Henkel *et al. BMC Biology* (2020) 18:174
https://doi.org/10.1186/s12915-020-00905-1

BMC Biology

METHODOLOGY ARTICLE Open Access

Genome-scale CRISPR screening at high sensitivity with an empirically designed sgRNA library

Luisa Henkel[†], Benedikt Rauscher[†], Barbara Schmitt, Jan Winter and Michael Boutros[*]

Abstract

Background: In recent years, large-scale genetic screens using the CRISPR/Cas9 system have emerged as scalable approaches able to interrogate gene function with unprecedented efficiency and specificity in various biological contexts. By this means, functional dependencies on both the protein-coding and noncoding genome of numerous cell types in different organisms have been interrogated. However, screening designs vary greatly and criteria for optimal experimental implementation and library composition are still emerging. Given their broad utility in functionally annotating genomes, the application and interpretation of genome-scale CRISPR screens would greatly benefit from consistent and optimal design criteria.

Results: We report advantages of conducting viability screens in selected Cas9 single-cell clones in contrast to Cas9 bulk populations. We further systematically analyzed published CRISPR screens in human cells to identify single-guide (sg) RNAs with consistent high on-target and low off-target activity. Selected guides were collected in a novel genome-scale sgRNA library, which efficiently identifies core and context-dependent essential genes.

Conclusion: We show how empirically designed libraries in combination with an optimized experimental design increase the dynamic range in gene essentiality screens at reduced library coverage.

Keywords: CRISPR/Cas9, Genetic screens, sgRNA design, Gene essentiality, Functional genomics

Background

Over the past decades, genetic screens have been used extensively to interrogate gene function in an unbiased manner [1–3]. In recent years, CRISPR/Cas9 [4, 5] has emerged as a scalable method to introduce targeted gene knockouts with unprecedented efficiency and specificity. Genetic screens with the CRISPR/Cas9 system are now applied to probe the protein-coding and noncoding genomes of hundreds of cell types in different organisms [6–11]. These experiments have led to new insights into many biological processes. Progress has been made especially in the field of cancer genetics, where genome-wide CRISPR screens have resulted in gene essentiality maps for hundreds of tumor cell lines [12, 13].

Despite these advances, currently no generally accepted design principles for sgRNA libraries [14] and large-scale CRISPR screens exist and published experiments vary substantially in design and performance. This is particularly important as we reason that screens conducted with high CRISPR editing efficiency based on high Cas9 and sgRNA functionality allow screens at reduced coverage, which substantially lowers experimental efforts and costs. Furthermore, higher knockout efficiency is likely to increase the dynamic range of CRISPR screens, which in turn further allows to separate weak signals from screening noise and thus hit calling. In recent years, the design of sgRNAs and CRISPR libraries was improved based on

[*] Correspondence: m.boutros@dkfz.de
[†]Luisa Henkel and Benedikt Rauscher contributed equally to this work.
German Cancer Research Center (DKFZ), Division Signaling and Functional Genomics and Heidelberg University, BioQuant and Medical Faculty Mannheim, D-69120 Heidelberg, Germany

design rules derived from comparing the nucleotide composition of active and non-active guides [15–17], training models to identify predictive features of efficiently editing sgRNAs [18, 19] and more recently also deep learning algorithms [20]. While these efforts have greatly improved library performance, we reasoned that sequence predictions might not always fully reflect knockout performance and that many factors that influence sgRNA efficacy likely remain unknown and have thus not been considered in previous library designs. Potential further parameters might be DNA accessibility [21–23] or the presence of functional protein domains [24]. Therefore, we consider using consistent and strong viability phenotypes of sgRNAs previously used in a diverse range of cell lines to be a powerful predictor of their functionality.

In this study, we generated a new library, termed the Heidelberg CRISPR library as a new tool for pooled genome-wide CRISPR/Cas9 screens. For its design, we selected guides with consistent high on-target and low off-target activity based on phenotypes in previously published CRISPR screens as recorded in the GenomeCRISPR database [25]. We show that empirical selection of sgRNAs also prioritizes guides with high sequence scores according to different design rules. Moreover, we show that our library efficiently identifies core and context-dependent essential genes by screening for essential genes in a cell line that was not considered for library design, the human HAP1 cell line. Screening in Cas9 single-cell clones increased depletion phenotypes of essential genes compared to a Cas9 bulk population, increasing the overall dynamic range. Interestingly, the heterogeneity of editing efficiency of individual clones in Cas9 bulk populations seems to more strongly interfere with editing efficiency compared to ploidy, since editing in haploid and diploid HAP1 cells was in a similar range. Furthermore, screening in selected single-cell clones allowed hit calling at reduced library coverage, while essential genes were overall similar in all cell populations.

We believe that empirically designed libraries will be a useful extension to the current CRISPR/Cas9 toolkit. Our results further suggest that clonal cell populations with high Cas9 activity are attractive models for CRISPR/Cas9 screens at minimal library coverage, especially when no prior information is available to inform hit calling.

Results

Data mining of sgRNA-associated phenotypes allows empirical sgRNA design

We hypothesized that the very large number of results from previously published CRISPR screens could be utilized to systematically identify sgRNAs with high activity. Specifically, we reasoned that sgRNAs with strong and consistent effects across multiple experiments would intrinsically combine all known and unknown characteristics that guarantee high on-target activity. In addition, we assumed that we could avoid sgRNAs with off-target activity by comparing their phenotypes to other sgRNAs targeting the same gene. To this end, we analyzed 439 genome-scale fitness screens (negative selection for viability) from GenomeCRISPR, a database that contains sgRNA phenotypes from CRISPR screens in human cells [6–8, 10, 12, 16, 25–30]. We excluded all sgRNAs that could not be mapped to a protein-coding transcript region of the latest (GRCh38.p10) human reference genome [31]. In addition to the sgRNA phenotypes, we annotated each guide sequence with additional information including the number of targeted transcripts, the number of times the sgRNA was screened, and the number of predicted off-targets (see “Methods”) and their GC content.

We next aimed to identify sgRNAs with high on-target activity (Fig. 1). To this end, we reanalyzed each of the selected 439 screens in GenomeCRISPR using the BAGEL software [32]. BAGEL uses reference sets of core essential and nonessential genes [17] to compute Bayes factors that indicate whether a gene is more likely to be essential or nonessential. To exclude low-quality screens from further analysis, we generated precision-recall curves for each screen to quantify how well core and nonessential reference genes could be separated based on the BAGEL-derived Bayes factors [33]. We retained all screens for which the area under the precision-recall curve (AUC) was greater than 0.9 (406 out of 439; Additional file 1: Fig. S1 A-B). We then determined essential genes in each of these screens at 5% false discovery rate (FDR). We labeled each sgRNA as active if its target gene was essential according to BAGEL and if the sgRNA was among the 20% sgRNAs with the strongest fitness phenotypes in a screen (Fig. 1 and Fig. 2a, b).

Next, we identified and excluded sgRNAs with gene-independent toxic phenotypes [19], which can among other things be based on copy number amplifications. Copy number-induced phenotypes at off-target sites are most likely sporadic effects that occur only in individual cell lines. In order to avoid the selection of sgRNAs based on sporadic phenotypes, for example due to copy number amplifications at off-target sites, we required that an sgRNA has to have a phenotype in at least 5% of the screens in which it was used for it to be considered for phenotype-based selection (see “Methods” section). GenomeCRISPR provides an “sgRNA effect” score that indicates how strong an observed sgRNA phenotype was in comparison to all other sgRNAs in the same screen [25]. We grouped sgRNAs by their target genes and scaled and centered their effect scores to identify sgRNAs whose phenotypes strongly deviated from the phenotypes of other sgRNAs targeting the same gene (|sgRNA effect *z*-score| > 1.25). We labeled these sgRNA

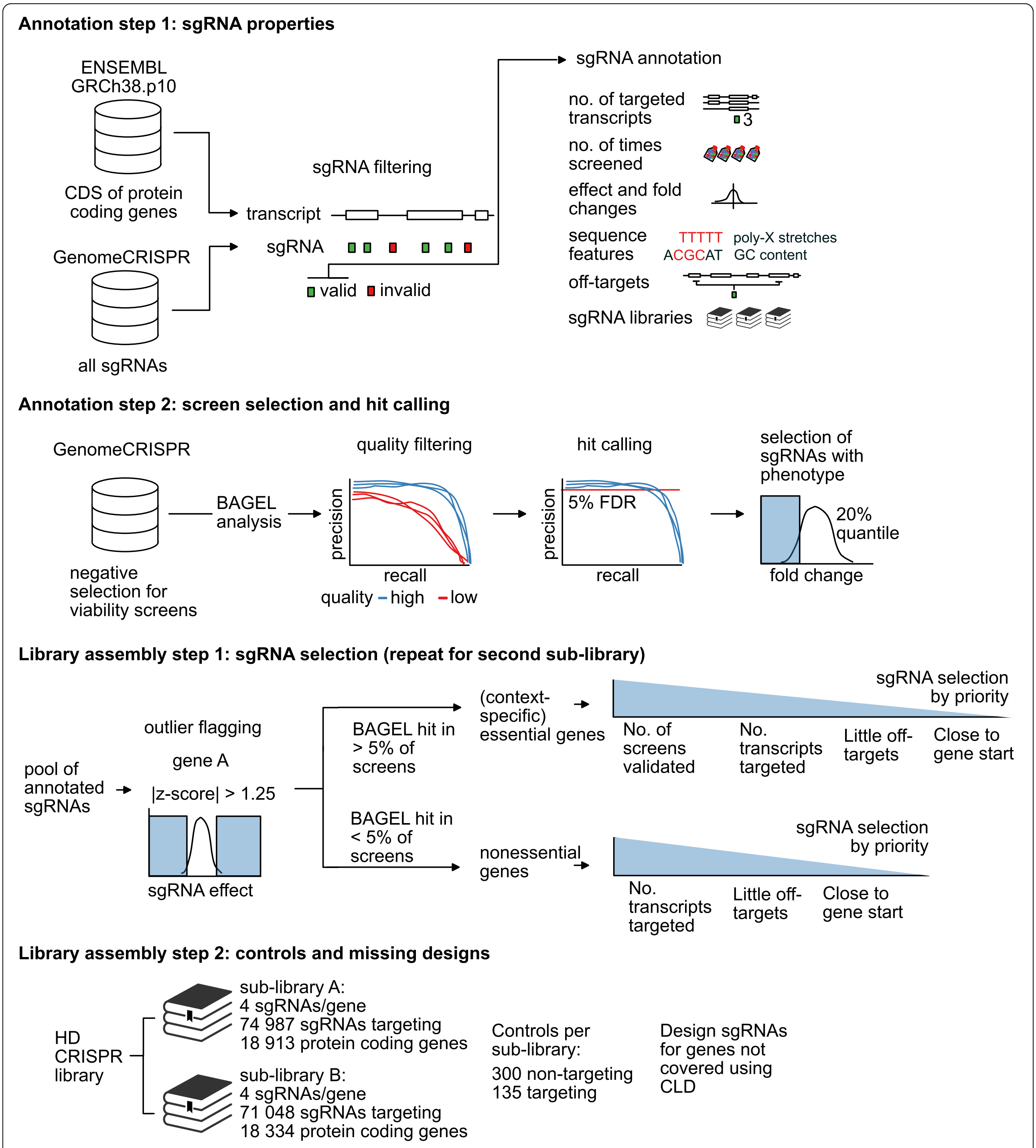

Fig. 1 Empirical design of the HD CRISPR library. Schematic illustration of the HD CRISPR library design process. sgRNA sequences that were previously used in negative selection screens contained in the GenomeCRISPR database were annotated with information on rationally selected sequence and phenotype features. All negative selection screens were reanalyzed with BAGEL. High-quality experiments were selected based on how well reference core and nonessential gene sets could be separated in those screens. sgRNAs with high on-target activity were then determined as sequences that both target an essential gene (as determined by BAGEL) and that rank among the 20% most strongly depleted sequences in the screen. Next, sgRNAs that showed unexpected phenotypes compared to other sgRNAs targeting the same gene were flagged as outliers with potential off-target effects. sgRNA sequences were then selected from the resulting pool of sequences to design a genome-wide library consisting of two mutually exclusive sub-libraries A and B, prioritizing sgRNAs with high on-target and low off-target activity

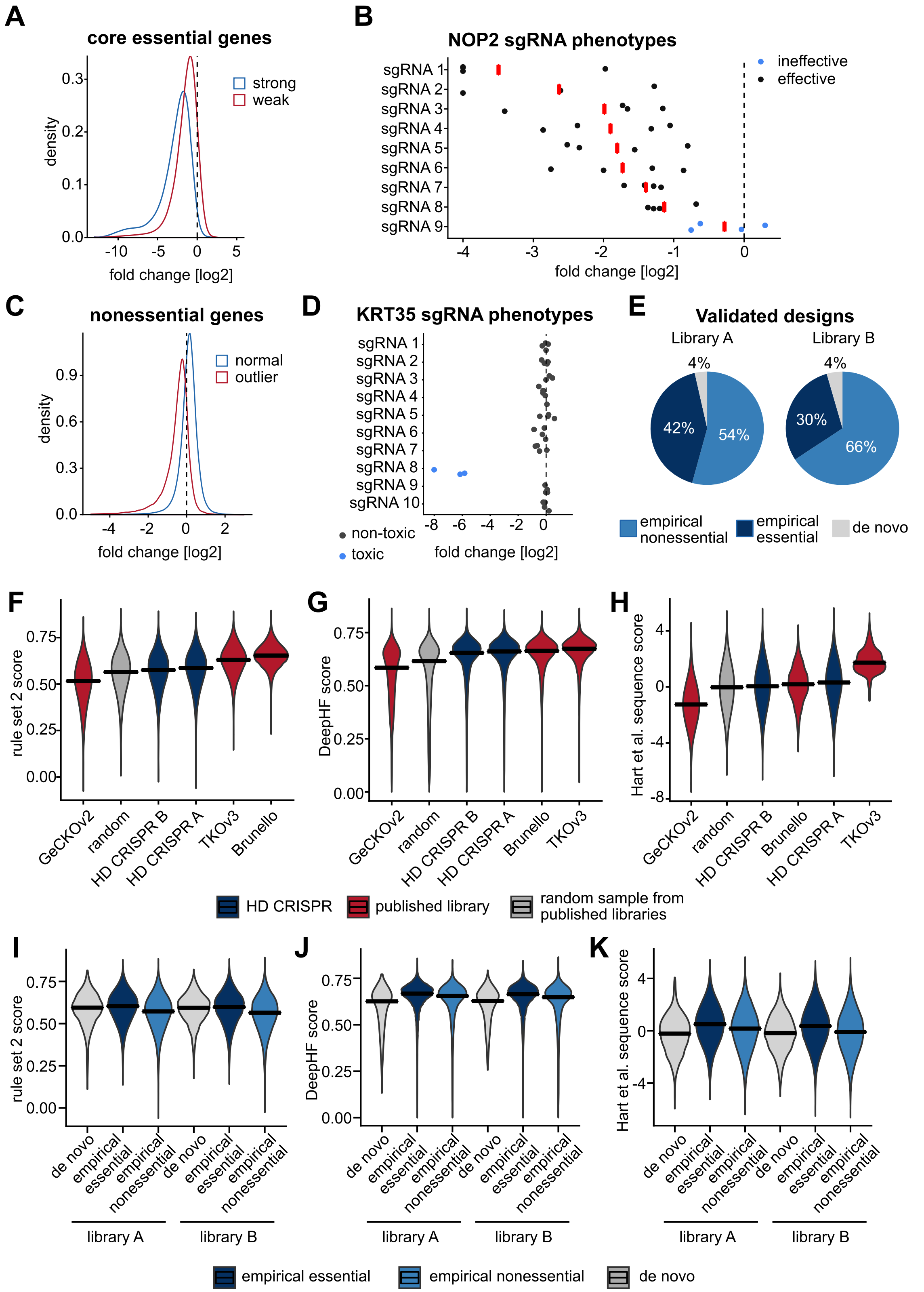

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Empirically selected sgRNAs in the HD CRISPR library. **a** Distribution of log2 fold changes for sgRNAs targeting core essential reference genes. The blue curve represents sequences with strong on-target phenotypes. The red curve represents sgRNAs with weak on-target phenotypes (see "Methods"). **b** Phenotypes of sgRNAs targeting the core essential gene NOP2 for 4 screens performed with the library described in Wang et al. [27]. Compared to other NOP2-targeting sgRNAs, sgRNA 9 showed unexpectedly weak depletion in these experiments and was thus labeled "ineffective". **c** Distribution of log2 fold changes for sgRNAs targeting nonessential reference genes. The blue curve represents sequences with low off-target activity. The red curve represents sgRNAs that led to unexpectedly strong phenotypes. **d** Phenotypes of sgRNAs targeting the nonessential gene KRT35 in 4 screens performed with the library described in Wang et al. [27]. Unlike other KRT35-targeting sgRNAs, sgRNA 8 consistently displayed toxic phenotypes in these experiments and was therefore marked as "toxic". **e** Percentage of sgRNA sequences in sub-libraries A (left) and B (right) that could be selected based on empirical evidence from published screening experiments. Empirical essential sgRNAs were selected based on inducing a viability phenotype, empirical nonessential sgRNAs based on the absence of a toxic phenotype. **f**–**h** Calculated sequence scores applying either the rule set 2 [19], the DeepHF [20] or the Hart et al. [17] algorithms. Score performance of the HD CRISPR sub-libraries A and B was benchmarked against the libraries whose design is based on respective scores (Brunello for rule set 2, TKOv3 for Hart et al.) if available as well as the GeCKOv2 library and a random sample of sgRNAs from published libraries. The DeepHF score was used as an independent measure none of the investigated libraries was designed on. **i**–**k** Comparison of sgRNA scores for empirically and de novo-designed sgRNAs within the HD CRISPR sub-libraries

sequences as potentially off-targeting and excluded them from the library design (Fig. 2c, d). When less than 8 sgRNAs meeting these criteria were available from published libraries to target a gene, we used the CRISPR library designer (CLD [34];) to design new sgRNAs (Additional file 1: Fig. S1 C).

We then selected sgRNAs for a new library, named the Heidelberg (HD) CRISPR library (Fig. 1; Additional file 2: Table S1). This library consists of two independent, mutually exclusive sub-libraries A and B that each contains 4 sgRNAs per gene, targeting 18,913 and 18,334 protein-coding genes, respectively (Fig. 1, Additional file 3: File S1, Additional file 4: File S2). For selection, we prioritized sgRNAs that showed high on-target activity (as determined above) in a large number of screens. If no information about sgRNA on-target activity was available (e.g., for nonessential genes), we picked sgRNAs that target constitutive exons with little predicted off-target effects (see "Methods") close to the transcription start site (Additional file 1: Fig. S1 D-E). These selection criteria are based on the assumption that the majority of algorithms do not inherently enrich for guides with off-target effects, which we, however, can also not exclude. Genes and their respective sgRNAs were categorized as essential if respective sgRNAs got depleted in at least 5% of the screens analyzed. In total, 42% of the sgRNA for libraries A and 30% of the sgRNAs for library B met these criteria. An additional 54% of sgRNAs for library A and 35% of sgRNAs for library B could be empirically selected based on their uniform phenotype for presumably mainly nonessential genes. For only 4% of sgRNAs, de novo design was necessary (Fig. 2e). Sub-library A contains sgRNAs that ranked best according to our design criteria to enable high-quality screens at a low library coverage. Sub-library B contains second-tier sgRNAs and can be used to supplement sub-library A when a higher sgRNA coverage is desired. After sgRNA prefiltering, we could select from on average 25 different sgRNAs per gene (Additional file 1: Fig. S1 F). In the filtered guide dataset, sgRNA phenotypes targeting the same gene were diverse, resulting in a large difference of the maximal and the minimal sgRNA effect score for the majority of genes. In contrast, our filtering criteria for guides with high on-target and low off-target activity resulted in a narrow distribution of guide phenotypes targeting the same gene (Additional file 1: Fig. S1 G). As an independent evidence of sgRNA performance, we further aimed to benchmark sgRNAs selected for the HD CRISPR library based on commonly used sgRNA design scores. For comparison, we used sgRNAs from the GeCKOv2, TKOv3, and Brunello libraries and also generated a sample of 70,000 randomly selected sgRNAs from published libraries [6, 8, 16–19, 27–29, 35]. For scoring, we applied the rule set 2 design rules, based on which the Brunello library was designed [19], as well as the sequence score developed for the design of the TKOv3 library [17]. As an independent metric, we also evaluated sgRNAs using the more recently published DeepHF score [20], an sgRNA activity prediction score based on deep learning algorithms, which has not been used for the design of either of the evaluated libraries. For each of the selected scores, the two HD CRISPR sub-libraries outperformed the GeCKOv2 library and the randomly picked sample of published sgRNAs (Fig. 2f–h). The performance for the rule set 2 and Hart et al. sequence score were slightly lower for the HD CRISPR library (median 0.59 and 0.58 for sub-libraries A and B, respectively) in comparison to the Brunello and TKOv3 libraries (median 0.65 for Brunello and 0.63 for TKOv3). Since the HD CRISPR library has solely been designed by relying on sgRNA-associated phenotypes but not by considering any of these design rules, this is not surprising. However, using the independent deep learning DeepHF scoring system, all three empirically designed libraries (Brunello, HD CRISPR, and TKOv3) performed similar and outperformed GeCKOv2 and the random sample (median scores: HD CRISPR A = 0.66; HD CRISPR B = 0.66; Brunello = 0.66; TKOv3 = 0.67; GeCKOv2 = 0.59; Random = 0.62) (Fig. 2g). Interestingly, sgRNAs empirically selected based on the induction of a viability phenotype scored better than sgRNAs selected for nonessential genes as well as de novo designed sgRNAs, for which only common design rules were applied (Fig. 2i–k).

Cas9-editing efficiency can be improved by pre-selection of highly editing single-cell clones

Next, we further addressed how Cas9 activity could be increased to improve the sensitivity of CRISPR screens. We asked whether knockout efficiency could be enhanced in large screening settings by pre-selecting single-cell clones with high Cas9 editing efficiency. Bulk populations of transduced cells are often heterogeneous in their transgene expression, largely due to differences in lentiviral integration, epigenetic silencing of exogenous DNA or the selection process during transfection or transduction [36]. This heterogeneity can impact downstream processes such as Cas9-editing efficiency. Indeed, a bulk population of Cas9 expressing cells has also previously been shown to contain clones with rare to even absent Cas9 editing efficiency, which was suggested to be based on an APOBEC3 mutational signature in the Cas9 coding sequence [16]. Recently, an in-depth comparison of editing efficiency of 96 clonal Cas9 mouse embryonic stem cell lines revealed very diverse knockout efficiencies from 0 to 100%, while the bulk population showed about 84% knockout efficiency and only few clones with editing above 95% could be identified [37]. However, to our knowledge, no direct comparison of pooled CRISPR viability screening performance in clonal and bulk Cas9 cell lines has been done so far.

In order to be able to assess such potential differences in more detail, we sorted single cells from the HAP1 Cas9 bulk population and measured Cas9 editing efficiency for several clones (Fig. 3a). We define editing efficiency as the combination of on-target DNA cutting efficiency and the subsequent induction of mutations by error-prone DNA damage repair. From the sorted clones, we selected two single-cell clones, henceforth referred to as SCC11 and SCC12, with high Cas9 editing activity. For our knockout experiments, we generated an sgRNA expression vector harboring an improved sgRNA scaffold [16, 38]. The vector further features a stuffer sequence encoding GFP and a lentiviral backbone similar to the previously published pLCKO vector [8]. Cloning of sgRNA sequences results in removal of the GFP stuffer (Additional file 5: Fig. S2 A), which allows to estimate the extent of remaining non-digested vector backbone in library preparations and thus cloning efficiency (Additional file 5: Fig. S2 B-C), while upon transduction the stuffer does not interfere with assays where GFP is used as a readout (Additional file 5: Fig. S2 B). The functionality of the HDCRISPRv1 vector was confirmed by knockout of the surface marker gene *CD81* in HAP1 Cas9 bulk cells (Additional file 5: Fig. S2 D, Additional file 6: Table S2). Using this vector to compare knockout efficiency in Cas9 bulk and single-cell clones, knockout of the core essential *POLR2E* gene led to much stronger depletion of viable cells in SCC11 and SCC12 compared to the Cas9 bulk population (Fig. 3b). Similarly, when addressing the amount of remaining surface protein upon knockout of *CD46* using two individual guides, depletion was 20 to 30% stronger in single-cell clones compared to the bulk population (Fig. 3c; Additional file 6: Table S2). We considered it possible that editing efficiency might be affected by the ploidy of the different cell lines. While HAP1 cells are a haploid cell line [39], their haploid state tends to be unstable. In a mixed population of haploid and diploid HAP1 cells, diploid cells have been shown to enrich over time due to a proliferative advantage, while the haploid state can be prolonged in clonal populations starting from a single haploid cell [40]. In line with this, we identified a larger proportion of clearly diploid cells in the HAP1 Cas9 bulk population compared to the two single-cell clones (~ 11.5% definite diploid in HAP1 Cas9 bulk vs. ~ 1.5% in HAP1 Cas9 SCC11 and ~ 6.3% in HAP1 Cas9 SCC12) (Additional file 7: Fig. S3 A). To rule out that ploidy has a major impact on editing efficiency in the HAP1 cell line, we sorted enriched haploid (68.2% for SCC11 and 66.8% for SCC12) and diploid populations (43.2% and 34.2% for SCC11 and SCC12, respectively) from the HAP1 Cas9 SCC11 and SCC12 cell lines. Subsequently, we directly compared editing efficiency in enriched haploid and diploid populations originating from the same single-cell clone. While we indeed observed a slightly lower editing efficiency in diploid cells compared to haploid controls, this difference was smaller than on average 6% for both cell lines and sgRNAs (Additional file 7: Fig. S3 B-C). This argues for other factors than ploidy to be the main drivers of differences in editing efficiency.

The HD CRISPR library identifies core and nonessential genes at high precision

In order to address screening performance of the HD CRISPR library, the sub-libraries A and B were cloned independently into the corresponding HDCRISPRv1 vector. Quality controls of the resulting plasmid preparations revealed a narrow sgRNA distribution with negligible background of non-digested vector backbone (Additional file 8: Fig. S4). Both libraries were screened in parallel in the HAP1 Cas9 bulk population and the two Cas9 single-cell clones in two independent replicates for 14 days to further assess the impact of Cas9 editing efficiency and the extent of clonality effects on hit calling (Fig. 4a; Additional file 9: Fig. S5 A). Since our reference set of published CRISPR screens used for library design did not include any screen conducted in HAP1 cells, we

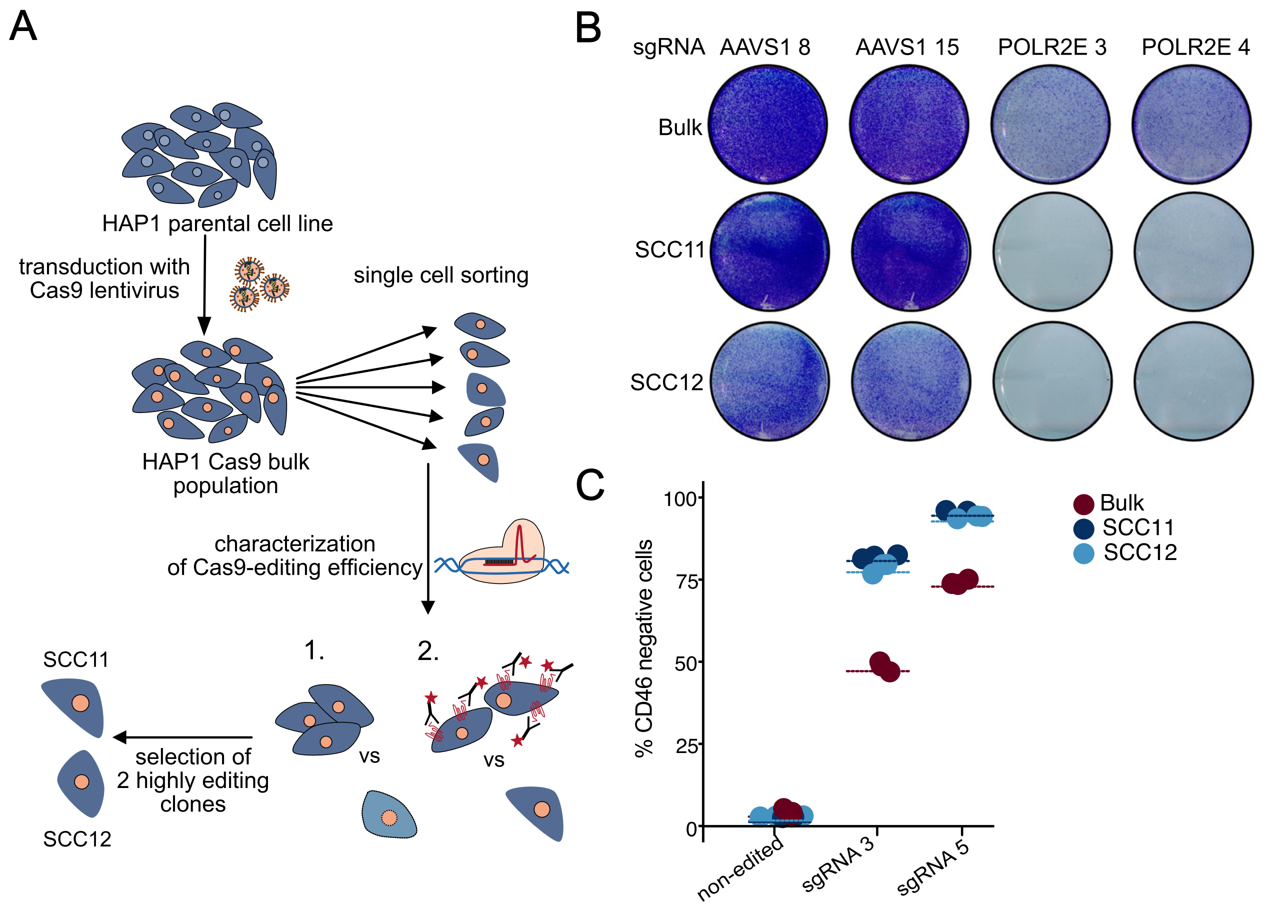

Fig. 3 Selected Cas9-expressing single cell clones show stronger editing efficiency compared to a Cas9 bulk population. **a** Workflow for the selection of Cas9 single-cell clones (SCCs). SCCs were sorted from the HAP1 Cas9 bulk population and further characterized. Cas9 editing was assessed by cell surface marker knockout followed by FACS staining and cell viability upon knockout of a core essential gene. Two highly editing single-cell clones (SCC11 and SCC12) were selected for further experiments. **b** HAP1 Cas9 bulk, Cas9 SCC11, and Cas9 SCC12 cells were transfected with the HDCRISPRv1 vector encoding an sgRNA targeting either the safe harbor locus *AAVS1* as a control or the core essential gene RNA Polymerase 2 subunit E (*POLR2E*). Editing efficiency based on cell viability of sgPOLR2E-transfected cells in comparison to sgAAVS1 control cells was addressed by crystal violet staining. The number of surviving cells was strongly reduced in cells transfected with an sgRNA directed against *POLR2E* ($n = 3$ for each cell line and sgRNA). **c** Editing efficiency was furthermore assessed upon transduction of HAP1 Cas9 bulk, Cas9 SCC11, and Cas9 SCC12 cells with the HDCRISPRv1 vector expressing sgRNAs targeting the surface marker *CD46*, followed by FACS staining of residual CD46 protein to address knockout efficiency. Antibody staining of the non-edited cell lines was used as a control. Lines represent the mean of independent measurements ($n = 3$ for each cell line and condition)

considered this cell line to be a suitable neutral model to evaluate library performance.

Comparing fold changes of sgRNAs targeting genes comprising a core essential and a nonessential reference gene set [17] confirmed strong loss of sgRNAs targeting essential genes over the course of screening for both libraries and all cell lines, while the representation of nonessential genes remained nearly unchanged to the plasmid library, validating our sgRNA selection strategy (Fig. 4b; Additional file 10: Table S3). As expected, depletion of sgRNAs targeting essential genes was stronger in selected single-cell clones compared to the bulk population and screening in single-cell clones furthermore strongly improved replicate correlation in comparison to screening in HAP1 Cas9 bulk cells (Additional file 9: Fig. S5 A). This confirms our assumption that a better resolution can be achieved upon screening in single-cell clones pre-selected for strong Cas9 editing efficiency. To further address library performance, we computed the area under the curve (AUC) for the empirical cumulative distribution function using core essential, nonessential [17, 33], and non-targeting reference gene sets (Fig. 4c). Effective screens are supposed to have an AUC value ≥ 0.5 for sgRNAs targeting essential genes and an AUC value ≥ 0.5 for sgRNAs targeting nonessential genes or designed as non-targeting controls [14], indicating that respective sgRNAs either get preferentially depleted (AUC ≥ 0.5) or remain (AUC ≥ 0.5) over the course of

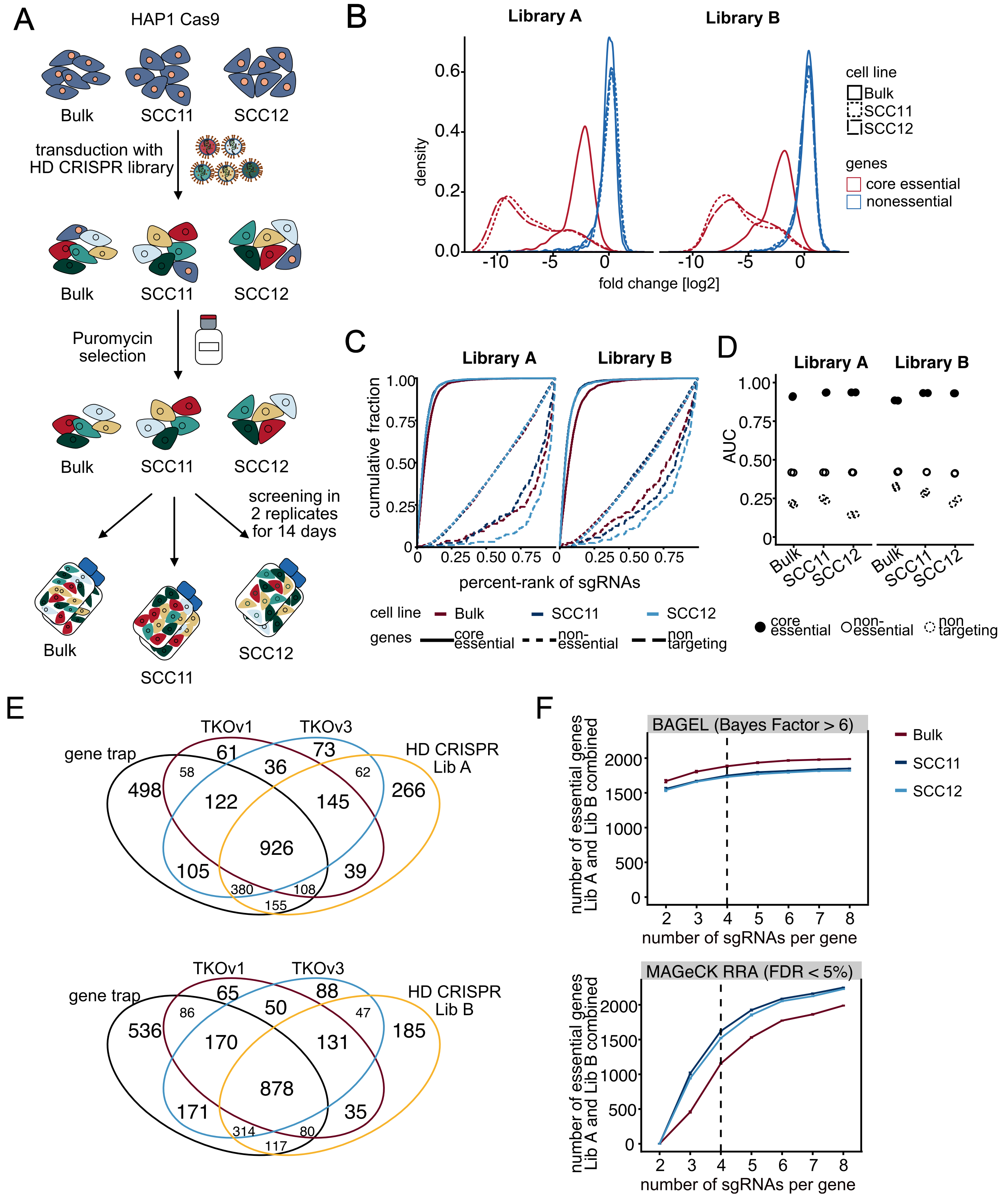

Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 The HD CRISPR library efficiently identifies core, non- and context-dependent essential genes. **a** Workflow of a pilot screen conducted with the HD CRISPR library in HAP1 cells. The screen was performed in parallel in the Cas9-expressing bulk population and two highly editing single cell clones for both libraries independently. Successfully transduced cells were selected with puromycin for 48 h and then split into two independent replicates. The screen was performed for a duration of 14 days. **b** Core essential genes were strongly depleted over the course of screening with either of the two libraries, HD CRISPR libraries A and B, in contrast to nonessential genes. Stronger depletion was observed in the two single cell clones with high Cas9 editing efficiency. **c** Empirical cumulative distribution function for viability screens conducted with different HAP1 Cas9 cell lines and both HD CRISPR sub-libraries. Shown are results for sgRNAs targeting genes from core essential or nonessential gene sets or representing non-targeting controls. **d** Area under the curve (AUC) values for individual replicates of the empirical cumulative distribution functions shown in **c**. **e** Comparison of HAP1 essential genes as identified with the HD CRISPR library, a gene trap screen by Blomen et al. [41] and two CRISPR screens using either the TKOv1 or TKOv3 library by Hart et al. [17]. **f** Number of essential genes detected with increasing number of sgRNAs per gene using BAGEL (left; BF > 6) or MAGeCK RRA (right; FDR < 5%). sgRNAs were subsampled from the combined HD CRISPR library (sub-libraries A and B). Each data point represents the average of 5 samples. Error bars are ±1 s.e.m

screening. We obtained AUC values close to 1 for both independent replicates of all screens when analyzing sgRNAs targeting essential genes, and AUC values below 0.5 for sgRNAs designed for nonessential genes and as non-targeting controls (Fig. 4d), indicating that our selection strategy for core and nonessential genes was successful. Moreover, differences in the maximal and the minimal $\log_2$ fold change for guides targeting the same gene were small with a median difference less than 1 for all individual screens (Additional file 9: Fig. S5 B). To benchmark our results with other screens conducted in HAP1 cells, we compared genes identified to be essential in the HAP1 Cas9 bulk population using either the HD CRISPR sub-libraries A or B with published data from gene essentiality screens conducted in HAP1 cells using either the TKOv1 or TKOv3 libraries (Hart et al., [8]) or a gene trap screen [41]. The intersection of hits from all screens was 926 genes for the HD CRISPR library A and 878 genes for the HD CRISPR library B (Fig. 4e), and computed Bayes factors for the HAP1 Cas9 bulk screens were highly correlated with those calculated for the TKOv1 and TKOv3 CRISPR screens conducted in HAP1 cells (Additional file 9: Fig. S5 D-E), suggesting that hits from previous screens—including HAP1 context-dependent essential genes—can be re-discovered using the HD CRISPR library. Moreover, precision-recall analysis for reference core essential and nonessential genes sets (Hart et al. [17]) was comparable to screens conducted using the TKOv3 library and slightly better to the TKOv1 library in HAP1 cells (Additional file 9: Fig. S5 C).

We furthermore compared hit calling of essential genes using different analysis methods, BAGEL and MAGeCK RRA. Using BAGEL, differences in the number of essential genes identified were marginal for both, increasing numbers of sgRNAs per gene (2–8) and screening in either the Cas9 bulk population or one of the two single-cell clones (Fig. 4f). This is likely due to the fact that BAGEL uses prior knowledge for gene essentiality hit calling. In contrast, for analysis software, which does not require prior information such as MAGeCK RRA or gscreend [42, 43], an increasing number of sgRNAs per gene also allowed the identification of more hits and single-cell clones with an enhanced editing efficiency were superior for hit calling in comparison to the bulk population (Additional file 11: Fig. S6 A-C). Overall, comparing genes identified to be essential using either BAGEL or MAGECK RRE revealed strong agreement with few essential genes private to each software (Additional file 11: Fig. S6 D).

In a final step, we asked how differences in editing efficiency, the number of sgRNAs per gene, and potential clonality effects might affect the determination of gene essentiality. We were especially interested in differences in gene essentiality between the Cas9 bulk cell line and the two Cas9 single-cell clones, since the generation of single-cell clones from bulk populations forces cells to go through a genetic bottleneck, which might favor certain genetic alterations [44] and thus dependencies. Therefore, we used BAGEL [32] to analyze depleted genes in each HAP1 cell line using the combined HD CRISPR library (8 sgRNAs per gene) and the individual sub-libraries A and B. We applied a strict Bayes factor cutoff of BF > 6 [17] to discriminate between essential and nonessential genes. This analysis revealed highly overlapping sets of essential genes for each cell population. Using the combined library, we found 2096 essential genes in the Cas9 bulk population, 1938 essential genes in SCC11, and 1925 essential genes in SCC12. Out of these, 1755 were shared between the three lines (84%, 91%, and 91% of total essential genes in each cell line). Only 58 (3% of total) essential genes were private to SCC11 and 46 (2.5% of total) essential genes were private to SCC12 (Additional file 12: Fig. S7 A). We observed similar overlap using only sub-library A or B (Additional file 12: Fig. S7 C-D). Accordingly, quantitative comparison revealed high Bayes factor correlation between bulk Cas9 population and single-cell clones (0.915 for SCC11 and 0.925 for SCC12; Additional file 12: Fig. S7 B). Overall, we could not observe major differences in hit calling between selected Cas9 single-cell clones and a bulk population when addressing general gene essentiality. Importantly, gene set enrichment analysis revealed no enrichment for genes involved in DNA damage repair pathways among SCC11- or SCC12-specific essential genes [45, 46]. Still, we think that caution is required when screening in more complex settings as, e.g., when addressing drug resistance mechanisms.

CRISPR screens at high sensitivity allow predictions about sgRNA cutting efficiency

CRISPR-induced DNA double-strand breaks at target loci are known to induce a DNA damage response, which significantly affects cell proliferation in comparison to negative controls [10]. As a result, non-targeting sgRNAs are likely to reflect wild-type growth, while sgRNAs targeting nonessential genes and targeting controls lead to slightly impaired growth [17]. Differences between targeting and non-targeting controls are usually subtle. However, small but distinct shifts in the fold change distributions of non-targeting and targeting control sgRNAs were visible in all our screens in the HAP1 cell line, further indicating that we were able to select efficiently cutting sgRNAs also for nonessential genes, for which empirical design due to the absence of a strong phenotype is more challenging (Additional file 13: Fig. S8 A). This shift was especially pronounced in the screens conducted in the HAP1 Cas9 SCC12 cell line (Fig. 5a, Additional file 13: Fig. S8 A). We reasoned that this effect could be explained by the increased sensitivity that can be achieved by using a single-cell clone selected for high Cas9 activity. We hypothesized that it might be possible to exploit this increased resolution to assess the activity of sgRNAs targeting nonessential genes. Using a Gaussian mixture model [47], we divided all control sgRNAs in the SCC12 screen into three groups: (1) sgRNAs with strong viability phenotypes, (2) sgRNAs targeting nonessential loci (mild phenotype due to induction of DNA double-strand breaks), and (3) inactive sgRNAs (Fig. 5a; Additional file 13: Fig. S8 B-C). We then classified all sgRNAs in the HD CRISPR library into one of these groups. To avoid bias, we excluded reference core essential genes [17] from the analysis. sgRNAs that could not be assigned to any of the groups with a probability of at least 80% were labeled “undetermined” (Fig. 5a). As expected, most sgRNAs targeting essential genes (determined using MAGeCK analysis [42] of the combined HD CRISPR library) were classified as group 1. This effect was more pronounced for sgRNAs selected based on previous phenotypes compared to other sgRNAs (82% and 60% of sgRNAs targeting essential genes classified as group 1, respectively; Additional file 13: Fig. S8 D). In addition, most sgRNAs targeting nonessential genes were classified as group 2 (likely cutting) whereas only a small fraction of sgRNAs were classified as group 3 (likely not cutting). Again, sgRNAs selected based on their phenotypes in previous screens appeared favorable compared to other sgRNAs with a larger fraction classified as likely cutting (46% compared to 39%) and a smaller fraction classified as likely not cutting (2% compared to 5%; Fig. 5b; Additional file 13: Fig. S8 E). These observations suggest that sgRNA phenotypes in past screens are in fact predictive of whether these sgRNAs will be effective in future screens and thus motivate empirical design as a viable strategy for sgRNA selection—not only for core essential genes but also for context-specific essential genes.

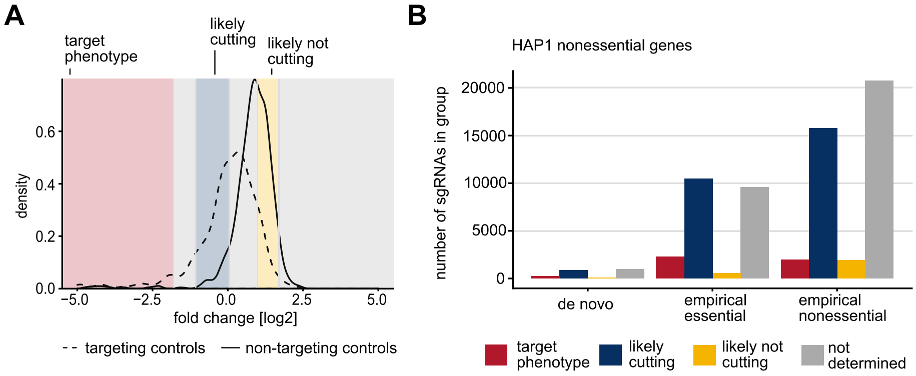

Fig. 5 CRISPR screens conducted at a high dynamic range predict the cutting efficiency of sgRNAs based on mild viability phenotypes. **a** A mixture model was used to divide control sgRNAs of the HD CRISPR library A in the single cell clone SCC12 screen into three groups: (1) sgRNAs with a target-dependent viability phenotype (red), (2) sgRNAs with a small target-independent phenotype likely caused by a double-strand break (blue), and (3) sgRNAs with no phenotype due to a lack of DNA cutting (yellow). Log2 fold change distributions of targeting and non-targeting control sgRNAs are indicated as dashed and solid curves, respectively. **b** Number of sgRNAs targeting nonessential HAP1 genes associated with each phenotype group. Nonessential genes were determined using MAGeCK which requires no prior knowledge for analysis. sgRNAs are stratified based on their design: “empirical essential” sgRNAs target context-specific essential genes and were selected for the HD CRISPR library based on their previous on-target phenotypes. “Empirical nonessential” sgRNAs are part of previously published libraries and target broadly nonessential genes. They were selected based on their lack of outlier phenotypes. De novo sgRNAs were designed using the software cld [34]

Identification of HAP1 context-dependent essential genes

Selecting sgRNAs with consistent knockout phenotypes across several independent screens is most conclusive for broadly essential genes. Therefore, we examined whether the HD CRISPR library could also identify context-dependent essential genes and as such genes that are specifically essential in HAP1 cells. We compared the list of hits identified in our screens to essential genes reported in other screens and cell lines in the GenomeCRISPR database. We identified a set of genes (including *ARGFX*, *CNIH4*, *TCEANC*, and *ZNF676*), which appeared to be essential in more than 75% of published screens we used for analysis, but not in our screens in HAP1 cells using the HD CRISPR library (Fig. 6a). These might reflect genes essential in most cell lines, but not HAP1, or might reflect genes that have been frequently wrongly annotated as essential. When comparing Bayes factors (BF) as a measurement for gene essentiality for these genes across different libraries, we observed that these genes were identified as essential in screens with the Avana sgRNA library [18], but not with other published genome-scale libraries (Additional file 14: Fig. S9). This suggests that the unexpected and broadly observed essentiality of these genes might reflect artifacts associated with the respective sgRNAs present in the Avana library. Such off-targets are likely found in each CRISPR library and only become apparent when many different cell lines are screened with the same library as is the case for the Avana library. Importantly, this implies that our selection strategy, although based on choosing sgRNAs with strong phenotypes in many screens, does not enrich for sgRNAs with strong off-target effects.

While the near-haploid karyotype of HAP1 cells renders them particularly amenable for genetic perturbation studies [41], relatively little is known about their identity and cell line-specific dependencies. Thus, we addressed HAP1 context-dependent essential genes as identified in our screen in more detail. In general, roughly 2000 genes are considered to be essential in cultured human cells [48], which is in accordance with our results (Fig. 4f). With ~700 genes comprising the core essential gene set [17], ~1300 of context-dependent essential genes can be considered to be identified. HAP1 cells originate from experiments to induce pluripotency in the leukemia KBM7 cell line by transduction with *KLF4*, *POU5F1* (Oct4), *SOX2*, and *MYC* [39, 49]. Interestingly, our screening results indicate that this treatment seemed to have rendered HAP1 cells strongly dependent on

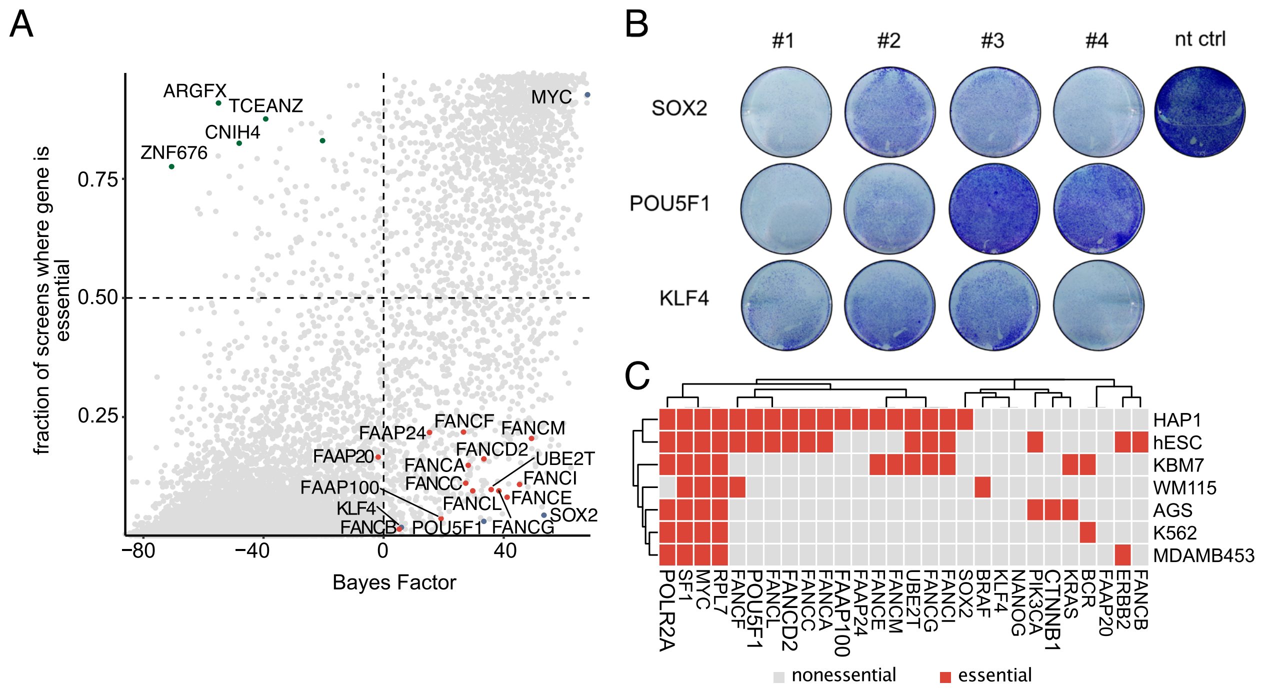

Fig. 6 HAP1 cells are highly dependent on Yamanaka factors and the Fanconi anemia pathway. **a** HAP1 context-dependent essential genes are enriched for genes comprising the Fanconi anemia pathway as well as factors known to induce pluripotency. The *y*-axis provides an estimate of the percentage of essentiality in other cell lines previously screened. Fanconi anemia pathway genes are highlighted in red, Yamanaka factors are shown in blue. HAP1-specific nonessential genes are highlighted in green. nt ctrl = non-targeting control. **b** Crystal violet staining of HAP1 cells treated with various siRNAs targeting *SOX2*, *POU5F1*, and *KLF4* and a non-targeting control. Shown is a representative result of three independent experiments. **c** Essentiality of known cancer drivers, Fanconi anemia pathway genes, and pluripotency factors across cell lines representing different cancer cell types in comparison to HAP1 cells and human embryonic stem cells

POU5F1 and *SOX2* and to a lesser extent also on *KLF4.* In contrast, *KLF4*, *POU5F1*, and *SOX2* were only identified to be essential in less than 5% of other cell lines screened so far (Fig. 6a). These genes were not identified as essential in former gene trap screens in HAP1 cells (Additional file 15: Fig. S10 A), which might be explained by the fact that HAP1 cells likely carry several copies of *KLF4*, *POU5F1*, and *SOX2* due to their initial transduction. Hence, we further wanted to exclude that the observed essentiality was due to multiple integrations of the lentiviral expression vector and thus an increased DNA damage response upon several CRISPR-induced cuts [10, 29]. Therefore, we validated the dependency of HAP1 cells on *KLF4*, *POU5F1*, and *SOX2* using siRNAs as an orthologous approach (Fig. 6b, Additional file 16: Table S4). In addition, we identified genes comprising the core complex of the Fanconi anemia pathway to be essential (BF > 6) in HAP1 cells, while often dispensable in other cell lines (Fig. 6a; Additional file 15: Fig. S10 A). This dependency was recapitulated for individual Fanconi anemia-associated genes in either the gene trap screens conducted in HAP1 cells (Blomen et al. [41]) or the CRISPR screens conducted in HAP1 cells using either the TKOv1 or TKOv3 sgRNA libraries (Hart et al. [17]) (Additional file 15: Fig. S10 A). However, while published screens only identified individual and different Fanconi anemia-associated genes to be essential in HAP1 cells, we could identify all members of the core complex and associated genes involved in interstrand crosslink repair to be essential in HAP1 cells (Additional file 15: Fig. S10 A). This effect was overall consistent for both HD CRISPR sub-libraries and in the HAP1 Cas9 bulk cell line as well as both Cas9 single-cell clones, thereby excluding that the stronger Cas9 editing efficiency and a potentially stronger DNA damage response in the Cas9 single-cell clones resulted in this dependency. Nevertheless, we aimed to again validate the context-specific gene essentiality of HAP1 cells on genes involved in the Fanconi anemia pathway using an approach independent of a DNA damage response and therefore again performed an siRNA knockdown for four genes of the Fanconi anemia core complex, *FANCE*, *FANCM*, *FANCG*, and *FANCL* (Additional file 15: Fig. S10 B-C, Additional file 16: Table S4). We used four different siRNAs per gene and observed the strongest loss of cell viability upon knockdown of *FANCE* and *FANC M*, while the reduction in cell viability was milder for *FANCL* and *FANCG.* Again, no difference was observed for the different HAP1 Cas9 cell lines. Since the generation of HAP1 *FANCG* and *FANCM* knockout cell lines has also been reported recently [50, 51], loss of expression of Fanconi anemia-associated genes might rather slow down cellular proliferation than directly inducing cell death.

The dependency on pluripotency factors prompted us to compare gene essentiality of HAP1 cells with previously published CRISPR data for gene essentiality of the original KBM7 cell line, as well as of human embryonic stem cells (hESCs) [27, 52]. The comparative analysis revealed that loss of *POU5F1* and some of the components of the Fanconi anemia core complex also resulted in reduced cell viability in hESCs. In contrast, KBM7 cells displayed a dependency on several genes associated with the Fanconi anemia pathway, but not for either of the Yamanaka factors except for the core essential gene *MYC* (Fig. 6c). Interestingly, viability of HAP1 cells did not seem to be affected by the knockout of known tumor type-specific oncogenes such as *BCR*, *PIK3CA*, *KRAS*, *CTNNB1*, or *BRAF* (Fig. 6c). Especially the lack of BCR dependency is surprising, since both HAP1 cells and their parental cell line KBM7 harbor a BCR-ABL fusion gene, which, however, is only required for proliferation in KBM7 cells. Based on their shared dependencies with hESCs while missing an addiction to known tumor type-specific oncogenes, we suggest that the initial transduction with pluripotency factors has rendered HAP1 cells to adopt stem cell-like characteristics.

Discussion

Over the last few years, pooled genome-scale CRISPR-Cas9 screens have quickly become an established and indispensable tool to functionally interrogate the human genome. Still, much remains to be learnt about optimal experimental conditions and sgRNA design to optimize especially the dynamic range and phenotype detection in CRISPR screens, while minimizing expenses and workload.

To date, large-scale CRISPR-Cas9 screens have been conducted in hundreds of human cell lines [8, 10, 12, 13, 27, 28]. These data now enable us to draw conclusions about the efficacy of hundreds of thousands of sgRNAs in diverse experimental contexts. Here, we introduced a strategy to leverage data from published CRISPR screens for sgRNA design. Our goal was to prioritize sgRNAs with consistently high on-target activity, while simultaneously avoiding sgRNAs with off-target effects. Here, our fundamental assumption is that sgRNA design algorithms do not naturally enrich for guides with strong off-target effects and that sgRNAs targeting the same gene but featuring different sgRNA sequences are unlikely to share the same toxic off-target. We were able to select high-quality sgRNAs and to assemble them into the Heidelberg (HD) CRISPR library. In total, 30–42% of sgRNAs for both sub-libraries were selected based on empirical essentiality phenotypic evidence, meaning that the selected sgRNA displayed a viability phenotype in at least 5% of the screens analyzed. This number is limited primarily by the fact that currently available CRISPR screening data are mostly derived from viability screens for gene essentiality. Therefore, we expect this number to grow quickly once screens for additional phenotypes become available. Furthermore, we could select non-toxic guides for the remaining 50 to 60% of protein-coding genes, generating libraries with overall 96% of empirically designed guides. We could show that guides in this library were enriched for high sequence scores according to different design rules, although these were not initially taken into account for library design. To evaluate the performance of this library, we conducted a genome-scale screen for gene essentiality in HAP1 cells and confirmed that our library was able to distinguish between core and nonessential reference genes with high precision and accuracy. Since it is desirable to minimize library size for CRISPR screens in order to reduce expenses and experimental efforts, we assessed how the number of sgRNAs per gene would affect the detection of essential genes. Therefore, we compared screens with the combined HD CRISPR library (8 sgRNAs per gene) and screens in individual sub-libraries A and B (4 sgRNAs per gene). Screening with only one of the two sub-libraries reduces experimental efforts by ~ 3.5×10^7 cells per replicate and split when screening with a 500-fold coverage.

To further analyze how these results would be affected by editing efficiency and clonality effects, we performed screens in a HAP1 Cas9 bulk population as well as two Cas9 single-cell clones. Overall, essential genes were highly consistent between the bulk population and the single-cell clones, while phenotypes were persistently stronger in the single-cell clones. This could not be explained by differences in ploidy of Cas9 single-cell clones and the bulk population, since enriched diploid populations of the two HAP1 Cas9 single-cell clones still showed similar editing efficiency to enriched haploid populations and remained superior in comparison to Cas9 bulk populations. This indicates that interpopulation heterogeneity is a stronger determinant of editing efficiency than ploidy. Our results suggest that single-cell clones with high Cas9 activity are attractive models to screen for gene essentiality at small library sizes, especially since the signal-to-noise ratio has recently been shown to correlate to the frequency of core essential gene dropout [37]. In-depth characterization of one Cas9 single-cell clone for general properties of the parental cell line is, however, advisable and especially worth the effort when several CRISPR screens with different experimental setups are planned to be done in the same cell line. Enhanced CRISPR editing efficiency in this case is likely to facilitate hit detection, while follow-up studies could be considered to be done including the parental cell line to exclude clonality effects. While we did not observe strong clonality effects between single-cell clones for detecting essential genes, it is likely that other phenotypes might be affected more severely by clonal differences. Therefore, caution is especially required when using single-cell clones in screens for phenotypes such as pathway activity or cell morphology. A middle-way could be to pool several independent single-cell clones with similar proliferation rates into a pseudo bulk population that can be used for screens. By this means, clonality effects may be minimized while still maintaining high editing efficiency.

Haploid cell lines such as HAP1 cells are an attractive model for genetic perturbation studies since the presence of a single allele implies enhanced knockout efficiency [41] compared to polyploid cancer cell lines where a full knockout requires out-of-frame editing of all alleles of a gene [53]. HAP1 cells with specific gene knockouts have therefore been applied in many studies addressing various research questions [54–59]. The exact cellular context of this cell line is, however, not fully understood, especially since HAP1 cells do not share major characteristics with their parental chronic myeloid leukemia (CML) cell line KBM7. In contrast to KBM7 cells, HAP1 cells grow adherent and are not dependent on *BCR* (Fig. 5c), while Bcr-Abl inhibitors represent the first-line therapy in CML [60]. When specifically assessing HAP1 context-dependent gene essentiality, we identified known and so far unreported HAP1 context-specific essential genes. In particular, we describe a strong HAP1-specific dependency on the Yamanaka factors *POU5F1*, *SOX2*, and to a lesser extent also *KLF4* [49]. These genes have been transduced into KBM7 cells during the generation of the HAP1 cell line [41]. Further, HAP1-specific essential genes were enriched for components of various complexes of the Fanconi anemia pathway. The Fanconi anemia pathway is mainly known to be responsible for the repair of stalled replication forks that occur as a consequence of interstrand crosslinks [61]. Recently, it has been reported that *FANCD2* localizes to sites of Cas9-induced DNA double-strand breaks where it supports CRISPR-mediated homology directed repair and in particular single-stranded template repair [62]. Interestingly a dependency on certain Fanconi anemia regulators can also be observed in two other haploid cell lines. These include the near-haploid HAP1 parental cell line KBM7 [27] and a haploid human embryonic stem cell line [52]. Since regulators of the Fanconi anemia pathway are rarely essential in other cell lines, we speculate that the occurrence of interstrand crosslinks might be more detrimental in cells featuring only one copy of any given gene.

In conclusion, we show that the available data on sgRNA phenotypes in large-scale CRISPR essentiality screens can be used to inform the empirical design of sgRNA libraries. We provide the HD CRISPR library, a new library for genome-wide CRISPR-Cas9 screens with high on-target and low off-target activity. We further show evidence that single-cell clones are powerful models to conduct gene essentiality screens at a low library coverage of sgRNAs per gene and can improve screening quality over screening in Cas9 bulk populations. These findings might guide experiment design for the next generation of CRISPR-Cas9 screens in mammalian cell culture.

Conclusions

We conclude that conducting CRISPR/Cas9 viability screens using next-generation empirically designed sgRNA libraries and strongly editing Cas9 single-cell clones improves the resolution of CRISPR screens and to differentiate also subtle viability phenotypes, which in turn allows hit detection with a smaller number of sgRNAs per gene. We show that empirical design of sgRNA libraries based on phenotypic evidence from previous screens is a suitable predictor for sgRNA cutting efficiency and allows selection of highly active guides, which we assembled in a genome-scale CRISPR library termed “HD CRISPR library”.

Methods

Heidelberg CRISPR library design

Raw sgRNA count data were downloaded from the GenomeCRISPR database (November 5, 2017 [25]. Only negative selection screens performed using humanized *Streptococcus pyogenes* Cas9 were retained for downstream analysis. At the time of sgRNA design, no published screens were available with the TKOv3 [17] and Brunello [19] libraries. Therefore, these sgRNA sequences were added to the GenomeCRISPR-derived sgRNA pool. Next, all targeted transcripts and protein-coding exons were annotated for each sgRNA. For this purpose, genomic information was derived from the ENSEMBL database (GRCh38 [31] using biomaRt [63]. In order to determine putative off-target effects for each sgRNA, sgRNA sequences were mapped to the genome using bowtie2 [64]. Specifically, local alignment was performed using bowtie2’s “very-sensitive-local” setting allowing for up to 3 mismatches. This strategy for computational off-target prediction was motivated by previous studies [34, 65].

To additionally avoid sgRNAs with off-target effects, sgRNAs were grouped by their target genes and their GenomeCRISPR effect scores [25] were z-normalized. This was done to identify sgRNAs whose phenotypes strongly deviate from the phenotypes of other sgRNAs targeting the same gene. Increasing depletion of sgRNAs targeting nonessential genes and lack of depletion of sgRNAs targeting core essential genes was detected at effect scores smaller than − 1.25 and larger than 1.25, respectively (Additional file 1: Fig. S1 C). Therefore, sgRNAs with an absolute effect score of greater than 1.25 were flagged.

Next, sgRNA sequences containing consecutive stretches of the same nucleotide (4 or more A’s/T’s, 5 or more G’s/C’s) were flagged since these sequences have been shown to create problems during polymerase transcription and PCR amplification. In addition, sequences containing BbsI restriction sites (GAAGAC and reverse complement) and sequences with a strong GC bias (GC greater than 75% or less than 20%) were excluded. To determine sgRNA performance, all negative selection (for viability) screens in GenomeCRISPR (in total 488) were analyzed using BAGEL v0.91 [32] with the CEGv2 and NEGv1 core and nonessential reference gene sets [17, 33]. To evaluate the quality of each screen, precision-recall curves (PR curves) were generated for each experiment using the ROCR R package [66]. These PR curves evaluate how well reference core and nonessential genes could be separated based on the sgRNA depletion phenotypes in the screen. Screens for which the area under the PR curve was less than 0.9 were excluded from further analysis. For all other screens, genes were categorized as essential and nonessential at 5% false discovery rate (FDR). An sgRNA was then determined as active in a screen if (a) it was determined to target an essential gene (in the screened cell line) and (b) its depletion phenotype (quantified as sgRNA count fold change compared to a T_0/plasmid sample) was among the 20% strongest sgRNA phenotypes in the screen. In a genome-wide CRISPR screen for cell proliferation, approximately 12% of all protein-coding genes are expected to be essential [12, 67] and therefore 12% of all sgRNAs are expected to have an on-target phenotype. However, in order to not miss sgRNAs with potential subtle phenotypes, we selected a lenient threshold of 20% to determine active sgRNAs. sgRNAs for the HD CRISPR library were finally selected from the remaining pool of sequences. sgRNAs that were determined as active in a large number of screens were prioritized. However, to avoid selecting sgRNAs based on spurious effects, sgRNA activity was only considered as selection criteria if the sgRNA was determined as active in at least 5% of the screens in which they were used. Otherwise, sgRNAs targeting exons present in many transcript isoforms, sgRNAs targeting exons close to the transcription start site, and sgRNAs with a low number of predicted off-target effects were prioritized. In total, a genome-wide library consisting of two mutually exclusive sub-libraries A and B was assembled. Each sub-library contains 4 sgRNAs per gene targeting 18,913 and 18,334 protein-coding genes, respectively. In cases where less than 4 sgRNAs were available for a gene based on the filtered pool of sequences described above, missing sgRNAs were designed with the CRISPR library designer [34]. Further 300 non-targeting controls derived from published libraries [6, 8, 27] and 135 targeting controls targeting intronic regions or the AAVS1 safe harbor locus [65] were added to each sub-library. Control sgRNAs are identical between the sub-libraries. In total, this resulted in sub-libraries of sizes 74,987 (library A) and 71,048 (library B) that can be combined into a library of size of 146,035 sgRNAs for increased detection power.

Cloning of the HD CRISPR sgRNA vector

For generation of the HDCRISPRv1 vector, an insert encoding the human U6 promoter, an eGFP stuffer, part of the improved sgRNA scaffold [38], the cPPT/CTS, the human PGK promoter, a puromycin N-acetyltransferase, and the WPRE next to stuffer sequences was ordered as a GeneArt Synthetic Gene (Life Technologies) and PCR-amplified. The bacterial backbone of the pLCKO vector (kind gift from the Moffat lab, Addgene #73311) was linearized from the 3′LTR to the RRE by PCR, and both PCR products were fused using In-Fusion Cloning (Takara Bio) and transformed into One Shot Stbl3 Chemically Competent *E. coli* (Life Technologies). A correct cloning product was identified by control restriction enzyme digestion using AarI (Thermo Fisher Scientific) and subsequent Sanger sequencing. The initial plasmid only contained half of the improved sgRNA scaffold due to different possible sgRNA cloning strategies. To clone the complete improved sgRNA scaffold inside, the plasmid was digested using AarI (Thermo Fisher Scientific) according to the manufacturer's recommendations. The larger fragment was gel purified and used for In-Fusion cloning with a gBlock (Integrated DNA Technologies) encoding overlapping overhangs of the human U6 promoter and the sgRNA scaffold, the eGFP stuffer, and the missing part of the improved sgRNA scaffold. The In-Fusion cloning product was again transformed into One Shot Stbl3 Chemically Competent *E. coli* (Life Technologies) and the correct plasmid verified by Sanger sequencing.

Cell lines and cell culture

Wild-type HAP1 C631 cells were ordered from Horizon Discovery and maintained at 37 °C, 5% CO_2 in IMDM (Life Technologies) supplemented with 10% FCS. For stable Cas9 expression, the parental cell line was transduced at an MOI of ~ 0.5 with lentivirus generated using the Lenti Cas9-2A-Blast plasmid (kind gift of the Moffat lab, Addgene #73310) and selected with 20 μg/ml blasticidin (InvivoGen) for at least 1 week. Out of this bulk population, single cells with a small cell size (indicative for a haploid genotype) were sorted in 96-well plates, expanded, and again selected with 20 μg/ml blasticidin. Blasticidin-resistant clones were further characterized for Cas9 editing efficiency. Mycoplasma contamination was periodically assessed for all cell lines. The HAP1 cell line was authenticated using Multiplex Cell Authentication by Multiplexion (Heidelberg, Germany) as described recently [68]. The SNP profile was unique.

Sorting of haploid and diploid HAP1 populations

Enrichment of haploid and diploid HAP1 Cas9 SCC11 and SCC12 populations was achieved using flow cytometry as described previously [40]. In brief, respective HAP1 cells were trypsinized and a sub-sample (~ 8×10^5 cells) was stained using 10 μg/ml Hoechst 33342 for 30 min at 37 °C. Stained cells were analyzed by flow cytometry and DNA Hoechst intensity peaks allowed for back gating of the haploid and diploid populations of interest in the FSC and SSC, as haploid cells are smaller in size. Respective gates in the FSC/SCC were used for sorting haploid and diploid HAP1 cells from unstained samples.

Lentivirus production and MOI determination

Low-passage (< 15) HEK293T cells were seeded to reach 70–80% confluency on the day of transfection. The lentiviral packaging vector psPAX2 (kind gift from the Didier Trono lab, Addgene #12260) and the lentiviral envelope vector pMD2.G (kind gift from the Didier Trono lab, Addgene #12259) were co-transfected with the respective lentiviral expression vector (~ 1:1:3 M ratio) using Trans-IT (VWR) and OptiMEM (Gibco). Roughly 16 h post transfection, the medium was replaced by fresh culture medium and lentiviral supernatant was harvested 48 h post transfection by filtration through a 0.45μm PES membrane, aliquoted, and stored at − 80 °C until transduction. For multiplicity of infection (MOI) determination, target cells were transduced with different amounts of lentiviral supernatant in the presence of 8 μg/ml polybrene. Transduced cells were selected for 48 h with 2 μg/ml puromycin (Biomol) starting 24 h post transduction, and the number of surviving cells was compared to a non-transduced control sample.

Single sgRNA cloning

Single sgRNA sequences were cloned into the HDCRISPRv1 vector as described previously [35]. In brief, the HDCRISPRv1 vector was sequentially digested with BfuAI (NEB) and BsrGI-HF (NEB), followed by dephosphorylation using CIP (NEB). The digested backbone was gel purified using the Macherey&Nagel NucleoSpin Gel and PCR Clean-up kit. sgRNA inserts were designed as two complementary oligos encoding the sgRNA target region as well as the cloning specific overhangs and were ordered as standard desalted oligos from Eurofins Genomics. Oligos were phosphorylated and annealed using T4 PNK (NEB) in 10X T4 Ligation Buffer (NEB). Diluted oligo duplexes were ligated into the digested vector using Quick Ligase (NEB) and transformed into Stbl3 recombination-deficient bacteria. A Midi- or Maxiprep (QIAfilter plasmid kit) was performed for lentivirus production or cell transfection.

Plasmid transfection

To address functional sgRNA expression from the HDCRISPRv1 vector, sgRNAs targeting the essential gene RNA Polymerase II Subunit E (POLR2E) and control sgRNAs designed to target the safe harbor locus AAVS1 were cloned into the HDCRISPRv1 vector as described above and transfected into HAP1 Cas9 bulk, HAP1 Cas9 SCC11, and HAP1 Cas9 SCC12 cells using Fugene transfection reagent (Promega). Then, 24 h post transfection, successfully transfected cells were selected using 2 μg/ml puromycin (Biomol) for 48 h. Dead cells were removed by washing with PBS and viable and still attached cells were stained using 0.5% Crystal violet (Sigma) staining solution supplemented with 20% methanol.

siRNA transfection

Cells were reverse transfected with 20 nM of the indicated siRNAs using Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific) and OptiMEM Reduced Serum Media (Thermo Fisher Scientific). Cell viability was analyzed 48 h post transfection.

Flow cytometry analysis

SgRNA sequences targeting surface markers were cloned into the HDCRISPRv1 vector as described above. Lentivirus was generated from all expression vectors. 1×10^5 HAP1 Cas9 bulk, HAP1 Cas 9 SCC11, and HAP1 Cas9 SCC12 cells supplemented with 8 μg/ml polybrene were reversely transduced using 5 μl viral supernatant, and successfully transduced cells were selected with 2 μg/ml puromycin (Biomol) 24 h post transduction. Five days after transduction, cells were harvested and the respective surface markers stained with an APC anti-human CD46 antibody (Biozol Diagnostica, Cat. No. 352405, RRID AB_2564356) or an APC anti-human CD81 antibody (Biozol Diagnostica, Cat. No BLD-349510, RRID AB_2564021). The percentage of CD46 and CD81 knockout was determined as the percentage of APC-negative single cells in three independent replicates.

Library cloning

Oligos encoding the HD CRISPR library A and B sgRNA sequences and flanking regions were ordered as an oligo pool from Twist Biosciences. Oligos were amplified using the KAPA HiFi HotStart ReadyMix (Roche) using flanking primers in 12 PCR cycles. The resulting PCR product was purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey&Nagel), and the correct fragment size was confirmed using a High Sensitivity Bioanalyzer DNA Kit (Agilent). For cloning, the HDCRISPR v1 vector was digested with BfuAI (NEB) and BsrGI-HF (NEB) overnight and dephosphorylated using CIP (NEB). The resulting linearized ~ 7000 bp vector missing the eGFP stuffer was gel purified again using the NucleoSpin Gel and PCR Clean-up kit (Macherey&Nagel) and used for library cloning in a one-step digestion-ligation-reaction [69]. The cloning product was purified by isopropanol precipitation and transformed into Endura ElectroCompetent Cells (Biocat), and individual transformation reactions were pooled and plated on freshly prepared LB-carbencillin plates. A 1:10,000- and 1:100,000-fold dilution was used to estimate the total number of colonies and thus the library coverage during transformation, which was aimed for to be at least 250-fold. Bacteria were incubated at 30 °C for maximal 16 h and harvested by scraping. The resulting plasmid pool was purified with the QIAfilter Plasmid Mega Kit (Qiagen). Representation and distribution of sgRNAs was analyzed by next-generation sequencing.

Analysis of non-digested backbone contaminations in plasmid preparations

To address cloning efficiency and background of remaining non-digested vector in HD CRISPR library preparations, plasmid purifications were transfected into adherent HAP1 C631 cells using Fugene (Promega). Remaining eGFP stuffer in the non-digested vector will lead to GFP expression upon transfection and thus serves as an indicator for pooled cloning efficiency. To assess the resulting fraction of GFP-positive cells upon transfection of a known proportion of undigested vector backbone, equally concentrated plasmid purifications of the non-digested HDCRISPRv1 vector and a plasmid purification of the HDCRISPRv1 vector expressing an AAVS1-targeting sgRNA were mixed at the indicated ratios. Successfully transfected cells were selected 24 h post transfection using 2 μg/ml puromycin (Biomol) for 48 h. Subsequently, cells were either imaged on an IN Cell Analyzer 6000 to address GFP expression or harvested and washed with PBS, and GFP expression was analyzed by flow cytometry.

Nicoletti ploidy stain

For DNA content analysis, HAP1 cells were harvested by trypsination and 5×10^5 cells were washed with cold PBS by centrifugation. Cells were subsequently resuspended in 400 μl Nicoletti buffer (0.1% sodium citrate, 0, 1% Triton X-100, 0.5 units/ml RNase A, 50 μg/ml propidium iodide (PI)) and incubated for 2 h at 4 °C under rotation. Haploid, diploid, and a 1:1 mixture of haploid and diploid HAP1 cells were prepared as controls. DNA content analysis was performed by measuring PI intensity using flow cytometry. Controls were used to set PI intensity peaks at 50 K for G1-phase haploid cells and 100 K for G1-phase diploid cells.

Pooled CRISPR depletion screen

Prior to screening, each cell line was re-selected with 20 μg/ml blasticidin (InvivoGen) for at least 1 week. A total of 225×10^6 HAP1 Cas9 bulk, HAP1 Cas9 SCC11, and HAP1 Cas9 SCC12 cells were transduced with the lentiviral HD CRISPR library A or B to achieve an initial library coverage of at least 300–500-fold upon infection at an MOI of ~ 0.1–0.3, including non-transduced control flasks and accounting for variation in transduction efficiency on a large scale. Then, 24 h post transduction, successfully transduced cells were selected with 2 μg/ml puromycin (Biomol). One plate was cultured in the absence of puromycin for MOI determination. Then, 48 h post selection, cells were harvested and the MOI was determined by calculating the ratio of living cells in the presence and absence of puromycin selection. Cells were re-seeded at a 500-fold library coverage in two independent replicates for each cell line and cultured for 14 days with regular cell splitting every 2 to 3 days. A 500-fold library coverage was maintained throughout the screen for each replicate, and 45×10^6 cells representing a ~ 500-fold library coverage were collected at every passage. Genomic DNA was isolated from the final passage for genomic DNA extraction and sequencing.

Genomic DNA extraction, library preparation, and sequencing

Genomic DNA was isolated from frozen cell pellets using the QIAAmp DNA Blood and Tissue Maxi Kit (Qiagen) and purified by ethanol precipitation. The sgRNA spanning region was amplified from purified genomic DNA at a 100-fold library representation. Illumina adapters and indices were added in the same one-step PCR reaction using the KAPA HiFi HotStart ReadyMix (Roche). The PCR product was purified using the QIAQuick PCR Purification Kit (Qiagen) and eluted on MiniElute columns (Qiagen) upon library preparation from a plasmid pool or further gel purified to remove genomic DNA contamination using the NucleoSpin Gel and PCR Clean-up kit (Macherey&Nagel) in case of library preparation for screening samples. The correct fragment size was confirmed with a High Sensitivity Bioanalyzer DNA Kit (Agilent). Sequencing was performed on an Illumina NextSeq 550 system with a High-Output Kit (75 cycles) (Illumina). Low heterogeneity was either addressed by using a custom sequencing primer binding immediately upstream of the sgRNA targeting sequence or by using the standard Illumina sequencing primer and adding 20% PhiX.

Calculation of sgRNA sequence scores

For the calculation of Ruleset 2 scores [19], flanking regions for each sgRNA were retrieved using the Bioconductor BSgenome package (human genome version hg38) [70]. Ruleset 2 scores were then calculated using the previously published software [19]. Similarly, DeepHF scores were computed using published software [20] according to the instructions on the corresponding GitHub page (https://github.com/izhangcd/DeepHF). Finally, optimized sgRNA scores according to Hart et al. were calculated based on the position weight matrix published with the corresponding manuscript [17].

Initial data processing and analysis

The MAGeCK software version 0.5.7 [42] was used to quantify sgRNA abundance from sequencing data. A pseudo count was added for each sgRNA in each sample. To adjust for differences in sequencing depth, samples were then normalized by dividing sgRNA counts to the median count of the targeting controls. sgRNA depletion phenotypes were quantified as $\log_2$ fold changes were calculated for each sgRNA as $fc_{\mathrm{sgRNA}} = \log_2(\frac{\mathrm{rc}_{\mathrm{sample}}}{\mathrm{rc}_{\mathrm{plasmid}}})$ where $\mathrm{rc}_{\mathrm{sample}}$ are normalized read counts of samples after 14 days of selection and $\mathrm{rc}_{\mathrm{plasmid}}$ are normalized counts of the plasmid library. Reproducibility between replicates was assessed using Pearson and Spearman correlation coefficients. To combine the screens in sub-libraries A and B into a combined library dataset, raw counts for each sample were first normalized to the same library depth through division by the median count in each sample and multiplication by the median count across all samples. Normalized counts for screens in sub-libraries A and B were then combined into a combined library count file. Normalized counts were further rounded to the closest count. To determine knockout phenotypes of house-keeping (core essential) genes compared to nonessential genes, previously reported gold standard lists of core (CEGv2) and nonessential (NEGv1) genes were used [17, 33].

Classification of essential genes

To evaluate screen performance, BAGEL v0.91 [32] was used to classify genes as essential and nonessential [71]. Specifically, a gene was classified as essential if the Bayes factor (BF) determined by BAGEL was greater than 6 (similar cutoffs were chosen as in [17]). Precision-recall curves and statistics to quantify how well core and nonessential reference genes could be separated based on each screen's depletion phenotypes were determined using the ROCR R package [66]. To compare essential gene detection power of individual sub-libraries compared to the combined library, the MAGeCK RRA [42] and gscreend [43] algorithms were used in addition to BAGEL to determine essential genes at 5% FDR. MAGeCK RRA was used with default parameters. For gscreend analysis, the number of permutations used to calculate the ρ_0 parameter for gene ranking was set to 10,000.

Analysis of cutting and non-cutting sgRNAs

A Gaussian mixture model with 4 components was fit to the fold change distributions of all targeting ($n = 270$) and non-targeting control ($n = 600$) for each screen separately using an expectation maximization algorithm implemented in the R package "mixtools" [47]. Here, two components were used to capture the fold change distributions of targeting and non-targeting controls (henceforth referred to as components 1 and 2, respectively) and two additional mixture components were used to capture moderate and several viability phenotypes of toxic controls (components 3 and 4; Additional file 13: Fig. S8 B). The resulting models were used to classify each additional sgRNA in the library as follows: each sgRNA was mapped to the control with the most similar phenotype. If that control was assigned to components 3 or 4 with a probability of at least 80%, the sgRNA was assumed to have a target-dependent viability phenotype. If the associated control was assigned to component 1 or 2 with a probability of at least 80%, an sgRNA was considered "likely cutting" or "likley not cutting", respectively. If a control could not confidently be assigned to any of the mixture components (probability for all components < 80%), then all sgRNAs with similar phenotypes were labeled "undetermined".

Influence of library coverage on essential gene detection

To investigate how library coverage affects the number of essential genes that can be detected, 2 to 8 sgRNAs per gene were sampled from the combined HD CRISPR library. Five independent samplings were performed for each library coverage. BAGEL v0.91 [32] and MAGeCK RRA v0.5.7 [42] were then used to classify essential genes at BF > 6 (for BAGEL) and FDR < 5% (for MAGeCK) in both the bulk population and the single-cell clones. CEGv2 and NEG1 core and nonessential reference gene sets were used for essential gene detection with BAGEL.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-020-00905-1.

Additional file 1: Figure S1. HD CRISPR library composition. (A) Precision recall curves for differentiating reference core and nonessential genes based on BAGEL Bayes Factors determined for published fitness screens in GenomeCRISPR. Blue curves indicate screens with an area under the curve (AUC) greater than 0.9. Red screens with an AUC of less than 0.9 were excluded for HD CRISPR sgRNA design. (B) Log2 fold change distributions of core essential (red) and nonessential (blue) reference genes for a high quality (left) and a low quality (right) example screen. (C) Library composition of the HD CRISPR sub-libraries A and B. Horizontal bars on the left indicate the number of designs used from different previously published libraries. The panel on the bottom right shows combinations of libraries, that include designs selected for the HD CRISPR library. The bars above this panel quantify the number of selected sgRNAs for each of these combinations. (D) Distribution of exon ranks targeted by the sgRNAs in the HD CRISPR library. (E) Distribution of the predicted off-target counts (see Materials & Methods) for sgRNAs in the HD CRISPR library. (F) Number of sgRNAs, which remained per gene after pre-filtering and were considered for library design. (G) Phenotypic deviation of published sgRNA phenotypes targeting the same gene. For each gene the difference between the GenomeCRISPR effect scores of the sgRNAs with the smallest and the largest effect scores was calculated. This process was repeated for each library using only those sgRNAs included in that library. Guides selected for the HD CRISPR libraries A and B show a narrow phenotypic deviation in published screens from which they were selected.

Additional file 2: Table S1. Annotated sgRNA sequences of the HD CRISPR Library.

Additional file 3: File S1. sgRNA sequences of the HD CRISPR Library A.

Additional file 4: File S2. sgRNA sequences of the HD CRISPR Library B

Additional file 5: Figure S2. Features and performance of the HDCRISPRv1 vector. (A) Composition of the lentiviral HD CRISPR sgRNA expression vector. (B) sgRNA cloning efficiency can be addressed upon transfection of the HDCRISPRv1 vector, since residual GFP stuffer in non-digested vector backbone leads to GFP expression (B.I) (*n* = 2). Complete removal as achieved when cloning single sgRNAs abolishes GFP expression (B.II) (n = 2), while remaining stuffer in 10% of the plasmid pool still leads to a substantial amount of GFP positive cells (B.III) (n = 2). Transduction of the non-digested vector still containing the GFP-stuffer does not result in GFP-expressing cells (B.IV) (*n* = 1). Scale bar = 100 μM (C) FACS analysis of GFP expression upon transfection of the non-digested HDCRISPRv1 vector (I) (n = 2) or the HDCRISPRv1 vector expressing an sgRNA (II) (*n* = 3). A mixture of GFP positive and negative cells can be observed upon transfection of a mixture of stuffer and sgRNA-containing vector (III and IV) (n = 3 for III and n = 2 for IV). (D) Editing efficiency was furthermore assessed upon transduction of HAP1 Cas9 cells with the HDCRISPR v1 vector expressing sgRNAs targeting the surface proteins CD81, followed by FACS staining of residual CD81 protein to address knockout efficiency. Antibody staining of the non-edited cell line was used as a control. Lines represent the mean of three independent experiments for each condition.

Additional file 6: Table S2. sgRNA sequences used in this study.

Additional file 7: Figure S3. DNA content analysis to determine ploidy of various HAP1 Cas9 populations. (A) HAP1 Cas9 bulk and HAP1 Cas9 SCC11 and SCC12 cells were stained for DNA content using Nicoletti buffer and FACS analysis. The percentage of G1 haploid cells and G2 diploid cells are indicated for each cell population (n = 2 for each condition). (B-C) Enriched haploid and diploid populations of the HAP1 Cas9 SCC11 (B) and Cas9 SCC12 (C) cell lines were obtained by FACS sorting. Subsequently, haploid and diploid populations were independently transduced with the HDCRISPRv1 vector expressing sgRNAs targeting the surface marker *CD46* and editing efficiency was directly compared in the haploid and diploid population of the same cell line. Non-edited samples of the respective cell lines served as a control. Lines represent the mean of three independent experiments for each condition.

Additional file 8: Figure S4. Cloning quality control of the HD CRISPR library. (A) Distribution of sgRNA read counts for the HD CRISPR plasmid library preparations. Skew ratios were determined as the quotient of the top 10 quantile divided by the bottom 10 quantile. (B) FACS analysis of GFP expression upon transfection of the HD CRISPR Library A and B plasmid pools to address the presence of remaining GFP stuffer (n = 3 for each condition).

Additional file 9: Figure S5. Reproducibility of negative selection screens with the HD CRISPR library. (A) Scatter plots showing the reproducibility of sgRNA phenotypes across biological replicates in screens with the HD CRISPR library. Each column includes screens performed in a bulk cell population (left) or in selected single cell clones with high Cas9 activity (middle and right). The top and bottom rows include screens with the HD CRISPR sub-libraries A and B, respectively. (B) Boxplot representing the distribution of the differences of the maximal and the minimal $\log_2$ fold change of guides targeting the same gene in individual screens. For each gene the difference between the maximal and the minimal sgRNA $\log_2$ fold change was calculated. This process was repeated for both HD CRISPR sublibraries using the phenotypes derived from screens in bulk population and single cell clones. Guides targeting the same gene result in similar $\log_2$ fold changes with a median difference of the maximal and the minimal $\log_2$ fold change smaller 1 for all screens. (C) Precision-recall-curve analysis for reference core essential and nonessential gene sets (Hart et al., 2015, Hart et al., 2017) of screens conducted in the HAP1 Cas9 bulk population using either the HD CRISPR Library A or B and two published CRISPR screens conducted in HAP1 cells using either the TKOv1 or TKOv3 library (Hart et al., 2017) as a reference. (D) Hit calling of the HD CRISPR Libraries A and B in comparison with a CRISPR screen conducted in HAP1 cells by Hart et al. (2017) using the TKOv1 library. (E) Hit calling of the HD CRISPR Libraries A and B in comparison with a CRISPR screen conducted in HAP1 cells by Hart et al. (2017) using the TKOv3 library. PCC = Pearson Correlation Coefficient, SCC = Spearman Correlation Coefficient.

Additional file 10: Table S3. BAGEL scores for individual genes in individual screens.

Additional file 11: Figure S6. Hit detection in screens with the HD CRISPR library. (A) Number of hits determined using BAGEL [32] at a strict Bayes factor cutoff (BF > 6) in different screens conducted with the HD CRISPR library. (B) Number of essential genes determined using MAGeCK RRA [42] at 5% FDR in different screens conducted with the HD CRISPR library. (C) Number of essential genes determined using gscreend [43] at 5% FDR in different screens conducted with the HD CRISPR library. (D) Venn diagrams showing the overlap between essential genes determined using either BAGEL or MAGeCK RRA for each screen.

Additional file 12: Figure S7. Essential genes are highly consistent between HAP1 Cas9 bulk population and single cell clones. A) Venn diagram showing essential gene overlap between a HAP1 bulk Cas9 population and two single cell clones that were selected for high Cas9 activity. Gene essentiality was determined using BAGEL with a Bayes Factor cutoff of 6 (see [17]). The combined HD CRISPR library with 8 sgRNAs per gene was used for essential gene inference. (B) Quantitative comparison of BAGEL Bayes Factors for each gene between the HAP1 bulk Cas9 population and selected single cell clones SCC11 and SCC12. Each dot represents a gene in the HD CRISPR library. Red dots indicate essential genes that are private to a single cell clone. The dashed diagonal is the identity line**.** (C) Venn diagram (left) and scatter plots (middle and right) showing essential gene overlap between a HAP1 bulk Cas9 population and two single cell clones in screens using the HD CRISPR sub-library A. (D) Venn diagram (left) and scatter plots (middle and right) showing essential gene overlap between a HAP1 bulk Cas9 population and two single cell clones in screens using the HD CRISPR sub-library A.

Additional file 13: Figure S8. Prediction of sgRNA DNA cutting activity based on control phenotypes. (A) Log2 fold change phenotype distributions for sgRNAs targeting nonessential genes (red) as well as targeting (blue) and non-targeting control sgRNAs (green) across different screens conducted with the HD CRISPR library. The screen with the HD CRISPR library A in HAP1 single cell clone SCC12, which was used for subsequent analyses, is highlighted in red. (B) Fit of a Gaussian mixture model with 4 components for screens in SCC12. Components 1 (yellow) and 2 (blue) represent non-targeting and targeting sgRNAs, respectively. Components 3 and 4 capture the phenotypes of sgRNAs with moderate and severe viability phenotypes. (C) Comparison of true fold change distributions of targeting and non-targeting sgRNAs (solid line) to the distributions estimated by the mixture model components (dashed lines) for both HD CRISPR libraries A and B. (D) Number of sgRNAs associated with each phenotype group targeting essential genes according to MAGeCK analysis. For this representation components 3 and 4 are combined in the red group 'target phenotype'. sgRNAs are stratified based on their design: 'empirical essential' sgRNAs target context-specific essential genes and were selected for the HD CRISPR library based on their previous on-target phenotypes. 'Empirical nonessential' sgRNAs are part of previously published libraries and target broadly nonessential genes. They were selected based on their lack of outlier phenotypes. De novo sgRNAs were designed using the software cld [34]. (E) Similar plot as D showing sgRNAs of the HD CRISPR library B associated with each phenotype group.

Additional file 14: Figure S9. HD CRISPR Library design strategy does not enrich for sgRNAs with strong phenotypes presumably caused by off-target effects. Bayes Factor analysis of selected HAP1 context-dependent nonessential genes across different screens conducted with various genome-scale CRISPR libraries in cancer cell lines.

Additional file 15: Figure S10. The identified dependency of HAP1 cells on pluripotency genes and the Fanconi anemia pathway is only partially detected in other published screens conducted in HAP1 cells. (A) Essentiality of Yamanaka factors and Fanconi anemia pathway members in individual HD CRISPR HAP1 and previously published TKO HAP1 CRISPR screens. Red boxes indicate that the gene was found essential and gray indicates non-essentiality. White boxes represent genes that are not targeted by the respective library. (B) siRNA knockdown of the *FANCL* and *FANCE* expression using four individual siRNAs each in the HAP1 Cas9 bulk and HAP1 Cas9 SCC11 and SCC12 cell lines. A pool of non-targeting siRNAs was used as a control. Surviving cells were stained with crystal violet solution. (C) Same as (B) for *FANCG* and *FANCM* expression. Representative images from four independent experiments conducted for each condition are shown. nt ctrl = non targeting control.

Additional file 16: Table S4. siRNA sequences of siRNAs used in this study.

Acknowledgements

We thank Fillip Port for help with the design of the cloning strategy and Josephine Bageritz for help with FACS analysis. We further thank the members of the Boutros laboratory for hands-on help of large-scale library cloning experiments and helpful discussions and feedback, especially Fillip Port, Katharina Imkeller, and Florian Heigwer for helpful comments on the manuscript. We are grateful for the excellent support of the DKFZ Genomics and Proteomics Sequencing core facility in sequencing screening samples, especially to Nicolle Diessl, as well as the DKFZ Flow Cytometry core facility. We further thank the ZMBH FACS core facility (Monika Langlotz) for help with sorting single-cell clones and the EMBL Genomics Core Facility, especially Vladimir Benes, for excellent sequencing support. Furthermore, we thank Matthias Hinterndorfer and Johannes Zuber for providing CD46 sgRNA sequences and the scientific community for sharing CRISPR screening data.

Authors' contributions

L.H., B.R., and M.B. designed the study. L.H. designed and implemented experiments. B.R. designed the HD CRISPR library and performed bioinformatic analyses. B.S. helped with experiments. J.W. helped with the design of the HDCRISPRv1 vector. L.H., B.R., and M.B. wrote the paper. All authors read and approved the final manuscript.

Funding

B.R. was supported by the BMBF-funded Heidelberg Center for Human Bioinformatics (HD-HuB) within the German Network for Bioinformatics Infrastructure (de. NBI). L.H. was supported by an Annemarie Poustka Fellowship of the Helmholtz Graduate School for Cancer Research. The project has been supported in part by an NCT Basic Science Grant. Work in the lab of M.B. was supported by an ERC Advanced Grant. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

These new lentiviral vectors and libraries described in this article will be made available to the scientific community through Addgene.
All data generated or analyzed during this study are included in this published article, its supplementary information files and publicly available repositories.
sgRNA sequences of the HD CRISPR libraries A and B are available in the Additional file 3: File S1 and Additional file 4: File S2. All annotated sgRNA sequences used for library design and their corresponding information and initial library can be accessed in Additional file 2: Table S1. sgRNA sequences for individual surface marker knockout experiments are listed in Additional file 6: Table S2. siRNA sequences and their ordering information are listed in Additional file 16: Table S4. BAGEL scores for individual genes in individual cell lines and HD CRISPR sub-libraries are listed in Additional file 10: Table S3. Raw sequencing data generated in this study are available from the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under the accession number PRJEB35190. Documented computer code to reproduce all figures presented in this study is available through GitHub and Figshare at https://github.com/boutroslab/Supplemental-Material/tree/master/Henkel%26Rauscher_2019.
Published CRISPR screening datasets used for the design of the HD CRISPR library were derived from GenomeCRISPR (http://genomecrispr.org) [25]. Data from published CRISPR screens in HAP1 cells with the TKOv1 and TKOv3 libraries were obtained from Hart et al. [17]. Data for the gene trap screen in HAP1 cells were obtained from Blomen et al. [41]. Essential genes in hESCs were derived from data published in Yilmaz et al. [52].

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 12 August 2020 Accepted: 20 October 2020
Published online: 23 November 2020

References

1. Grimm S. The art and design of genetic screens: mammalian culture cells. Nat Rev Genet. 2004;5:179–89.
2. Boutros M, Ahringer J. The art and design of genetic screens: RNA interference. Nat Rev Genet. 2008;9:554–66.
3. Doench JG. Am I ready for CRISPR? A user's guide to genetic screens. Nat Rev Genet. 2018;19:67–80.
4. 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.
5. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.
6. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343:80–4.
7. 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.
8. 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.
9. Chen S, Sanjana NE, Zheng K, Shalem O, Lee K, Shi X, et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell. 2015;160:1246–60. https://doi.org/10.1016/j.cell.2015.02.038.
10. 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. 2016;6:914–29.
11. John Liu S, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science. 2017;355:eaah7111.
12. 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.
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. 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.
15. Chen C-H, Xiao T, Xu H, Jiang P, Meyer CA, Li W, et al. Improved design and analysis of CRISPR knockout screens. Bioinformatics. 2018;34:4095–101.
16. 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.
17. 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;7:2719–2727.
18. 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.
19. 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.
20. Wang D, Zhang C, Wang B, Li B, Wang Q, Liu D, et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nat Commun. 2019;10:4284.
21. Isaac RS, Jiang F, Doudna JA, Lim WA, Narlikar GJ, Almeida R. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. Elife. 2016;5 https://doi.org/10.7554/eLife.13450.
22. 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. 2016;5 https://doi.org/10.7554/eLife.12677.
23. Daer RM, Cutts JP, Brafman DA, Haynes KA. The impact of chromatin dynamics on Cas9-mediated genome editing in human cells. ACS Synth Biol. 2017;6:428–38.
24. 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.
25. Rauscher B, Heigwer F, Breinig M, Winter J, Boutros M. GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. Nucleic Acids Res. 2017;45:D679–86.
26. Rauscher B, Valentini E, Hardeland U, Boutros M. Phenotype databases for genetic screens in human cells. J Biotechnol. 2017;261:63–9.
27. 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.
28. Wang T, Yu H, Hughes NW, Liu B, Kendirli A, Klein K, et al. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. Cell. 2017;168:890–903.e15.
29. 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.
30. Steinhart Z, Pavlovic Z, Chandrashekhar M, Hart T, Wang X, Zhang X, et al. Genome-wide CRISPR screens reveal a Wnt–FZD5 signaling circuit as a druggable vulnerability of RNF43-mutant pancreatic tumors. Nature Medicine. 2017;23:60–8.
31. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019;47:D745–51.
32. Hart T, Moffat J. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics. 2016;17:164.
33. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014;10:733.
34. Heigwer F, Zhan T, Breinig M, Winter J, Brügemann D, Leible S, et al. CRISPR library designer (CLD): software for multispecies design of single guide RNA libraries. Genome Biol. 2016;17:55.
35. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods. 2014;11:783–4.
36. Kaufman WL, Kocman I, Agrawal V, Rahn H-P, Besser D, Gossen M. Homogeneity and persistence of transgene expression by omitting antibiotic selection in cell line isolation. Nucleic Acids Res. 2008;36:e111.
37. Michlits G, Jude J, Hinterndorfer M, de Almeida M, Vainorius G, Hubmann M, et al. Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. Nat Methods. 2020;17:708–16.
38. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li G-W, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell. 2013;155:1479–91.
39. Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, et al. Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. Nat Biotechnol. 2011;29:542–6.
40. Olbrich T, Mayor-Ruiz C, Vega-Sendino M, Gomez C, Ortega S, Ruiz S, et al. A p53-dependent response limits the viability of mammalian haploid cells. Proc Natl Acad Sci U S A. 2017;114:9367–72.
41. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350:1092–6.
42. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 2014;15:554.
43. Imkeller K, Ambrosi G, Boutros M, Huber W. gscreend: modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection. Genome Biol. 2020;21:53.
44. Giuliano CJ, Lin A, Girish V, Sheltzer JM. Generating single cell-derived knockout clones in mammalian cells with CRISPR/Cas9. Curr Protoc Mol Biol. 2019;128:e100.
45. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
46. 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.
47. Benaglia T, Chauveau D, Hunter D, Young D. mixtools: An R package for analyzing finite mixture models. J Stat Softw. 2009;32:1–29.

48. Rancati G, Moffat J, Typas A, Pavelka N. Emerging and evolving concepts in gene essentiality. Nat Rev Genet. 2018;19:34–49.
49. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006;126:663–76.
50. Velimezi G, Robinson-Garcia L, Muñoz-Martínez F, Wiegant WW, Ferreira da Silva J, Owusu M, et al. Map of synthetic rescue interactions for the Fanconi anemia DNA repair pathway identifies USP48. Nat Commun. 2018;9:2280.
51. Moder M, Velimezi G, Owusu M, Mazouzi A, Wiedner M, Ferreira da Silva J, et al. Parallel genome-wide screens identify synthetic viable interactions between the BLM helicase complex and Fanconi anemia. Nat Commun. 2017;8:1238.
52. Yilmaz A, Peretz M, Aharony A, Sagi I, Benvenisty N. Defining essential genes for human pluripotent stem cells by CRISPR-Cas9 screening in haploid cells. Nat Cell Biol. 2018;20:610–9.
53. Van Campenhout C, Cabochette P, Veillard A-C, Laczik M, Zelisko-Schmidt A, Sabatel C, et al. Guidelines for optimized gene knockout using CRISPR/Cas9. Biotechniques. 2019;66:295–302.
54. Davis EM, Kim J, Menasche BL, Sheppard J, Liu X, Tan A-C, et al. Comparative haploid genetic screens reveal divergent pathways in the biogenesis and trafficking of glycophosphatidylinositol-anchored proteins. Cell Rep. 2015;11:1727–36.
55. Schick S, Rendeiro AF, Runggatscher K, Ringler A, Boidol B, Hinkel M, et al. Systematic characterization of BAF mutations provides insights into intracomplex synthetic lethalities in human cancers. Nat Genet. 2019;51:1399–410.
56. Lenk GM, Park YN, Lemons R, Flynn E, Plank M, Frei CM, et al. CRISPR knockout screen implicates three genes in lysosome function. Sci Rep. 2019;9:9609.
57. Smits AH, Ziebell F, Joberty G, Zinn N, Mueller WF, Clauder-Münster S, et al. Biological plasticity rescues target activity in CRISPR knock outs. Nat Methods. 2019; https://doi.org/10.1038/s41592-019-0614-5.
58. Gerhards NM, Blomen VA, Mutlu M, Nieuwenhuis J, Howald D, Guyader C, et al. Haploid genetic screens identify genetic vulnerabilities to microtubule-targeting agents. Mol Oncol. 2018;12:953–71.
59. Baggen J, Thibaut HJ, Hurdiss DL, Wahedi M, Marceau CD, van Vliet ALW, et al. Identification of the cell-surface protease ADAM9 as an entry factor for encephalomyocarditis virus. MBio. 2019;10 https://doi.org/10.1128/mBio.01780-19.
60. Rossari F, Minutolo F, Orciuolo E. Past, present, and future of Bcr-Abl inhibitors: from chemical development to clinical efficacy. J Hematol Oncol. 2018;11:84.
61. Ceccaldi R, Sarangi P, D'Andrea AD. The Fanconi anaemia pathway: new players and new functions. Nat Rev Mol Cell Biol. 2016;17:337–49.
62. Richardson CD, Kazane KR, Feng SJ, Zelin E, Bray NL, Schäfer AJ, et al. CRISPR-Cas9 genome editing in human cells occurs via the Fanconi anemia pathway. Nat Genet. 2018;50:1132–9.
63. Drost H-G, Paszkowski J. Biomartr: genomic data retrieval with R. Bioinformatics. 2017;33:1216–7.
64. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
65. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods. 2014;11:122–3.
66. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940–1.
67. 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.
68. Castro FV, McGinn OJ, Krishnan S, Marinov G, Li J, Rutkowski AJ, et al. 5T4 oncofetal antigen is expressed in high risk of relapse childhood pre-B acute lymphoblastic leukemia and is associated with a more invasive and chemotactic phenotype. Leukemia. 2012;26:1487–98.
69. Engler C, Kandzia R, Marillonnet S. A one pot, one step, precision cloning method with high throughput capability. PLoS One. 2008;3:e3647.
70. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21.
71. Zhan T, Boutros M. Towards a compendium of essential genes - From model organisms to synthetic lethality in cancer cells. Crit Rev Biochem Mol Biol. 2016;51:74–85.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.