Published in final edited form as:

Nat Methods. 2017 August 31; 14(9): 831–832. doi:10.1038/nmeth.4423.

GUIDES: sgRNA design for loss-of-function screens

Joshua A. Meier[1,2,3,4], **Feng Zhang**[1,2], and **Neville E. Sanjana**[3,4,†]

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA

[2]McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]New York Genome Center, 101 Avenue of the Americas, New York, NY 10013

[4]Department of Biology, New York University, 100 Washington Square East, New York, NY 10003

Main text

Genome-scale CRISPR-Cas9 knockout libraries have emerged as powerful tools for unbiased, phenotypic screens[1]. These libraries contain a fixed number of Cas9 single-guide RNAs (sgRNAs) targeting each gene in the genome and typically require large numbers of cells (>10^8) to maintain genome-scale representation. However, there are many applications where it would be preferable to design a custom library targeting specific genes sets (e.g. kinases, transcription factors, chromatin modifiers, the druggable genome) with higher coverage for these specific genes. To address this need, we developed Graphical User Interface for DNA Editing Screens (GUIDES), a web application that designs CRISPR knock-out libraries to target custom subsets of genes in the human or mouse genome (available at http://guides.sanjanalab.org/ and https://github.com/sanjanalab/GUIDES).

After providing a list of genes (as gene symbols, Ensembl IDs, or Entrez IDs), GUIDES creates a library with multiple sgRNAs to target each gene (Fig. 1). To pick optimal sgRNAs, GUIDES integrates tissue-specific RNA expression, protein structure prediction, Cas9 off-target prediction/avoidance, and Cas9 on-target local sequence preferences in an integrated multi-stage pipeline.

For each gene, GUIDES first identifies coding regions using the Consensus CoDing Sequence (CCDS) database. For the human genome, GUIDES can use tissue-specific RNA-sequencing gene expression data from the GTEx Consortium (v6, 8,555 tissue samples from

[†]Correspondence should be addressed to: neville@sanjanalab.org (N.E.S).

Author contributions

N.E.S. conceived of the library design tool. J.A.M. wrote the code and performed the experiments. N.E.S. and J.A.M. analyzed the experiments. J.A.M., F.Z., and N.E.S. wrote the manuscript. N.E.S. and F.Z. supervised the work.

Competing financial interests

A patent has been filed relating to the described work. F.Z. is a founder and scientific advisor for Editas Medicine, and a scientific advisor for Horizon Discovery.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

544 donors) to target Cas9 preferentially to exons with higher expression[2] (Supplementary Fig. 1). Targeting sgRNAs to exons constitutively expressed in the target cell type/tissue can be important, as mutations in alternatively-spliced exons may not result in protein knock-out[3].

For each exon, GUIDES first prioritizes potential Cas9 target sites using the established cutting frequency determination (CFD) score[4]. For each 20 bp target site, GUIDES identifies all sequences with 1, 2 or 3 base mismatches present in the exome and assigns an off-target score by summing individual CFD scores. Target sites with perfect matches elsewhere in the exome are selected only in cases where no other sgRNAs exist to target the gene. By using exome-wide CFD scoring during design of a library with ~2,000 genes, the percentage of designed sgRNAs with predicted off-targets decreases from ~43% to ~4% (Supplementary Fig. 2).

Saturation mutagenesis screens tiling over entire genes have shown increased knock-out efficiency when targeting protein functional domains[5], presumably due to in-frame mutations in regions tolerant to mutations. To take advantage of this, GUIDES includes an option to preferentially choose sgRNAs that target functional protein domains identified in the Protein Family (Pfam) database (v30, 16,306 protein families)[6]. This can have a significant impact on library design since 90% of protein-coding genes in the human genome contain at least one Pfam-annotated domain[6].

After identifying exons and protein domains to target, GUIDES uses a previously validated boosted regression tree classifier to score Cas9 target sites based on local sequence preferences learned from saturation mutagenesis screens and adds the highest-scoring sgRNAs to the library[4]. When targeting the same sets of genes, sgRNAs designed with this criterion have a ~30% higher on-target efficiency score (Supplementary Fig. 3). Other CRISPR library design tools have also used on-target efficiency scoring to help automate sgRNA design, but do not include RNA expression or protein domain identification to sgRNA targeting (Supplementary Table 1). Some existing library design tools use a command-line interface, whereas GUIDES is a graphical, web-based tool that allows fine-tuning of sgRNA selection directly in the web browser (Fig. 1).

To benchmark the performance of GUIDES-selected sgRNAs in genome-scale screens, we tested whether sgRNAs designed by GUIDES have consistently higher/lower activity using a meta-analysis of 77 pooled CRISPR screens from the GenomeCRISPR database[7]. By examining sgRNAs targeting essential genes, we found that GUIDES-generated sgRNAs were more depleted by approximately one 10%-quantile (with sgRNAs given a percentage rank within each pooled screen) than a size-matched control set of sgRNAs targeting the same gene (Supplementary Fig. 4) (n = 403 genes with 8 ± 6 sgRNAs per gene, $p = 5\text{x}10^{-7}$, t = -5.1, df = 409, two-sample paired t-test).

GUIDES also manages several practical aspects of library design, including eliminating sgRNAs with homopolymer repeats that are difficult to synthesize, alerting the user when the sgRNA targets the last exon which may escape nonsense-mediate decay of mRNA[8], eliminating sgRNAs with Pol3 transcriptional terminators, creating synthesis-ready

oligonucleotides with flanking sequences for PCR-based cloning and adding in non-targeting sgRNAs for calculating false-discovery rates in pooled CRISPR screens. By taking advantage of several algorithmic optimizations, run time is linear with respect to gene count (Supplementary Fig. 5). For example, GUIDES takes ~15 seconds to design a library targeting 500 genes involved in chromatin regulation with 6 sgRNAs per gene (Intel i7 3Ghz, 16 GB RAM).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank B. Cummings for help with GTEx data processing and the entire Sanjana laboratory for support and advice. F.Z. is supported by the NIH through NIMH (5DP1-MH100706 and 1R01-MH110049); NSF; the New York Stem Cell Foundation; the Allen Distinguished Investigator Program, through The Paul G. Allen Frontiers Group; the Simons and Vallee Foundations; the Howard Hughes Medical Institute; the Skoltech-MIT Next Generation Program; James and Patricia Poitras and the Poitras Center for Affective Disorders; Robert Metcalfe; and David Cheng. F.Z. is a New York Stem Cell Foundation-Robertson Investigator. N.E.S. is supported by the NIH through NHGRI (R00-HG008171) and a Sidney Kimmel Scholar Award.

References

1. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015; 16:299–311. [PubMed: 25854182]
2. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
3. Murphy D, Singh R, Kolandaivelu S, Ramamurthy V, Stoilov P. Alternative Splicing Shapes the Phenotype of a Mutation in BBS8 To Cause Nonsyndromic Retinitis Pigmentosa. Mol Cell Biol. 2015; 35:1860–1870. [PubMed: 25776555]
4. Doench JG, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016; 34:184–191. [PubMed: 26780180]
5. Shi J, et al. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol. 2015; 33:661–667. [PubMed: 25961408]
6. Mistry J, et al. The challenge of increasing Pfam coverage of the human proteome. Database J Biol Databases Curation. 2013; 2013:bat023.
7. 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–D686. [PubMed: 27789686]
8. Popp MWL, Maquat LE. Organizing principles of mammalian nonsense-mediated mRNA decay. Annu Rev Genet. 2013; 47:139–165. [PubMed: 24274751]

Figure 1. GUIDES design environment

Screenshot of interactive designer for adding and deleting genes and sgRNAs.