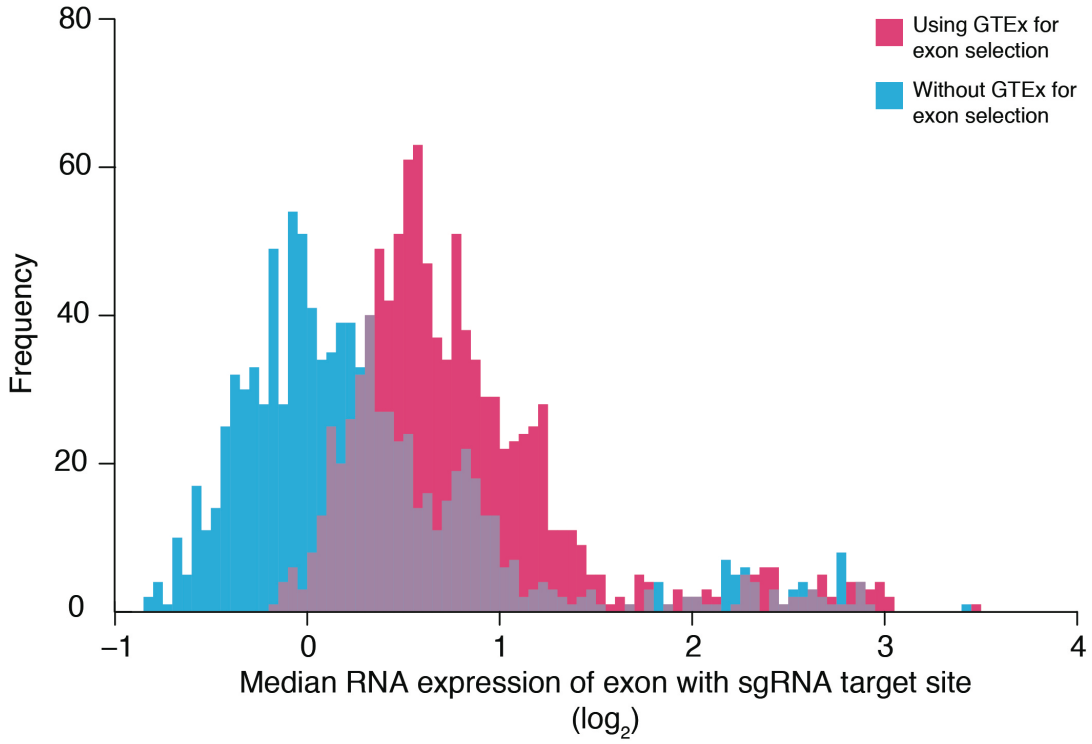
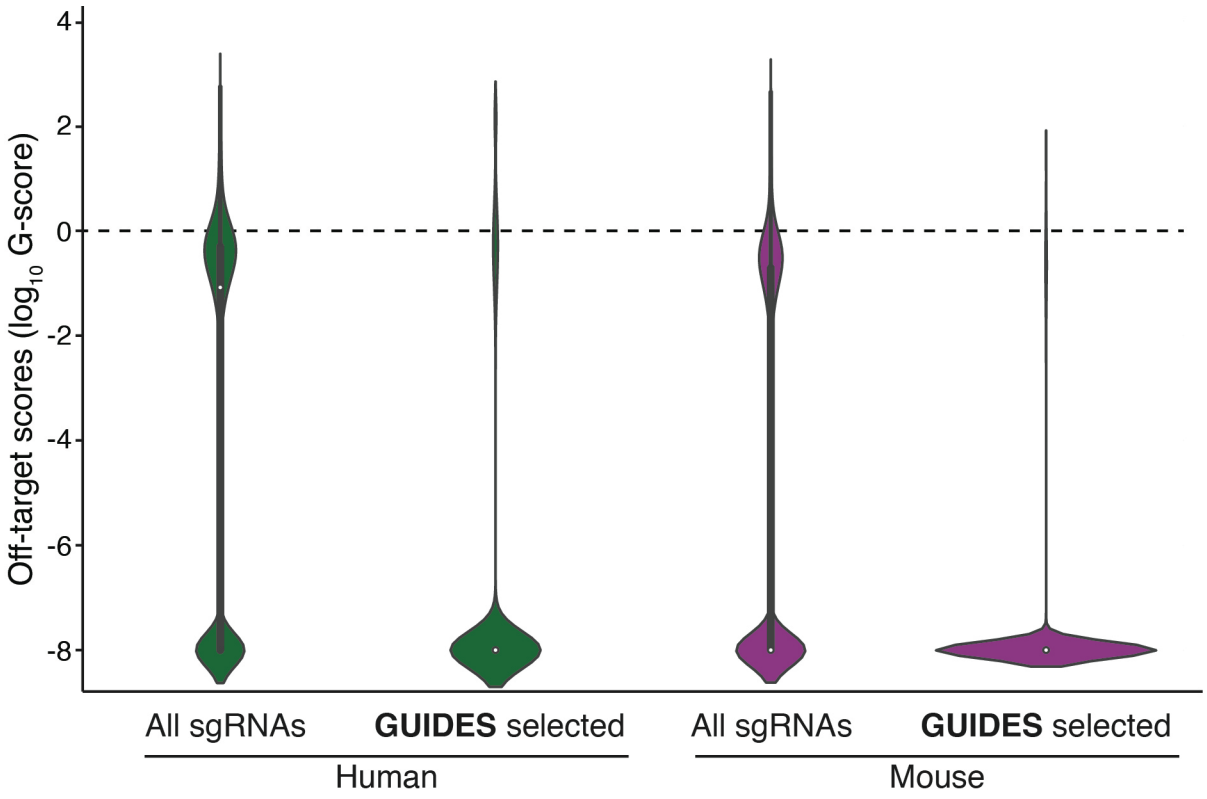


Supplementary Figure 1

Integration of GTEx data shifts sgRNA targeting toward highly expressed exons.

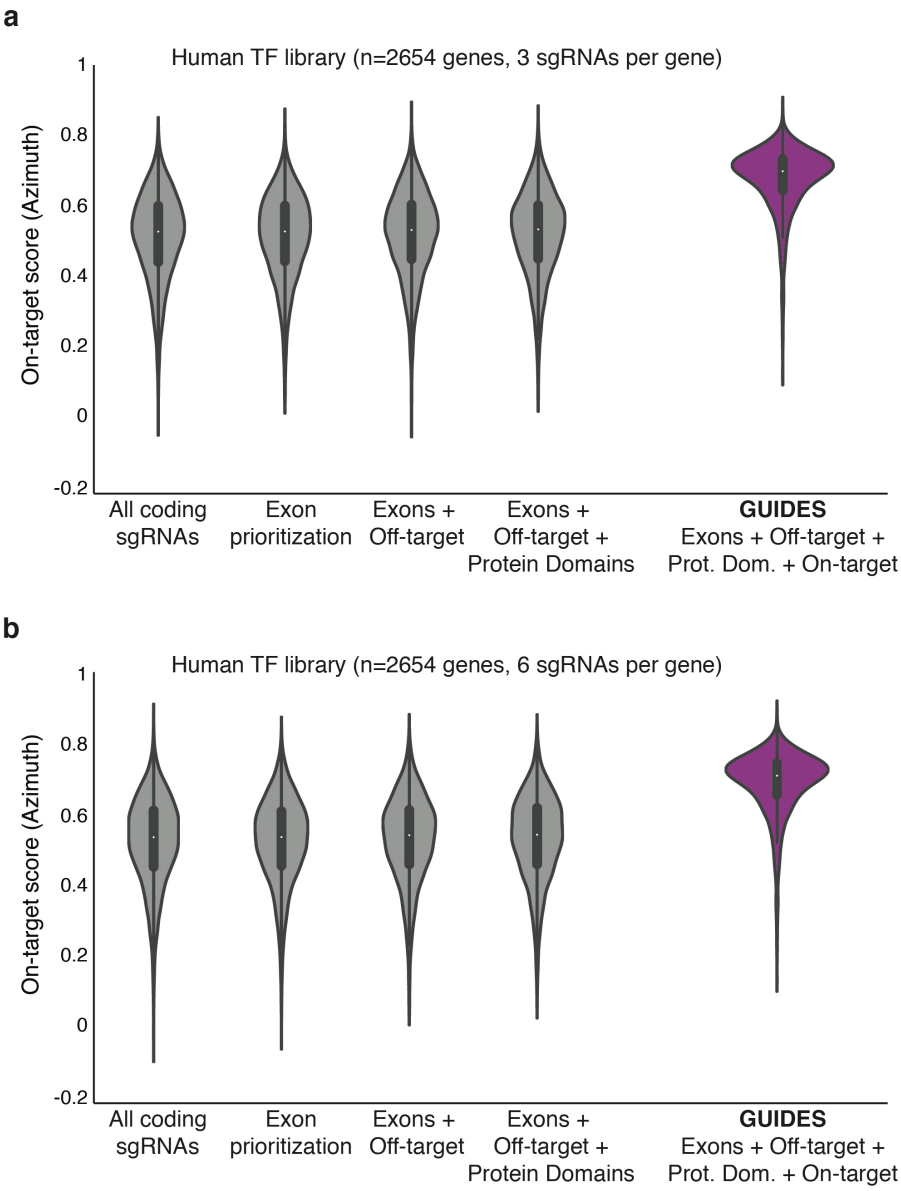
GUIDES-generated libraries using data for median expression across all GTEx tissues (red) or without using GTEx data (blue) to choose exons to target. For each library, 500 genes were selected at random (n = 1000 randomized libraries) from the human genome and GUIDES was instructed to design 5 sgRNAs per gene in the selected exons. On average, incorporation of GTEx data increases average expression of targeted exons by a factor of 1.5.



Supplementary Figure 2

Aggregate cut-frequency determination (CFD) off-target score reduces predicted off-target sites for selected sgRNAs.

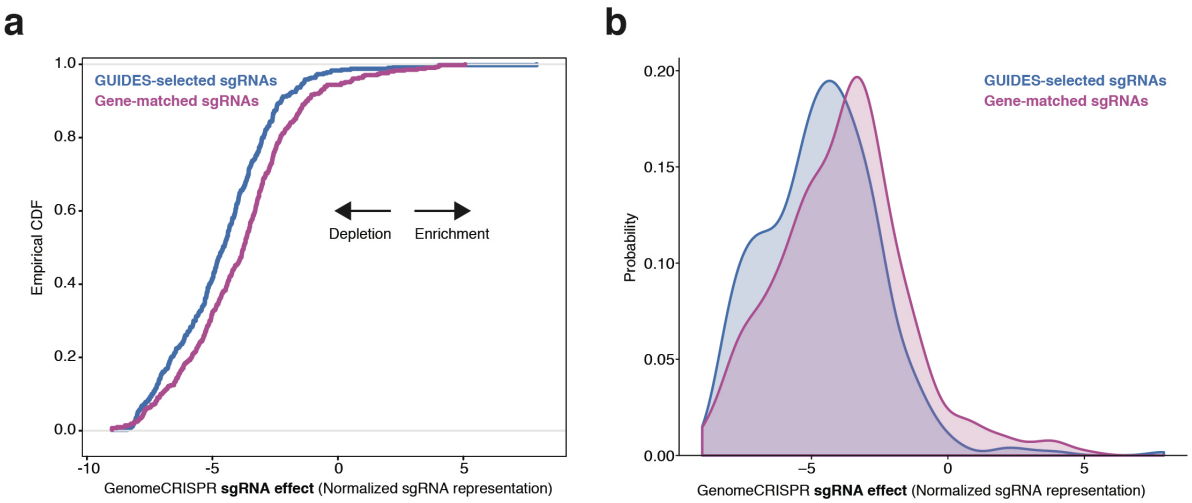
GUIDES calculates the sum of the cut-frequency determination (CFD) score[4] for all 0-3 bp mismatches in the human/mouse exome ("G-score"). When optimizing sgRNAs for off-target avoidance/specificity, designed human and mouse libraries (n = 2,000 genes) have fewer sgRNAs with a high G-score (i.e. fewer sgRNAs with 0-3 bp potential exome off-targets).



Supplementary Figure 3

On-target scores at each stage of the GUIDES pipeline.

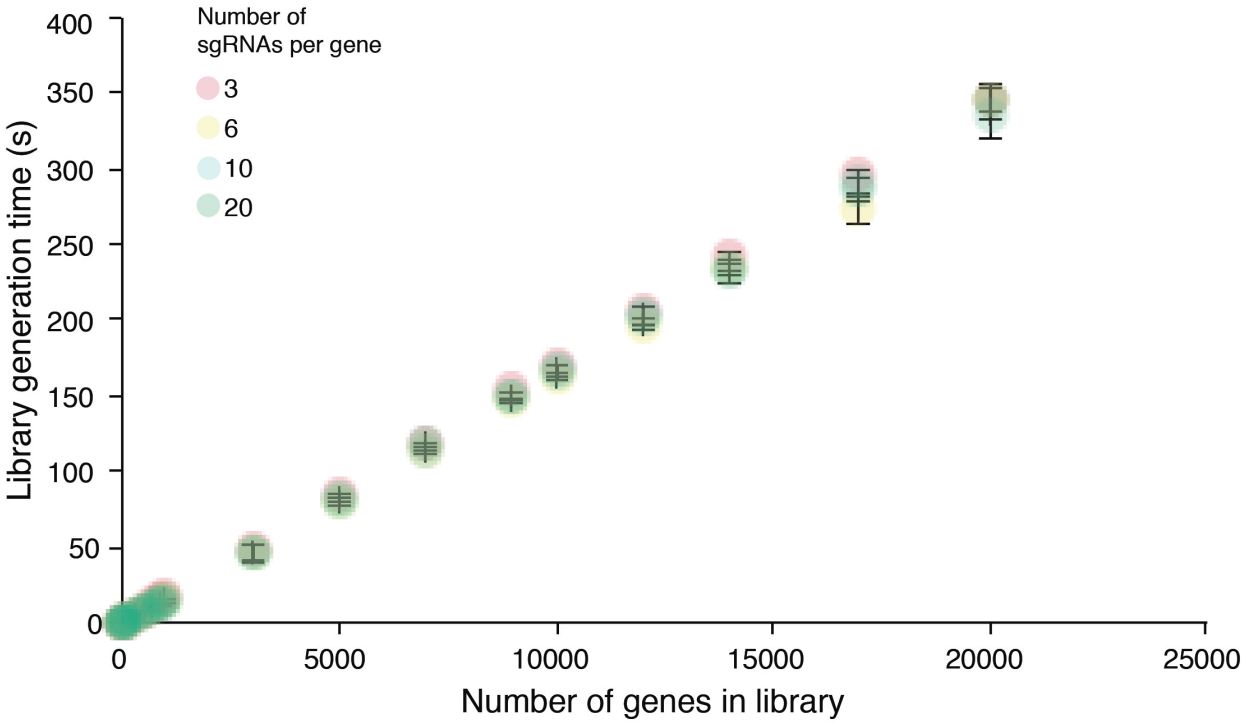
Average on-target scores for a sgRNA library targeting 2,654 transcription factors in the human genome with 3 sgRNAs per gene (**a**) or 6 sgRNAs per gene (**b**). On-target (efficiency) scores were calculated using the Microsoft Azimuth algorithm as in ref. 4. Despite on-target optimization being the last stage of the GUIDES pipeline (see *Supplementary methods)*, on-target score optimization increases the average on-target score of chosen sgRNAs even after other optimizations such as off-target- and protein domain-based prioritization.



Supplementary Figure 4

GUIDES-selected sgRNAs targeted essential genes display greater depletion in a meta-analysis of 77 pooled CRISPR screens.

(**a**) Cumulative density function of GenomeCRISPR sgRNA effect scores for GUIDES-selected sgRNAs versus a matched-size sample of sgRNAs targeting the same genes randomly chosen from the GenomeCRISPR database. (**b**) Probability density function for the per-gene data shown in (a). The average increase in depletion by using GUIDES-generated sgRNAs over the size-matched randomly selected sets was 0.73 sgRNA effect (~10% increased depletion, *n* = 403 genes examined in 77 genome-scale screens using 61 different cell lines), which is significantly greater depletion (*p* = 5e-07, *t* = -5.1, *df* = 409, two-sample paired t-test).



Supplementary Figure 5

Library generation time scales linearly with number of genes targeted.

The indicated number of genes was selected from the human genome and the time required for library generation was tracked as a function of the number of sgRNAs per gene. Gene count and generation time are linearly correlated (r^2>0.99 over a range of 5 - 20,000 genes targeted for 3, 6, 10, and 20 sgRNAs/gene, n = 100 library generation runs, error bars indicate s.d.). Since all potential sgRNAs for each gene are precomputed, generation times are not affected by the number of sgRNAs. For benchmarking, GUIDES was run on a computer with a 2.5 GHz Intel Core i7 processor and 16 GB of memory running Linux (Ubuntu 14).

Supplementary Table 1 | Comparison of CRISPR sgRNA library design tools.

Software tool	**GUIDES**	**GuideScan**	**CLD**	**E-CRISP**	**MIT CRISPR tool**	**CHOP-CHOP**	**CRISPR scan**
Website to run/download software	http://guides.sanjanalab.org/	http://www.guidescan.com/	https://github.com/boutroslab/cld	http://www.e-crisp.org/E-CRISP/	http://crispr.mit.edu/	https://chopchop.rc.fas.harvard.edu/	http://www.crisprscan.org/
Can design sgRNA libraries? (i.e. accepts multiple genes)	**Yes**	**Yes**	**Yes**	No	No	No	No
Can input gene names?	**Yes**	No	**Yes**	**Yes**	No	No	**Yes**
Prioritizes exons to target?	**Yes, by expression**	No	No	No	No	No	No
Targets protein functional domains?	**Yes**	No	No	No	No	No	No
Includes control sgRNAs in library?	**Yes**	No	No	No	No	No	No
Graphical user interface (GUI)	**Yes**	**Yes**	No	**Yes**	**Yes**	**Yes**	**Yes**
sgRNA library selection in GUI	**Yes**	No	No	No	No	No	No
Runtime for library design (500 genes, human genome)	**15 sec**	3 min[1]	90 min[2]	n/a	n/a	n/a	n/a
Reference	This manuscript	*Perez et al. (2017)*	*Heigwer et al. (2016)*	*Heigwer et al. (2014)*	*Hsu et al. (2013)*	*Montague et al. (2014)*	*Moreno-Mateos et al. (2015)*

[1] Requires manual conversion of gene list to genomic coordinates. These steps were not included in the runtime cited.
[2] Requires database setup and computationally intensive initial processing step.

Supplementary Methods

GUIDES algorithms and workflow

We implemented several algorithmic optimizations to generate large (e.g. genome-scale) libraries without excessive user wait times. For search processes during initial server startup and during GUIDES library generation, we employed data structures which reduce search from linear to logarithmic time. Specific instances of these optimizations are described in detail below.

Data sources for reference genomes, gene expression and protein structure

Human and mouse genome sequences are from Ensembl (http://www.ensembl.org, human genome GRCh37, mouse genome GRCm38) and tissue-specific gene expression data for each exon was obtained from the Genotype-Tissue Expression (GTEx) Consortium (http://gtexportal.org/, v6 dataset)[1]. Pfam data was obtained from the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables, table ucscGenePfam for hg19 and mm10) [2]. Since GUIDES needs to query which protein domains exist at specific genomic loci, we transformed BED-format intervals into interval trees, which reduces search time for sgRNA overlap with protein domains from linear to logarithmic in the number of domains.

User input to generate a new GUIDES library (runtime)

In the web interface, a user provides the following parameters to the tool:

1. Target genome: ***Human*** or ***mouse***
2. Library complexity: ***m*** sgRNAs/gene
3. List of genes: Either ***gene symbols***, ***Ensembl Gene IDs,*** or ***Entrez IDs*** can be used
4. *Human libraries only:* Consider GTEx expression data? (***Yes***/***No***)
    a. If ***Yes***, the user can average over tissues or specify one or more tissues with gene expression data from the list given in the next paragraph
5. Consider Pfam protein domains? (***Yes***/***No***)

Design pipeline overview

In GUIDES, the exonic region of a gene is defined as the union of Consensus CoDing Sequence (CCDS) regions associated with the gene[3]. Specifically, the CCDS coordinates are used by GUIDES to lookup sequence data for each gene's coding exons (initiating ATG codon to stop codon). Upon initial provisioning, GUIDES discards all non-exonic genomic sequences and saves the sequences encoding each exon in separate files. Then, GUIDES iterates through each CDS exon and determines all Cas9-targetable sites by identifying the PAM sequence (`NGG` for *Sp*Cas9) on either DNA strand.

By default, the resulting guides are ranked first by off-target avoidance/specificity; second by presence/absence of a Pfam protein domain in the cutting region; and third by on-target efficiency. Thus, sgRNAs with low on-target efficiency scores that target a Pfam protein domain are given a higher rank than those with higher on-target efficiency scores that do not target a Pfam protein domain. For most genes, there are sufficiently many sgRNAs available without exonic off-targets, thus making it possible to avoid selection of sgRNAs with predicted off-target sites in the exome (Supplementary Fig. 2). Although on-target efficiency prioritization occurs at a later stage in the GUIDES pipeline (e.g. after off-target avoidance or Pfam domain), it plays a crucial role in sgRNA choice, resulting in chosen guides having a higher on-target score (Supplementary Fig. 3). Also, in the web interface, the user can re-rank the list of sgRNAs for a specific gene by any of these criteria (on-target, off-target, protein domain, exon number) to allow for additional flexibility in the design.

Selection of highly expressed exons using GTEx data

For GTEx data, GUIDES uses the `exon_reads` file[1], which contains RNA-sequencing expression values (as read counts): `GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_exon_reads.txt.gz`. GUIDES groups individual GTEx data samples (patient-tissue pairs) by tissue and computes the median expression value for each exon in each tissue (computed over all samples with the same GTEx `SMTS` tissue type). The median read counts are then normalized by exon size (base pairs) and then saved into a gene-specific file, where rows denote different exons (by Ensembl Gene ID and exon number) and columns denote different tissues (from the list in the previous paragraph).

GUIDES supports specific gene expression profiles for the following GTEx dataset tissues: Adipose Tissue, Adrenal Gland, Bladder, Blood, Blood Vessel, Bone Marrow, Brain, Breast, Cervix Uteri, Colon, Esophagus, Fallopian Tube, Heart, Kidney, Liver, Lung, Muscle, Nerve, Ovary, Pancreas, Pituitary, Prostate, Salivary Gland, Skin, Small Intestine, Spleen, Stomach, Testis, Thyroid, Uterus, Vagina. When the user generates a library, exon expression values are computed based on the user-selected subset of tissues. For each exon, the expression value displayed in the GUIDES interface is the median of the exon's expression across selected tissue samples.

On-target specificity analysis

On-target scores for all sgRNA target sites in the exome were computed using the Azimuth 2.0 Python package from Microsoft Research (https://github.com/MicrosoftResearch/Azimuth)[4]. For each target site, on-target efficiencies are computed using gradient-boosted regression trees from Microsoft Azimuth on

the surrounding 30 bp. On-target scores are generated with the `azimuth.model_comparison.predict` function. During initial provisioning, GUIDES computes a table of on-target scores for all sgRNAs in the exome (~$4x10^6$ sgRNAs over human and mouse CCDS).

Off-target specificity analysis (G-score)

Aggregate off-target scores (G-scores) for all sgRNA target sites in the exome were computed as follows. G-scores utilize the Cutting Frequency Determination (CFD) scoring algorithm[4], which computes the likelihood of a sgRNA cutting at a particular off-target site based on experimental data from ~10,000 sgRNAs with mismatches, insertions and deletions. For a given sgRNA, GUIDES finds all potential off-targets with up to 3 mismatches in the human/mouse exome (followed by a `NGG` motif) and calculates the CFD between the given sgRNA and the potential off-target. The sum of these scores is linearly weighted by the number of times the potential off-target occurs in the exome and the result is returned as the G-score for the given sgRNA.

Mathematically, we define the G-score of sgRNA ***j*** as:

$$\text{G-Score}(j) = \sum_{i=1}^{N} c_i \cdot \text{CFD}(i, j)$$

where ***N*** is the number of potential off-targets with up to 3 mismatches, ***i*** ranges over these mismatches, and $\boldsymbol{c_i}$ is the number of times the mismatch occurs in the exome. For figures with the G-score, we added a small value (10^{-8}) to each G-score to avoid *log*(0).

Removal of homopolymer repeats and Pol III terminators

Since homopolymeric regions can be difficult to synthesize and sequence accurately[5], GUIDES removes any sgRNA guide sequence containing stretches of 5 or more of the same base (`A`, `T`, `C` or `G`). Furthermore, GUIDES excludes any sgRNA guide sequence with 4 or more sequential `T` bases which can result in premature termination of Pol III transcription[6].

Design of non-targeting (negative control) sgRNAs

After library generation, GUIDES prompts the user to also include non-targeting (negative control) sgRNAs in the library. By default, GUIDES suggests adding a pool of non-targeting controls of size equivalent to 5% of the number of targeting sgRNAs in the library (up to a maximum of 1000 non-targeting sgRNAs). For example, for a GUIDES library with 1000 gene-targeting sgRNAs, GUIDES will suggest adding 50 additional non-targeting sgRNAs. Using a slider or text entry box, the user can

customize further to specify any number of non-targeting sgRNAs (between 1-1000) or decline to add any at all.

The non-targeting sgRNAs are those designed not to target in the respective genome (human or mouse) and are taken from the 1000 non-targeting human and mouse guide sequences in the GeCKOv2 libraries[7]. Briefly, we generated 10,000 random 20mer sequences and aligned them to a target genome (human or mouse) using a short-read aligner (`bowtie`), allowing alignment with up to 3 mismatches[8]. From this output, we selected non-targeting guide sequences as those that do not align to the target genome with 0, 1, 2 or 3 mismatches. Several studies using the GeCKOv2 human and mouse libraries have set a false-discovery (background enrichment/depletion) rate with these non-targeting sgRNAs[9–12].

Scaffolds for synthesis-ready oligonucleotides

GUIDES also produces full-length, synthesis-ready oligonucleotides that flank the sgRNA guide sequence with overhangs for Gibson cloning into an appropriate screening vector (e.g. lentiCRISPRv2 or lentiGuide-Puro)[7]. These flanking sequences include the end of the U6 primer on the 5' side of the guide sequence and the beginning of the sgRNA scaffold on the 3' side of the guide sequence. In this format, the GUIDES output can be sent directly for synthesis to common pooled oligonucleotide synthesis service providers (e.g. Twist Bioscience, CustomArray, Agilent). The user can select between either the full-length sgRNA scaffold (with the 85-nt tracrRNA) or a modified version ("E+F modification") with an `A-U` flip to prevent early Pol III termination and a 5 bp (10-nt) extension of the first stem loop[13]. The flanking sequences used for synthesis-ready oligonucleotides and appropriate primers for PCR (T_a = 63C) and Gibson cloning for each scaffold are:

Full-length scaffold	GGAAAGGACGAAACACCGXXXXXXXXXXXXXXXXXXXXGTTTT
73-nt including guide	AGAGCTAGAAATAGCAAGTTAAAATAAGGC
Full-length cloning F	TAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTG GAAAGGACGAAACACCG
Full-length cloning R	ACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCT ATTTCTAGCTCTAAAAC
E+F modified scaffold	GGAAAGGACGAAACACCGXXXXXXXXXXXXXXXXXXXXGTTTA
63-nt including guide	AGAGCTATGCTGGAAACAGC
E+F cloning F	TAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTG GAAAGGACGAAACACCG

E+F cloning R	GACTAGCCTTATTTAAACTTGCTATGCTGTTTCCAGCATAGCT CTTAAAC

Procedure to generate a GUIDES library

For selection of sgRNAs to return to the user, we developed a data structure which enables processing in linear (instead of linearithmic) time. In this data structure, each gene is considered independently.

For each gene, GUIDES keeps track of two different lists of sgRNAs: the ***primary*** list and the ***secondary*** list. Each list (***primary*** or ***secondary***) contains all sgRNAs from a particular group of exons and each exon can be on only one list at a time. The ***primary*** list is unordered (guides from all exons on the list are grouped together) whereas the ***secondary*** list is ordered (guides from higher-rank exons will be considered before guides from lower-rank exons).

During initial provisioning, a sorted list of sgRNAs is generated for each exon. The optimized, linear-time data structure keeps track of the top-ranked guide from each exon. During guide selection, GUIDES selects the highest-ranked sgRNA (see *Design pipeline overview* section for details of guide ranking) from the ***primary*** list. If the ***primary*** list is empty or all remaining guides have exact matches elsewhere in the exome, then the highest-ranked exon from the ***secondary*** list is moved into the primary list. This continues until all sgRNAs are selected or both lists are empty.

For a gene containing N exons, when GTEx-based exon selection is disabled, the ***primary*** list contains exons 2 to N-1. The ***secondary*** list (ordered) contains exon 1 followed by exon N. The last coding exon is given last priority since mRNA may escape nonsense-mediated decay when mutations are in the last exon[14]. When GTEx is enabled, the ***primary*** list contains the top M exons by RNA expression in the selected tissues (default: $M = 4$) and the ***secondary*** list (ordered) contains the remaining exons (with exon 1 and exon N always placed second-to-last and last, respectively, in the ***secondary*** list).

In addition to returning the top ***i*** sgRNAs (where ***i*** is the number of sgRNAs per gene requested by the user), GUIDES also returns the next ***j*** (default: ***j*** = 10) guides as "unselected" to the front-end to allow the user to further fine-tune the library. During sgRNA selection, real-time updates are provided to the front-end to provide an accurate indication of remaining library generation time.

Implementation and software framework details

Back-end services for GUIDES were implemented in Python using the Flask web framework with Eventlet-enabled concurrent network operations. The application was deployed on the Gunicorn HTTP Server. The interactive front-end visualization scheme was written in coffeescript using the AngularJS framework, with Asynchronous Javascript and XML (AJAX) for real-time interfacing with the server. We implemented a Redis-based storage system so that users can navigate away from GUIDES during longer library runs.

Each library design is processed in parallel using a Celery message broker. This generates the library in the background on top of the Redis store, while making current progress accessible to the front-end. When the front-end observes that the current routine has completed, it loads the finalized results via AJAX. The results are returned in JavaScript Object Notation (JSON) for front-end display. Additionally, the Celery broker uses smtplib to notify the user of completion via email. We designed interactive charts for sgRNA visualization using the open-source Chart.js library and HTML5 canvas element.

In order to speed up gene expression computations, we divided the GTEx data into a separate pandas dataframe per gene and serialized using cPickle. This allowed us to transition the overhead of gene lookups to the Linux filesystem, which is significantly faster than Pandas. We experimented on various encodings for serializing the list of precomputed and presorted sgRNAs, finding that binary serialization using MessagePack enables the fastest decoding. All other serializations were performed in pickle (for Python objects) or cPickle (for text-only objects).

Analysis of GUIDES-selected sgRNAs in genome-scale screens

To quantify 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. To do this, we analyzed depletion screen results from the GenomeCRISPR database, which compiles data from multiple genome-scale CRISPR screens[15]. In this database, each sgRNA is normalized by depletion/enrichment within each screen using percentile rank to allow for relative comparison across the entire dataset (`sgRNA effect`). Using the previously computed `sgRNA effect` scores[15], we sought to test if sgRNAs chosen by GUIDES have consistently higher/lower `sgRNA effect` scores than a size-matched control set chosen from all sgRNAs targeting the same gene.

To first obtain a set of universally-essential genes, we combined results from two recent studies that measured depletion using genome-scale CRISPR loss-of-function screens across multiple different cell types[16,17]. Hart *et al.* identify 829 genes as essential in all 5 cell lines (from diverse tissues) that they examined. We retrieved the same number of top-ranked genes from Wang *et al.* (ranked by average depletion *p* value across the 4 cell lines) and then computed the intersection of these 2 lists to find genes in common between these studies.

For each gene, we used GUIDES to generate a list of the top 50 sgRNAs per gene (GUIDES parameters: GTEx expression data enabled using the average of all tissues, Pfam protein domain targeting enabled). For these sgRNAs, we then searched for as many of these 50 sgRNAs as possible in GenomeCRISPR depletion experiments (mean ± s.d.: 8 ± 6 sgRNAs per gene found in GenomeCRISPR). For each gene, we also randomly selected the same number of sgRNAs from the GenomeCRISPR database (i.e. with no preference for GUIDES-ranking).

Over all genes, the average increase in depletion using GUIDES-generated sgRNAs compared to the size-matched randomly selected sgRNAs was 0.73 `sgRNA effect` ($n = 403$ genes examined in 77 genome-scale screens using 61 different cell lines, $p = 5 \times 10^{-7}$, $t = -5.1$, $df = 409$, two-sample paired t-test). In the GenomeCRISPR database, `sgRNA effect` ranges from -9 to +9 and assigns sgRNAs into 10%-quantiles based on a within-screen percentage rank. Thus, the increase in depletion with GUIDES-generated sgRNAs is approximately one 10%-quantile. The increased depletion can be visualized by examining the cumulative distribution of depletion scores (Supplementary Fig. 4), where negative values of `sgRNA effect` indicate greater depletion.

Comparison with existing sgRNA design tools

Several tools already exist for sgRNA design and we present a comparison with six tools in Table 1. Many of these tools address certain aspects of sgRNA design such as calculating on-target and off-target scores but they do not include several features unique to GUIDES, such as prioritizing exons to target by expression and tissue-specific library design, targeting protein functional domains inside genes, minimal user wait times for design of large libraries, flexible gene input (HUGO, Entrez or Ensembl IDs), and allowing the user to dynamically edit the sgRNA selection for each gene in a graphical user interface.

Supplementary References

1. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
2. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279-285 (2016).
3. Harte, R. A. *et al.* Tracking and coordinating an international curation effort for the CCDS Project. *Database J. Biol. Databases Curation* **2012,** bas008 (2012).
4. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34,** 184–191 (2016).
5. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355,** 950–954 (2017).
6. Bogenhagen, D. F. & Brown, D. D. Nucleotide sequences in Xenopus 5S DNA required for transcription termination. *Cell* **24,** 261–270 (1981).
7. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11,** 783–784 (2014).
8. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
9. Golden, R. J. *et al.* An Argonaute phosphorylation cycle promotes microRNA-mediated silencing. *Nature* **542,** 197–202 (2017).
10. Erb, M. A. *et al.* Transcription control by the ENL YEATS domain in acute leukaemia. *Nature* **543,** 270–274 (2017).
11. Parnas, O. *et al.* A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162,** 675–686 (2015).
12. Chen, S. *et al.* Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160,** 1246–1260 (2015).

13. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155,** 1479–1491 (2013).
14. Popp, M. W.-L. & Maquat, L. E. Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.* **47,** 139–165 (2013).
15. Rauscher, B., Heigwer, F., Breinig, M., Winter, J. & Boutros, M. GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. *Nucleic Acids Res.* **45,** D679–D686 (2017).
16. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350,** 1096–1101 (2015).
17. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163,** 1515–1526 (2015).