reported in the lower triangle of the grid. Plots in the diagonal represent the distribution of the respective metric.*



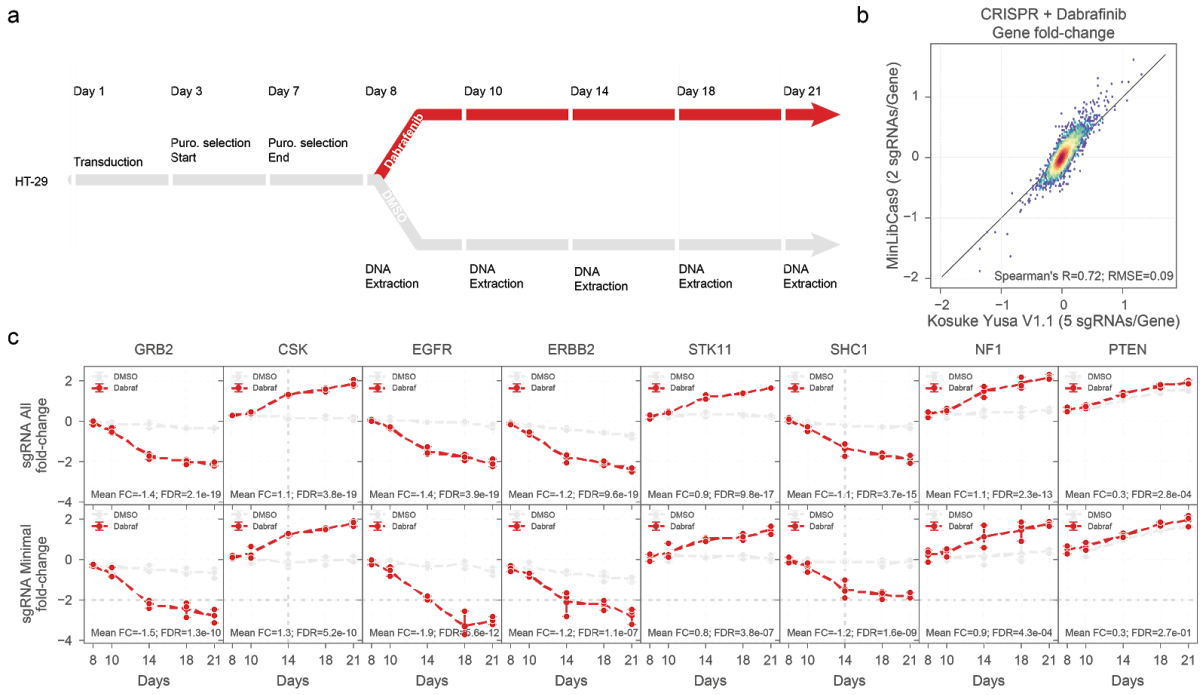
Supplementary Figure 3. Downsample analysis of top n sgRNAs ranked using KS and JACKS metrics. *Combined score discards sgRNAs with a JACKS score outside the range of [0, 2] and then selects the top n sgRNAs according to the KS score (descending order, stronger KS scores and thereby stronger absolute fold-changes). Essential/Non-essential AROCs are the area under the ROC curve (at 20%FDR) using known essential and non-essential genes. Precision and recall rates are calculated between the sets of significant gene-level dependencies (at 1% FDR) estimated using the original library and the downsampled library. Each box-and-whisker plot show 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians (n=245 cancer cell lines).*



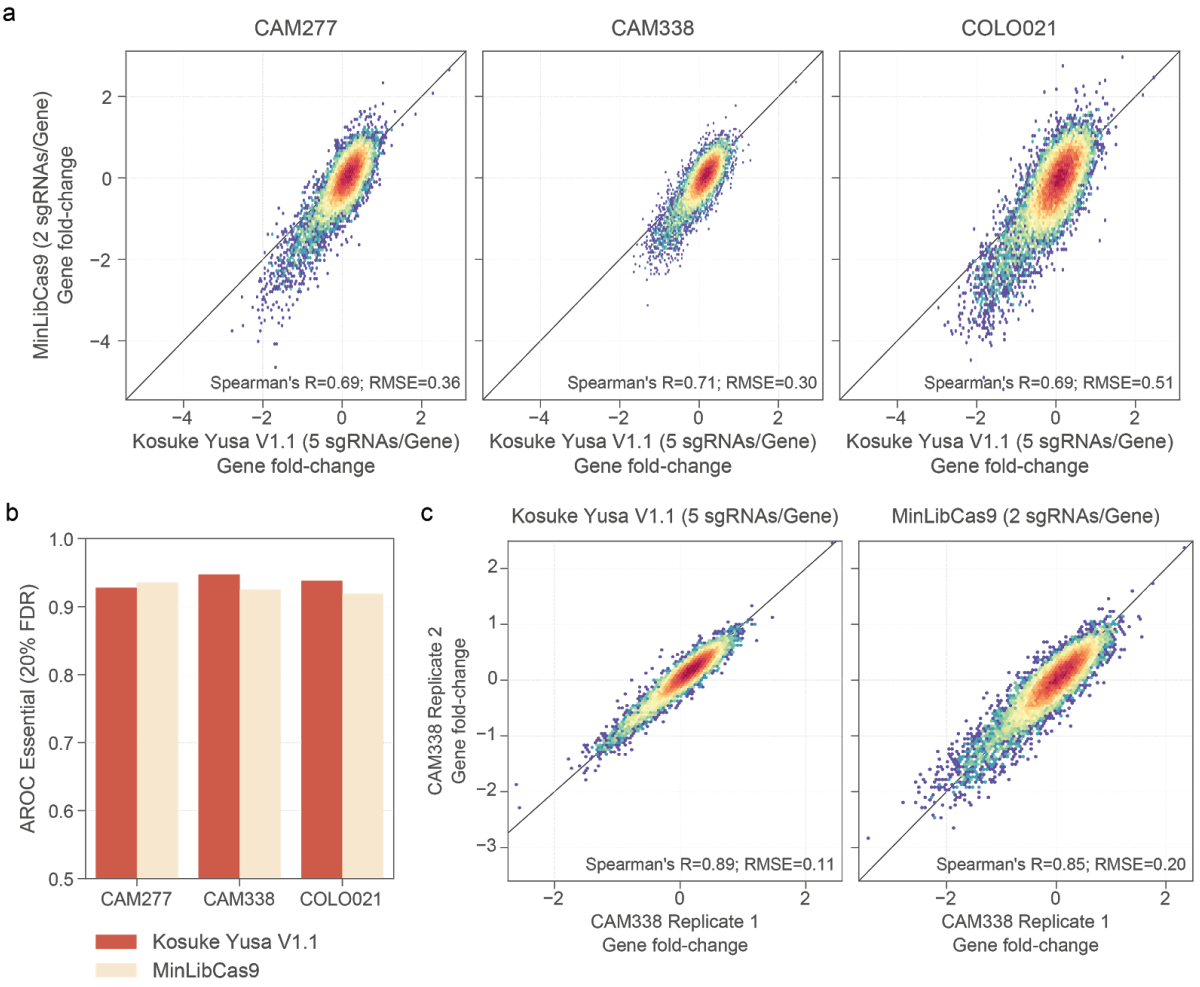
Supplementary Figure 4. Minimal library benchmark across CRISPR-Cas9 screens from 245 cancer cell lines. ***a****, AROC of essential genes obtained with the full library versus the minimal library.* ***b****, fold-change threshold identified using a 1% FDR of essential versus nonessential genes.* ***c****, cumulative number of dependencies (at 1% FDR) identified in 245 cancer cell lines for each gene with both the full original library and the minimal library.* ***d****, scaled fold-changes (median essential genes fold-change = 1; median non-essential genes fold-change = 0) of dependencies recapitulated (n=251,387) and missed (n=36,996) with the minimal library (two-sided Welch's t-test p-value < 0.001). Box-and-whisker plots show 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians.* **e**, *fold-change thresholds of significant dependencies and respective average precision (AP) score between MinLibCas9 and original Project Score library at different FDR thresholds. Top 2 genes with strongest disagreement in the number of dependencies found with the minimal library showing. Error bars present 1 x standard deviation.* ***f****, more, and* ***g****, less dependencies than the original library. Off-target (OT) summaries with the number of alignments to the GRCh38 build with 0nt (OT 0) and 1nt (OT 1) mismatches are provided.*



Supplementary Figure 5. sgRNA library coverage analysis in KM-12 cancer cells. ***a****, AROC of essential/non-essential genes at different guide coverage levels.* ***b****, technical replicates correlation. Comparisons are made between the original Yusa v1.1 library and the in silico minimal library.*



Supplementary Figure 6. CRISPR-Cas9 dropout screens upon treatment with Dabrafenib. ***a****, diagram of the experimental setup.* ***b****, gene fold-changes averaged across the different time points (day 8, 10, 14, 18 and 21) obtained using the original library compared to the in silico MinLibCas9.* ***c****, time-series fold-changes of the top significantly essential hits (compared to control experimental arm, DMSO) obtained with both full library and minimal library, technical triplicates are represented.*



Supplementary Figure 7. CRISPR-Cas9 loss-of-fitness screens in 3D organoids. ***a****, Comparison of gene fold-changes obtained using the in silico minimal and original library in a colon carcinoma organoid (COLO021) and two oesophageal adenocarcinoma organoids (CAM277 and CAM338).* ***b****, AROC of essential/non-essential genes of each organoids obtained with both libraries.* ***c****, CAM338 technical replicates correlation.*

Supplementary tables

Supplementary Table 1. *Median number of sgRNAs per gene and library size of currently available human genome-wide CRISPR-Cas9 libraries.*

Supplementary Table 2. *Reference CRISPR-Cas9 library containing sgRNAs originating from multiple libraries with standardised genomic annotation and guide efficiency metrics.*

Supplementary Table 3. *Genome-wide minimal human CRISPR-Cas9 library, MinlibCas9.*

Supplementary Table 4. *KS scores estimated for sgRNAs of Project Score and Avana libraries.*

Supplementary Table 5. *Raw counts of the CRISPR-Cas9 screens at different guide coverage performed in KM-12 cancer cell line.*

Supplementary Table 6. *sgRNA counts of the CRISPR-Cas9 screens followed by drug treatment with dabrafenib in HT-29 cells.*

Supplementary Table 7. *CRISPR-Cas9 raw counts for three different organoids derived from cancer samples.*