nature biotechnology

GuideScan software for improved single and paired CRISPR guide RNA design

Alexendar R Perez[1–4], Yuri Pritykin[1,4], Joana A Vidigal[2,4], Sagar Chhangawala[1,3], Lee Zamparo[1], Christina S Leslie[1] & Andrea Ventura[2]

We present GuideScan software for the design of CRISPR guide RNA libraries that can be used to edit coding and noncoding genomic regions. GuideScan produces high-density sets of guide RNAs (gRNAs) for single- and paired-gRNA genome-wide screens. We also show that the trie data structure of GuideScan enables the design of gRNAs that are more specific than those designed by existing tools.

CRISPR–Cas9 genome editing[1,2] can be used for both straightforward disruption of single protein-coding genes and genome-wide loss-of-function screens[3–6]. However, disruption of noncoding RNA or DNA elements using CRISPR–Cas9 remains limited[6–9] because it usually requires the concomitant expression of two gRNAs to engineer deletions[8,10]. Although there has been progress toward making the construction of paired-gRNA vectors scalable[11], the absence of tools to design paired-gRNA libraries has hampered progress toward genome-wide non-coding genetic screens. To overcome this limitation we developed GuideScan, an open-source software package that allows users to construct comprehensive and fully customizable gRNA databases for any genome or CRISPR endonuclease and design paired- and single-gRNA libraries (**Fig. 1a**).

The first step in building a gRNA database is identifying the targetable genome space. To accommodate different CRISPR endonucleases with distinct specificity requirements, GuideScan allows the user to define target sequences by setting the values of three parameters: the protospacer-associated motif (PAM), the PAM position relative to the gRNA binding sequence, and gRNA length[12] (**Fig. 1a**, middle panel). Non-canonical PAMs can also be specified, since they can be recognized and cleaved by CRISPR proteins with some efficiency and thus may contribute to off-target cutting.

Next, gRNA sequences that can lead to cleavage of multiple genomic loci need to be identified and removed. Because CRISPR endonucleases can tolerate single mismatches in gRNA–DNA pairing[13], gRNAs that have less than two mismatches to off-target loci are typically avoided[14]. Third-party alignment tools are often used to identify potential off-target loci in the genome, but we and others[12] have found that they are unable to return all mismatch neighbors (data not shown) and thus consistently underestimate the number of potential off-target sites for a given gRNA[15].

Rather than using an alignment tool, GuideScan uses a retrieval tree (trie) data structure, which efficiently and precisely enumerates all targetable sequences present in a given genome (**Fig. 1a**, right panel). Traversals of the trie allow for the computation of sequence mismatch neighborhoods, which are used to construct databases of gRNAs whose target sites are unique in the genome up to a user-defined number of mismatches (M). Additionally, for gRNAs in the database, more degenerate target sequences – up to Q (Q > M) mismatches – can be correctly enumerated, and stored in the

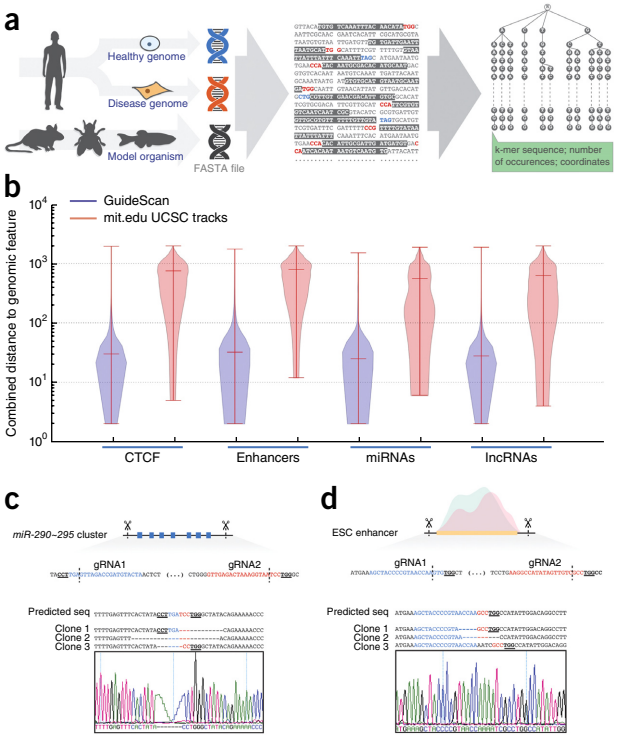

Figure 1 The GuideScan gRNA design tool. (**a**) Overview of GuideScan. Left, GuideScan takes as input a FASTA file containing any genome of choice. Middle, targetable sequences are defined by choosing the PAM sequence(s) (Cas9's canonical PAM, red; non-canonical PAM, blue), its position relative to the gRNA, and the length of the gRNA (gray box). Right, targetable sequences are indexed in a retrieval tree (trie), and associated information is stored at leaf nodes. R, trie root node. (**b**) Distributions of combined distance of flanking gRNA-pairs to the boundaries of selected noncoding genomic features using GuideScan (blue) or mit.edu genome-wide tracks (red). (**c**,**d**) Example deletions of genomic regions containing RNA (**c**) and DNA (**d**) noncoding elements using pairs of gRNAs designed by GuideScan. gRNA sequences, blue and red; PAM sequences, bold underlined. The predicted sequence after deletion, the sequences of three edited alleles, and a representative chromatogram are shown for each targeted locus.

[1]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [2]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [3]Weill Cornell Graduate School of Medical Sciences of Cornell University, New York, New York, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.V. (venturaa@mskcc.org), C.S.L. (cleslie@cbio.mskcc.org) or J.A.V. (decampoj@mskcc.org).

Received 22 August 2016; accepted 27 January 2017; published online 6 March 2017; doi:10.1038/nbt.3804

database as potential off-target loci. Each gRNA can be further annotated with additional information, including on-target efficiency scores[13] and the genomic feature(s) it overlaps with (for example, exon, intron, intergenic region). Once constructed, the database allows efficient individual or batch queries of genomic coordinates for single or paired gRNA designs.

To validate GuideScan, we generated a murine Cas9 database with M = 1. This database contains over 1.7×10^8 gRNAs, with an average

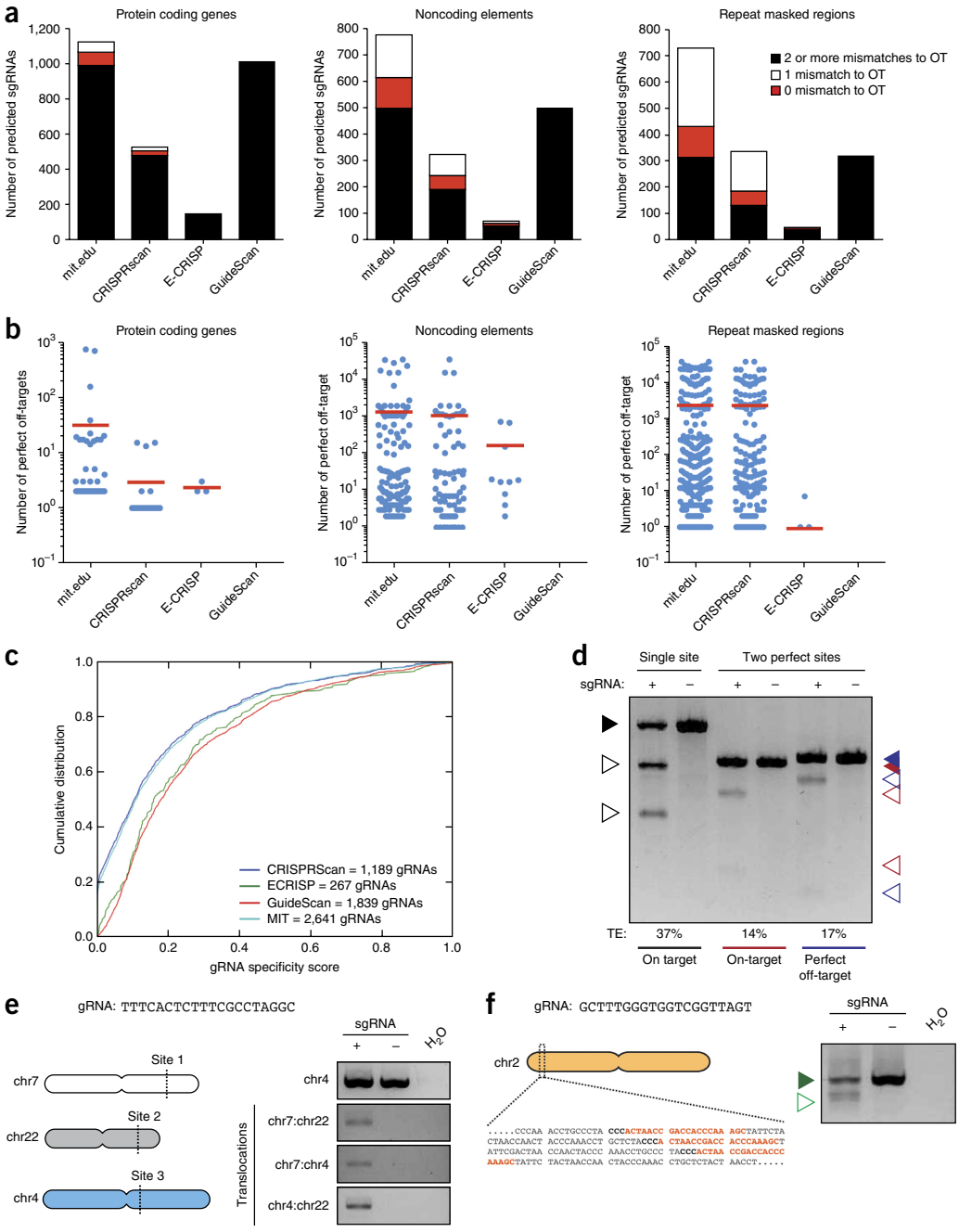

Figure 2 GuideScan correctly enumerates off-target sequences and filters out promiscuous gRNAs. (**a**) Number of murine gRNAs (20 mers) designed by each tool for a random sample of protein-coding genes, noncoding elements, and repetitive regions. Number of gRNAs with off-target sites within at least two mismatches from gRNA (black), within a single mismatch (white), and with perfect off-target sites (red). OT, off-target. (**b**) Number of perfect off-target sites for the gRNAs designed by each tool. Each dot represents a gRNA (mean, red line). (**c**) Cumulative distribution of specificity scores for the gRNAs designed by each tool. (**d**) T7 cleavage assay for gRNAs having a single (black, on-target site) or multiple (red, on-target site; blue, perfect off-target site) perfect matches in the genome. Position of the cleavage substrates, filled triangles; position of cleavage products, open triangles. Estimated total editing (TE) at each site is shown below the corresponding lane. (**e**) Left, schematic representation of the chromosomal locations of three perfect target sites of a gRNA labeled highly specific by competitor tools (mit.edu score = 78). Right, PCR-based identification of chromosomal translocations between all three targets. +, gRNA; – empty plasmid. (**f**) Left, schematic representation of the position of three perfect target sites—all within chromosome 2—of a gRNA labeled highly specificly by competitor tools (mit.edu score = 89). Genomic sequence: target sites, red; PAM sequence, bold. Right, PCR-based identification of chromosomal deletions between target sites. Position of the wild-type amplicon, filled triangle; position of deletion amplicons, open triangle. +, gRNA; –, empty plasmid. Gels in **d** and **e** were cropped from full-length versions shown in **Supplementary Figure 2**.

distance of 15.5 nucleotides between gRNAs designed over autosomes, and is two orders of magnitude larger than the genome-wide gRNA database, provided by mit.edu as a UCSC genome browser track (**Supplementary Fig. 1a** and **Supplementary Note**).

To test if using Guidescan translated into a better practical outcome in terms of designing gRNAs for precise genomic deletions in the noncoding genome, we took the coordinates of mouse CTCF binding sites, enhancers, miRNAs, and lncRNAs and used GuideScan or the mit.edu genome-wide UCSC track gRNA database to design pairs of gRNAs against each target site. We found that the combined median distances of the GuideScan gRNA pairs to the feature boundaries were 29, 31, 24, and 27 base pairs (bp) respectively, compared with 781, 783, 716 and 774 bp when using gRNAs from the mit.edu genome-wide track (**Fig. 1b**). Thus, databases generated by our tool are suitable for the engineering of precise genomic deletions and the generation of loss-of-function alleles for noncoding regulatory elements (**Fig. 1c,d**).

To further benchmark GuideScan, we compared it to three widely used gRNA design tools: the mit.edu web interface[14] (**Supplementary Note**), CRISPRscan[10], and E-CRISP[16]. Unlike GuideScan these tools do not provide direct access to the underlying databases, so we queried 150 regions in the mouse genome overlapping randomly selected protein-coding genes, noncoding elements, and repetitive regions. All tools returned distinct but overlapping sets of gRNAs (**Fig. 2a** and **Supplementary Table 1**), and surprisingly all except for GuideScan returned a substantial fraction of guides having more than one perfect target site (**Fig. 2a** and **Supplementary Table 2**). In the most extreme cases, individual gRNAs had more than 30,000 perfect off-target sites (**Fig. 2b**) that were missed by the corresponding design tool (**Supplementary Table 1**) and consequently were not considered when calculating the gRNA's specificity score. In fact, gRNAs with more than one perfect target site were roughly equally distributed between low (26%), medium (37%), and high (37%) specificity score categories according to the output of the mit.edu web interface[11] (**Supplementary Fig. 1b**).

In addition, the competitor tools that we tested underreported the number of off-target sites with single mismatches to the gRNA (**Supplementary Table 1**). Based on these observations we predicted that gRNAs returned by GuideScan should have, on average, greater specificity than gRNAs reported by the other tools. To test this prediction, we used GuideScan's trie function to enumerate all potential off-target loci ($Q = 3$) for the gRNAs designed by all tools, and calculated the likelihood of cleavage at these sites using a recently reported metric that takes into account the number, position, and nature of mismatches[13]. We then used these values to calculate the aggregate specificity score for each gRNA, as previously described[14]. We found that gRNAs designed by GuideScan had on average significantly higher specificity scores than those returned by the mit.edu web interface ($P < 2.2 \times 10^{-16}$; $D = 0.22$; $Z = 7.38$) or by CRISPRScan ($P < 2.2 \times 10^{-16}$; $D = 0.25$; $Z = 6.59$) (**Fig. 2c**). E-CRISP and GuideScan had similar distribution of specificity scores, but E-CRISP returned an order-of-magnitude fewer gRNAs, some of which had multiple perfect or near-perfect matches in the genome (**Fig. 2a,b** and **Supplementary Table 2**).

Failure to discard gRNAs with multiple perfect target sites has important implications in the design and interpretation of gene editing experiments. The number of double-strand breaks induced by a gRNA in a single cell correlates well with gene-independent gRNA depletion, and is a major source of noise in negative-selection screens[17]. Thus, the correct identification and filtering of promiscuous gRNAs is crucial for designing effective CRISPR libraries. Furthermore, unknowingly using gRNAs with multiple perfect target sites can result in highly efficient gene editing at undesired sites (**Fig. 2d**) as well as the generation of chromosomal rearrangements[18] such as translocations (**Fig. 2e**) and deletions (**Fig. 2f**) that include the desired cleavage site and the unknown additional sites.

To facilitate the construction of single and paired gRNA libraries, we have released a web interface that includes access to pre-computed genome-wide Cas9 and Cpf1 gRNA databases for many model organisms (http://www.guidescan.com/). This website allows users to input coordinates of genomic features in batch, to choose between designing single internal gRNAs or pairs of flanking gRNAs, and retrieve for each genomic coordinate a pre-defined number of gRNAs or gRNA pairs.

Collectively, these data show that GuideScan is a substantial improvement compared with existing gRNA design tools. We expect that this tool will facilitate ongoing efforts aimed at deducing functions of coding and the noncoding parts of genomes.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the Ventura and the Leslie laboratories for comments and suggestions. We thank L. Fairchild and R. Pelossof for providing source code for the SplashRNA web server to serve as the backbone for the GuideScan website. This work was supported in part by NIH: grants P30-CA008748 (MSK Core), U01-HG007033 (C.S.L.), U01-HG007893 (C.S.L.), and by grants from the Geoffrey Beene Cancer Research Foundation (A.V.), the Uniting Against Lung Cancer Foundation (A.V.), the Cycle for Survival Foundation (A.V.), the Pershing Square Sohn Cancer Research Alliance (A.V.), and the Lung Cancer Research Foundation (J.A.V.). The GuideScan source code and all associated documentation are deposited at guidescan.com.

AUTHOR CONTRIBUTIONS

J.A.V., C.S.L., and A.V. conceived and supervised the project. Y.P. and A.R.P. developed the GuideScan algorithm with input from C.S.L.; A.R.P. and Y.P. implemented the GuideScan software package; A.R.P. performed the computational experiments; J.A.V. performed the wet-lab experiments; A.R.P. and S.C. implemented the web-server; L.Z. provided expertise in software development and helped improve the website user experience; J.A.V. drafted the manuscript with contributions from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Doudna, J.A. & Charpentier, E. *Science* **346**, 1258096 (2014).
2. Hsu, P.D., Lander, E.S. & Zhang, F. *Cell* **157**, 1262–1278 (2014).
3. Shalem, O. *et al. Science* **343**, 84–87 (2014).
4. Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera, Mdel C. & Yusa, K. *Nat. Biotechnol.* **32**, 267–273 (2014).
5. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. *Science* **343**, 80–84 (2014).
6. Korkmaz, G. *et al. Nat. Biotechnol.* **34**, 192–198 (2016).
7. Rajagopal, N. *et al. Nat. Biotechnol.* **34**, 167–174 (2016).
8. Zhu, S. *et al. Nat. Biotechnol.* **34**, 1279–1286 (2016).
9. Canver, M.C. *et al. Nature* **527**, 192–197 (2015).
10. Moreno-Mateos, M.A. *et al. Nat. Methods* **12**, 982–988 (2015).
11. Vidigal, J.A. & Ventura, A. *Nat. Commun.* **6**, 8083 (2015).
12. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. *Nat. Biotechnol.* **32**, 279–284 (2014).
13. Doench, J.G. *et al. Nat. Biotechnol.* **34**, 184–191 (2016).
14. Hsu, P.D. *et al. Nat. Biotechnol.* **31**, 827–832 (2013).
15. Tsai, S.Q. *et al. Nat. Biotechnol.* **33**, 187–197 (2015).
16. Heigwer, F., Kerr, G. & Boutros, M. *Nat. Methods* **11**, 122–123 (2014).
17. Aguirre, A.J. *et al. Cancer Discov.* **6**, 914–929 (2016).
18. Maddalo, D. *et al. Nature* **516**, 423–427 (2014).

ONLINE METHODS

Command line gRNA database generation. To generate a customized genome-wide gRNA database for a given CRISPR system the user supplies the target genome as a FASTA file and specifies parameters such as desired gRNA length, PAM position with respect to gRNA, canonical and alternative PAM sequences, Hamming distance (M) for which gRNAs are required to have a unique target site in the genome, and Hamming distance (Q) for which potential off-target sites will be enumerated. Like other methods, the algorithm scans a genome for canonical and alternative PAM sequences and identifies all k-mers associated with them[19]. The full universe of these k-mers and PAM sequences with their coordinates are written out to a temporary file along with a count of how often a particular k-mer occurs in the genome. At the next stage, the algorithm analyzes the list of k-mers and determines which of these k-mers constitute potential gRNAs. A trie of all k-mers is constructed. The trie stores a user-defined maximum number of k-mer coordinates. If and only if a k-mer occurs with a canonical PAM uniquely in the genome then it is initially labeled as a candidate gRNA. However, if the k-mer occurs more than once in the genome or occurs with an alternative PAM, then it is labeled as a non-candidate gRNA. These non-candidate gRNAs are written out to 'blacklist' files, which define why a k-mer was rejected as a candidate gRNA. Furthermore, the algorithm ensures that each gRNA has a unique target in the genome up to M mismatches, thereby forcing all candidate gRNAs to be distinct from one another by at least M mismatches. This task is accomplished through trie traversals, where for each candidate gRNA sequence a mismatch neighborhood (which in this case is the set of inexact matches to the candidate gRNA up to M mismatches) is assessed. If even one mismatch neighbor sequence is found for a given candidate gRNA, then this gRNA is not unique in the genome up to M mismatches and is therefore relabeled as a non-candidate gRNA and written to the 'blacklist' file. The gRNAs that are unique in the genome up to M mismatches can have off-target sites completely enumerated up to Q (where $Q > M$) mismatches. Off-target information for candidate gRNAs is again determined by trie traversals and computing the mismatch neighborhood for a given gRNA sequence. By construction, the gRNAs undergoing off-target enumeration have no off-target sites within M mismatches, and so if an inexact match to the gRNA sequence with P mismatches is found, such that $M < P \leq Q$, then the number of times this off-target sequence occurs, its Hamming distance P, and coordinates for a user-defined number of off-target are recorded. Candidate gRNAs are written from the trie to a file in sequence alignment map (SAM) format[20]. This file has the gRNA sequence as a unique identifier as well as off-target information for the gRNA stored in a hex-byte array. SAM tags record a maximum count of off-target sites, the distance for which off-target sites were searched, and the hex-byte array of off-target sequences and their coordinates. This SAM file is then converted to a BAM file and indexed so that it can be quickly accessed using Samtools.

Database cutting efficiency scores. Cas9 guideRNA databases can be enriched with additional information such as on-target cutting efficiency scores. We adopt the scores defined by Doench *et al.*[13]. Rule Set 2[13] Scores are computed for each gRNA and added to the previously constructed database. This rule set predicts on-target gRNA cutting efficiency using a learned boosted regression tree model. The input for this model requires a 30-mer sequence, only assesses cutting efficiency for 20-mer gRNAs, and only predicts cutting efficiency for gRNAs with the canonical Cas9 PAM sequence (NGG).

Database cutting specificity scores. A database BAM file, composed of Cas9 gRNAs with 20-mer complementary region and NGG PAM, can be supplemented to include gRNA specificity scores. Specificity scores for gRNAs are based on the Doench *et al.* CFD model[13], which computes the likelihood of a gRNA cutting at each individual off-target site using an experimentally derived mutation matrix. Specifically, for a given gRNA, GuideScan enumerates all its neighbors up to Q mismatches, calculates the CFD score for each neighbor, and then multiplies that score by the number of times the neighbor occurs in the genome. It then aggregates the CFD values into a single composite score using the formula used by Hsu *et al.*[14]:

$$\text{Specificity Score} = \frac{1}{\sum_{i=1}^{n} CFD_i * q_i}$$

Here, n represents the number of unique targetable sites within up to z mismatches. The desired on-target site ($z = 0$) is included in this computation and will give a CFD value of one. The value q_i represents the number of times the i^{th} neighbor occurs in the genome. For a unique target site up to z mismatches, the specificity score would be 1 since CFD = i = n = q_i = 1. The resulting composite specificity score is written out to a text file along with information such as the target sequence and target coordinates. These scores are then added to the BAM file as a new SAM tag. Importantly, because the GuideScan algorithm is general and allows the construction of gRNA databases for distinct enzymes and distinct gRNA lengths, specificity scores are not automatically computed. If a user chooses to compute these scores, GuideScan first determines whether the database conforms to the parameters required for CFD scoring, and if so computes the aggregate specificity score for each gRNA in that database. For the specificity scores in the pre-computed Cas9 databases on the GuideScan web-interface, the parameters $z = Q = 3$ were used.

Database query and gRNA annotations. The algorithm allows for the direct query of the BAM file database using PySam (python interface to samtools; https://github.com/pysam-developers/pysam) and allows for the lookup of gRNAs by genomic coordinates. The output contains the gRNAs in the queried region, off-target information, predicted Rule Set 2 cutting efficiency score and specificity score if appropriate, as well as the exon annotation of on-target gRNA and off-target cut sites. The exon annotation relies on the creation of an interval tree constructed from a BED file containing genome-wide exon coordinates. Consequently, the exon annotation for gRNAs is done at the time of database query.

Code availability. The code is freely available at guidescan.com, in **Supplementary Code** and at bitbucket (bitbucket.org/arp2012/crispr-project/overview). Version v0.0.4 of the code was used to generate the data presented in this manuscript.

Web server implementation. The algorithm was run on a select set of model organisms, as well as on human genome, to produce gRNA databases where all constituent gRNAs are unique in the genome up to two mismatches. These databases are accessible through a web interface that allows for batch queries by genomic coordinates. The website allows the user to find gRNAs within a genomic region or flanking a genomic region and allows the output to be sorted according to either number of off-target sites, distance to target site, cutting efficiency score, or target specificity score. The website utilizes the database query ability of our algorithm to rapidly report gRNAs for user defined regions and achieves its functionality through CherryPy, a python web framework (http://www.cherrypy.org/).

Tool comparisons. The comparisons shown in **Figure 1b** were done using the coordinates of mouse CTCF binding sites, enhancers, miRNAs, and lncRNAs retrieved from the following publications[19,21–24]. The outputs of GuideScan, crispr.mit.edu web portal[14], CRISPRscan[10], and E-CRISP[16] were compared on sequences overlapping randomly chosen protein-coding genes, noncoding genomic elements, and repeat masked regions. We limited our test to a total of 150 sequences (50 protein-coding genes, 50 noncoding elements, 50 repeat masked regions) from mm10 (the most recent assembly of *M. musculus* genome) because the input limits (file size/ number of sequences) associated with some of these tools made larger scale comparisons impractical. Similarly, we limited the size of the test sequences to 150 base pairs due to input limits of some of the tools. The genomic coordinates of the sequences used in these comparisons are provided in **Supplementary Table 3**. Importantly, the CRISPRscan tool provides both 19-mer and 20-mer gRNA designs. For the purpose of this experiment we limited our analysis to 20 mers because they could be directly compared to the gRNA designs provided by the remaining softwares. The number and type of off-target sites for each gRNA was determined by querying GuideScan's software package intermediate kmers file and independently through a R query of BSgenome mm10 version.

Statistical methods. A one-sided Kolmogorov–Smirnov test was used to compare the specificity scores of 1,839 GuideScan gRNAs against 267 E-CRISP gRNAs, 1,189 CRISPRScan gRNAs, and 2,641 MIT gRNAs; P values were

doi:10.1038/nbt.3804

computed using the ks.test function in R. Additionally, the effect sizes (Z) of these comparisons were reported using the formula

$$Z = D\sqrt{\frac{n_1 * n_2}{n_1 + n_2}}$$

where D is the Kolmogorov–Smirnov statistic equaling the maximum difference between empirical CDF functions of the two samples in the comparison, and n_1 and n_2 are the sample sizes.

DNA constructs. Paired-gRNA vectors to generate deletions of DNA and RNA noncoding elements were cloned as previously described[11]. Briefly, s/mU6 oligos carrying the sequence corresponding to two gRNAs for each locus (miR-290~295: gRNA1, TAGTACATCGGTCTAACTCA; gRNA2, GTTGAGACTAAAGGTAATCC. Enhancer element: gRNA1, AGCTACCCCGTAACCAAGTG; gRNA2, AAGGCCATATAGTTGTCGCC) were PCR amplified and cloned into BsmBI-digested lentiCRISPRv1 vector (Addgene #49535) using pDonor_sU6 intermediate vector (Addgene #69351). Single gRNAs were cloned into BsmBI-digested lentiCRISPRv1 vector using standard oligo cloning protocols.

Cell culture and detection of genomic editing. The V6.5 Mouse Embryonic Stem cells (obtained from Rudolf Jaenisch) were tested for germline transmission and for the absence of mycoplasma infection. These cells were grown on a monolayer of irradiated mouse embryo fibroblasts at 37 °C (5% CO_2) in KnockOut DMEM media (Gibco) supplemented with 15% FCS, L-glutamine (2 mM), penicillin (100 U ml^{-1}), streptomycin (100 µg ml^{-1}), and LIF (10^3 U ml^{-1}). To generate genomic deletions, paired-gRNA vectors were transfected using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocols. Cells were collected in lysis buffer (100 mM Tris-HCl pH8.5, 200 mM NaCl, 5 mM EDTA, 0.2% SDS and 100 ng ml^{-1} proteinase K) 6 d after transfection, and genomic DNA extracted using phenol-chloroform followed by ethanol precipitation. Genomic deletions were detected by PCR using primers flanking the gRNA cut sites (miR_fwd: AGGGAGGAACGAGCCTATGT, miR_rev: GCATGCCTAAATCCCAAGAG; enh_fwd: GTGGCTCAGTGTTTCCCATT, enh_rev: CAGGCAAACTCTCCCATGTT). PCR bands corresponding to the genomic deletions were cloned into the Topo Blunt II vector (Invitrogen) and plasmid DNA from individual bacterial clones subject to Sanger sequencing.

To test cleavage and rearrangements produced by single gRNAs, 293T cells (obtained from American Type Culture Collection ATCC, cultured under standard conditions, and tested negative for mycoplasma contamination) were plated on 12-well-plates (Corning). Constructs expressing individual gRNAs were transfected into cells the following day using lipofectamine 2000 (Invitrogen) and transfected cells selected with puromycin (4 µg ml^{-1}) for 2 d. Four days following transfection, genomic DNA was extracted as above and used to determine total cleavage and generation of genomic rearrangements. Total editing at perfect target sites was determined using a T7 endonuclease assay. Briefly, target sites in a pool of transfected cells were amplified by PCR, and 200 ng of resulting amplicon were used to generate DNA heteroduplexes. The resulting molecules were incubated with 10 U of T7 endonuclease (NEB) for 15 min, and the product of the digestion run on a 2% agarose gel. Total editing estimates were calculated as previously described[25]. Sequences or gRNAs and primers used in these experiments are shown in **Supplementary Table 4**.

19. Pliatsika, V. & Rigoutsos, I. *Biol. Direct* **10**, 4 (2015).
20. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
21. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
22. Whyte, W.A. *et al. Cell* **153**, 307–319 (2013).
23. Kozomara, A. & Griffiths-Jones, S. *Nucleic Acids Res.* **39**, D152–D157 (2011).
24. Harrow, J. *et al. Genome Res.* **22**, 1760–1774 (2012).
25. Lin, S., Staahl, B.T., Alla, R.K. & Doudna, J.A. *eLife* **3**, e04766 (2014).