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], Mamta Sharma[1], David A. Jackson[1], Stacey Price[1], Charlotte M. Beaver[1], Oliver Dovey, David Parry-Smith[1], Francesco Iorio[1,3], 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
3. Human Technopole, Via Cristina Belgioioso 147, 20157 Milano - Italy

* Corresponding author: mg12@sanger.ac.uk

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

Abstract

CRISPR guide-RNA libraries have been iteratively optimised to provide increasingly efficient reagents, although their large size is a barrier for many applications. We designed an optimised minimal genome-wide human CRISPR-Cas9 library (MinLibCas9), by mining existing large-scale gene loss-of-function datasets, resulting in a greater than 42% reduction in size compared to other libraries while preserving assay sensitivity and specificity. MinLibCas9 increases the dynamic range of CRISPR-Cas9 loss-of-function screens and extends their application to complex models and assays.

Manuscript

CRISPR-Cas9 loss-of-function screens have been used in a variety of model organisms, including human cells [1,2]. All broadly used genome-wide libraries have at least 4 single-guide RNA (sgRNA) per gene and contain over 65,000 sgRNAs [2–13] (Figure 1a) (Supplementary Table 1). *In silico* down-sampling analyses have shown that 2 sgRNAs per gene can recover previously defined essential genes [12,14] (Supplementary Figure 1). Genome-wide CRISPR-Cas9 sgRNA libraries have been iteratively optimised to reduce off-target activity and increase on-target efficiency [8,9,11,12,15,16], mostly using nucleotide sequence-based sgRNA efficacy prediction algorithms [9,11,15,16]. The recent availability of data from CRISPR-Cas9 knockout screens performed in hundreds of cell lines [13,17] makes it now possible to empirically improve library design through selection of sgRNA with strong and consistent biological effects across diverse contexts, taking in consideration additional factors that might influence guide efficacy [18,19]. Smaller genome-wide CRISPR libraries are more cost effective and increase feasibility when assaying complex models (e.g. primary cultures, organoids, co-cultures, *in vivo* screens), measuring complex phenotypic endpoints (e.g. scRNAseq or perturbations), and when probing genetic interactions using multiplexed CRISPR-Cas9 libraries. We therefore assembled a standardised resource that harnesses large-scale CRISPR-Cas9 screens, together with multiple efficiency metrics to evaluate and rank over 300,000 unique sgRNAs originating from the most broadly adopted libraries. These were used to design an optimised minimal genome-wide library with two sgRNAs per gene (MinLibCas9).

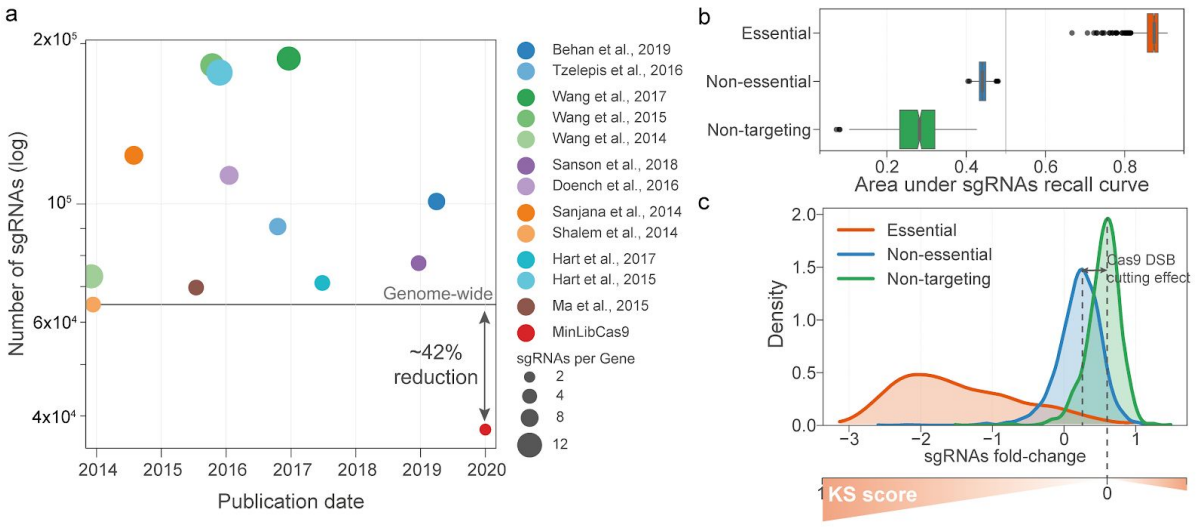


Figure 1. Human CRISPR-Cas9 sgRNA libraries. a*, number of sgRNAs in each CRISPR-Cas9 library since the first reported genome-wide screens, excluding Wang et al. 2014 which targets 7,114 genes.* ***b****, area under the recall curve of sgRNAs targeting known essential (n=1,469) and non-essential (n=3,251) genes, and non-targeting guides (n=997). Recall curves were calculated for each replicate of Project Score [13] (n=663) and represented by the cumulative distribution of each sgRNA group across all sgRNAs sorted by ascending fold-changes. Box-and-whisker plots show 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians.* ***c****, fold-change distribution, based on Project Score data-set of the different sgRNAs groups. Diagram depicting how the KS scores vary across the CRISPR-screens fold-change range. KS scores are calculated by testing if the distribution of each sgRNA across cell lines is drawn from that of the non-targeting sgRNAs using a two-sided Kolmogorov-Smirnov distribution.*

We began by compiling multiple CRISPR-Cas9 sgRNA libraries, namely Project Score (Kosuke Yusa V1.1), Avana, Brunello and TKOv3 [8,9,11–13], to provide standardized annotation for 300,167 unique sgRNAs with a median of 19 sgRNA per gene (Supplementary Table 2). This included updated mapping to GRCh38 and off-target summaries using the CRISPR genome editing database WGE [20], and where possible multiple guide efficacy metrics (JACKS [14], Rule Set 2 [9], FORECasT [21], and scores exported from CRISPOR [22] such as MIT specificity [23] and CrisprScan [24]). This reference library provides a single standardized resource to select guides based on multiple user defined criteria, and was used to assemble a minimal genome-wide human CRISPR-Cas9 library (MinLibCas9) (Supplementary Table 3). Preference was given to guides from the Project Score or Avana libraries, as these have been robustly and empirically tested across hundreds of cancer cell lines. Additionally, we prioritised Project

Score guides [13] to preserve library consistency [25] and to allow *in silico* down-sampling benchmarking of MinLibCas9 .

To construct MinLibCas9, we first minimised potential sgRNA off-target activities. Updated sgRNA off-target summaries in the reference library were used to exclude non-selective guides [20]. In addition, JACKS scores [14] were used to identify sgRNAs with fitness profiles dissimilar to the mean of all sgRNAs targeting the same gene, thereby empirically excluding sgRNAs with outlier profiles suggestive of off-target or reduced on-target activity. We next sought to prioritise guides with maximal on-target activity. Approximately one third of all protein-coding genes induce a cellular loss-of-fitness effect upon knockout [13,17], thus for the remaining two thirds of genes it is challenging to distinguish between efficient and 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 [26–28]. The Project Score library included 997 expressed non-targeting sgRNAs that do not align to any region in the human genome [13]. These non-targeting sgRNAs were positively enriched across all samples, which demonstrates a detectable relative growth advantage in the absence of a double-stranded break (Figure 1b and 1c). Thus, to empirically identify optimal on-target sgRNAs, we performed a non-parametric Kolmogorov-Smirnov test (KS score) comparing the distribution of the fitness fold-changes of every sgRNA to that of the non-targeting guides (Figure 1c). Guides with high KS scores (values closer to 1) have strong negative or positive median fold-changes, whereas those with low KS scores are more likely to have weak or no activity, similar to non-targeting controls (Supplementary Figure 2). We expanded this approach to the Avana library and estimated KS scores for all sgRNAs using hundreds of CRISPR-Cas9 screens (DepMap19Q2) [17,29,30].

No strong association between different sgRNA design metrics (i.e. KS, JACKS [14], Rule Set 2 [9] and FORECasT percentage of in-frame deletions [21]) was observed, suggesting that they inform on different aspects of guide efficiency and provide complementary information (Supplementary Figure 2). A combined strategy using JACKS to exclude guides with outlier effects, followed by ranking the remaining guides using KS scores, improved recall rates of gene dependencies identified with the original library (Supplementary Figure 3). Notably, the top 2 selected sgRNAs using this approach performed similarly compared to using the complete library, and limited improvement was observed when considering more than 2 guides. Based on these selection criteria, MinLibCas9 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 independently validate our guide selection procedure we used sgRNA selectivity (MIT specificity [23]) and efficiency (CrisprScan [24]) metrics from CRISPOR [22] that were not used for library design. MinLibCas9 outperformed or had similar scores to existing libraries (Supplementary Figure 4a). Poly-T stretches ≥ 4 were largely absent from the reference library and T-stretches ≤ 4 displayed minimal to no impact on guide expression and efficacy (Supplementary Figure 4b). Overall, MinLibCas9 library targets an additional 964 protein-coding genes compared to the original Project Score library. Notably, it is 41.7% to 79.9% smaller in size compared to any currently publicly available genome-wide CRISPR-Cas9 human library, and specifically 62.7% and 66.7% smaller than Project Score and Avana libraries, respectively (Figure 1a).

Prioritising the selection of the Project Score library guides ensures that we could *in silico* benchmark our MinLibCas9 by subsampling the full set of sgRNAs across 245 cancer cell lines (90.6% of the sgRNAs originated from Project Score library). Overall, MinLibCas9 preserved the ability to identify known essential genes (Figure 2a) and recovered the majority of significant dependencies found with Project Score library, with an average precision greater than 89.8% in at least 80% of the 245 cancer cell lines (Figure 2b, Supplementary Figure 4c). The selected optimal two guides provided gene-level fold-changes largely concordant with the original library (mean Spearman's R=0.77) and a greater fold-change dynamic range (Supplementary Figure 4d). 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 4e). Dependencies not identified with MinLibCas9 had on average weaker fold-changes (two-sided Welch's t-test p-value < 0.001) (Supplementary Figure 4f). A total of 107 genes had discordant significant dependencies, primarily due to sub-groups of sgRNAs with very distinct fold-change profiles, we therefore replaced these guides with others from the reference library (Supplementary Figure 4e). Different levels of sgRNA coverage (25x, 50x, 75x, 100x and 500x) during screening revealed, for both the minimal and full library, lower coverage has no impact on identifying essential genes (Supplementary Figure 5a). Replicate correlation at the gene-level is lower with the minimal library, this is to be expected when considering the lower number of sgRNAs per gene (Supplementary Figure 5b).

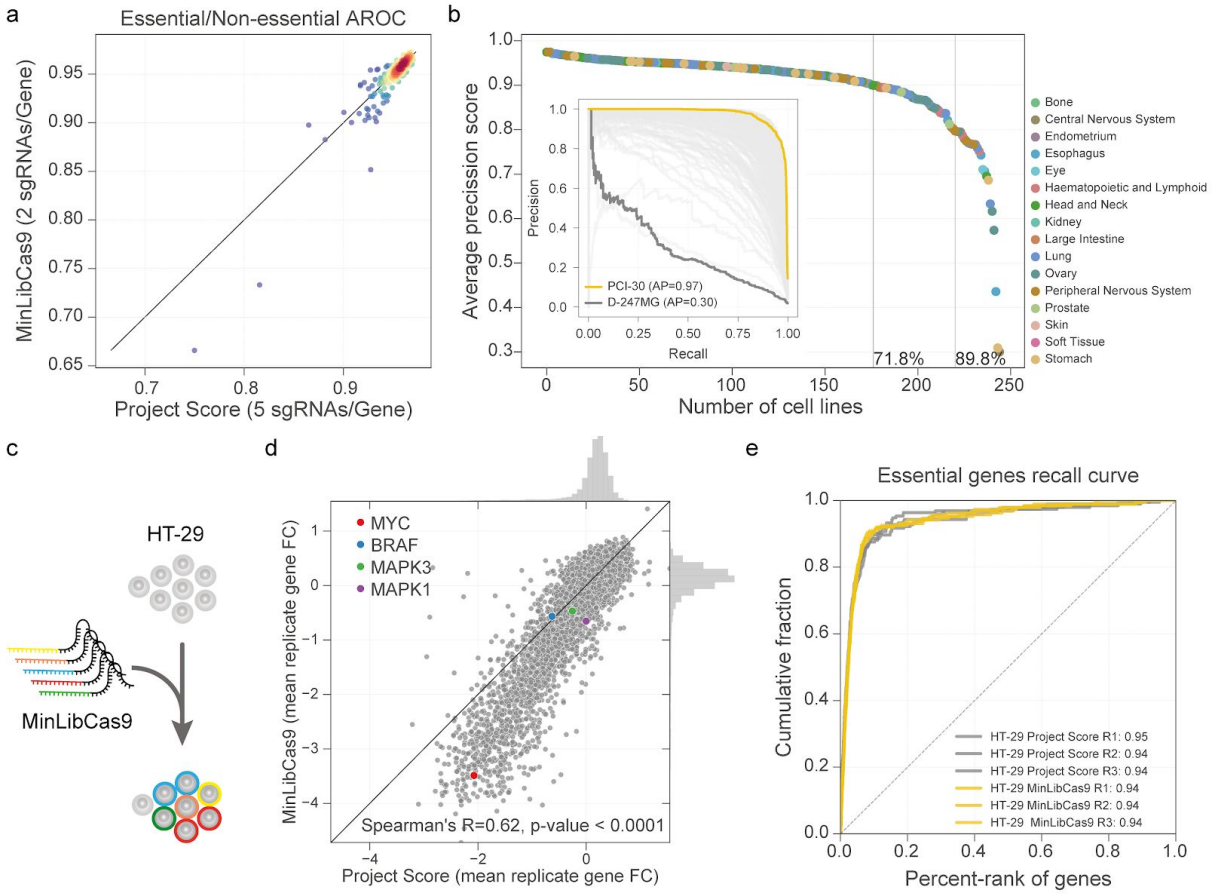


Figure 2. Benchmarking of minimal CRISPR-Cas9 library. a*, Standardized area under the receiver operating characteristic curve for 245 cell lines at 20% FDR for the essential genes calculated using the minimal and original full library.* ***b****, Average Precision (AP) scores to classify significant gene dependencies identified at 1% FDR in Project Score 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.* ***c****, CRISPR-Cas9 screen performed on HT-29 cancer cells using MinLibCas9 library.* ***d****, correlation between gene level fold-changes obtained with the two libraries.* ***e****, recall of essential genes across all replicates for HT-29 performed with MinLibCas9 and Project Score libraries.*

To further benchmark MinLibCas9, we assessed if it could recapitulate dependencies in more complex models and assays. We began by analysing CRISPR-Cas9 screens used to identify genes that enhance or suppress sensitivity to a BRAF inhibitor (dabrafenib) in a partially-sensitive BRAF-mutant colorectal cancer cell line (HT-29) (Supplementary Figure 6a). Using *in silico* down-sampling analysis, gene fold-changes with both libraries were strongly correlated (Spearman's R=0.72, p-value < 0.001) and the time-series profiles of top dependencies 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 Project Score and *in silico* down-sampled minimal library were strongly correlated (average Spearman's R=0.70), confirming the minimal library provides similar replicates, correlation and capacity to identify known essential genes (Supplementary Figure 7).

Finally, we synthesised and cloned the MinLibCas9 library and re-screened the HT-29 colorectal cancer cell line (Figure 2c, Supplementary Figure 8a and 8b). Compared to HT-29 screens performed in Project Score [13], we observed good correlation between gene-level fold-changes (Figure 2d, Supplementary Figure 8c) and similar recall of essential genes (Figure 2e, Supplementary Figure 8d). Consistent with the subsampling analysis, MinLibCas9 showed an increased dynamic range with stronger overall fold-changes, which improved the identification of previous cancer dependencies for this particular cell line, including MYC copy-number amplification, BRAF V600E gain-of-function mutation, and ERK1/2 (MAPK1/2) sensitivity to drug inhibition [31–33] (Figure 2d and Supplementary Figure 8e). These results confirm that MinLibCas9 can identify cancer cell dependencies and provides improved dynamic range, which is important to distinguish more subtle cancer vulnerabilities.

In summary, we designed an optimised minimal genome-wide human CRISPR-Cas9 library (MinLibCas9) using previously reported experimental data to select and rank sgRNAs, and validated its utility in multiple experimental settings. MinLibCas9 is at least 42% smaller than most currently used libraries and preserves the sensitivity and specificity required to identify gene dependencies. Our reference guide library, encompassing the most broadly accepted libraries with comprehensive and standardised sgRNA efficiency metrics, is a valuable resource to support the selection of optimised CRISPR-Cas9 reagents. Underlying the importance smaller CRISPR libraries, recent studies have proposed to either exploit large-scale CRISPR-Cas9 data-sets to empirically improve library design [34], build smaller complementary libraries from single genome-wide library [35], or by utilising multiplexed CRISPR-Cas9 systems [36]. Uniquely, MinLibCas9 combines some of these strategies to design a minimal and optimised sgRNA library, which unlocks the application of genome-wide screens to complex models currently limited to the delivery of libraries focused on predefined and small gene sets. 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 [37,38]) and to build large-scale genetic interaction libraries.

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 compared 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). Next, 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 down-sampling 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 [6,30]. Similarly to Allen et al. [14], 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 and using the gene fold-changes obtained with MinLibCas9. Curves were summarized using average precision (AP) scores, defined as follows:

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

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 [39] 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, whereas 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 [13] and for 73,911 sgRNAs across 1,257 samples (562 unique cancer cell lines) of the Broad DepMap19Q2 data-set [17,29] (Supplementary Table 4).

Reference Master CRISPR-Cas9 library

All the sgRNAs described in the Project Score [13], Avana [9], Brunello [9,12] and TKOv3 [11] libraries were assembled into a single reference master library containing a total of 354,715 sgRNAs with a median of 19 guides per gene (Supplementary Table 2). The location of each guide and PAM on the GRCh37 assembly was confirmed for the Project Score, Avana and TKoV3 libraries using CasOffinder [40], before being lifted-over to the GRCh38 assembly. Unique CRISPR_IDs and off-target summaries were extracted from the WGE CRISPR database [20] for all sgRNAs using the GRCh38 genomic coordinates and confirmed by sequence identity, with any inconsistent matches manually verified. Where possible guides were complemented with efficiency scores from Rule Set 2 [9,15], JACKS [14], FORECasT in-frame indels [21]. Guides from Avana and Project Score libraries were also annotated with KS scores estimated from large-scale screens available for each library [13,17,29]. Additionally, CRISPOR resource [22] was used to import off-target MIT-specificity scores [23] and on-target efficiency CRISPR-scan scores [24]. All CRISPOR scores were extracted from the crisprAllTargets dataset in the UCSC genome browser using the TableBrowser function.

Although ≥ 4 poly-T stretched can act as RNA Polymerase (Pol III) terminators for the often used U6 promoter [41], we determined that the majority of sgRNAs containing poly-T

stretches of 4 and 5 (no sgRNA in the master library had more than 5 long T-stretches) are in fact expressed. Most poly-T containing sgRNAs display consistent effects with other non-poly-T containing sgRNAs targeting the same gene and have good sgRNA metrics. The effect varies depending on the position of the poly-T within the sgRNA, with decreased efficiencies observed closer to the PAM motif (Supplementary Figure 4b). This is consistent with previous studies, which specify that full termination is only achieved with a poly-T stretch ≥ 6 [41]. Furthermore, substitution of the first base of the sgRNA to a G is recommended to improve guide expression [16] and therefore reduces the number of 4 T-stretches and no 5 T-stretch is considered in the reference library. Annotation is provided if a 4 or 5 poly-T stretch is contained within the sgRNA sequence (excluding PAM motif).

CRISPR-Cas9 sgRNA coverage

KM-12 colorectal carcinoma cancer cell lines were CRISPR-Cas9 screened with Project Score library similarly to Behan, et al. [13] 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. [13] 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 on 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 [42] 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 Project Score 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, ATCC identifier HCM-SANG-0270-C20), 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 [43]. 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 [13]. 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 as pellets 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.

Design of minimal genome-wide CRISPR-Cas9 library

A minimal genome-wide library was assembled from the master reference library by ranking sgRNAs that minimise off-target and maximise on-target effects. Of the 354,715 total guides, 497 sgRNAs that did not match any position in GRCh38 and/or targeted any gene were removed. Additionally, 738 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

(Project Score 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. [16], with no filter based on guide efficacy metric, similar to amber sgRNAs.

For all protein-coding genes defined in HGNC [44], 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 Project Score 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 the Project Score library.

The assembled minimal library covers 18,761 protein-coding genes with 37,522 sgRNAs (33,986 Project Score; 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 Project Score 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).

Plasmid construction for minimal genome-wide CRISPR-Cas9 library

All plasmids (including the Human sgRNA MinLibCas9) are in the process of being deposited with Addgene. MinLibCas9 sgRNA sequences can be found in Supplementary Table 3 and primer sequences used in the construction of the MinLibCas9 sgRNA library can be found in Supplementary Table 9.

The Human sgRNA MinLibCas9 was constructed as previously described [8]. With the following exception; a modified version of the vector backbone pKLV2-U6gRNA(BbsI)-PGKpuro2ABFP-W [8] was used to improve downstream pooled oligonucleotide cloning. Briefly, the ccdB resistance gene cassette was PCR amplified from pKLV1-fl-U6gRNA(BbsI)-ccdB-PGKpuro2ABFP (a kindly provided by E. Metzokapian) with the following oligos, Gibson_pKLV2-ccdb_Fwd and Gibson_pKLV2-ccdb_Rev, which included appended oligonucleotide sequences complementary to the pKLV2-U6gRNA(BbsI)-PGKpuro2ABFP-W backbone and BbsI restriction enzyme sites. The amplified ccdB gene product was cloned into the pKLV2-U6gRNA(BbsI)-PGKpuro2ABFP-W plasmid, linearized by BbsI digestion, by Gibson Assembly Cloning (New England Biolabs) to generate the modified pKLV2-U6gRNA(BbsI)-ccdB-PGKpuro2ABFP-W plasmid vector. This was subsequently transformed into ccdB Survival E.coli (Invitrogen).

A single stranded oligonucleotide pool containing all guide RNA sequences was synthesized by TWIST Bisocience. Sequenced oligonucleotides included primer appends for generating double stranded oligonucleotides compatible for Gibson assembly into the pKLV2-U6gRNA(BbsI)-ccdB-PGKpuro2ABFP-W plasmid vector (i.e. Pool_PCR_Fwd and Pool_PCR_Rev). Pools were amplified using Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs), PCR cycling conditions: 98 °C for 30s, 98 ° for 10s, 67 ° for 10s, and 72 °C for 15s, for 14 cycles with a final extension of 72 °C for 2 min. Amplicons were PCR-purified using (DNA clean and concentrator kit, Zymo) and cloned into 100ng of BbsI linearized pKLV2-U6gRNA(BbsI)-ccdB-PGKpuro2ABFP-W vector by Gibson Assembly Cloning (New England Biolabs) following the manufacturer's instructions. Multiple Gibson assemblies were pooled, ethanol precipitated and transformed into 200l of electrocompetent E.coli (Lucigen Endura™ ElectroCompetent Cells, Lucigen). 4 x 250ml LB preps (supplemented with 100 µg mL−1 carbenicillin) were inoculated and grown at 37 °C for 16h. Plasmid DNA (pDNA) was extracted using a Qiagen maxi prep kit.

Illumina sequencing of sgRNAs and analysis of guide distribution was performed as follows. 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. For the plasmid library, total read counts per sgRNA were calculated using the count_spacers python script [45]. Illumina sequencing of the plasmid library identified 88.3% of reads were perfect matches when compared to predicted sgRNA sequences. All guides were detected. The read distribution of each sgRNA was determined using the Ineq package

in R (version 3.5.3) to calculate both the Lorenz-curve and Gini-coefficient (Supplementary Figure 8 a and b).

Lentiviral vectors and packaging plasmids (psPax2 and pMD2.G, Addgene) were transfected into 60% confluent HEK293 cells using Lipofectamine LTX (Gibco) at the following ratio: 7.5 µg lentiviral vector (MinLibCas9 plasmid DNA), 18.5 µg psPax2 and 4 µg pMD2.G per 15 cm dish.

CRISPR-Cas9 MinLibCas9 HT-29 screens

Titrations of the lentiviral packaged sgRNA library were performed to identify the volume of supernatant to achieve 30% transduction efficiency, cells were analysed for stable BFP expression using flow cytometry 72 hours post transduction. Using this volume viral transduction of 12.5 $x10^6$ (100X coverage of the sgRNA library, consistent with what was used for the original Project Score dataset) Cas9 expressing HT-29 cells was performed in technical triplicate. 72 hours post viral transduction 30% transduction efficiency was confirmed by flow cytometry and 1 week selection with 2ug/ml puromycin started. At 14 days post library transduction cells were harvested and $2.5x10^7$ cells were pelleted; library coverage was also tested by analysis of BFP expression and found to be 62%. A minimum of 1.9 $x10^7$ cells were maintained throughout the 2 week screen. Genomic DNA extraction and amplification was performed as per Behan et al. [13], and sequencing was performed on a HiSeq2500.

References

1. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.

2. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343:84–7.

3. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. science.sciencemag.org; 2014;343:80–4.

4. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014;11:783–4.

5. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350:1096–101.

6. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al.

High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163:1515–26.

7. Ma H, Dang Y, Wu Y, Jia G, Anaya E, Zhang J, et al. A CRISPR-Based Screen Identifies Genes Essential for West-Nile-Virus-Induced Cell Death. Cell Rep. 2015;12:673–83.

8. Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. 2016;17:1193–205.

9. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34:184–91.

10. Park RJ, Wang T, Koundakjian D, Hultquist JF, Lamothe-Molina P, Monel B, et al. A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. Nat Genet. 2017;49:193–203.

11. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 . 2017;7:2719–27.

12. Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, et al. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun. 2018;9:5416.

13. Behan FM, Iorio F, Picco G, Gonçalves E, Beaver CM, Migliardi G, et al. Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. Nature. 2019;568:511–6.

14. Allen F, Behan F, Khodak A, Iorio F, Yusa K, Garnett M, et al. JACKS: joint analysis of CRISPR/Cas9 knockout screens. Genome Res. 2019;29:464–71.

15. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014;32:1262–7.

16. Koike-Yusa H, Li Y, Tan E-P, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32:267–73.

17. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49:1779–84.

18. Shi J, Wang E, Milazzo JP, Wang Z, Kinney JB, Vakoc CR. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol. 2015;33:661–7.

19. Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, et al. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. Elife [Internet]. 2016;5. Available from: http://dx.doi.org/10.7554/eLife.12677

20. Hodgkins A, Farne A, Perera S, Grego T, Parry-Smith DJ, Skarnes WC, et al. WGE: a CRISPR database for genome engineering. Bioinformatics. 2015;31:3078–80.

21. Allen F, Crepaldi L, Alsinet C, Strong AJ, Kleshchevnikov V, De Angeli P, et al.

Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol [Internet]. 2018; Available from: http://dx.doi.org/10.1038/nbt.4317

22. Haeussler M, Schönig K, Eckert H, Eschstruth A, Mianné J, Renaud J-B, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. 2016;17:148.

23. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31:827–32.

24. Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, Fernandez JP, Mis EK, Khokha MK, et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods. 2015;12:982–8.

25. Dempster JM, Pacini C, Pantel S, Behan FM, Green T, Krill-Burger J, et al. Agreement between two large pan-cancer CRISPR-Cas9 gene dependency data sets. Nat Commun. 2019;10:5817.

26. Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang C-Z, Ben-David U, et al. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discov. American Association for Cancer Research; 2016;6:914–29.

27. Munoz DM, Cassiani PJ, Li L, Billy E, Korn JM, Jones MD, et al. CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discov. 2016;6:900–13.

28. Gonçalves E, Behan FM, Louzada S, Arnol D, Stronach EA, Yang F, et al. Structural rearrangements generate cell-specific, gene-independent CRISPR-Cas9 loss of fitness effects. Genome Biol. 2019;20:27.

29. DepMap B. DepMap 19Q2 Public [Internet]. 2019. Available from: https://figshare.com/articles/DepMap_19Q2_Public/8061398

30. Dempster JM, Rossen J, Kazachkova M, Pan J, Kugener G, Root DE, et al. Extracting Biological Insights from the Project Achilles Genome-Scale CRISPR Screens in Cancer Cell Lines [Internet]. bioRxiv. 2019 [cited 2019 Aug 15]. p. 720243. Available from: https://www.biorxiv.org/content/10.1101/720243v1

31. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 2012;483:570–5.

32. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–61.

33. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell. Elsevier; 2016;166:740–54.

34. Henkel L, Rauscher B, Schmitt B, Winter J, Boutros M. Pooled CRISPR screening at high sensitivity with an empirically designed sgRNA library [Internet]. bioRxiv. 2020 [cited 2020 Apr 27]. p. 2020.04.25.061135. Available from: https://www.biorxiv.org/content/10.1101/2020.04.25.061135v1?%3Fcollection=

35. DeWeirdt PC, Sangree AK, Hanna RE, Sanson KR, Hegde M, Strand C, et al. Genetic screens in isogenic mammalian cell lines without single cell cloning. Nat Commun. 2020;11:752.

36. Peets EM, Crepaldi L, Zhou Y, Allen F, Elmentaite R, Noell G, et al. Minimized double guide RNA libraries enable scale-limited CRISPR/Cas9 screens [Internet]. bioRxiv. 2019 [cited 2020 Apr 29]. p. 859652. Available from: https://www.biorxiv.org/content/10.1101/859652v1

37. Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, et al. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. Cell. 2016;167:1867–82.e21.

38. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167:1853–66.e17.

39. Jones E, Oliphant T, Peterson P, Others. SciPy: Open source scientific tools for Python [Internet]. 2016. Available from: http://www.scipy.org/

40. Bae S, Park J, Kim J-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics. 2014;30:1473–5.

41. Gao Z, Herrera-Carrillo E, Berkhout B. Delineation of the Exact Transcription Termination Signal for Type 3 Polymerase III. Mol Ther Nucleic Acids. 2018;10:36–44.

42. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

43. Li X, Francies HE, Secrier M, Perner J, Miremadi A, Galeano-Dalmau N, et al. Organoid cultures recapitulate esophageal adenocarcinoma heterogeneity providing a model for clonality studies and precision therapeutics. Nat Commun. 2018;9:2983.

44. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017;45:D619–25.

45. Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, et al. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc. 2017;12:828–63.

Data availability

All data is contained in Supplementary Information.

Code availability

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

Acknowledgements

We acknowledge Joel Rein for helpful comments on the integration of WGE annotations. We would also like to thank the core support provided by the Cellular Operations teams, namely Cellular Generation and Phenotyping (CGaP), Gene Editing (GE), Flow Cytometry (CCF), Cellular RnD, with specific thanks to Lucy Kitchin and Claire Hardy. 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., M.S., D. A. J., C. M. B., S. P., O.D.

Formal analysis: E.G.

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

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

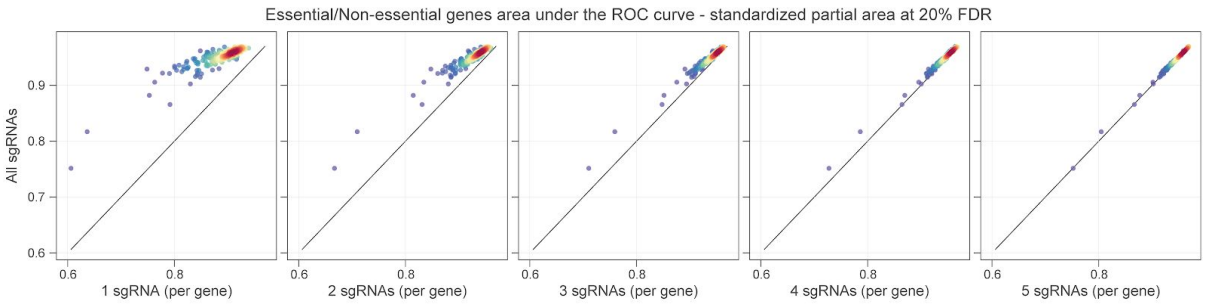
Writing - reviewing and editing: All authors

Visualisation: E.G.

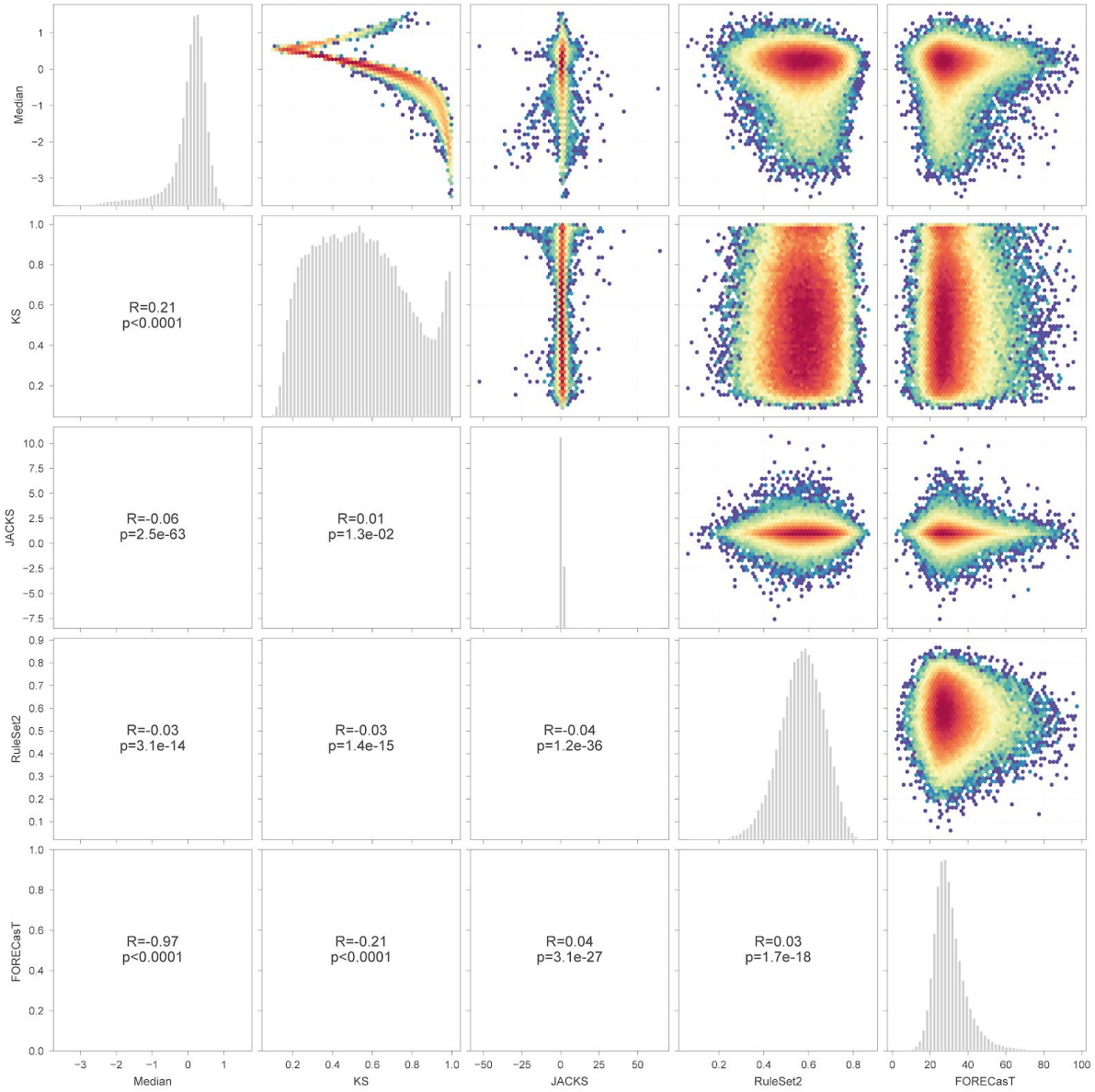
Supervision: O.D., D.P.S., F.I., L.P., K.Y., M.G.

Funding acquisition: M.G.

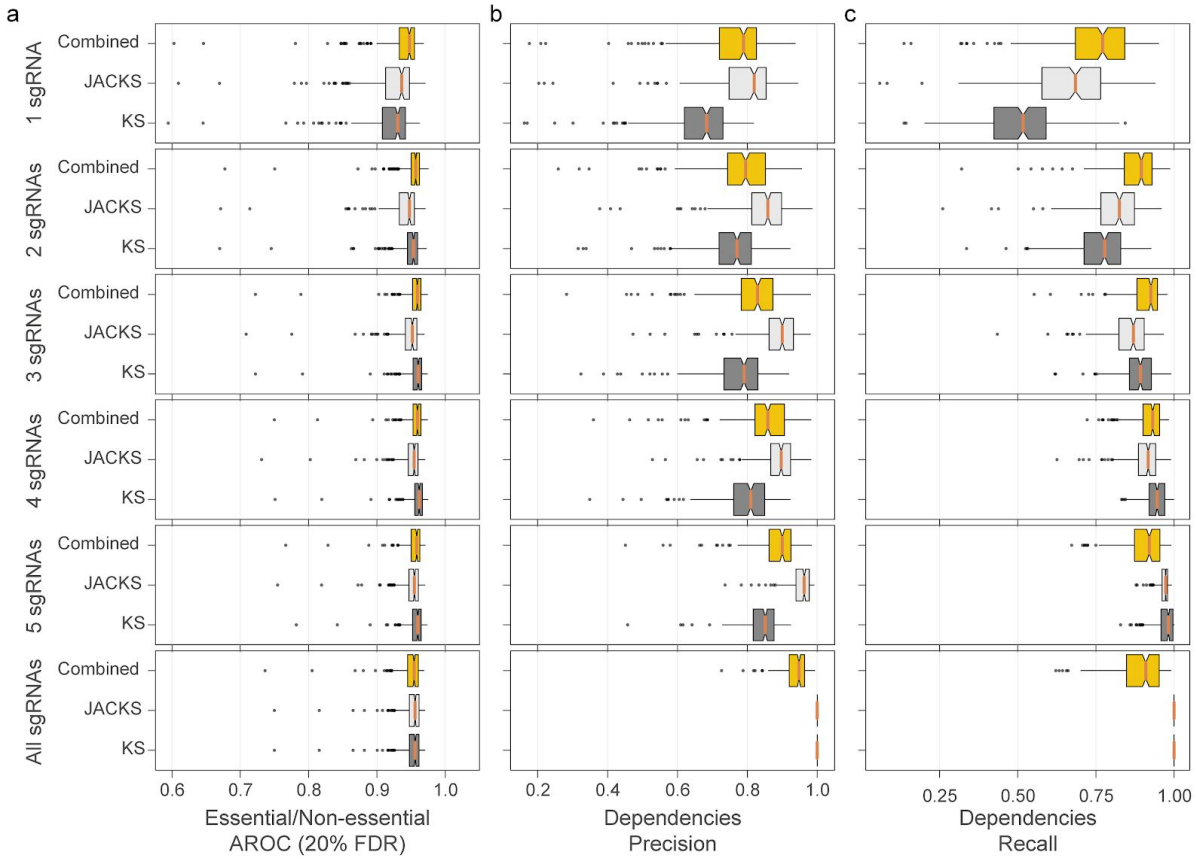
Supplementary figures



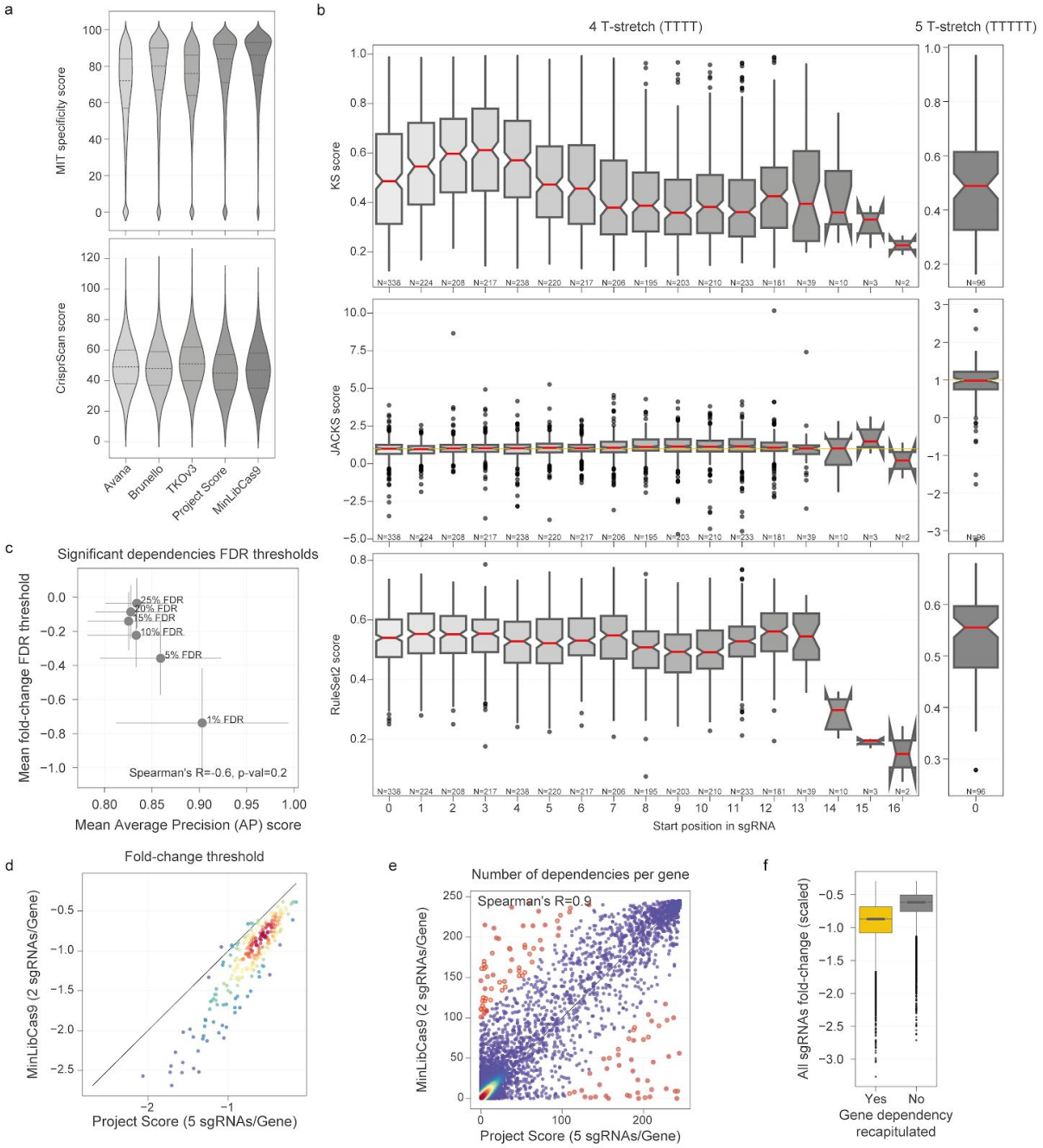
Supplementary Figure 1. Randomised selection of n sgRNAs per gene. *ROC curve derived from previously defined sets of essential and non-essential genes [6]. Standardized partial area under the ROC curve (AROC) calculated per cell line over the range of maximum false discovery rate of 20%. AROCs are compared between down-sampled 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 sgRNAs - KS metric (Kolmogorov Smirnov test) comparison to non-targeting guides, JACKS scores [14], Rule Set 2 [9] 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.*

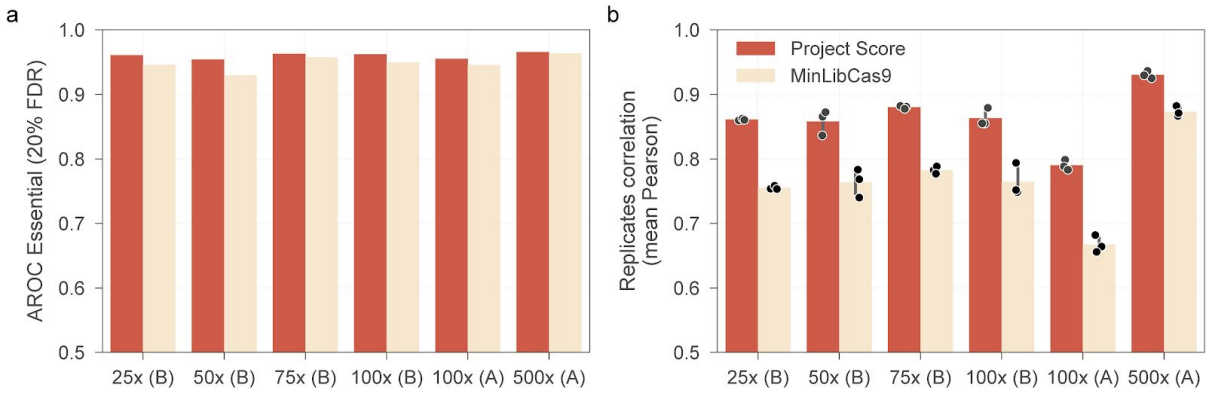


Supplementary Figure 3. Down-sample analysis of top n sgRNAs ranked using KS and JACKS metrics. *Combined score discards sgRNAs with a JACKS score outside the range of [0, 2] and then selects the top n sgRNAs according to the KS score (descending order, stronger KS scores and thereby stronger absolute 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 down-sampled library. Each box-and-whisker plot shows 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians (n=245 cancer cell lines).*

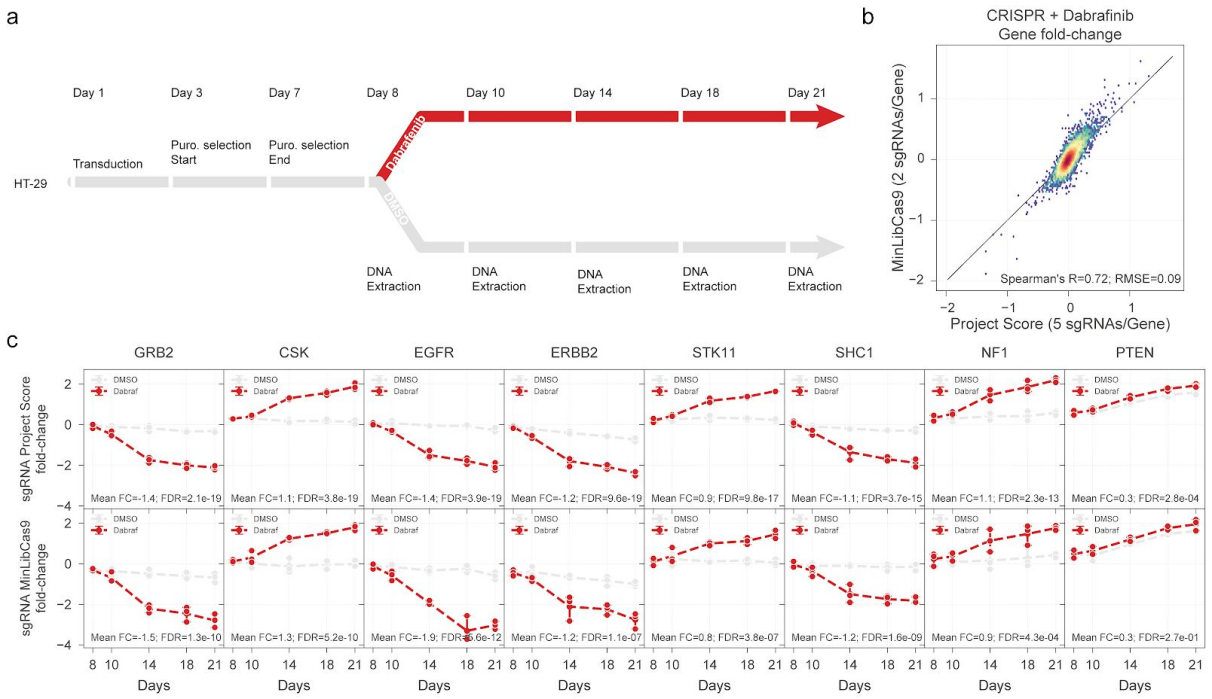


Supplementary Figure 4. MinLibCas9 benchmarking. a*, distributions of CRISPOR sgRNA scores.* ***b****, distributions of KS, JACKS and RuleSet2 scores for sgRNAs from Project Score library that contain T-stretches of 4 and 5.* ***c****, significant dependencies identified using different FDR rates between essential and nonessential genes (1%, 5%, 10%, 15%, 20% and 25%). MinLibCas9 fold-change thresholds at each FDR rate and the average precision (AP) of the gene dependencies identified in MinLibCas9 that are also found with Project Score library . Error bars represent standard deviation.* ***d****, MinLibCas9 fold-change threshold for each cell line identified at 1% FDR of essential versus non-essential genes plotted against those identified with Project Score library . Colour represents density.* ***e****, cumulative*

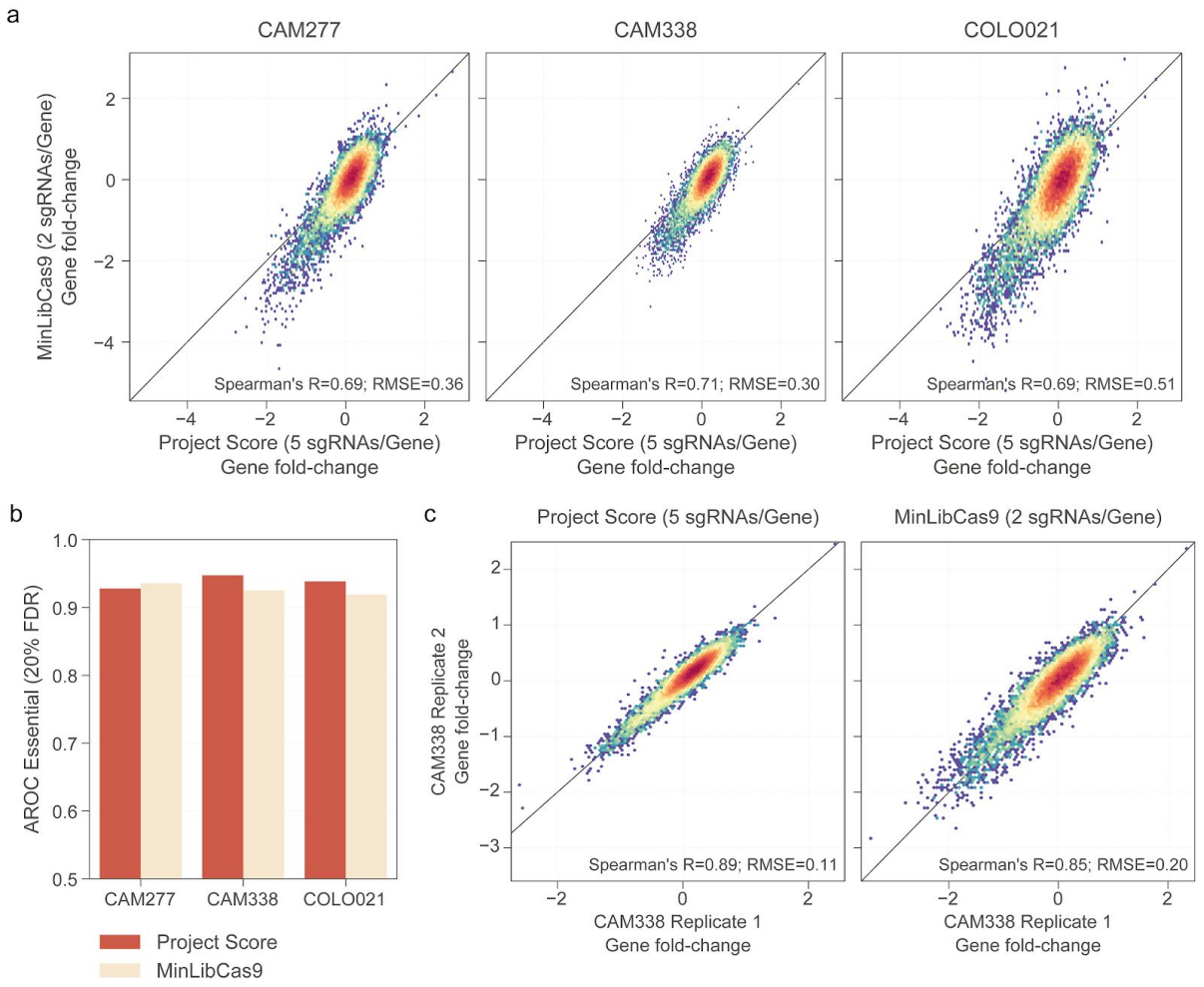
number of dependencies (at 1% FDR) identified in 245 cancer cell lines for each gene with both the full original library and the minimal library. ***f****, scaled fold-changes (median essential genes fold-change = 1; median non-essential genes fold-change = 0) of dependencies recapitulated (n=251,387) and missed (n=36,996) with the minimal library (two-sided Welch's t-test p-value < 0.001). Each box-and-whisker plot shows 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians.*



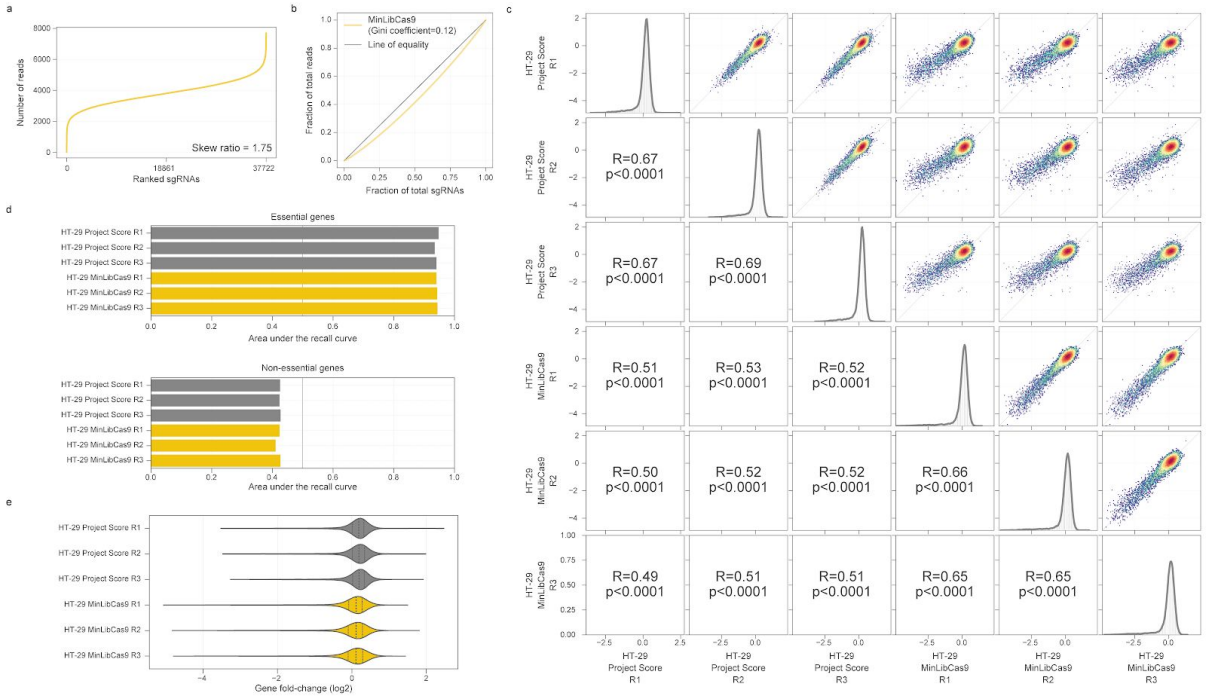
Supplementary Figure 5. sgRNA library coverage analysis in KM-12 cancer cells. a*, AROC of essential/non-essential genes at different guide coverage levels.* ***b****, technical replicates correlation. Comparisons are made between the original Project Score library and the in silico minimal library.*



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 Project Score library compared to the in silico MinLibCas9. Colour represents point density.* ***c****, time-series fold-changes of the top significantly essential hits (compared to control experimental arm, DMSO) obtained with Project Score and MinLibCas9 library, the values of three technical replicates are represented with the mean and by the minimum and maximum values as error bars.*



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. Colour coding represents point density.*



Supplementary Figure 8. MinLibCas9 screens in HT-29 cancer cell line. a*, Distribution of guides within MinLibCas9 library plasmid calculated to be comparatively even. Skew ratio of top 10% vs bottom 10% guides = 1.75 (minimum read count = 30, maximum read count 7698).* ***b****, lorenz curve of the MinLibCas9 plasmid library, Gini coefficient = 0.12.* ***c****, correlation of gene level fold-changes of the three technical replicates of MinLibCas9 screens versus the three technical replicates of HT-29 screens performed in Project Score. Spearman's Rho correlations are reported on the lower diagonal, fold-changes distributions are displayed in the diagonal, and upper diagonal shows the scatter plots between the two samples fold-changes colored by density.* ***d****, recall of previously defined essential (upper panel) and non-essential (lower panel) genes quantified by the area under the recall curve.* ***e****, distributions of gene level fold-changes for each sample. Box-and-whisker plot shows 1.5 x interquartile ranges and 5–95th percentiles, centres indicate medians.*

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.*

Supplementary Table 8. *MinLibCas9 raw counts for three technical replicates of HT-29.*

Supplementary Table 9. *Oligonucleotide and primer sequences.*