ARTICLES
https://doi.org/10.1038/s41592-019-0473-0

nature methods

Corrected: Publisher Correction

Combinatorial mutagenesis en masse optimizes the genome editing activities of SpCas9

Gigi C. G. Choi[1], Peng Zhou[1], Chaya T. L. Yuen[1], Becky K. C. Chan[1], Feng Xu[1], Siyu Bao[2], Hoi Yee Chu[2], Dawn Thean[1], Kaeling Tan[3,4], Koon Ho Wong[3,5], Zongli Zheng[2,6,7] and Alan S. L. Wong[1,8*]

The combined effect of multiple mutations on protein function is hard to predict; thus, the ability to functionally assess a vast number of protein sequence variants would be practically useful for protein engineering. Here we present a high-throughput platform that enables scalable assembly and parallel characterization of barcoded protein variants with combinatorial modifications. We demonstrate this platform, which we name CombiSEAL, by systematically characterizing a library of 948 combination mutants of the widely used *Streptococcus pyogenes* Cas9 (SpCas9) nuclease to optimize its genome-editing activity in human cells. The ease with which the editing activities of the pool of SpCas9 variants can be assessed at multiple on- and off-target sites accelerates the identification of optimized variants and facilitates the study of mutational epistasis. We successfully identify Opti-SpCas9, which possesses enhanced editing specificity without sacrificing potency and broad targeting range. This platform is broadly applicable for engineering proteins through combinatorial modifications en masse.

Protein engineering has proven to be an important strategy for generating enzymes, antibodies and genome-editing proteins with new or enhanced properties[1–7]. Combinatorial optimization of a protein sequence relies on strategies for creating and screening a large number of variants, but current approaches are limited in their ability to systematically and efficiently build and test variants with multiple modifications in a high-throughput fashion[8–11]. Conventional site-directed mutagenesis based on structural and biochemical knowledge facilitates the generation of functionally relevant mutants, but using a one-by-one approach to screen combination mutants lacks throughput and scalability. Gene synthesis technology can be deployed to make combination mutants in pooled format, but it typically produces between one and ten errors per kilobase synthesized[12,13] and is prohibitively expensive if the mutations being introduced are scattered over different regions of a protein. Methods such as combinatorial DNA assembly[14,15] and recombination and shuffling[16] create combination mutants by fusing multiple mutated sequences together to assemble the entire protein sequence; however, subsequent genotyping and characterization of the mutations requires selection of clonal isolates or long-read sequencing, and neither of these methods can feasibly be used to track a large number of mutants. Mutagenesis via error-prone PCR and mutator strains for directed evolution allows positive selection of beneficial mutations, but it suffers from selection bias toward a subset of amino acids owing to the rare occurrence of two or more specific nucleotide mutations in a single codon. Even if a great diversity of protein variants could be achieved with sequence randomization, the very limited throughput resulting from genotyping and analyzing selected hits individually would represent a major obstacle in protein engineering. Furthermore, pinpointing the exact mutations that confer a desired phenotype from passenger mutations could be useful for accelerating the combinatorial optimization process.

Here we devised a new cloning method, which we term CombiSEAL, to couple seamless combinatorial DNA assembly with the barcode concatenation strategy used in combinatorial genetics en masse (CombiGEM)[17–19] and allow pooled assembly of barcoded combination mutants that can be easily tracked by high-throughput short-read sequencing (Fig. 1). CombiSEAL works by modularizing the protein sequence into composable parts, each comprising a repertoire of variants tagged with barcodes specifying predetermined mutations at defined positions. Type IIS restriction enzyme sites are used to flank the barcoded parts to create overhangs following digestion originating from the protein-coding sequence, thereby achieving seamless ligation when segments are fused with preceding parts. Unique barcodes are concatenated and appended to each protein-coding-sequence variant in the resultant library after iterative pooled cloning of the parts. This method is advantageous over other strategies as it circumvents the need to perform long-read sequencing over the whole protein-coding region to cover multiple mutations, and instead offers a cost-effective way to quantitatively track each variant in a pool by high-throughput sequencing of short (for example, 50-base-pair) barcodes without the need to select clonal isolates (Supplementary Fig. 1). In addition, pooled characterization of variants allows head-to-head comparisons under the same experimental condition, and facilitates the study of mutational epistasis. Unlike CombiGEM, which only allows combinatorial assembly of discrete genetic components, CombiSEAL does not leave behind a fusion scar sequence, but it can seamlessly link consecutive sequences (for example, different segments of proteins).

[1]Laboratory of Combinatorial Genetics and Synthetic Biology, School of Biomedical Sciences, The University of Hong Kong, Hong Kong, China. [2]Ming Wai Lau Centre for Reparative Medicine, Karolinska Institutet, Hong Kong, China. [3]Faculty of Health Sciences, University of Macau, Macau, China. [4]Genomics, Bioinformatics and Single Cell Analysis Core, Faculty of Health Sciences, University of Macau, Macau, China. [5]Institute of Translational Medicine, University of Macau, Macau, China. [6]Department of Biomedical Sciences, City University of Hong Kong, Hong Kong, China. [7]Biotechnology and Health Centre, City University of Hong Kong Shenzhen Research Institute, Shenzhen, China. [8]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China. *e-mail: aslw@hku.hk

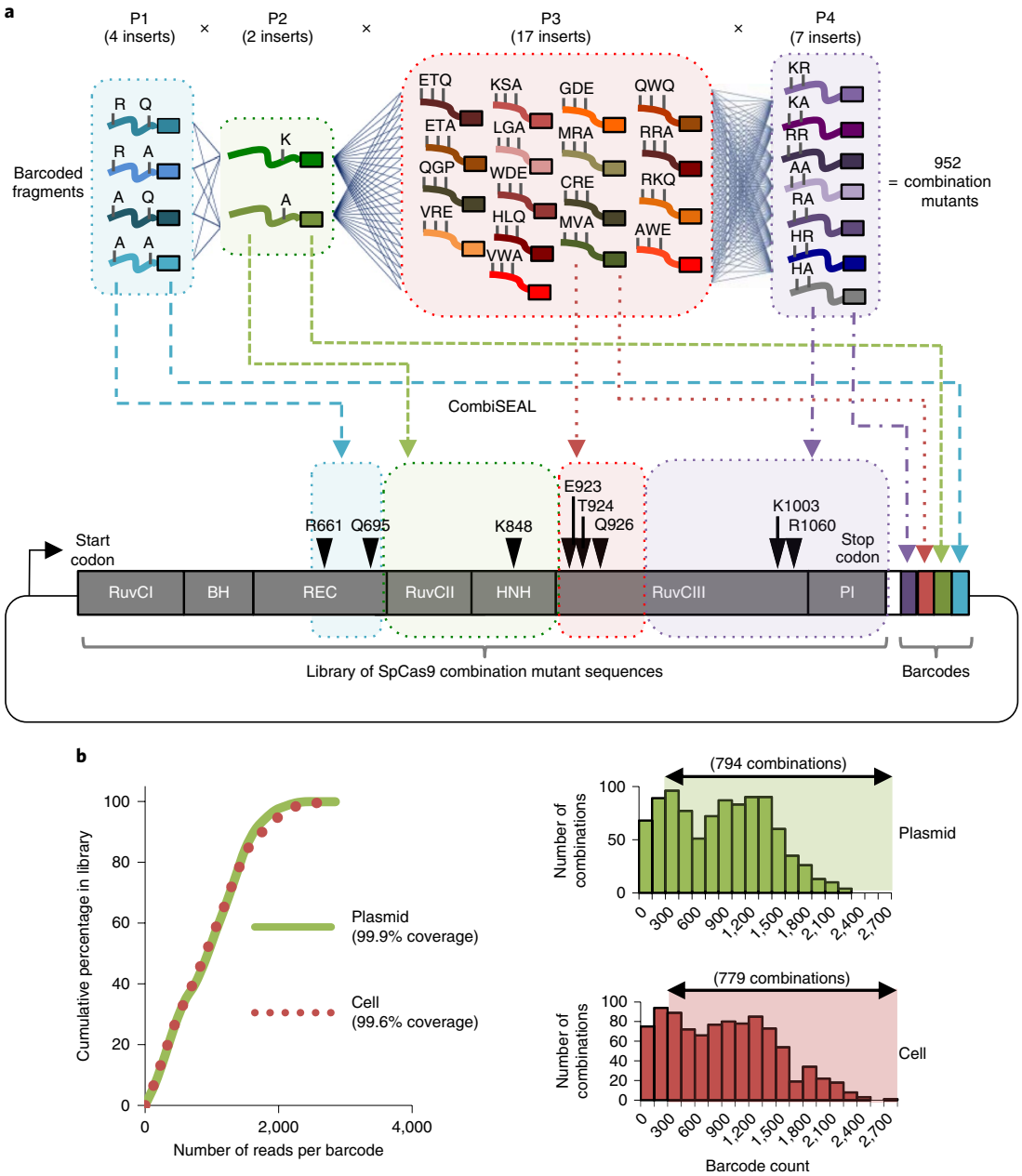

Fig. 1 | Generation of a high-coverage library of combination mutants of SpCas9 and efficient delivery of the library to human cells. a, Strategy for assembling a library of combination mutants for SpCas9. The coding sequence of SpCas9 was modularized into four composable parts (P1 to P4), each comprising a repertoire of barcoded fragments encoding predetermined amino acid substitutions at defined positions. A library of 952 SpCas9 variants was assembled by consecutive rounds of one-pot seamless ligation of the parts, and concatenated barcodes that uniquely tagged each variant were generated (Supplementary Fig. 2). **b**, Cumulative distribution of sequencing reads for the barcoded library of combination mutants in the plasmid pool extracted from *E. coli* and infected OVCAR8-ADR cell pools. High coverage of the library within the plasmid and infected cell pools (~99.9% and ~99.6%, respectively) was detected from ~0.8 million reads per sample, and most combinations were detected with at least 300 absolute barcode reads (shaded areas).

Therefore, this new platform has tremendous potential for protein engineering.

Results

High-throughput screening of SpCas9 combination mutants. We applied CombiSEAL to assemble a library of combination mutants for SpCas9, a CRISPR nuclease that is widely used for genome engineering[20–23], with the aim of identifying optimized variants with high editing specificity and activity. Previously, SpCas9 nucleases carrying specific combinations of mutations, including eSpCas9(1.1) (ref. [3]), SpCas9-HF1 (ref. [4]), HypaCas9 (ref. [5]) and evoCas9 (ref. [6]), have been engineered to minimize off-target editing. However, these variants have fewer targetable sites owing to their incompatibility with guide RNAs (gRNAs) starting with a mismatched 5′ guanine[3–6,24–27]. A limited number of combination mutants have been generated and tested to date (Supplementary Table 1), and thus a more systematic exploration of other SpCas9 variants with better compatibility with gRNAs bearing an additional 5′ guanine is necessary.

Using CombiSEAL, we modularized the SpCas9 sequence into four parts, and barcoded inserts, comprising different random and

specific mutations in individual parts, were cloned into storage vectors (Fig. 1a and Supplementary Fig. 2a,b; see Methods for details). A combinatorial barcoded library (with $4\times2\times17\times7=952$ SpCas9 variants and including the wild-type (WT) SpCas9 and eSpCas9(1.1) sequences) was then pooled and assembled in a lentiviral vector. The individual parts and assembled constructs in the library were sequenced to confirm the highly accurate assembly of barcoded variants (Methods). We detected high coverage for the library within both the plasmid pools stored in *Escherichia coli* (951 of 952 variants) and infected human cell pools (948 of 952 variants) (Fig. 1b), in addition to highly reproducible representation between the plasmid and infected cell pools, as well as between biological replicates of infected cell pools (Supplementary Fig. 2c).

To search for robust and specific SpCas9 variants, we established a reporter system using monoclonal human cell lines that stably expressed red fluorescent protein (RFP) and a gRNA targeting the *RFP* gene sequence (referred to as RFPsg5-ON and RFPsg8-ON here inafter; Fig. 2a). Unlike in previous screens, which primarily used 20-nucleotide gRNAs starting with a 5′ guanine[3–6], we used gRNAs carrying an additional 5′ guanine in our reporter system to look for compatible SpCas9 variants for which targeting range was not diminished. Cells were infected with the SpCas9 variant library and sorted into bins on the basis of RFP fluorescence 14 d after infection. Loss of RFP fluorescence reflects DNA cleavage and indel-mediated disruption of the target site, and thus cells harboring active SpCas9 variants would be enriched in the sorted bin with low RFP fluorescence. Using Illumina HiSeq to track the barcoded SpCas9 variants, a subpopulation of variants was found to be enriched by more than 1.5-fold in the sorted bin that encompassed the ~5% of the cell population with the lowest level of RFP fluorescence (that is, bin A) as compared to the unsorted population (Fig. 2b and Supplementary Fig. 3). WT SpCas9 was enriched in bin A in both RFPsg5-ON and RFPsg8-ON reporter systems, while eSpCas9(1.1) was enriched in this bin in the RFPsg8-ON system. To facilitate parallel characterization of the on- and off-target activities of SpCas9 variants, we also generated cell lines harboring synonymous mutations in *RFP*, such that targeting of the mismatched site would indicate the off-target activity of the SpCas9 variant (referred to as RFPsg5-OFF5-2 and RFPsg8-OFF5 here inafter; Fig. 2a). WT SpCas9, but not eSpCas9(1.1), was enriched in bin A in both the RFPsg5-OFF5-2 and RFPsg8-OFF5 reporter systems (Fig. 2b and Supplementary Fig. 3).

We ranked and plotted the on- and off-target activities for the library of SpCas9 variants on the basis of enrichment in the sorted bin relative to the unsorted population, and found that a majority of the mutants were impaired in both their on- and off-target activities (Fig. 3a). We defined activity-optimized variants as those with enrichment ratios that were at least 90% of the ratio for WT SpCas9 in both the RFPsg5-ON and RFPsg8-ON reporter lines and less than 60% of the WT value in both the RFPsg5-OFF5-2 and RFPsg8-OFF5 reporter lines. One variant (here inafter referred to as Opti-SpCas9) met these criteria and was selected for further characterization (Supplementary Table 2). We also identified a variant with high fidelity, named OptiHF-SpCas9, on the basis of this variant having an enrichment ratio of at least 50% of the WT ratio for both the RFPsg5-ON and RFPsg8-ON lines and less than 90% of the WT value for both the RFPsg5-OFF5-2 and RFPsg8-OFF5 lines (Supplementary Table 2). We verified the efficiency and specificity of Opti-SpCas9 and OptiHF-SpCas9 by performing individual validation assays to measure on- and off-target activity. Using multiple cell lines each expressing a gRNA that targeted the matched or mismatched *RFP* site, we confirmed that, in comparison to WT SpCas9, Opti-SpCas9 exhibited equivalent on-target activity (94.6% of WT activity, averaged from three matched sites) and substantially reduced off-target activity (1.7% of WT activity, averaged from three mismatched sites), while OptiHF-SpCas9 showed reduced activities at both on-target (63.6% of WT activity, averaged from two matched sites) and off-target (2.0% of WT activity, averaged from two mismatched sites) sites (Fig. 3b).

Studying mutational epistasis for the editing efficiency of SpCas9. Systematic construction of protein variants by CombiSEAL allows us to classify sets of amino acid substitutions as neutral, beneficial or deleterious and explore their hard-to-predict epistatic interactions. Using the enrichment ratio as an index for the editing activity of SpCas9 (Supplementary Fig. 4), we constructed heat maps presenting the on- and off-target activities conferred by the combinations of mutations and the epistatic interactions involved (Fig. 4 and Supplementary Fig. 5). We found that the number and type of substitutions introduced in the amino acid sequence of SpCas9 at residues predicted to interact with the target and non-target DNA strands (such as R661, Q695, K848, Q926, K1003 and K1060) governed the optimal balance between maximizing on-target efficiency and minimizing off-target activity. The activity-optimized variant Opti-SpCas9 differs from WT SpCas9 by two substitutions at DNA-contacting residues (R661A and K1003H). A comparison of the effects of a conservative change to one of three basic residues (lysine, arginine or histidine) at amino acid position 1003 of SpCas9 revealed that K1003H was the preferred substitution, exhibiting a positive epistatic interaction with the R661A substitution and conferring Opti-SpCas9 with high editing efficiency at on-target sites (Fig. 4). Addition of the Q926A substitution, which has been shown to confer higher specificity for SpCas9-HF1 (ref. [4]), to Opti-SpCas9 slightly decreased the off-target effect (from 1.0% of WT activity for Opti-SpCas9 to 0.2% for Opti-SpCas9 Q926A; averaged from three mismatched target sites), and considerably reduced the on-target activity by 21.6%, 62.4% and 99.9% across the three matched sites tested (Fig. 3b). Moreover, we found that most SpCas9 variants bearing three or more mutations at the DNA-contacting residues generated fewer edits at both on- and off-target sites (Fig. 4). These results are consistent with previous findings that excessive numbers of alanine substitutions at these DNA-contacting residues severely reduce the editing activity of SpCas9 (ref. [25]). However, additional substitutions introduced at residues responsible for conformational control of the HNH and RuvC nuclease domains of SpCas9 (ref. [28]) (such as E923M and T924V, and E923H and T924L, at residues located in the linker region connecting the two domains) restored on-target editing at the RFPsg5-ON site for some of the SpCas9 variants carrying three or more mutations at the DNA-contacting residues (Fig. 4). The high-fidelity variant OptiHF-SpCas9 also contains E923M and T924V substitutions in addition to Q695A, K848A and Q926A substitutions, and it showed slightly higher on-target activity at the RFPsg8-ON site than the variant with only Q695A, K848A and Q926A substitutions (Fig. 4). Our data support the model that the DNA binding and cleavage activities of SpCas9 are functionally coupled to determine editing specificity and efficiency[5,29], and highlight the potential to program the editing performance of SpCas9 by modifying linker residues.

Characterizing the optimized SpCas9 variants. In the design and construction of gRNAs, a 5′ guanine is commonly included or added to the start of a gRNA sequence to facilitate efficient transcription from the U6 promoter. WT SpCas9 is compatible with gRNAs that have an additional 5′ guanine that is mismatched to the protospacer sequence. In contrast, eSpCas9(1.1), SpCas9-HF1, HypaCas9 and evoCas9 lose their editing efficiency when a 20-nucleotide gRNA bearing an additional 5′ guanine (that is, GN_{20}) or lacking a starting guanine (that is, HN_{19}) is used[4,6,24–27]. The use of gRNAs with a 5′ guanine matched to the protospacer sequence could reduce the number of editable sites in the human genome by approximately 4.3-fold given the availability of $GN_{19}NGG$ sites as compared to $N_{20}NGG$ sites (Supplementary Fig. 6). We further characterized the editing activity of Opti-SpCas9 with gRNAs carrying an

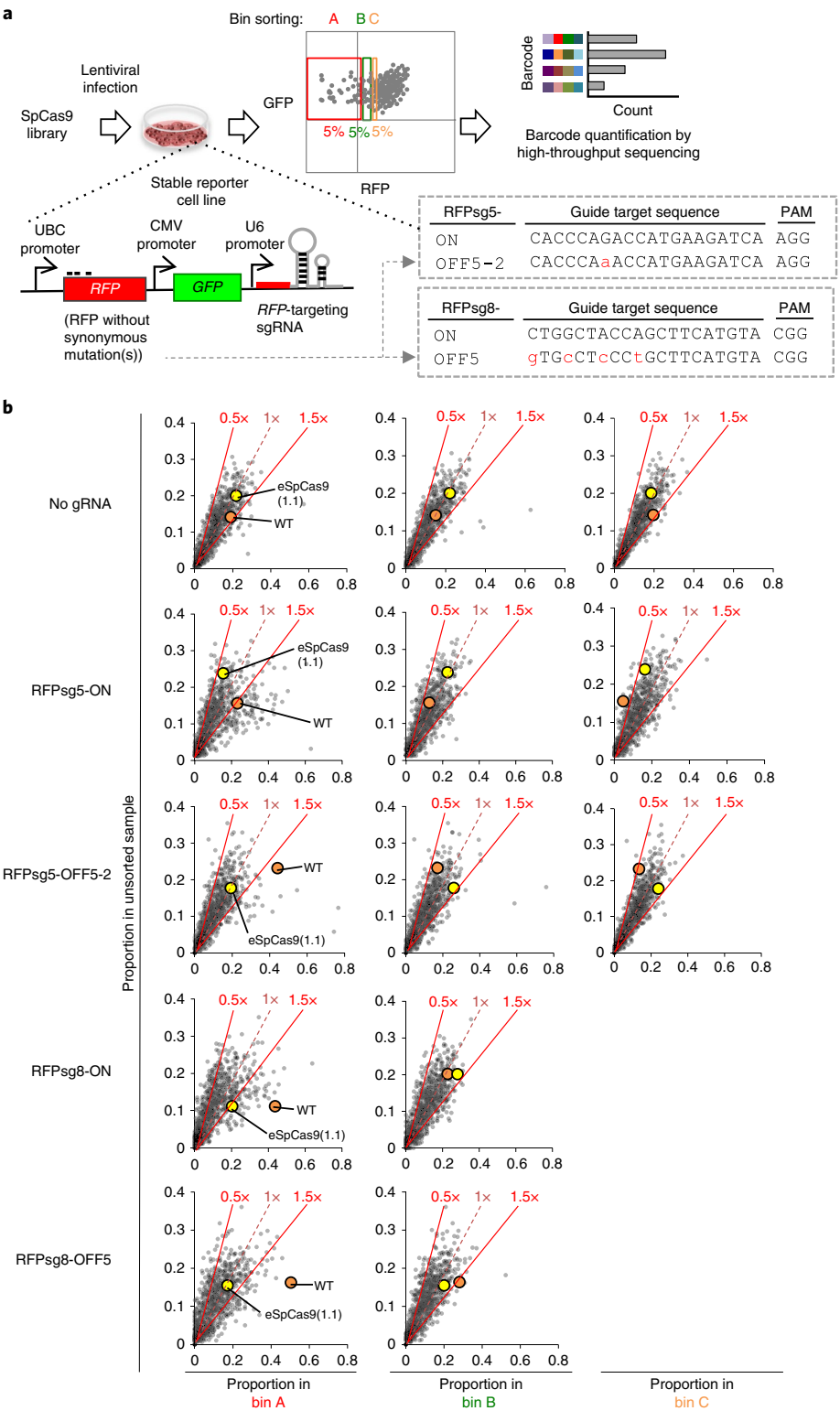

Fig. 2 | Strategy for the profiling of on- and off-target activities of SpCas9 variants in human cells. a, The SpCas9 library was delivered via lentiviruses at a multiplicity of infection of ~0.3 to OVCAR8-ADR reporter cell lines in which the *RFP* and *GFP* genes are expressed from UBC and CMV promoters, respectively, while in tandem a gRNA targeting *RFP* (RFPsg5 or RFPsg8) is expressed from a U6 promoter. RFP and GFP expression were analyzed by flow cytometry. The on-target activity of SpCas9 was measured using reporter systems in which the gRNA spacer sequence completely matched the *RFP* target site, while its off-target activity was measured using reporter systems in which the *RFP* target site harbored a synonymous mutation. Cells with an active SpCas9 variant were expected to lose RFP fluorescence. Cells were sorted into bins each encompassing ~5% of the population on the basis of RFP fluorescence, and their genomic DNA was extracted for quantification of the barcoded SpCas9 variant by Illumina HiSeq. **b**, Scatterplots comparing the barcode count of each SpCas9 variant between the sorted bins (A, B and C) and the unsorted population. Each dot represents an SpCas9 variant, and WT SpCas9 and eSpCas9(1.1) are labeled. Solid reference lines denote 1.5-fold enrichment and 0.5-fold depletion in barcode counts, and the dotted reference line corresponds to no change in barcode count in the sorted bin as compared to the unsorted population.

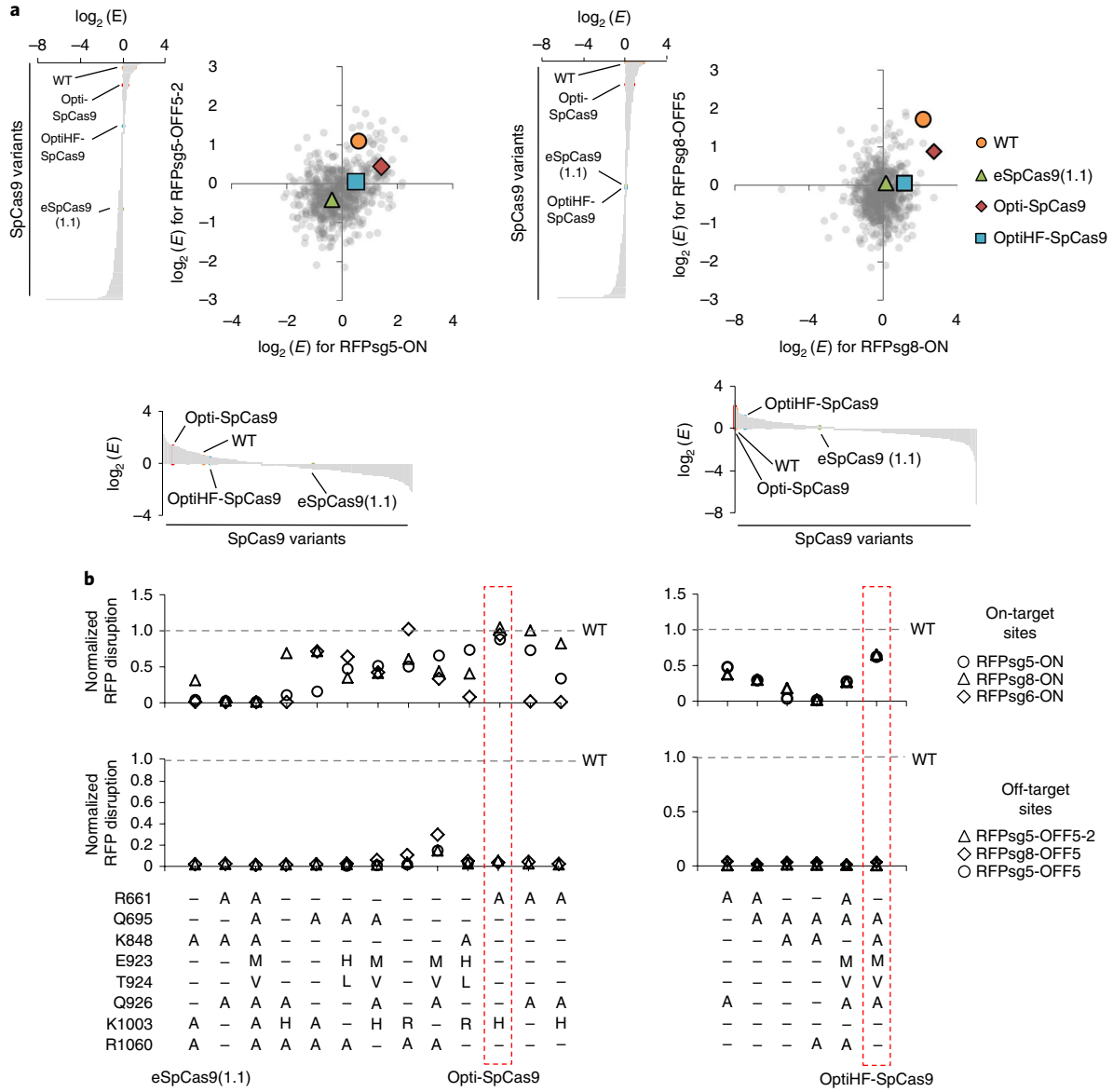

Fig. 3 | High-throughput profiling reveals the broad-spectrum specificity and efficiency of SpCas9 combination mutants. a, Combination mutants of SpCas9 were ranked by their $\log_2$-transformed enrichment ratio ($\log_2(E)$) representing their relative abundance in the sorted RFP-depleted cell population for each of the on-target (*x* axis) and off-target (*y* axis) reporter cell lines, using profiling data from two biological replicates (Supplementary Table 2; see Methods for details). Each dot in the scatter plots represents an SpCas9 variant, and WT SpCas9, eSpCas9(1.1), Opti-SpCas9 and OptiHF-SpCas9 are labeled. More than 99% of the combination mutants had a lower $\log_2(E)$ than WT SpCas9 in the two off-target reporter lines RFPsg5-OFF5-2 and RFPsg8-OFF5, while 16.2% and 2.5% of the mutants had a higher $\log_2(E)$ than WT SpCas9 in the two on-target reporter lines RFPsg5-ON and RFPsg8-ON, respectively. **b**, OVCAR8-ADR cells harboring reporter constructs with on-target (top) and off-target (bottom) sites were infected with lentiviruses encoding individual SpCas9 combination mutants. The editing efficiency of the SpCas9 variants was measured as the percentage of cells with depleted RFP fluorescence and compared to the efficiency for WT SpCas9.

additional 5′ guanine, and found that it exhibited on-target DNA cleavage activity comparable to that of WT SpCas9 (95.1% of WT activity) by assaying endogenous loci that we and others have previously studied[3–5,18,30], while eSpCas9(1.1) and HypaCas9 exhibited greatly reduced activity (32.4% and 25.6% of WT activity, respectively) (Fig. 5a and Supplementary Fig. 7). The reduced editing was not due to decreased protein expression levels of the two SpCas9 variants (Supplementary Fig. 8). These results corroborate the on-target activities observed for these variants in our screening systems in which gRNAs bearing an additional 5′ guanine were used (Figs. 2 and 3a), as well as the on-target activities observed in independent validation experiments using green fluorescent protein (GFP) disruption assays (Fig. 3b and Supplementary Fig. 9). In addition, Opti-SpCas9, eSpCas9(1.1) and HypaCas9 exhibited editing activity comparable to that of WT SpCas9 (109.1%, 103.3% and 106.8% of WT activity, respectively) when 20-nucleotide gRNAs starting with a matched 5′ guanine were used (Fig. 5a). We further compared Opti-SpCas9 with OptiHF-SpCas9 and the more recently characterized high-fidelity variants evoCas9 (ref. [6]) and Sniper-Cas9 (ref. [31]), and found that OptiHF-SpCas9, evoCas9 and Sniper-Cas9 generated fewer on-target edits than Opti-SpCas9 (60.7%, 99.8% and 51.7% fewer, respectively, when expressed with gRNAs carrying an additional 5′ guanine, and 40.1%, 87.7% and 63.9% fewer, respectively, when using gRNAs starting with a matched 5′ guanine) (Fig. 5b and Supplementary Figs. 7 and 8). Altogether, the requirement for a matched 5′ guanine as the first base of the 20-nucleotide

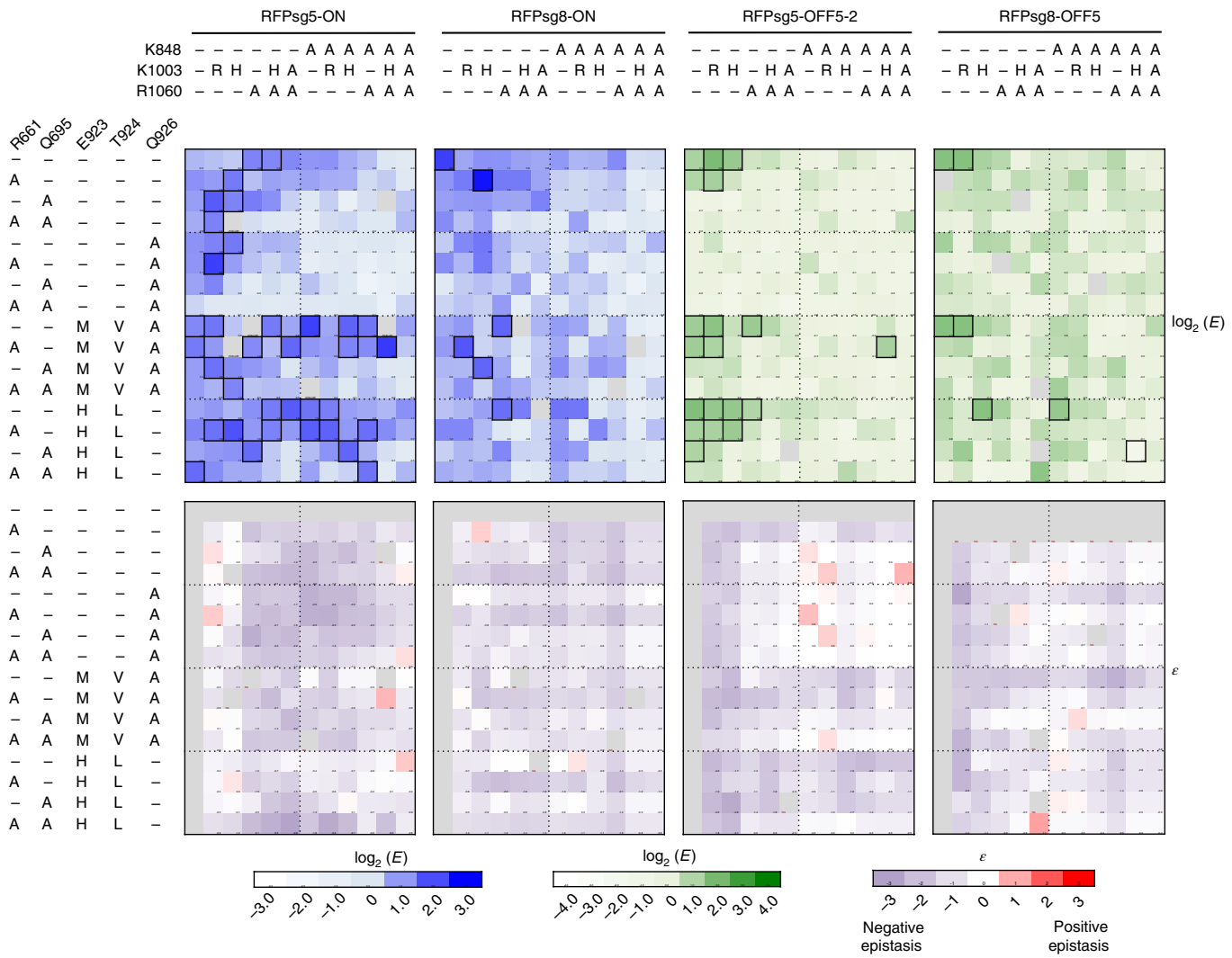

Fig. 4 | Heat maps depicting editing efficiency and epistasis for on- and off-target sites. Editing efficiency (top; measured by $\log_2(E)$) and epistasis (bottom; ε) scores were determined for each SpCas9 combination mutant as described in the Methods. Amino acid residues that are predicted to make contact with the target DNA strand or are located in the linker region connecting the HNH and RuvC domains of SpCas9 are grouped on the *y* axis, while those predicted to interact with the non-target DNA strand are presented on the *x* axis, to aid visualization. We computed the *P* value for $\log_2(E)$ of each combination by comparing the $\log_2(E)$ value with the whole population obtained from two independent biological replicates using a two-sample two-tailed Student's *t* test (MATLAB function ttest2). Adjusted *P* values (that is, *Q* values) were calculated on the basis of the distribution of *P* values (MATLAB function mafdr) to correct for multiple-hypothesis testing. A $\log_2(E)$ value was considered as statistically significant relative to the entire population at $Q < 0.1$ (significant combinations are outlined). The full heat maps are presented in Supplementary Fig. 5. The combinations for which no enrichment ratio or epistasis score was measured are in gray.

gRNA sequence for transcription from the U6 promoter, which limits the practical usefulness of other previously engineered SpCas9 proteins with improved specificity, does not apply to Opti-SpCas9, which is compatible with gRNAs carrying an additional 5′ guanine. Our findings demonstrate that engineered SpCas9 proteins do not necessarily have to sacrifice targeting range for specificity.

We further examined the off-target activity of the different SpCas9 variants. Eight potential off-target loci that are edited by WT SpCas9 when using gRNAs targeting *VEGFA* site 3 and *DNMT1* site 4 were amplified[3–5,30], and genomic indels induced by WT SpCas9 were detected at four of these sites (*VEGFA* OFF1, *VEGFA* OFF2, *VEGFA* OFF3 and *DNMT1* OFF1) in OVCAR8-ADR cells. When Opti-SpCas9, eSpCas9(1.1) and HypaCas9 were used instead of WT SpCas9, we detected off-target edits only at the *VEGFA* OFF1 site (Supplementary Fig. 10). Among the four variants, Opti-SpCas9 showed the highest ratio of on- to off-target activity at this site (Supplementary Fig. 10). To compare the mismatch tolerance of the different SpCas9 variants, we generated gRNAs containing mismatches of one to four bases in comparison to the reporter gene target (a genomically integrated *GFP* gene sequence). The mismatched bases spanned across different positions of the gRNA spacer sequence. We measured the loss of GFP fluorescence as a proxy for DNA cleavage and indel-mediated disruption of the target site. We found that Opti-SpCas9 was largely intolerant to gRNAs with two or more mismatched bases, although a relatively low level of activity (3.5% for Opti-SpCas9 versus 73.2% for WT SpCas9) was detected in one of the eight sites carrying two base mismatches (Supplementary Fig. 11). We noticed that eSpCas9(1.1) and HypaCas9 produced fewer edits at both the on-target site (reduced by more than 60%) and the off-target sites in our reporter systems (Supplementary Fig. 11). Although a similar level of on-target activity was observed for WT SpCas9 and Opti-SpCas9 (97.6% of WT activity), Opti-SpCas9 showed a higher specificity than WT SpCas9, as indicated by the generation of significantly fewer off-target edits at 13 of the 20 sites

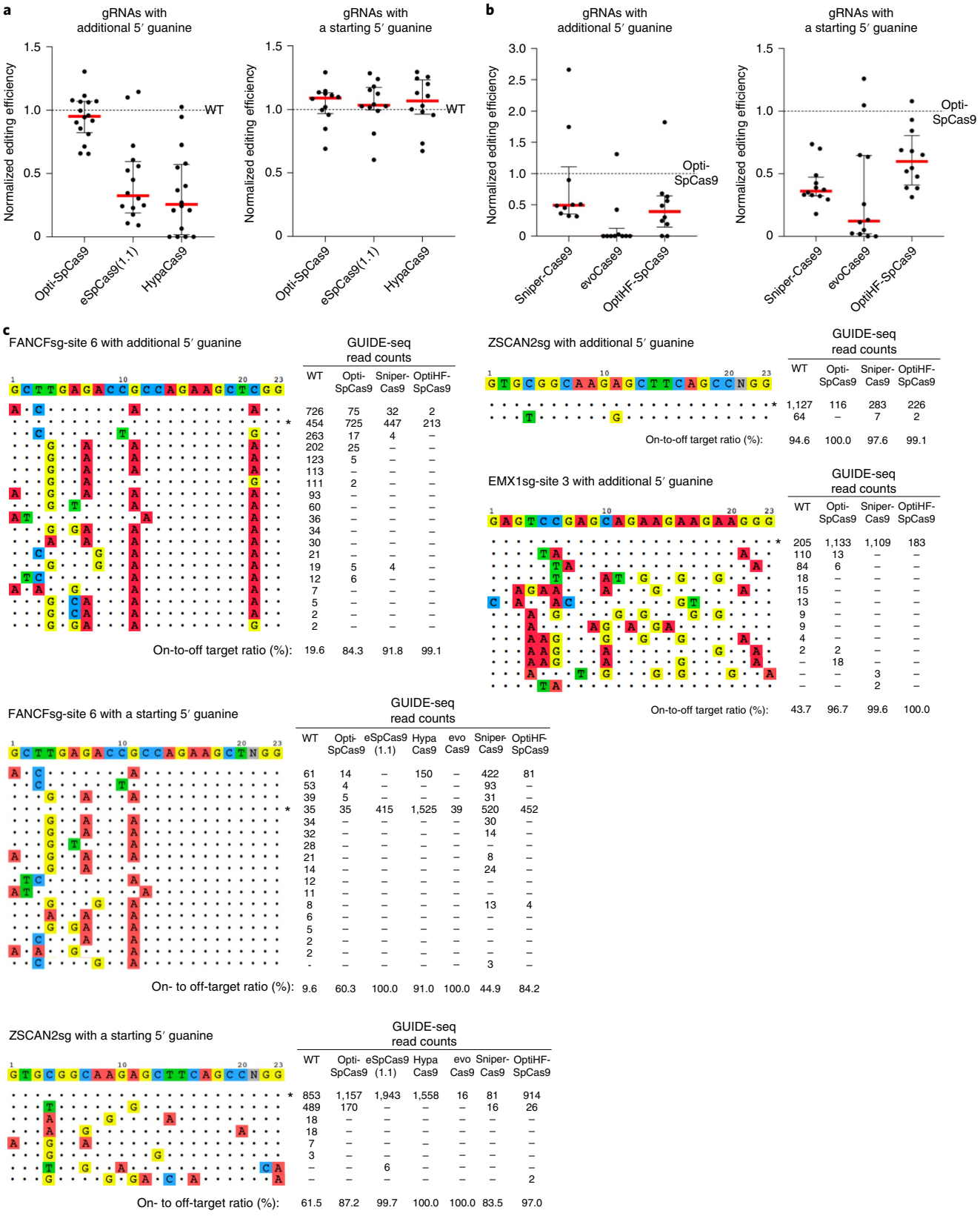

Fig. 5 | Opti-SpCas9 exhibits robust on-target and reduced off-target activities. a,**b**, Assessment of SpCas9 variants for efficient on-target editing with gRNAs targeting endogenous loci. The percentage of sites with indels was measured using a T7 endonuclease I (T7E1) assay. The ratio of the on-target activity of SpCas9 variants to the acitivity of WT SpCas9 (**a**) and Opti-SpCas9 (**b**) was determined, and the median and interquartile range for the normalized percentage of indel formation are shown for the 10–16 loci tested. Each locus was measured once or twice; the full dataset is presented in Supplementary Fig. 7. **c**, GUIDE-seq genome-wide specificity profiles for the panel of SpCas9 variants paired with the indicated gRNAs. Mismatched positions in off-target sites are colored, and GUIDE-seq read counts were used as a measure of the cleavage efficiency at a given site. The list of gRNA sequences used is presented in Supplementary Table 5.

containing a single-base mismatch, despite the fact that a considerable number of off-target edits were detected (Supplementary Fig. 11). Others have also reported editing activity at sites with single-base mismatches using eSpCas9(1.1), SpCas9-HF1, HypaCas9, evoCas9 and Sniper-Cas9 (refs. [3,5,6,31]). Nevertheless, a majority of the in silico predicted off-target sites in the genome contains two or more mismatches in comparison to the gRNA sequence[32], and thus tolerance toward single-base mismatches should not limit the ability of SpCas9 to achieve accurate genome editing. Unbiased identification of double-strand breaks enabled by sequencing (GUIDE-seq) was performed to look at the genome-wide cleavage activities of Opti-SpCas9 and other engineered SpCas9 variants. Our results indicated that Opti-SpCas9 generated substantially less off-target cleavage than WT SpCas9, and OptiHF-SpCas9 showed an increased ratio of on- to off-target activity, which was comparable to that of other reported high-fidelity variants, such as eSpCas9(1.1), HypaCas9, evoCas9 and Sniper-Cas9 (Fig. 5c and Supplementary Table 3). As compared to eSpCas9(1.1) and HypaCas9, Opti-SpCas9 exhibited better compatibility when used with truncated gRNAs (Supplementary Fig. 12), which could offer a complementary strategy to improve the editing specificity of Opti-SpCas9 (ref. [33]).

Discussion

We have established CombiSEAL, a simple yet extremely powerful platform, to address the unmet need for rapid and simultaneous profiling of high-order combinatorial mutations for protein engineering. This strategy uses a pooled assembly approach to bypass the laborious steps of building individual combination mutants one by one, and exploits barcoding tactics to allow parallel experimentation on, and identification of, the top performers from a large number of protein variants to facilitate protein engineering. Furthermore, the method can be applied to map epistatic relationships between mutations. Using the CombiSEAL method, we successfully identified Opti-SpCas9 and OptiHF-SpCas9, which are new variants of SpCas9 with superior genome-editing efficiency and specificity across a broad range of endogenous targets in human cells (Supplementary Table 3). The CombiSEAL pipeline can be readily applied to build even more Cas9 variants to broaden the search for variants with other properties, such as increased protospacer adjacent motif (PAM) flexibilty[7] and enhanced compatibility with ribonucleoprotein delivery[34]. We envision that CombiSEAL will accelerate the engineering of CRISPR enzymes (including SaCas9 (ref. [35]) and Cpf1 (ref. [36])) and their derivatives (for example, base editors[37–40]) for precise editing of the genome. The generalizability of this approach will also expand our scope to systematically engineer diverse proteins, as well as other biological molecules and systems including synthetic DNAs and genetic regulatory circuits, which are relevant to many biomedical and biotechnology applications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0473-0.

Received: 27 July 2018; Accepted: 3 June 2019;
Published online: 15 July 2019

References

1. Bornscheuer, U. T. et al. Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
2. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
3. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
4. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
5. Chen, J. S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).
6. Casini, A. et al. A highly specific SpCas9 variant is identified by in vivo screening in yeast. *Nat. Biotechnol.* **36**, 265–271 (2018).
7. Hu, J. H. et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).
8. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).
9. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
10. Gasperini, M., Starita, L. & Shendure, J. The power of multiplexed functional analysis of genetic variants. *Nat. Protoc.* **11**, 1782–1787 (2016).
11. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
12. Ma, S., Saaem, I. & Tian, J. Error correction in gene synthesis technology. *Trends Biotechnol.* **30**, 147–154 (2012).
13. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
14. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).
15. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
16. Trudeau, D. L., Smith, M. A. & Arnold, F. H. Innovation by homologous recombination. *Curr. Opin. Chem. Biol.* **17**, 902–909 (2013).
17. Wong, A. S., Choi, G. C., Cheng, A. A., Purcell, O. & Lu, T. K. Massively parallel high-order combinatorial genetics in human cells. *Nat. Biotechnol.* **33**, 952–961 (2015).
18. Wong, A. S. et al. Multiplexed barcoded CRISPR–Cas9 screening enabled by CombiGEM. *Proc. Natl Acad. Sci. USA* **113**, 2544–2549 (2016).
19. Cheng, A. A., Ding, H. & Lu, T. K. Enhanced killing of antibiotic-resistant bacteria enabled by massively parallel combinatorial genetics. *Proc. Natl Acad. Sci. USA* **111**, 12462–12467 (2014).
20. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR–Cas9. *Science* **346**, 1258096 (2014).
21. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR–Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
22. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).
23. Barrangou, R. & Horvath, P. A decade of discovery: CRISPR functions and applications. *Nat. Microbiol.* **2**, 17092 (2017).
24. Kim, S., Bae, T., Hwang, J. & Kim, J. S. Rescue of high-specificity Cas9 variants using sgRNAs with matched 5′ nucleotides. *Genome Biol.* **18**, 218 (2017).
25. Kulcsar, P. I. et al. Crossing enhanced and high fidelity SpCas9 nucleases to optimize specificity and cleavage. *Genome Biol.* **18**, 190 (2017).
26. Zhang, D. et al. Perfectly matched 20-nucleotide guide RNA sequences enable robust genome editing using high-fidelity SpCas9 nucleases. *Genome Biol.* **18**, 191 (2017).
27. Kato-Inui, T., Takahashi, G., Hsu, S. & Miyaoka, Y. Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 with improved proof-reading enhances homology-directed repair. *Nucleic Acids Res.* **46**, 4677–4688 (2018).
28. Sternberg, S. H., LaFrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR–Cas9. *Nature* **527**, 110–113 (2015).
29. Singh, D. et al. Mechanisms of improved specificity of engineered Cas9s revealed by single-molecule FRET analysis. *Nat. Struct. Mol. Biol.* **25**, 347–354 (2018).
30. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR–Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
31. Lee, J. K. et al. Directed evolution of CRISPR–Cas9 to increase its specificity. *Nat. Commun.* **9**, 3048 (2018).
32. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).
33. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR–Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
34. Vakulskas, C. A. et al. A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).
35. Ran, F. A. et al. In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
36. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system. *Cell* **163**, 759–771 (2015).
37. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

38. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729 (2016).
39. Gaudelli, N. M. et al. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
40. Li, X. et al. Base editing with a Cpf1–cytidine deaminase fusion. *Nat. Biotechnol.* **36**, 324–327 (2018).

Acknowledgements

We thank members of the Wong lab for helpful discussions, and Z. Dong, L. Qin, N. Shirgaonkar and L. Pardeshi from the Genomics, Bioinformatics and Single Cell Analysis Core of the Faculty of Health Sciences at the University of Macau for their technical support. We thank J. Chan for support at the High Performance Computing Cluster (HPCC) of ICTO of the University of Macau. We thank T. Ochiya for OVCAR8-ADR cells. We thank the Faculty Core Facility at the LKS Faculty of Medicine of The University of Hong Kong for providing and maintaining the equipment needed for flow cytometry analysis and cell sorting. This work was supported by The University of Hong Kong start-up and internal funds, the Croucher Foundation Start-up Allowance and the Hong Kong Research Grants Council (ECS-27105716, GRF-17104619 and TRS-T12-710/16-R) (to A.S.L.W.); the Swedish Research Council (2016-02830) and the National Natural Science Foundation of China (81672098) (to Z.Z.); and the Science and Technology Development Fund of Macau S.A.R. (FDCT 085/2014/A2), the Research Services and Knowledge Transfer Office of the University of Macau (MYRG2016-00211-FHS and MYRG2018-00017-FHS), and the Start-up fund from the Faculty of Health Sciences, University of Macau (to K.H.W. and K.T.).

Author contributions

G.C.G.C. and A.S.L.W. conceived the work. G.C.G.C., P.Z., C.T.L.Y., B.K.C.C., F.X., D.T. and A.S.L.W. designed and performed the experiments and interpreted and analyzed the data. G.C.G.C., C.T.L.Y., K.T., K.H.W. and A.S.L.W. performed computational analyses on next-generation sequencing data for CombiSEAL experiments. G.C.G.C., P.Z., A.S.L.W., S.B., H.Y.C. and Z.Z. performed GUIDE-seq experiments and analyzed the data. G.C.G.C. and A.S.L.W. wrote the paper.

Competing interests

A.S.L.W. and G.C.G.C. have filed a patent application that is based on this work.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0473-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.S.L.W.

Peer review information: Lei Tang and Nicole Rusk were the primary editors on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Construction of DNA vectors. The vectors used in this study (Supplementary Table 4) were constructed using standard molecular cloning techniques, including PCR, restriction enzyme digestion, ligation and Gibson assembly. Custom oligonucleotides were purchased from Integrated DNA Technologies and Genewiz. The vector constructs were transformed into *E. coli* strain DH5α, and 50 μg ml^{-1} of carbenicillin–ampicillin was used to isolate colonies harboring the constructs. DNA was extracted and purified using Plasmid Mini (Takara) or Midi (Qiagen) kits. Sequences of the vector constructs were verified with Sanger sequencing.

To create the lentiviral expression vector encoding eSpCas9(1.1), HypaCas9 or SpCas9-HF1, together with zeocin as the selection marker, the SpCas9 sequences were amplified and/or mutated from pAWp30 (Addgene, 73857), eSpCas9(1.1) (Addgene, 71814) and VP12 (Addgene, 72247) by PCR using Phusion DNA polymerase (New England Biolabs) and cloned into the pFUGW lentiviral vector backbone using Gibson Assembly Master Mix (New England Biolabs). Lentiviral expression vectors encoding evoCas9, Sniper-Cas9 and xCas9(3.7) were created by amplifying their SpCas9 sequences from Addgene constructs 107550, 113912 and 1803380, respectively, and cloning them into the pFUGW vector backbone. To construct a storage vector to drive expression of a gRNA that targeted a specific gene from a U6 promoter, oligonucleotide pairs with the gRNA target sequences were synthesized, annealed and cloned into a BbsI-digested pAWp28 vector (Addgene, 73850) using T4 DNA ligase (New England Biolabs) as previously described[18]. To find SpCas9 variants that were compatible with gRNAs carrying an additional 5′ guanine at the start of the 20-nucleotide spacer sequence to favor transcription from the U6 promoter, gRNAs containing an additional 5′ guanine were used in this study, except for some of those used in Fig. 5 and Supplementary Fig. 9. The gRNA spacer sequences are listed in Supplementary Table 5. To construct a lentiviral vector for expression of gRNA from a U6 promoter, U6-gRNA expression cassettes were prepared by digestion of the storage vector with BglII and MfeI enzymes (Thermo Fisher Scientific) and inserted into the pAWp12 (Addgene, 72732) vector backbone using ligation via the compatible sticky ends generated by digestion of the vector with BamHI and EcoRI enzymes (Thermo Fisher Scientific). To express the gRNAs together with the dual RFP and GFP reporters, the U6-driven gRNA expression cassettes were inserted into the pAWp9 (Addgene, 73851) lentiviral vector backbone (rather than the pAWp12 backbone) using the strategy described above.

Creation of barcoded DNA parts for SpCas9. Guided by the evidence available when we started this study, we focused on building a library of combination mutants with substitutions at amino acid residues that were predicted to make contacts with the target and non-target DNA strands of the gRNA-directed genomic sites (including those identified in SpCas9-HF1 (ref. [4]) and eSpCas9(1.1) (ref. [3]), respectively) or to control the conformational dynamics of the HNH and RuvC nuclease domains of SpCas9 for DNA cleavage[28]. Eight amino acid residues were selected and modified to harbor specified or randomly generated substitutions (Fig. 1a). The basic residues were mutated to alanine to evaluate the role of these charged residues. In addition to an alanine substitution at K1003, which was previously introduced to eSpCas9(1.1), we also mutated this residue to other positively charged residues (arginine and histidine) to minimize the impact of the substitution on protein stability. We hypothesized that specific combinations of these mutations in SpCas9 could maximize on-target editing efficiency and enhance compatibility with gRNAs, while minimizing undesirable off-target activity.

We modularized the SpCas9 sequence into four parts (P1, P2, P3 and P4) for building combination mutants, and created 4 inserts for P1, 2 inserts for P2, 17 inserts for P3 and 7 inserts for P4. Each of the inserts was amplified and mutated from pAWp30 (Addgene, 73857) or eSpCas9(1.1) (Addgene, 71814) by PCR using Phusion (New England Biolabs) or Kapa HiFi (Kapa Biosystems) DNA polymerase. To generate site-directed mutations at amino acid positions 923, 924 and 926 of SpCas9, the three original codon sequences were replaced with the degenerate codon NNS in the PCR primer. An 8-base-pair barcode unique to each DNA insert was added after cloning into the storage vector (pAWp61 or pAWp62). BsaI restriction enzyme sites were added to flank the ends, and BbsI sites and a primer-binding site for barcode sequencing were introduced in between the insert and the barcode for pAWp61 and pAWp62, respectively. Each pAWp61 and pAWp62 storage vector was thus configured as 'BsaI–insert–BbsI–BbsI–barcode–BsaI' and 'BsaI–insert–primer-binding site–barcode–BsaI', respectively. Sanger sequencing was performed to confirm the sequence identity of individual inserts and their barcodes. In cases where the engineered sequence of interest contained BsaI or BbsI sites, other type IIS restriction enzyme sites could be used instead of BsaI and BbsI, or synonymous mutations could be introduced to the protein-coding sequence to remove the restriction sites while encoding the same amino acid residues.

Creation of a barcoded combination mutant library for SpCas9. Storage vectors harboring the inserts for each part of SpCas9 were mixed at an equal molar ratio. Pooled inserts were generated by single-pot digestion reactions of the mixed storage vectors with BsaI. The destination vector (pAWp60) was digested with BbsI. The digested P1 inserts and vectors were ligated to create a pooled P1 library in the destination vector. The P1 library was digested again with BbsI and ligated with the digested P2 inserts to assemble the library with two-way combinations (P1×P2). Sequential rounds of ligation reactions were performed to generate the three-way (P1×P2×P3) and four-way (P1×P2×P3×P4) combination libraries. After the pooled assembly steps, the protein-coding parts of the inserts were seamlessly linked and localized to one end of the vector construct and their respective barcodes were concatenated at the other end. We built a four-way (4×2×17×7) combination library of 952 SpCas9 variants, each carrying one to eight mutations (except for WT) at amino acid residues that were predicted to interact with the target and non-target DNA strand of the gRNA-directed genomic site[3,4] or alter the conformational dynamics of the SpCas9 nuclease domains[28] (Fig. 1a). The combinatorial complexity could be expanded by introducing additional barcoded parts and scaled up to simultaneously study tens of thousands or even more combinatorial modifications. Sanger sequencing analysis was performed, and a majority of the assembled barcoded combination mutant constructs were verified to carry the expected mutations in the two-way (20 of 20 colonies), three-way (14 of 15 colonies) and four-way (8 of 8 colonies) libraries. Except for one three-way combination mutant construct that carried an unintended base substitution, no other random mutation was detected in the constructs. The final library was subcloned into the pFUGW lentiviral vector to express the SpCas9 variants together with the selection marker zeocin from an elongation factor 1 alpha short (EFS) promoter. Sanger sequencing of the full-length sequence of the barcoded SpCas9 variants assembled in the lentiviral vector (7 of 7 colonies sampled from the library) confirmed that only expected mutations, and no random mutations, were present.

Generation of SpCas9 variants for individual validation. Lentiviral vectors encoding individual SpCas9 variants, including Opti-SpCas9, were constructed with the same strategy that was used for the generation of the combination mutant library described above, except that the assembly was performed one by one with individual inserts and vectors.

Human cell culture. HEK293T cells were obtained from the American Type Culture Collection (ATCC). OVCAR8-ADR cells were a gift from T. Ochiya (Japanese National Cancer Center Research Institute)[41]. The identity of the OVCAR8-ADR cells was confirmed by a cell line authentication test (Genetica DNA Laboratories). Monoclonal stable OVCAR8-ADR cell lines were generated by transducing cells with lentiviruses containing *RFP* and *GFP* genes expressed from ubiquitin-C (UBC) and cytomegalovirus (CMV) promoters, respectively, and a tandem U6-promoter-driven expression cassette of a gRNA targeting sites in the *RFP* gene. RFPsg5-ON, RFPsg8-ON and RFP-sg6-ON lines harbor target sites in the *RFP* gene that completely match the gRNA spacer, while the RFPsg5-OFF5-2, RFPsg8-OFF5 and RFPsg5-OFF5 lines harbor target sites in the *RFP* gene carrying synonymous mutations and are mismatched to the gRNA spacer (Supplementary Table 6). HEK293T cells were cultured in DMEM supplemented with 10% heat-inactivated FBS and 1× antibiotic–antimycotic (Life Technologies) at 37 °C with 5% CO_2. OVCAR8-ADR cells were cultured in RPMI supplemented with 10% heat-inactivated FBS and 1× antibiotic–antimycotic (Life Technologies) at 37 °C with 5% CO_2.

Lentivirus production and transduction. Lentiviruses were produced in six-well plates with 2.5×10^5 HEK293T cells per well. Cells were transfected using FuGENE HD transfection reagents (Promega) with 0.5 μg of lentiviral vector, 1 μg of pCMV-dR8.2-dvpr vector and 0.5 μg of pCMV-VSV-G vector mixed in 100 μl of OptiMEM medium (Life Technologies) for 15 min. The medium was replaced with fresh culture medium 1 d after transfection. Viral supernatants were then collected every 24 h between 48 and 96 h after transfection, pooled together and filtered through a 0.45-μm polyethersulfone membrane. For transduction with individual vector constructs, 500 μl of filtered viral supernatant was used to infect 2.5×10^5 cells in the presence of 8 μg ml^{-1} polybrene (Sigma) overnight. For transduction of the pooled library into human cells (OVCAR8-ADR), lentivirus production was scaled up using the same experimental conditions. To produce a high-coverage library containing sufficient representation for most combinations, infection was carried out with a starting cell population containing ~300-fold more cells than the library size to be tested. Lentiviruses were titrated to a multiplicity of infection of ~0.3 to give an infection efficiency of ~30% in the presence of 8 μg ml^{-1} polybrene, such that the SpCas9 variant library was delivered at low copy numbers.

Cell sorting. Cell sorting was performed on a BD Influx cell sorter (BD Biosciences). Drop delay was determined using BD Accudrop beads. Cells were filtered through 70-μm nylon mesh filters before sorting through a 100-μm nozzle using 1.0 Drop Pure sorting mode. Cells were gated on GFP signal and sorted on the basis of fluorescence level of RFP into three bins (A, B and C) such that approximately 5% of cells in the population were collected in each bin encompassing cells with a lower RFP level. The percentage of cells in the population to be sorted into each bin could be adjusted to balance the trade-off between the representation of individual combinations in the sorted population and the sensitivity of detecting enrichment of variants between bins. About 0.2–0.3 million cells were collected for each sorted bin in each sample.

Sample preparation for barcode sequencing. For the combination mutant vector library, plasmid DNA was extracted from *E. coli* transformed with the vector library using the Plasmid Mini kit (Qiagen). For the human cell pools infected with the combination mutant library, genomic DNA of cells collected from various experimental conditions was extracted using the DNeasy Blood and Tissue kit (Qiagen). DNA concentrations were measured by Quant-iT PicoGreen dsDNA Assay kit (Life Technologies). PCR amplification of 393-base-pair fragments, each containing a unique barcode representing an individual combination mutant, Illumina anchor sequences and an 8-base-pair indexing barcode for multiplexed sequencing, was performed using Kapa HiFi Hotstart Ready-mix (Kapa Biosystems). The forward and reverse primers used were 5′-AATGATACGGCGACCACCGAGATCTACACGGAACCGCAACGGTATTC-3′ and 5′-CAAGCAGAAGACGGCATACGAGATNNNNNNNNGGTTGCGTCA GCAAACACAG-3′, where NNNNNNNN denotes a specific indexing barcode assigned for each experimental sample. To avoid bias in PCR that could skew the population distribution, we optimized PCR conditions to ensure amplification occurred during the exponential phase. The PCR amplicons were purified with two rounds of size selection using a 1:0.5 and 1:0.95 ratio of Agencourt AMPure XP beads (Beckman Coulter Genomics) before real-time PCR quantification using Kapa SYBR Fast qPCR Master Mix (Kapa Biosystems) with a StepOnePlus Real Time PCR system (Applied Biosystems). The forward and reverse primers used for quantitative PCR were 5′-AATGATACGGCGACCACCGA-3′ and 5′-CAAGCAGAAGACGGCATACGA-3′, respectively. The quantified samples were then pooled at the desired ratio for multiplexing, assessed using the high-sensitivity DNA chip (Agilent) on an Agilent 2100 Bioanalyzer and run on an Illumina HiSeq using a primer (5′-CCACCGAGATCTACACGGAACCGCAACGGTATTC-3′) and indexing barcode primer (5′-GTGGCGTGGTGTGCACTGTGTTTGCTGACGCAACC-3′).

Barcode sequencing data analysis. Barcode reads for each combination mutant were processed from sequencing data. Barcode reads representing each combination were normalized per million reads for each sample categorized by the indexing barcodes. Profiling was performed in two biological replicates. We measured the frequency of each combination mutant between sorted bin A and the unsorted population, and calculated the enrichment ratio (E) between them relative to the rest of the population. Bin A was selected because enrichment of variants was most obvious in this bin (Fig. 2b). The following equation was used to calculate the enrichment ratio:

$$E = \frac{(N_{\text{bin}}/N_{\text{unsorted}})}{(1-N_{\text{bin}})/(1-N_{\text{unsorted}})}$$

where N_{bin} represents the frequency of the combination mutant in the sorted bin and N_{unsorted} represents the frequency of the combination mutant in the unsorted bin.

The $\log_2$-transformed mean score determined from the replicates (that is, $\log_2(E)$) comparing the sorted bin A against the unsorted population was used as a measure of target editing activity. Only barcodes that gave more than 300 absolute reads in the unsorted population were analyzed to improve data reliability. The correlation between $\log_2(E)$ score determined from the pooled screen and individual validation data (Supplementary Fig. 4) could be improved by increasing the fold representation of cells per combination in the pooled screen to reduce the experimental noise[42]. We defined activity-optimized variants (Opti-SpCas9 identified in this study) as those with $\log_2(E)$ (for bin A versus the unsorted population) that were at least 90% of WT for both RFPsg5-ON and RFPsg8-ON, and less than 60% of WT for both RFPsg5-OFF5-2 and RFPsg8-OFF5. OptiHF-SpCas9 was identified as a variant with high fidelity on the basis of enrichment ratios of at least 50% of WT for both RFPsg5-ON and RFPsg8-ON, and less than 90% of WT for both RFPsg5-OFF5-2 and RFPsg8-OFF5. The full list is presented in Supplementary Table 2.

To determine epistasis, we applied a scoring system similar to ones previously described for protein fitness[43,44], and calculated epistasis (ε) scores for each combination in Fig. 4. The epistasis scores were determined as observed fitness − expected fitness, where the expected fitness for the combination $[X,Y]$ is $(\log_2(E_{[X]}) + \log_2(E_{[Y]}))$ according to the additive model. In general terms, combinations that exhibited better fitness than predicted were defined as positive epistasis, whereas combinations that were less fit than expected were defined as negative epistasis. In this work, the $\log_2(E)$ values for a lethal or nearly lethal combination mutant were set equal to those for a SpCas9 variant with eight mutations (R661A, Q695A, K848A, E923M, T924V, Q926A, K1003A and R1060A) for comparison, and our individual validation data confirmed its minimal activity in disrupting the target *RFP* sequences (Fig. 3b). The expected fitness was capped at the $\log_2(E)$ values for a lethal or nearly lethal combination mutant to minimize spurious epistasis values resulting from meaningless predicted fitness. In future work, it could be beneficial to include a nuclease-dead mutant of SpCas9 in the pooled screens as a lethal mutant for comparison.

Fluorescent protein disruption assay. Fluorescent protein disruption assays were performed to evaluate DNA cleavage and indel-mediated disruption at the target site of the fluorescent protein (GFP or RFP) produced by expression of SpCas9 and gRNA, which results in loss of cell fluorescence. Cells harboring an integrated *GFP* or *RFP* reporter gene together with SpCas9 and a gRNA were washed and resuspended with 1× PBS supplemented with 2% heat-inactivated FBS, and assayed with an LSR Fortessa analyzer (Becton Dickinson). Cells were gated on forward and side scatter. At least 1×10^4 cells were recorded per sample in each dataset.

Immunoblot analysis. Cells were lysed in 2× RIPA buffer supplemented with protease inhibitors (Gold Biotechnology, GB-108-2). Lysates were collected by scraping of the culture plate on ice, and then centrifuged at 15,000 r.p.m. for 15 min at 4 °C. Supernatants were quantified using the Bradford assay (Bio-Rad). Protein was denatured at 99 °C for 5 min before gel electrophoresis on a 10% polyacrylamide gel (Bio-Rad). Proteins were transferred to polyvinylidene difluoride membranes at 110 V for 2 h at 4 °C. The primary antibodies used were as follows: anti-Cas9 (7A9-3A3) (1:2,000; Cell Signaling, 14697) and anti-β-actin (1:10,000; Sigma, A2228). The secondary antibody used was horseradish peroxidase (HRP)-linked anti-mouse IgG (1:20,000; Cell Signaling, 7076). Membranes were developed by WesternBright ECL HRP substrate (Advansta, K-12045-D20).

T7 endonuclease I assay. A T7 endonuclease I assay was carried out to evaluate DNA mismatch cleavage at genomic loci targeted by the gRNAs. Genomic DNA was extracted from cell cultures using QuickExtract DNA extraction solution (Epicentre) or the DNeasy Blood and Tissue kit (Qiagen). Amplicons harboring the targeted loci were generated by PCR using the primers and PCR conditions listed in Supplementary Table 7, and then purified using Agencourt AMPure XP beads (Beckman Coulter Genomics). About 400 ng of the PCR amplicons was denatured, self-annealed and incubated with 4 U of T7 endonuclease I (New England Biolabs) at 37 °C for approximately 40 min. The reaction products were resolved by 2% agarose gel electrophoresis. Quantification was based on relative band intensities measured using ImageJ. Indel percentage was estimated by the formula $100 \times (1 - (1 - (b + c)/(a + b + c))^{1/2})$ as previously described[45], where a is the integrated intensity of the uncleaved PCR product, and b and c are the integrated intensities of each cleavage product.

GUIDE-seq detection of genome-wide off-target sites. Genome-wide off-target sites were accessed using the GUIDE-seq method[46]. For each GUIDE-seq sample, 1.5 million OVCAR8-ADR cells infected with SpCas9 variants and gRNAs were electroporated with 1,000 pmol freshly annealed GUIDE-seq end-protected double-stranded oligodeoxynucleotide (dsODN) using 100-µl Neon tips (Thermo Fisher Scientific) according to the manufacturer's protocol. The dsODN oligonucleotide sequences used were 5′-P-G*T*TTAATTGAGTTGTCATATGTTAATAACGGT*A*T-3′ and 5′-P-A*T*ACCGTTATTAACATATGACAACTCAATTAA*A*C-3′, where P represents 5′ phosphorylation and asterisks indicate a phosphorothioate linkage. Genomic DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen) 72 h after electroporation. Genomic DNA concentration was quantified by Qubit fluorometer double-stranded DNA HS assay (Thermo Fisher Scientific), and 400 ng was used for library construction following the GUIDE-seq protocol with minor modifications. In brief, DNA was enzymatically fragmented by KAPA Frag kit (KAPA Biosystems), followed by adaptor ligation and two rounds of heminested PCR enrichment for dsODN integration sequences. To unify Illumina sequencing workflows for obtaining dual-indexed data using a single-indexed sequencing workflow across various Illumina platforms, we redesigned the half-functional adaptors with the sample index (index 2) placed at the head of read 1, following the unique molecular index (Supplementary Table 8). Final sequencing libraries were quantified by KAPA Library Quantification Kits for Illumina and sequenced on an Illumina NextSeq 500 System. Data demultiplexing of index 1 was performed by bcl2fq v.2.19, followed by custom scripts for index 2 demultiplexing and formatting for analysis using the GUIDE-seq software[47].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Source data for the count matrices determined for SpCas9 variants on the basis of pooled characterization that are shown in Fig. 3 are provided with the online version of this paper. GUIDE-seq data are available from the European Nucleotide Archive under accession PRJEB32521.

Code availability

The custom scripts for data analysis are available at https://github.com/AWHKU/BC-analyzer.

References

41. Honma, K. et al. *RPN2* gene confers docetaxel resistance in breast cancer. *Nat. Med.* **14**, 939–948 (2008).
42. Kampmann, M., Bassik, M. C. & Weissman, J. S. Functional genomics platform for pooled screening and generation of mammalian genetic interaction maps. *Nat. Protoc.* **9**, 1825–1847 (2014).

43. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
44. Aakre, C. D. et al. Evolving new protein–protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
45. Guschin, D. Y. et al. A rapid and general assay for monitoring endogenous gene modification. *Methods Mol. Biol.* **649**, 247–256 (2010).
46. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
47. Tsai, S. Q., Topkar, V. V., Joung, J. K. & Aryee, M. J. Open-source guideseq software for analysis of GUIDE-seq data. *Nat. Biotechnol.* **34**, 483 (2016).

nature research

Corresponding author(s): Alan S. L. Wong

Last updated by author(s): May 21, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	High-throughput sequencing data were collected using Illumina HiSeq2000/2500 with HCS 2.2.68/RTA 1.18.66.3 software. Flow cytometry data were collected using the BD FACSDiva™ software.
Data analysis	A custom python code was used to retrieve and count barcode reads for each combination mutant from the Illumina sequencing data, and normalized them to per million reads for each sample categorized by the indexing barcodes. MATLAB R2018b was used to perform Student's t-test and calculate adjusted P-values. bcl2fq v2.19 and custom scripts were used for de-multiplexing of Index 1 and Index 2 of GUIDE-Seq samples, respectively. The GUIDE-Seq software was used for genome-wide off-targets analysis. FlowJo version 10.5.3 was used to analyze data generated from flow cytometry experiments.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The count matrices determined for SpCas9 variants based on the pooled characterization are provided in Source Data. GUIDE-Seq data is available on ENA via accession PRJEB32521.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen due to being able to show reproducibility and statistical significance. No methods were used to predetermine sample size. >300-fold more cells for lentiviral infection than the size of library being tested in genetic screens were used to ensure high fold-representation.
Data exclusions	It was previously reported that filtering of the genetic screen data to remove library members with low representation in the reference set resulted in a reduced false negative rate (Sim et al., Genome Biol. 2011; 12(10): R104). Yet, the exclusion criteria has not been standardized and thus were not pre-established. For Illumina sequencing data from the genetic screens in this study, only barcodes that gave more than 300 absolute reads in the unsorted population were analyzed to improve data reliability.
Replication	All data was reliably reproduced. Methods and materials used in our experiments were described in the manuscripts to facilitate replication of our studies. Transduction of the library of constructs into human cells was performed independently to produce biological replicates. Infected cell pools were sorted into bins independently to produce biological replicates for genomic DNA extraction for barcode sequencing. All biological replicates were analyzed independently and replicate numbers are provided in the text and figure legends.
Randomization	No randomization was used for samples as samples with particular genetic constituents were needed for the experiments. During construction of mutant library, cell culture, transfection, infection, cell sorting, sample preparation for barcode sequencing, sequencing, and data analysis, samples were not grouped in a way relating to the identity of the sample. Timing of when samples were ready determined the grouping of samples in sequencing runs.
Blinding	Blinding was not relevant to the studies as samples with particular genetic constituents were needed for the experiments. Labeling of samples was used to prevent mixed up of experimental samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Anti-Cas9: https://www.cellsignal.com/products/primary-antibodies/cas9-7a9-3a3-mouse-mab/14697
Supplier name: Cell Signaling Technology
Catalog number: #14697
Clone name: 7A9-3A3
Lot number: NA
Dilutin: 1:2000

Monoclonal anti-beta actin antibody
Supplier name: Sigma-Aldrich
Catalog number: A5316
Clone name: AC-74
Lot number: NA
Dilution: 1:10000

Anti-mouse IgG, HRP-linked antibody
Supplier name: Cell Signaling Technology
Catalog number: 7076
Clone name: NA
Lot number: 34
Dilution: 1:20000

Validation

Cas9 (7A9-3A3) mouse monoclonal antibody recognizes transfected levels of total Cas9 protein. The use of this antibody is recommended for:
• Western blotting
• Immunohistochemistry (paraffin)
• Immunofluorescence
• Flow cytomertry
(https://media.cellsignal.com/pdf/14697.pdf)

Monoclonal anti-beta actin antibody recognizes an epitope located on the N-terminal end of the β-isoform of actin. It specifically labels β-actin in a wide variety of tissues and species. The use of this antibody is recommended for Western blotting.
Species reactivity: human and etc.
(https://www.sigmaaldrich.com/catalog/product/sigma/a5316?lang=en&region=HK)

The use of anti-mouse IgG, HRP-linked antibody is recommended for Western blotting. This antibody is validated by the commercial vendor.
(https://www.cellsignal.com/products/secondary-antibodies/anti-mouse-igg-hrp-linked-antibody/7076).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293T cells were obtained from American Type Culture Collection (ATCC). OVCAR8-ADR cells were gifts from T. Ochiya (Japanese National Cancer Center Research Institute, Japan)
Authentication	The identity of the OVCAR8-ADR cells was authenticated by STR profiling (Genetica DNA Laboratories). HEK293T cells were authenticated by STR profiling by the commercial vendor.
Mycoplasma contamination	Mycoplasma contamination was tested and confirmed to be negative. All cell culture medium was supplemented with antibiotic-antimycotic solution to prevent bacterial and fungal contamination.
Commonly misidentified lines (See ICLAC register)	The identity of the OVCAR8-ADR cells was confirmed by a cell line authentication test (Genetica DNA Laboratories), and is not the misidentified MCF-7/AdrR (NCI/ADR-RES) cell line. No other commonly misidentified cell lines were used (e.g., HEK293T cells are not included in the ICLAC register).

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cell cultures were treated with trypsin and diluted in complete media or PBS for flow cytometry experiments.
Instrument	BD LSRFortessaTM was used for data collection. Cell sorting was performed on a BD Influx cell sorter.
Software	All cytometry data were analyzed by FlowJo.
Cell population abundance	Drop delay was determined using BD Accudrop beads. Cells were filtered through 70μm nylon mesh filters before sorting through a 100-μm nozzle using 1.0 Drop Pure sorting mode. About 0.2 - 0.3 million cells were collected for each sorted bin in each sample. Details are described in Methods Section.
Gating strategy	Viable and intact cells were gated from FSC/SSC for analysis. Within the population, infected cells that were selected for downstream analysis by gating cells expressing GFP. Details are described in Methods Section.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.