

Breaking-Cas—interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes

Juan C. Oliveros¹, Mònica Franch¹, Daniel Tabas-Madrid¹, David San-León¹,
Lluís Montoliu^{1,2}, Pilar Cubas¹ and Florencio Pazos^{1,*}

¹National Centre for Biotechnology (CNB-CSIC), c/ Darwin 3, 28049 Madrid, Spain and ²CIBERER, ISCIII, Madrid, Spain

Received January 30, 2016; Revised May 02, 2016; Accepted May 03, 2016

ABSTRACT

The CRISPR/Cas technology is enabling targeted genome editing in multiple organisms with unprecedented accuracy and specificity by using RNA-guided nucleases. A critical point when planning a CRISPR/Cas experiment is the design of the guide RNA (gRNA), which directs the nuclease and associated machinery to the desired genomic location. This gRNA has to fulfil the requirements of the nuclease and lack homology with other genome sites that could lead to off-target effects. Here we introduce the *Breaking-Cas* system for the design of gRNAs for CRISPR/Cas experiments, including those based in the Cas9 nuclease as well as others recently introduced. The server has unique features not available in other tools, including the possibility of using all eukaryotic genomes available in ENSEMBL (currently around 700), placing variable PAM sequences at 5' or 3' and setting the guide RNA length and the scores per nucleotides. It can be freely accessed at: <http://bioinfogp.cnb.csic.es/tools/breakingcas>, and the code is available upon request.

INTRODUCTION

A general method for introducing generic modifications in the genome sequence of any organism at a desired position with high precision has been a long standing goal since the beginnings of molecular biology. The most recent technology toward this objective is the CRISPR/Cas system (*clustered regularly interspaced palindromic repeats/CRISPR-associated*), which allows targeting a CRISPR-associated (Cas) nuclease to a desired genomic location with nucleotide precision (1). It is based on an ancient and conserved prokaryotic defensive system against foreign DNA that stores information on previous infections (pieces of the invading DNA (2)) so that it can respond to new ones of

the same pathogen (3,4). Specific nucleases whose substrate specificity is dictated by a bound RNA complementary to these foreign DNA fragments are used for cleaving forthcoming instances of the invading genetic material (5). These 'programmable' RNA-dependent DNA endonucleases are the basis of the CRISPR/Cas technology for genome editing. Among them, Cas9 from *Streptococcus pyogenes* is so far the most widely used, but other similar nucleases are being described and applied, including Cpf1 (6), SaCas9 (7) and other from different CRISPR/Cas class II systems (8).

Therefore, one of the critical steps in a CRISPR/Cas experiment is the design of the 'guide RNA' (gRNA) that will target the Cas nuclease to the desired genomic location. In the CRISPR/Cas9 system from *S. pyogenes*, gRNAs have in 3' a fixed sequence required for Cas9 recognition, and in 5' a 20-nucleotide (nt) sequence complementary to the target DNA. An additional constraint in gRNA design is that the target genomic sequence must have in 3' a short motif termed 'protospacer adjacent motif' (PAM) recognized by the nuclease. In the case of *S. pyogenes* Cas9, the PAM is usually NGG, but other PAM sequences have been reported for *Staphylococcus aureus* Cas9 (7) or Cpf1 (6) nucleases. Furthermore, Cpf1 requires the PAM to be located at 5'.

A second critical consideration when choosing the 20-nt motif is the possibility that it hybridizes to identical or highly similar DNA sequences in other genomic locations, as this would lead to undesired off-target effects. *In vitro* studies in mammalian cells have found that even sites differing in as many as 4–5 out of 20 nt with the intended motif can still be cleaved by the nuclease (9). This would lead to hundreds of potential off-targets for a generic 20 nt segment in a random DNA sequence the size of the human genome (1). Several authors have exhaustively explored which mismatches (within the 20-nt sequence) would still allow Cas recognition and DNA cutting. For this, they generated large collections of gRNAs with variations at every position and evaluated their effects on nuclease activity (e.g. (10,11)). These data allowed deriving matrices to score the feasibility of any given 20-nt motif to be an off-target.

*To whom correspondence should be addressed. Tel: +34 915854669; Fax: +34 915854506; Email: pazos@cnb.csic.es

Indeed, CRISPR/Cas off-targets is one of the hottest topics in the field due to the implications for the reliability/specificity of this technique and its safety in therapeutic applications, especially those eventually involving human germ-line modifications. Experimental evidence suggests that off-target effects can become a problem when the experiment involves continuous expression of Cas and gRNA (e.g. *in vitro* plasmid transfection), whereas its relevance is limited when the CRISPR/Cas machinery is transiently induced, as it is the case for *in vivo*, germline-type modifications (12). For example, in mice, the presence of unintended off-target modifications is rarely observed (12,13).

Numerous computational tools, both public and commercial, are already available to assist gRNA design for CRISPR/Cas experiments (e.g. (10,11,14–17)). See (18) for a recent review). They all take into account the above-mentioned sequence constraints, and each has its own advantages and shortcomings. The main general limitation is the reduced number of genomes they can work with, a few tens of model organisms in the best cases (18). Another important limitation is that most are designed for only one or a few number of nucleases.

We have developed the *Breaking-Cas* web server for multi-organism design of gRNAs. Our aim was to encompass the best concepts and ideas from many of these systems, as well as those derived from our own experience helping users designing specific gRNAs for their CRISPR/Cas experiments. This tool can be applied for designing gRNAs to target any eukaryotic genome available through ENSEMBL. It can be used beyond CRISPR/Cas9 since it includes parameters useful for a variety of recently-reported Cas-like nucleases, as well as a generic method for incorporating the parameters (mismatch allowance, PAM characteristics, etc.) of other newly discovered CRISPR/Cas class II systems (8).

THE BREAKING-CAS SYSTEM

Main features

The main features of *Breaking-Cas*, absent in other available tools are:

- (i) The possibility of designing experiments for any eukaryotic genome available in the ENSEMBL and ENSEMBLGENOMES databases (19). All other existing tools have only a limited number of model organisms available (18), while one of the main advantages of the CRISPR/Cas technology is the possibility of applying it to any organism or cell line, especially to those for which other genetic tools are limited or do not exist.
- (ii) The flexibility to customize the PAM motif sequence and its position relative to the guide oligonucleotide (5' or 3'). Cas9-based systems from numerous bacteria are being characterized (20,21). Therefore is of paramount relevance to develop a system that can be adjusted to other PAM requirements.
- (iii) A flexible method for introducing nuclease-specific scoring systems for assessing off-targets, including the possibility of using oligonucleotide sequences of different lengths (from 18 to 25 nucleotides), to accommodate recent reported results in the CRISPR/Cas field (6,8).

- (iv) A rich, user-friendly graphical interface for navigating the results, including a built-in mini genome browser that allows easy inspection of the genomic context of predicted targets and off-targets.

The input form

The *Breaking-Cas* web server is freely accessible on-line at: <http://bioinfogp.cnb.csic.es/tools/breakingcas>

The input form was designed with simplicity in mind. The user has to provide the name of the reference organism where the experiment is going to be performed, the characteristics of the Cas-like nuclease to be used, and the sequence(s) of the intended target genomic region (in FASTA format, with a limit of 20 000 nucleotides in total). It is possible to apply the tool to many sequences in a single run by providing a FASTA file with multiple entries. It is advised to use concise FASTA headers as the job identifiers will be based on them. As the system can work with hundreds of organisms (~700 eukaryotic genomes available in ENSEMBL and ENSEMBLGENOMES right now), special care was taken to facilitate the selection of the query species. Upon typing the first letters of a scientific or common species name, a list of matching genomes appears, from which the user can select that of interest. Alternatively, the user can open a link with a clickable alphabetic list of all available genomes. The species names are presented along with their corresponding NCBI taxonomy identification number.

The characteristics of the nuclease include the PAM motif, its position respect to the guide oligonucleotide (5' or 3'), the length of the oligonucleotide, the number of allowed mismatches, and a matrix of position weights to determine the feasibility of a given DNA segment to be an off-target depending on the position of the non-matching nucleotides. There are several pre-sets of these parameters for widely used nucleases (*Cas9* and *Cpf1* from various bacteria), but the user can also change them freely to adapt the query to other nucleases. In the case of SpCas9 the positions weights are those experimentally determined (10), while for Cpf1 this info is not yet available (6). The PAM motif can be freely introduced using standard IUPAC notation, including generic nucleotides (e.g. N, R) (e.g. NGG of *S. pyogenes* Cas9).

Finally, an email address can be optionally introduced for being informed on the completion of long runs.

Breaking-Cas workflow

The process starts by locating in the input sequences all fragments of the specified length with a compatible PAM at the specified end (i.e. 20 nucleotides and PAM at 3' in the case of SpCas9). These fragments (candidates for designing the gRNA) are stored in FASTQ format with all qualities set to 'I' (22).

Next, the fragments are aligned against the DNA sequence of the reference genome with BWA (23), retrieving all possible hits with up to 4 mismatches (or the specified number of mismatches introduced by the user). For each candidate, its set of homologous hits constitutes a preliminary list of putative off-targets. The 3' (or 5') end of every putative off-target is compared with the PAM motif and

only off-targets with a compatible PAM at 3' or 5' (depending the selected nuclease: Cas9, Cpf1 or other) are selected.

A score is assigned to each off-target based on the number and position of the mismatches respect to its original candidate 20-nucleotide fragment and using a formulation similar to that used in the 'CRISPR Design' tool (10) for the case of Cas9. The score use a matrix of mismatch weights, obtained empirically, that reflect the importance of each position in determining real off-targets when it presents mismatches (10). The score for a given off-target is calculated as:

$$S_{off} = \prod_{p \in M} (1 - W_p) \times \frac{1}{\frac{19-d}{19} \times 4 + 1} \times \frac{1}{m^2} \times 100$$

where the product runs for all positions with mismatches (M), W_p is the weight for the mismatch in position p , and m is the number of mismatches (0 to 4). Off-targets with no mismatches are scored with a value of 100. The average pairwise distance between mismatches (d) is approximated by the distance between the first and last position with mismatches over the number of mismatches minus 1:

$$d = \frac{\max_pos - \min_pos}{m - 1}$$

The higher the score is, the higher the probability of acting as a true secondary Cas9 site. In general, for *S. pyogenes* Cas9 mismatches at positions close to the PAM strongly decrease the offtarget's score.

In absence of experimental evidences of the positional influence of mismatches (i.e. right now for nucleases other than Cas9), the probability of being a true secondary target is estimated as above except that all positional weights are set to '0'. Similarly, the '19' in the formula is replaced by *oligosize-1*, for other nucleases.

It is also possible to introduce customized positional weights (values between 0 and 1, being '0' a totally allowed mismatch and '1' a totally forbidden mismatch) for nucleases with experimental values known by the user.

Finally, an aggregated score is calculated for each of the original 20nt candidates as:

$$S_{guide} = \frac{100}{100 + \sum_{i=1}^n S_{off(i)}}$$

where $S_{off(i)}$ are the scores of the n potential off-targets for that oligo.

The higher this aggregated score (S_{guide}), the better the oligo is for being used for the gRNA, as it has less predicted off-targets and/or with lower scores.

The next step is to annotate each potential off-target with the surrounding genomic elements (usually genes). In order to do that, the 'bedtools window' command (24) is used for comparing the coordinates of the off-targets (stored in BED format) with the official annotation file of the reference genome (in GTF format) as provided by ENSEMBL. Overlapping and nearby genes (at a maximum distance of 1000 bp, upstream or downstream) are identified.

All sequences, coordinates, scores and annotation details of every oligonucleotide candidate and off-target are imported into the *Breaking-Cas* interactive viewer.

Once a job is completed, the results page contains a link to this interactive viewer, as well as a downloadable compressed .ZIP file that allows locally storing and visualizing the results off-line later.

The interactive viewer

All results are presented in an interactive viewer compatible with any modern web browser. The viewer is divided in two main parts. The left panel contains a detailed list with all candidate oligonucleotides present in the input sequence(s), including their positions within them, DNA strand, sequence (including the PAM, at the selected relative position, 5' or 3'), number of perfect hits, number of hits with mismatches (potential off-targets), aggregated score and nearby genes (Figure 1). The list can be sorted by any of these parameters clicking on the corresponding column header. In these oligonucleotides, some sequence features not taken into account by the score but known to affect their faithfulness as good gRNAs in some circumstances are highlighted. For example, it is known that a consecutive stretch of four or more T's should be avoided in gRNA expressing plasmids with U6 promoters since its coincidence with a termination motif for RNA polymerase III might affect the transcription of the gRNA (25).

By clicking a given oligonucleotide, a panel on the right with details on the 100 best scored targets (perfect hits and potential off-targets) is presented, including their individual score, genomic coordinates, and 'affected' genes. Additional targets can be shown by clicking at the bottom of this list. A built-in genome mini-browser is also included showing the genomic region surrounding each target (Figure 1). The 'on-target/s' (perfect match/es) are highlighted with a yellow background. This genomic representation can be zoomed in/out with the controls provided, and double-clicking shows/hides additional genomic information, such as splicing variants. Additionally, the alignment between the initial oligo and the target is shown, highlighting the PAM motif and indicating the mismatches. By clicking any gene ID in this mini-browser, a full-featured external ENSEMBL browser opens on a separate window, so that all annotations available for that gene are accessible and the user can further navigate them. With all these panels, it is easy to see whether the predicted off-target is in a coding region, and if so in which gene, etc. All these novel features of *Breaking-Cas* help manually assessing the risk of a given off-target depending on the particular experiment performed, and complement its numeric score.

Finally, both lists, with the oligonucleotide candidates and with the targets, can be exported as text-tabulated files to be imported on a standard spreadsheet application.

CONCLUSIONS

Due to its relative simplicity, one of the main advantages of the CRISPR/Cas system for the fine editing of genomes is the possibility of applying it to any organism, including those for which no other genetic tools exist. In addition, the assessment of the potential off-targets associated to a given gRNA is crucial, especially for the promising therapeutic applications of this technique, e.g. (26). Together with ethi-

sms-2

Caenorhabditis elegans (roundworm)

Source files: Caenorhabditis_elegans.WBcel235.dna.toplevel.fa; Caenorhabditis_elegans.WBcel235.84.gtf

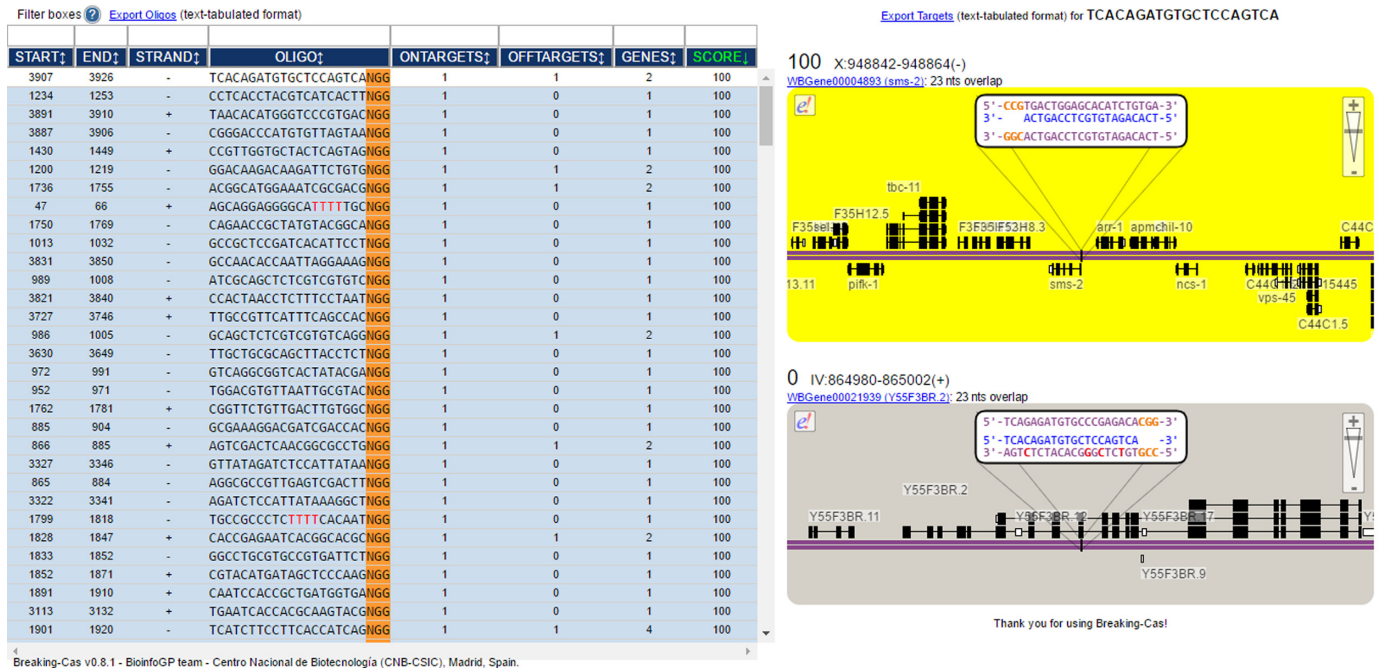


Figure 1. *Breaking-Cas* interactive results page. On the left, a list of the oligonucleotides with their associated parameters is shown, highlighting the neighbouring PAM and other interesting features, such as the poly-T stretches. Upon clicking any of these oligonucleotides, a list of its potential off-targets is presented on the right panel, as well as their intended 'on-target' (yellow background). This list is shown in the form of a mini genome browser, thus allowing the user to inspect the genomic neighbourhoods to better assess them.

cal issues, the safety of genetic modifications in humans (especially those involving germ-line heritable changes) is the key factor for eventually approving such procedures (27). Not surprisingly, the first attempt of genome-editing human embryos reported the existence of numerous off-target modifications (28). The importance of detecting off-targets suggests that simply scoring those predicted genomic sites according to some sequence-similarity-based criterion might not be sufficient. An interactive inspection of all the genomic features of the potential off-targets (genes/proteins involved, epigenetic factors, relationship with tissue specificity, etc.) is therefore crucial. Moreover, given the continuous movement in this field, programs for designing gRNAs should be flexible and adaptable for the new CRISPR/Cas systems and setups.

Consequently, we have developed the *Breaking-Cas* tool with these goals in mind, presently not available in other existing packages for the design of gRNAs. First, *Breaking-Cas* allows working with all eukaryotic genomes available in the ENSEMBL database, while other existing tools can only work with a few tens model genomes in the best cases (18). Second, it provides a user-friendly browser for a detailed inspection of the genomic neighbourhood of every single potential off-target so as to qualitatively assess them. Third, this web server allows specifying the characteristics and parameters of the nuclease to be used, although predefined set are available for several popular Cas proteins (including Cas9 and Cpf1). The system can be easily updated with

new genomes and versions released in ENSEMBL since it is based on standard file formats of this resource. Finally, the server also turned out to be 3–4 times faster than other widely used tools in the tests we made, although an exhaustive speed comparison was not performed.

While other experimental strategies for minimizing off-targets are being explored (e.g. nickases which cut concomitantly in two adjacent regions hence reducing the chances of off-targets (29)) this computational framework for predicting off-targets with the current technologies is undoubtedly useful. For the future we plan to implement in the *Breaking-Cas* workflow additional scoring systems for the oligonucleotides and off-targets, which are currently being generated based on experimental data. However, now they can be to some extent incorporated by manually modifying the parameters associated to the nuclease. We also plan to include information on restriction nucleases recognition motifs to facilitate the validation of efficiency and the identification of mutants by PCR-approaches combined with analytical restriction digestion of the targeted genomic loci. Finally, the incorporation of one of the existing approaches for evaluating the on-target efficiency would also be desirable.

ACKNOWLEDGEMENTS

We thank the members of the Computational Systems Biology for interesting discussions and suggestions.

FUNDING

Spanish Ministry for Economy and Competitiveness (MINECO) [BIO2011-25687, BIO2012-39980, BIO2014-57011-R].

Conflict of interest statement. None declared.

REFERENCES

- Sander, J.D. and Joung, J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, **32**, 347–355.
- Mojica, F.J. and Garrett, R.A. (2013) In: Rodolphe, B. and van der Oost J (eds). *CRISPR-Cas Systems. RNA-Mediated Adaptive Immunity in Bacteria and Archaea*. Springer, Berlin, Heidelberg, pp. 1–31.
- Wiedenheft, B., Sternberg, S.H. and Doudna, J.A. (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, **482**, 331–338.
- Fineran, P.C. and Charpentier, E. (2012) Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. *Virology*, **434**, 202–209.
- Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
- Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J. and Regev, A. (2015) Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, **163**, 759–771.
- Nishimasu, H., Cong, L., Yan, W.X., Ran, F.A., Zetsche, B., Li, Y., Kurabayashi, A., Ishitani, R., Zhang, F. and Nureki, O. (2015) Crystal structure of *Staphylococcus aureus* Cas9. *Cell*, **162**, 1113–1126.
- Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S. and Severinov, K. (2015) Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell*, **60**, 385–397.
- Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K. and Sander, J.D. (2013) High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.
- Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
- Chari, R., Mali, P., Moosburner, M. and Church, G.M. (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, **12**, 823–826.
- Iyer, V., Shen, B., Zhang, W., Hodgkins, A., Keane, T., Huang, X. and Skarnes, W.C. (2015) Off-target mutations are rare in Cas9-modified mice. *Nat. Methods*, **12**, 479.
- Seruggia, D., Fernández, A., Cantero, M., Pelczar, P. and Montoliu, L. (2015) Functional validation of mouse tyrosinase non-coding regulatory DNA elements by CRISPR-Cas9-mediated mutagenesis. *Nucleic Acids Res.*, **43**, 4855–4867.
- Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J. and Root, D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
- Montague, T.G., Cruz, J.M., Gagnon, J.A., Church, G.M. and Valen, E. (2014) CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.*, **42**, W401–W407.
- Heigwer, F., Zhan, T., Breinig, M., Winter, J., Brügemann, D., Leible, S. and Boutros, M. (2016) CRISPR library designer (CLD): software for multispecies design of single guide RNA libraries. *Genome Biol.*, **17**, 55.
- Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J. and Mateo, J.L. (2015) CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PLoS One*, doi:10.1371/journal.pone.0124633.
- Graham, D.B. and Root, D.E. (2015) Resources for the design of CRISPR gene editing experiments. *Genome Biol.*, **16**, 260.
- Kersey, P.J., Allen, J.E., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., Hughes, D.S., Humphrey, J., Kerhornou, A., Khobova, J. *et al.* (2013) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.*, **42**, D546–D552.
- Esvelt, K.M., Mali, P., Braff, J.L., Moosburner, M., Yang, S.J. and Church, G.M. (2013) Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods*, **10**, 1116–1121.
- Hou, Z., Zhang, Y., Propson, N.E., Howden, S.E., Chu, L.F., Sontheimer, E.J. and Thomson, J.A. (2013) Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 15644–15649.
- Cock, P.J., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–1771.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
- Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
- Arimbasseri, A.G., Rijal, K. and Marais, R.J. (2013) Transcription termination by the eukaryotic RNA polymerase III. *Biochim. Biophys. Acta*, **1829**, 318–330.
- Tabebordbar, M., Zhu, K., Cheng, J.K.W., Chew, W.L., Widrick, J.J., Yan, W.X., Maesner, C., Wu, E.Y., Xiao, R., Ran, F.A. *et al.* (2016) In vivo gene editing in dystrophic mouse muscle and muscle stem cells. *Science*, **351**, 407–411.
- Bosley, K.S., Botchan, M., Bredenoord, A.L., Carroll, D., Charo, R.A., Charpentier, E., Cohen, R., Corn, J., Doudna, J., Feng, G. *et al.* (2015) CRISPR germline engineering—the community speaks. *Nat. Biotechnol.*, **33**, 478–486.
- Liang, P., Xu, Y., Zhang, X., Ding, C., Huang, R., Zhang, Z., Lv, J., Xie, X., Chen, Y., Li, Y. *et al.* (2015) CRISPR/Cas9-mediated gene editing in human triploid zygotes. *Protein Cell*, **6**, 363–372.
- Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.