Chapter 4

Computational Prediction of CRISPR/Cas9 Target Sites Reveals Potential Off-Target Risks in Human and Mouse

Qingbo Wang and Kumiko Ui-Tei

Abstract

The clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated (Cas) system is a prominent genome engineering technology. In the CRISPR/Cas system, the RNA-guided endonuclease Cas protein introduces a DNA double-stranded break at the genome position recognized by a guide RNA (gRNA) based on complementary base-pairing of about 20-nucleotides in length. The 8- or 12-mer gRNA sequence in the proximal region is especially important for target recognition, and the genes with sequence complementarity to such regions are often disrupted. To carry out target site-specific genome editing, we released the CRISPRdirect (http://crispr.dbcls.jp/) website. This website allows us to select target site-specific gRNA sequences by performing exhaustive searches against entire genomic sequences. In this study, target site-specific gRNA sequences were designed for human, mouse, *Drosophila melanogaster*, and *Caenorhabditis elegans*. The calculation results revealed that at least five gRNA sequences, each of them having only one perfectly complementary site in the whole genome, could be designed for more than 95% of genes, regardless of the organism. Next, among those gRNAs, we selected gRNAs that did not have any other complementary site to the unique 12-mer proximal sequences to avoid possible off-target effects. This computational prediction revealed that target site-specific gRNAs are selectable for the majority of genes in *D. melanogaster* and *C. elegans*. However, for >50% of genes in humans and mice, there are no target sites without possible off-target effects.

Key words CRISPR/Cas9, Target site, Off-target gene, CRISPR direct

1 Introduction

Genome engineering using the clustered regularly interspaced short palindromic repeat/CRISPR-associated (CRISPR/Cas) system has been widely applied in recent years due to its simplicity and wide range of applicability [1–5]. The step-by-step mechanism of the CRISPR/Cas system, which is derived from the adaptive immune system of prokaryotes, is as follows [6–8]:

(a) The Cas protein, an RNA-guided endonuclease, interacts with a single-guide RNA (sgRNA) to form a Cas-sgRNA complex. The sgRNA is a short RNA artificially constructed by

Izuho Hatada (ed.), *Genome Editing in Animals: Methods and Protocols*, Methods in Molecular Biology, vol. 1630,
DOI 10.1007/978-1-4939-7128-2_4,

connecting CRISPR RNA (crRNA) with the trans-activating crRNA (tracrRNA).

(b) The Cas protein in the Cas-sgRNA complex recognizes a specific sequence motif called proto-spacer adjacent motif (PAM), 5′-NGG in the case of *Streptococcus pyogenes*, downstream of the target site. If the 20-mer sequence upstream of the PAM is homologous to the guide RNA (gRNA) spacer sequence, the gRNA spacer region pairs with the complementary strand upstream of the PAM (target sequence).

(c) After binding to the target sequence, the Cas protein in the Cas-sgRNA complex cleaves both DNA strands a few bases upstream of the PAM sequence to introduce a double-stranded break (DSB).

(d) The DSB site is repaired by error-prone non-homologous end joining (NHEJ) that often results in a small insertion or deletion. When there is a homologous template DNA, the site is repaired by homology-directed repair (HDR) that results in insertion of a specific DNA sequence.

Genome engineering using the CRISPR/Cas system allows for the introduction of DSBs to induce NHEJ or HDR in the intended genomic regions by designing a target site-specific gRNA sequence. Due to this, the CRISPR/Cas system is advantageous compared to previous genome engineering techniques based on protein engineering technology (such as transcription activator-like effector nucleases "TALENs" and zinc finger nucleases "ZFNs"), which require considerable efforts to design effective proteins [9–11].

However, the stringency of target sequence recognition by the Cas-sgRNA complex is not well understood. Previous studies have revealed that site recognition and cleavage by the Cas protein can occur even when there are gaps or mismatches between the gRNA spacer and target sequences [12–14]. A number of different mismatch patterns have been reported for such nonspecific cleavage ("off-target effect"), but the 8 or 12 nucleotides upstream of the PAM (seed region) are especially important for target site recognition. In many cases, off-target mutations happen at sites where the gRNA seed region has no mismatches but the non-seed region does [3, 15]. To knock out a specific region in the genome, such off-target effects should be avoided. Reducing such risk is especially important when we consider further application of the technology, including therapeutic applications.

In this study, we implemented a computational pipeline to design site-specific gRNAs. Our pipeline enabled evaluation of the off-target risk of each gRNA by calculating the number of seed-matched off-target candidate sequences in the entire genome and allowed for the design of off-target risk-reduced gRNAs. Using the

pipeline, the number of applicable gRNAs was gg (Subheading 3.4. also, *see* **Note 1**). The percentage of genes that had a certain number of such gRNAs was also calculated and compared between four different organisms (Subheadings 3.5 and 3.6).

2 Materials

2.1 Protein-Coding Sequence Preparation

1. All human (hg19, GRCh37 Genome Reference Consortium Human Reference 37), mouse (mm10, Genome Reference Consortium Mouse Build 38), *Drosophila melanogaster* (*D. melanogaster*) (dm3, BDGP Release 5), and *Caenorhabditis elegans* (*C. elegans*) (ce10, Washington University School of Medicine GSC and Sanger Institute WS220) mRNA sequences (protein-coding sequences) were downloaded from RefSeq, through the UCSC Genome Browser (https://genome.ucsc.edu/index.html) and stored as fasta files (Fig. 2, **step 1**).
2. The table describing the relationship between mRNA ID and gene name was also downloaded from the UCSC Table Browser by changing the "output format" to "all fields from selected table." This table was used to map each mRNA to the corresponding gene.

2.2 Software for Genome-Wide Specificity Check

The CRISPRdirect web server [16] (https://crispr.dbcls.jp) was used to conduct genome-wide investigation of off-target sites. This server investigates the entire genome of the organisms of interest for PAM-proximal 20-, 12-, or 8-mer-matched sequences, and lists all the possible gRNA sequences and the number of matched sequences.

2.3 Building Environment for Iteration and Visualization

Bash (GNU bash, version 3.2.51(1)-release) and Python 2.7.5 were used to construct a computational pipeline (Fig. 2) for iterative design of site-specific gRNAs against each gene. "Pandas" [17], an open-source tool for data analysis, was used to parse the data into tab-separated value (tsv) files. The calculation results were visualized using matplotlib [18], a python library for 2D plotting. The entire process, described in the next chapter (from gRNA search to graph visualization), was automated in a single pipeline under the environment.

3 Methods

3.1 Directory Structure Construction

Before beginning the calculation process, the directories for output file storage were structured as shown in Fig. 1 using bash scripts.

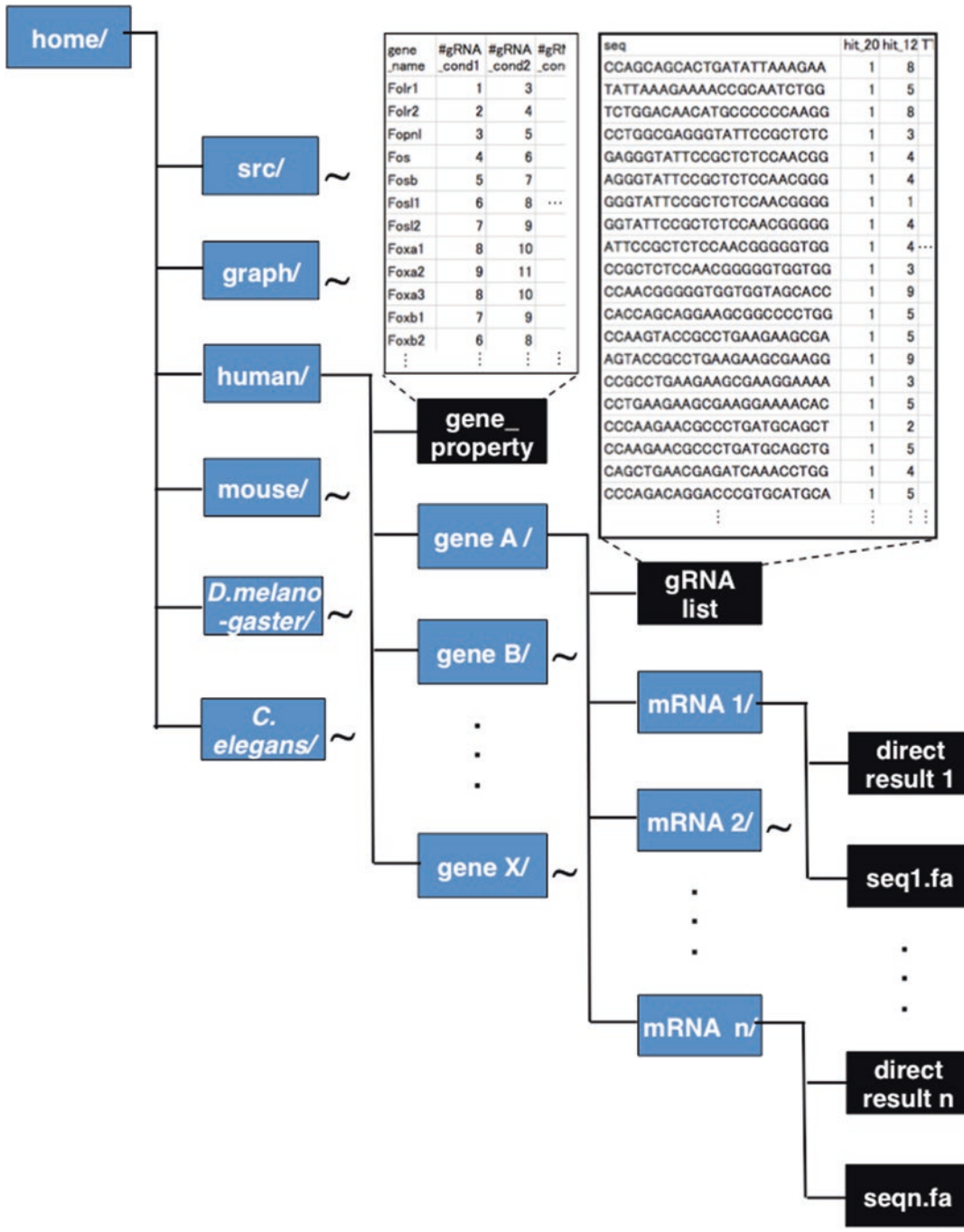

Fig. 1 The directory structure for storage of output files. The home directory contains src/ (where source codes are stored), graph/ (where the visualized results of the calculation are stored), and the directory of each organism. The directories for each of the genes were positioned under the directory of each organism, and followed by mRNA directories. The calculation summary for all the genes of a single organism (*gene_property*) is stored under the directory of each organism. Each directory of a gene contains a list of all the target site-specific gRNAs for the gene (*gRNA_list*). The results of off-target candidate search using CRISPRdirect software (*direct_result n*) and the genomic sequence of each mRNA (*seqn.fa*) are stored under the directory of each mRNA. *Blue boxes* indicate directories. *Black boxes*, tsv or fasta files. "~," the lower directories

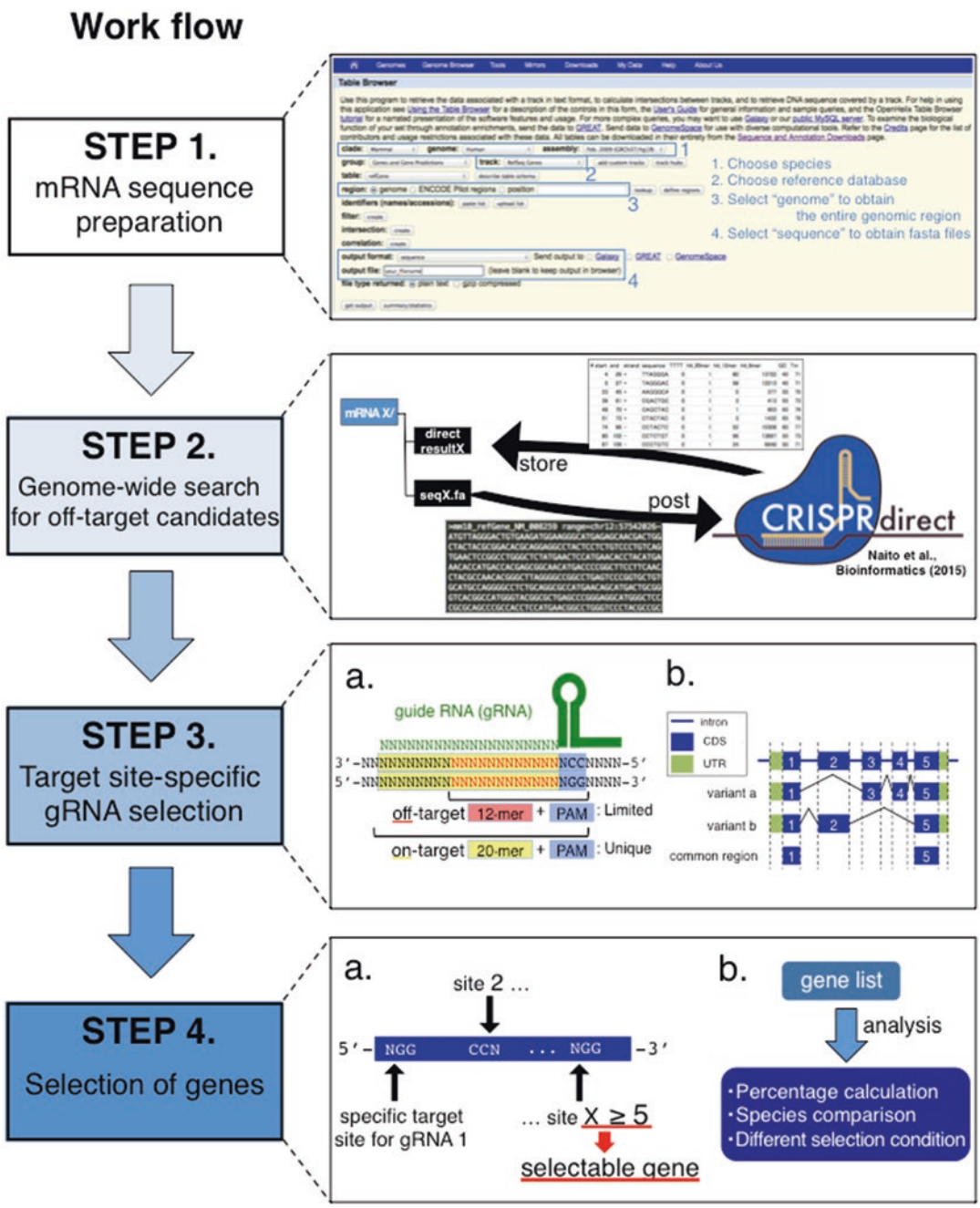

Fig. 2 Workflow of the target site-specific gRNA selection process with a minimum number of off-target hits, and genes that have more than a certain number of selectable gRNAs per gene. *STEP 1*: Preparation of mRNA sequences used. All mRNA sequences (protein-coding mRNA sequences) were downloaded through the UCSC Genome Browser (https://genome.ucsc.edu/index.html). *STEP 2:* Genome-wide search for off-target candidates using CRISPRdirect. Given the query sequence, CRISPRdirect investigates the entire genome for PAM-proximal 20-, 12-, or 8-mer-matched sequences. *STEP 3:* Target site-specific gRNA selection based on STEP2 results. (a) The schematic structure of 20-mer on-target region, 12-mer off-target region, and PAM region used for the evaluation of gRNA specificity. (b) The mRNA regions where gRNAs were designed in this study. *STEP 4:* Selection of genes containing a sufficient number of target site-specific gRNAs based on the *STEP 3* results. (a) A gene containing more than five gRNAs, for example. (b) The selection step of genes under different conditions

The workflow of the calculation pipeline is shown in Fig. 2, and the details of each step are shown below.

3.2 Genome-Wide Search for Off-Target Candidates

Each mRNA sequence, stored in a fasta format, was posted to our software, CRISPRdirect [16], using the API provided by the software (Fig. 2, **step 2**). The list of possible gRNA candidates with the specificity check results were then stored in the local directory as a tsv file.

3.3 Target Site-Specific gRNA Selection

The target site-specific gRNA candidates satisfying the following conditions were selected:

(a) Uniqueness of the target sequence among the entire genome: No perfect match other than the target site (20-mer sequence + PAM) was allowed (Fig. 2, **step 3a**).

(b) Limitation of the number of possible off-target sites with seed (PAM-proximal 12-mer sequence) complementarity among the entire genome: Only a limited number of seed-matched sites was allowed (Fig. 2, **step 3a**).

(c) Absence of a "TTTT" stretch: No "TTTT" stretch (more than three sequential Ts) was allowed in the gRNA sequence.

(d) The target site was positioned in the common exons among all the transcription variants of a target gene (Fig. 2, **step 3b**).

Conditions (a) and (b) were applied to reduce the risk of off-target cleavage, and (c) was applied to avoid the termination of gRNA transcription by RNA polymerase III (the "TTTT" stretch is a known RNA polymerase III termination site). Condition (d) was needed to disrupt the target gene expression regardless of the transcription pattern (*see* **Notes 2–4**).

The gRNAs that did not satisfy any one of four conditions shown above were removed using a filtering operation in Pandas, and the gRNAs that simultaneously satisfied all four conditions were defined as "target site-specific gRNA candidates."

3.4 Calculating the Number of Target Site-Specific gRNA Candidates per Gene

Given the list of target site-specific gRNA candidates selected in Subheading 3.3, the number of selectable target site-specific gRNA candidates for each gene was calculated in four organisms: human, mouse, *D. melanogaster*, and *C. elegans*. The cumulative fraction distribution of genes (*y* axis) as a function of the number of selectable gRNAs (*x* axis) was shown in two different cases: (1) the case where an unlimited number of seed-matched off-target candidate sites was allowed (Fig. 3a), and (2) no seed-matched site other than the target site was allowed (Fig. 3b). In the former case, *C. elegans* had relatively fewer gRNAs compared to the other three organisms (Fig. 3a). However, the cumulative curves for *C. elegans* and *D. melanogaster* were similar to each other but different from humans and mice when no seed-matched site other than the target site was allowed (Fig. 3b).

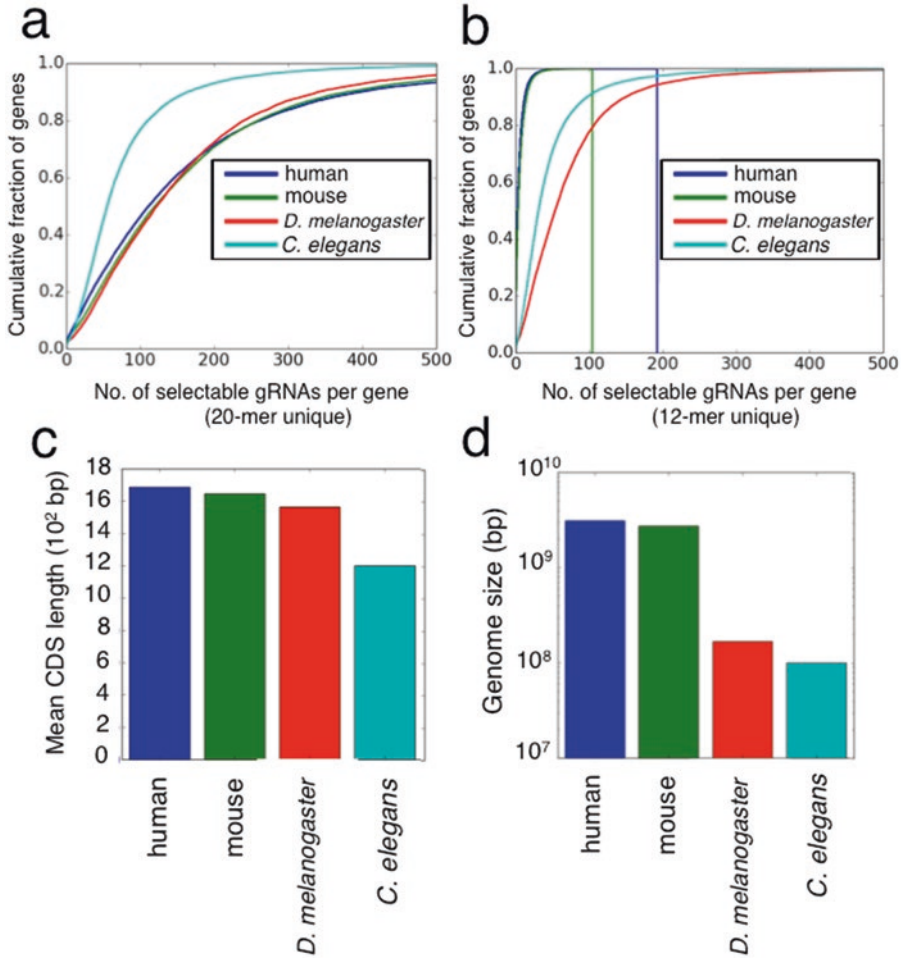

Fig. 3 Comparison of the number of target site-specific gRNA candidates per gene, mean CDS length, and genome sizes of human (*blue*), mouse (*green*), *D. melanogaster* (*red*), and *C. elegans* (*light blue*). (**a**), (**b**) Cumulative distributions of the genes as a function of the number of gRNAs with a 20-mer (**a**) or 12-mer (**b**) completely matched unique site per gene. The *x* axis indicates the number of designable gRNAs per gene, and the *y* axis is the cumulative fraction of genes. The *blue* and *green vertical lines* indicate the cumulative fractions of human and mouse genes reached to 1.0, respectively, in (**b**). The mean CDS length (c) and genome sizes (**d**) of four different organisms used in this research. For CDS length, the mean length of transcription variants was calculated for each gene, and then the mean CDS length was calculated for each species

The results indicated that at least two different parameters may contribute to specificity. First, in our selection conditions, since the target region that we designed gRNAs for was restricted to the coding sequence (CDS), the relatively short average CDS length (Fig. 3c) was likely the reason for the lower abundance of gRNAs in *C. elegans*, especially in the case where an unlimited number of seed-matched sites was allowed. Second, the genome sizes of these four organisms differ by double-digits, and range from 10^9 to 10^7 bp (Fig. 3d). Since the off-target search was performed on a

genome-wide scale, the off-target hits may increase based on genome size. The sharp decrease in abundance of target site-specific gRNAs in human and mouse genomes could be attributed to their large genome sizes. Thus, it may be difficult to select a large amount of target site-specific gRNAs without off-target candidates in organisms with large genomes.

3.5 Selection of Genes with More than Five Target Site-Specific gRNAs for Each Gene

Even when a target site-specific gRNA with a perfectly matched target sequence is selected, there is no guarantee that the target site would be cleaved by the gRNA [19, 20] since the conditions that determine gRNA sequence functionality are not well known (*see* **Note 5**). One of the possible procedures to efficiently induce DSB is to design multiple gRNAs that target different sites of the same gene. For example, in recent large-scale knock-out experiments, five gRNAs were designed for each gene, and numerous gene sets corresponding to fundamental biological processes in mammalian cells were identified [21, 22]. In an analogous way, we used the genes for which more than five target site-specific gRNAs were selectable, and defined them as "5 ≤ gRNA" genes (Fig. 2, **step 4a**). The percentage of such genes was calculated based on the results shown in Fig. 3a and b. Since gRNA abundance is highly dependent on the number of off-target hits that we allowed, we gradually changed the maximum number of seed-matched off-target hits per gRNA and counted the selectable gRNA abundance for each condition. The results revealed that at least five gRNA sequences with only one perfectly complementary site in the whole genome could be designed for more than 95% of genes for all four organisms when neglecting possible off-target risks with 12-mer matched sequence (Fig. 4a). However, the percentage of such genes decreased with the decreased number of possible off-target sites when 12-mer matched sequences were allowed, especially for human and mouse genomes. The results suggest that when considering seed-matched off-target risks, knock-out screening experiments using the CRISPR/Cas system would be relatively more feasible for *D. melanogaster* and *C. elegans*, compared to humans and mice.

3.6 Selection of "N ≤ gRNA" Genes

For the human genome, which had the lowest percentage, the minimum number of target site-specific gRNA candidates was changed for each selected gene (i.e., we selected "$N \leq$ gRNA" genes, where N is not restricted to 5, ranging from 1 to 10.) to examine the changes in percentage of selectable genes. The results (Fig. 4b) clearly show that there is a trade-off relationship between the minimum gRNA-abundance threshold N (y axis) and the maximum number of off-target candidate sites allowed per gRNA (x axis): If we strictly select gRNAs that have fewer 12-mer matched off-target candidate sites, we can only select a few gRNAs per gene. In other words, the N parameter must be relatively small to maintain a high percentage of selectable genes.

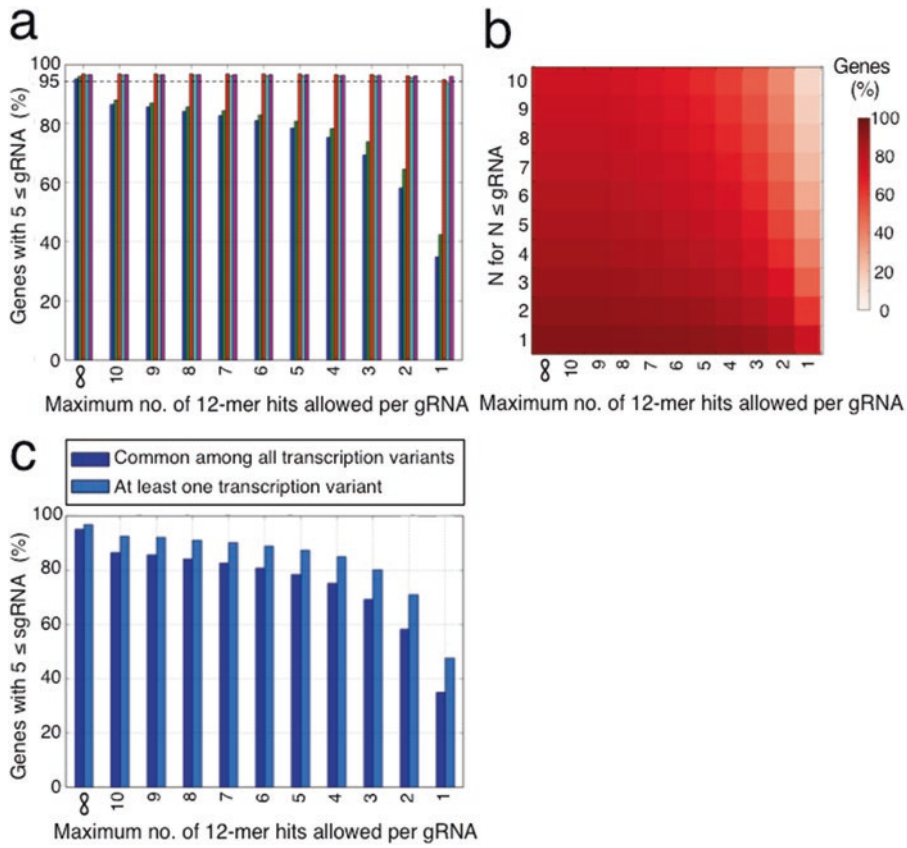

Fig. 4 The percentage of "$N \leq$ gRNA" genes. (**a**) The percentage of genes with more than five target site-specific gRNAs in different organisms (*y* axis), with different numbers of 12-mer matched sequences allowed (*x* axis). The *colors of the bars* correspond to specific organisms: human (*blue*), mouse (*green*), *D. melanogaster* (*red*), and *C. elegans* (*light blue*). Note that the "∞" column of (**a**) corresponds to the intersection of $x = 5$ and each curve in Fig. 3a, and the "1" column of (**a**) to that in Fig. 3b. (**b**) Heat map of the percentage of "$N \leq$ gRNA" genes (*N* ranging from 1 to 10) in the human genome, with different numbers of possible off-target sites with 12-mer matched sequences allowed. *Y* axis, *N* for "$N \leq$ gRNA" gene. *X* axis, the number of 12-mer matched sequences allowed. The *color gradient* indicates the percentage of genes in each condition. (**c**) Comparison of the percentage of genes with more than five target site-specific gRNAs, with and without considering the splice variants of each gene. The *dense blue bar* indicates the result from selecting gRNAs targeting common regions among splice variants, and the *light blue bar* is the result of selecting gRNA not limited to the common region

Overall, our study provided a computational pipeline to design target gene-specific gRNAs, and we quantitatively compared the number of selectable candidate gRNAs per gene among four different organisms under differing selection conditions. This pipeline may be applied to data sets of different organisms that were not analyzed in this research.

4 Notes

1. The time needed to complete the calculation depended mainly on the speed of web communication, rather than the algorithm itself. When the human genome (hg19, 41,845 mRNAs, and 19,132 genes) was used, the whol0e process required 48 hours to complete (Subheading 3.3, Fig. 2, **step 2**).
2. The gRNAs that target exon-exon junctions, the regions without continuous genomic DNA sequences, are eliminated by the CRISPRdirect algorithm (Subheading 3.2), and therefore were not included in this study.
3. In this research, the off-target candidate sites were estimated based on the complementarity to 12-mer, rather than 8-mer, seed sequences to increase prediction accuracy as shown in Subheading 3.3, **step (b)** (Fig. 2, **step 3a**). Although seed match is important for target recognition, different patterns of off-target hits are reported [13]. For example, cell type, chromatin state, or SNP presence are also important factors [23–25]. Revealing more in-depth mechanisms of target site recognition to reduce off-target effects could also be meaningful.
4. If it is not necessary to target the common regions of all transcription variants of a gene, it is possible to eliminate condition (d) in Subheading 3.3 (Fig. 2, **step 3b**). The alteration can be easily achieved by making a minor change in the Python script. Figure 4c shows the comparison of the percentages in the human genome with and without condition (d).
5. As described in Subheading 3.5, the major reason we designed multiple gRNAs for a single gene is because the editing efficiency of each gRNA in the CRISPR/Cas system is not known. Therefore, we did not exclude gRNAs that are likely to show no or weak genome engineering function in this research. Additional research may improve our understanding of both the efficiency and off-target risks of the CRISPR/Cas system, leading us to select truly target-specific and highly efficient gRNAs.

Acknowledgment

We thank Dr. Yuki Naito for valuable discussion and technical advice. The English in this document has been checked by at least two professional editors, both native speakers of English. This work was supported by the grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan to K.U.-T.

References

1. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM (2013) RNA-guided human genome engineering via Cas9. Science 339:823–826
2. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PE, Wu X, Jiang W, Marraffini LA, Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823
3. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J (2013) RNA-programmed genome editing in human cells. elife 2:e00471
4. Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineerging with CRISPR-Cas9. Science 346:1258096
5. Mohr SE, Hu Y, Ewen-Campen B, Housden BE, Viswanatha R, Perrimon N. (2016) CRISPR guide RNA design for research applications. FEBS J 283(17):3232-3238
6. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 506:62–67
7. Anders C, Niewoehner O, Duerst A, Jinek M (2014) Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature 513:569–573
8. Amital G, Sorek R (2016) CRISPR-Cas adaptation: insights into the mechanism of action. Nat Rev Microbiol 14:67–76
9. Gaj T, Gersbach CA, Barbas CF 3rd (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31:397–405
10. Sternberg SH, Richter H, Charpentier E, Qimron U (2016) Adaptation in CRISPR-Cas system. Mol Cell 61:797–808
11. Pattanayak V, Guilinger JP, Liu DR (2014) Determining the specificities of TALENs, Cas9, and other genome-editing enzymes. Methods Enzymol 546:47–78
12. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le JP, Aryee MJ, Joung JK (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol 33:187–197
13. Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, Wile BM, Vertino PM, Stewart FJ, Bao G (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res 42:7473–7485
14. Cradick TJ, Fine EJ, Antico CJ, Bao G (2013) CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. Nucleic Acids Res 41:9584–9592
15. Sender JD, Joung JK (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 32:347–355
16. Naito Y, Hino K, Bono H, Ui-Tei K (2015) CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics 31:1120–1123
17. McKinney W (2011) Pandas: a foundational python library for data analysis and statistics. http://www.slideshare.net/wesm/pandas-a-foundational-python-library-for-data-analysis-and-statistics
18. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95
19. Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS, Brown M, Liu XS (2015) Sequence determinants of improved CRISPR sgRNA design. Genome Res 25:1147–1157
20. Farasat I, Salis HM (2016) A biophysical model of CRISPR/Cas9 activity for rational design of genome editing and gene regulation. PLoS Comput Biol 12:e1004724
21. Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera MC, Yusa K (2014) Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol 32:267–273
22. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 343:84–87
23. Horlbeck MA, Witkowsky LB, Guglielmi B, Replogle JM, Gilbert LA, Villalta JE, Torigoe SE, Tijan R, Weissman JS (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. elife 5:e12677
24. Bae S, Kweon J, Kim HS, Kim JS (2014) Microhomology-based choice of Cas9 nuclease target sites. Nat Methods 11:705–706
25. Moreno-Mateos MA, Vejnar CE, Beaudoin JD, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ (2015) CRISPRscan: designing highly efficient sgRNAs for CRISPR/Cas9 targeting in vivo. Nat Methods 12:982–988