Online High-throughput Mutagenesis Designer Using Scoring Matrix of Sequence-specific Endonucleases

Dayong Guo[1,2*], Xiaojing Li[1], Pan Zhu[1], Yanrong Feng[1], Juan Yang[1], Zhihong Zheng[3], Wei Yang[3], Enuo Zhang[3], Yang Yu[3], Shenglai Zhou[3], Hongyu Wang[1,2]

[1]Syncbiotech Co. Ltd, 12 Xuefu Rd, High-tech District, Nanjing, China
[2]Model Animal Research Center, Nanjing University, Nanjing, China
[3]Key Laboratory of Transgenic Animal Research, Laboratory Animal Center of Liaoning Province, China Medical University, Shenyang, China

Summary

CRISPR Cas9 and other sequence-specific endonucleases are fundamental genome editors supporting gene knockout and gene therapy. A speedy and accurate computational allele designer is required for a high through-put gene mutagenesis pipeline using these new techniques.

An automatic system, Cas9 online designer (COD), was created to screen Cas9 targets and off-targets, as well as to provide gene knockout and genotyping strategies. A gene knockout rat model was successfully created and genotyped under the direction of this online system confirming its ability to predict real targets and off-targets. Gene knockout strategies to mutate 72 rat cytochrome P450 genes were designed instantly by the system to demonstrate its high-throughput efficiency. Also, the system used an off-target scoring matrix which can be applied to any sequence-specific genome editing tools besides Cas9.

The COD system (http://cas9.wicp.net) has established a speedy, accurate, flexible and high through-put computational gene knockout pipeline supporting the sequence-specific endonuclease induced mutagenesis.

1 Background

An RNA-guided DNA endonuclease system, CRISPR Cas9, has been one of the most popular genome editing tools practiced from bacteria to mammals, both in cultured cells and animal models [1-6]. The advantage of CRISPR Cas9 system lies in its customised sequence-specific endonuclease activity determined by a short guide RNA template containing a 5’ 17-20 nt fragment complementing its double stranded DNA substrate. However, off-targets slightly different from the designed sequence could also be cleaved [7-9]. The wild type CRISPR Cas9 cleaves both strands of its DNA substrate. Certain mutants of CRISPR Cas9 enzyme, Cas9 nickases, merely produce a single cleavage on one strand of its double strand DNA substrates inducing fewer off-target mutations [10-12]. It has been reported that both non-homologous end joint repair and homologous recombination are enhanced near the cut sites [13-17]. These discoveries unveiled very useful genome editing approaches leading to the ultimate goal of precise, rapid and high through-put knock-outs, knock-ins and other genetic modifications in medical and industrial pipelines.

After Cas9 treatment, mutations on the expected target or potential off-targets have to be confirmed by sequencing. Unidentified off-targets could severely compromise the phenotype and mechanism studies on the mutant cell or animal models. Next generation sequencing

[*] Corresponding Email: guodayong@gmail.com

(NGS) upon the whole genome or transcriptome could be a comprehensive solution but slow and costly [10]. A straightforward and economical approach is Sanger sequencing on the most likely targets. Several attempts have been done on the computational prediction of Cas9 targets in a background species in the past two years. Originally, selecting Cas9 targets was simply based on any downstream 5’ NGG 3’ photospacer adjacent motif (PAM) next to the 20 nt target sequence without evaluation of potential off-targets [3,18,19]. After the occurrence of off-targeting was discovered, BLAST-like query tools were created to pickup hits sorted by the number of mismatches [20,21]. Later, another form of Cas9 PAM, 5’ NAG 3’, was discovered resulting in an expanded off-target scope [11,22]. Further comprehensive assays quantitatively revealed a complicated, skewed correlation between each nucleotide of the guide RNA to the activity of the Cas9 endonuclease [7,9]. It appears that the matches on the 3’ side of the guide RNA are more stringent than on the 5’ side since more detected off-targets contained mismatches on the 5’ end. Also, it seems that off-targets fluctuate in experiments conducted on different cell or animal models, using distinctly prepared guide RNA with ranged concentrations or lengths. Besides CRISPR Cas9 system, other nucleic acid cleaving tools also showed strong genome-editing potentials, such as the Ago DNA-directed endonuclease family [23,24], chemistry-based artificial DNA cutter [25,26], zinc finger nucleases [27], TALE nucleases [28], or combination of these tools [29]. Each of these sequence-specific cleavers requires its own tailored targeting parameters to perform computational estimations. It is a great significance to create a one-size-fits-all scoring platform customizable to various lengths and nucleotide positions required by diverse endonucleases in different experiments.

Massive gene knockout studies are essential foundations of medical researches and biotechnologies. The multi-national Knockout Mouse Project (KOMP) was an iconic high through-put gene knockout pipeline based on the Nobel Prize winning technique of homologous recombination, as well as on an efficient computational allele designer [30,31]. KOMP and its following projects not only explored the function of each gene, but also proposed tremendous amount of novel therapeutic pathways and drug targets. However, restricted by the nature of bacterial artificial chromosome (BAC) homologous recombination, targeting murine genes with high SNPs was hardly successful, nor was targeting the clustered kindred genes sharing highly homologous flanking regions. On the other hand, embryonic stem cells (ESC) and BAC libraries of other species are not as abundant as mouse’s. For instance, rats have been the government-designated models for drug metabolism and toxicity tests because of the physiological similarity between rats and humans [32-34]. However, genetically modified rat models were less widely practiced than mouse models limited by the absence of rat’s resources to do homologous recombination, especially for the structurally complicated genes (e.g. the cytochrome P450 family). New genome editing techniques, such as CRISPR Cas9, TALE nucleases, and zinc finger nucleases, unlocked alternative approaches to modify these untouched genes and species despite their SNP, cluster, BAC or ESC conditions. However, it would be extremely tedious to manually select proper Cas9 targets for each gene, especially for those homologous ones. A speedy and accurate computational gene mutagenesis designer is absolutely required for a high through-put pipeline using these new techniques.

2 Methods

The COD (Cas9 Online Designer) system is a computational platform supporting Cas9 induced gene mutagenesis. Its computational core was built upon Perl 5.18.1 from www.perl.org, and Bioperl 1.6.1 from www.bioperl.org following their online instructions. Genomic sequences and annotations of transcripts, exons and SNPs were downloaded from

Ensembl [35] and UCSC genome browsers [36], and made into searchable databases by a locally installed BLAST+ 2.2.28 [37]. Gene ID conversions were performed using UCSC Table Browser [38], Biomart 0.7 [39] and Microsoft Access 2003. PCR and sequencing primers were designed using Primer-BLAST [40]. Sequence alignments were performed using CLUSTALW2 [41]. Web hosting was enabled by XAMPP 3.2.1 APACHE, on a local server (http://cas9.wicp.net) or in Amazon Web Services cloud (http://ec2-54-186-84-183.us-west-2.compute.amazonaws.com). The local workstation was configured as Intel Corei5-3337U CPU @ 1.8 GHz and 4 GB RAM with Microsoft Windows 8. The Amazon cloud server was configured as Intel Xeon CPU @ 2.5 GHz and 1 GB RAM with Microsoft Windows 2012 Server R2. The serial components of COD in Figure 1 are explicitly described as below.

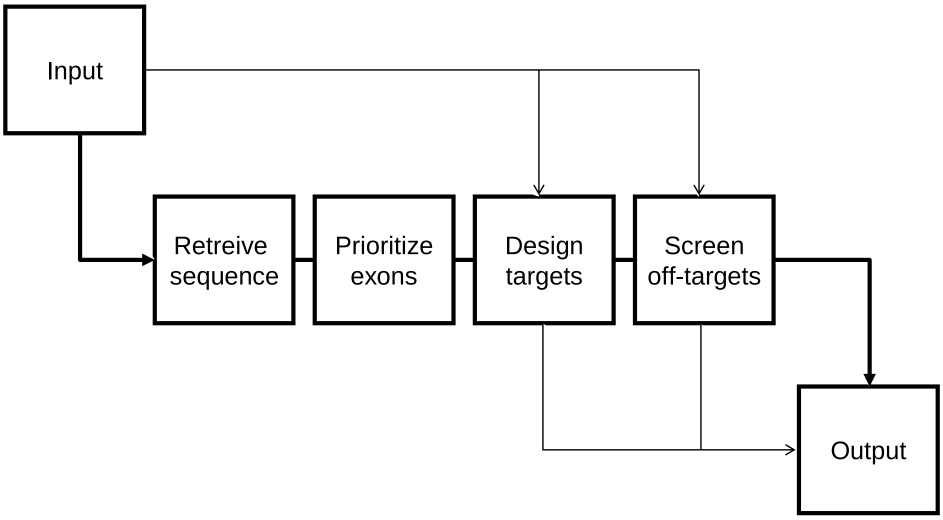


Figure 1: The workflow of COD system. The COD system provided an automatic pipeline of mutagenesis using sequence-specific endonucleases, e.g. Cas9. Given an Entrez gene ID, gene knockouts can be processed through the workflow indicated by the thick arrows. Also, any user defined DNA sequence can be processed to design Cas9 targets and analyze off-targets as indicated by the thin arrows.

2.1 Basic Cas9 designer

The first functional component created in COD was screening of Cas9 candidates from an input DNA sequence. Any fragment containing a downstream PAM was considered as a candidate. PAM were set to be either “NGG” or “both NGG and NAG” since the 5’ NGG 3’ was considered as the cardinal PAM while the 5’ NGG 3’ as the secondary PAM of lower activity [11,22]. The length of Cas9 targets excluding PAM was set to be 17 to 20 nt since recent studies suggested shorter targets were cleaved as well as 20 nt targets [42]. To estimate on- and off-targets, a BLAST search within a user-selected background genome was conducted. User’s off-target options were briefly proposed as 12 (the lowest stringency with the most noise), 15, 17 or 20 nt (the highest stringency with the least noise) identical to the 3’ part of the ideal target. This setting was consistent with the fact that more off-targets contained mismatches on the 5’ side [7,9,10]. Only the 3’ side identical BLAST hits were counted. It was computational more efficient to ignore the BLAST alignments with 3’ mismatches. The output files included a graphic view and a genbank file labelled with Cas9 targets as misc_feature. Each feature elucidated the number of off-target hits in genome, cut site position, and direction of the Cas9 candidate. Users can pick the candidates with the fewest off-targets as final choices. Detailed instructions on inputs, sample outputs, result interpretations, nickase selection strategy, and background genomes were listed online by clicking “Design Cas9” on the COD website (http://cas9.wicp.net or Amazon cloud).

2.2 Off-target screening

The second component accomplished by COD was the thorough prediction and analysis of Cas9 off-targets. Given a user defined 17 to 20 nt Cas9 target excluding PAM, its off-target analysis was processed in two steps. Firstly, BLAST alignments were performed using the user's input sequence as query against a selected species as genome database. Secondly, each BLAST hit was rated using a scoring matrix. The off-target scoring matrix accommodated both the number and the position of mismatches between a perfectly matched Cas9 target and its off-targets. Based on previously published correlations between the positions of mismatches versus the activity of Cas9 endonuclease, a set of off-target scoring matrix was used assuming the activity of a perfect target to be 100%. From 5' to 3' of a 23 nt input including the PAM of NGG, mismatches on position 1 to 23 were respectively weighed as (0.95 0.95 0.9 0.8 0.8 0.7 0.7 0.7 0.7 0.7 0.7 0.7 0.4 0.2 0.2 0.2 0.2 0.2 0.2 0.2 1 0 0) approximately derived from our empirical data, Hsu, Fu and Pattanayak's papers [7-9]. Each score can be understood as the remaining endonuclease activity in case of a mismatch on that position. Obviously, a score of 0 meant no mismatch allowed on the position; a score of 1 meant complete freedom of variation on that position. The scoring matrix highlighted flexibility on the 5' positions, and restriction on the 3' positions near the PAM. The Cas9 target containing a PAM of 5' NAG 3' was assumed about 40% as active as that of PAM 5' NGG 3' [11,22]. If shorter Cas9 targets (22~20 nt including PAM) were loaded, the scoring matrix automatically trimmed off the 1~3 scores on the 5' end. The overall score of a BLAST hit was the mathematical product of all scores on mismatched positions assuming the impacts of mutations on different positions were independent events. All potential targets and off-targets which scored higher than a user defined threshold were listed in a csv output table. Each row in the output table included an alignment between the off-target and the user's input, genome coordinates of the off-target, related outside links, and an off-target score. A higher score meant more likely to be cleaved by the user defined Cas9. The outside links pointed to the corresponding genomic sequence containing -/+ 300 bp flanking regions in genbank format, and a graphic map view of the sequence. All features of transcriptions, CDS, SNPs, repetitive regions, etc. can be retrieved to facilitate further genotyping confirmations, such as PCR and sequencing. Instructions on off-target analysis were listed on COD website by clicking "Off-targets".

A more advanced customized off-target analysis was adapted to any non-standard Cas9 experiment, or even applicable to sequence-specific endonuleases other than Cas9. The scoring matrix can be tailored into any user defined recognition site. Using an arbitrary sequence and a set of scores corresponding to each position of the sequence, a user can retrieve all potential off-targets in a given genome with score and genomic coordinates. The customized off-target analysis was posted on the COD website under "Customized Off-targets".

2.3 Cas9 gene knockout pipeline

The core function of COD was to generate Cas9 gene knockout designs automatically. The whole process was made of several steps assembling previously described components. 1) A genomic sequence was retrieved according to the user's Entrez gene ID; 2) Coding exons of the target gene were prioritized by their frequency of appearances among its various transcripts, starting from the 5' upstream to 3' downstream; 3) Prioritized coding exons were streamed through the Cas9 designer to generate output graphic views, Cas9 genbank files and summaries for user's review; 4) Off-targets were predicted and genotyping strategies were prepared for any selected Cas9 candidates. Customizable parameters included the minimum length of exons, the number of prioritized exons to design Cas9, plus previously described parameters during the automation. Currently, step 1) to 3) can be processed in a batch mode for multiple genes of several species. The whole workflow of automatic gene knockout pipeline was accessible online by clicking “Gene Knockout Pipeline” on the COD website.

2.4 Mutant rat model created by COD

Following the Cas9 target and off-target estimations analyzed by COD, a rat model of mutant *Tnfrsf1a* was generated and genotyped. A unique Cas9 target on rat genome was designed by COD, transcribed into gRNA, and co-injected with Cas9 mRNA into rat embryo as previously published [43-45]. The embryos were transplanted into surrogate females to deliver the F0 generation. Using COD automatically proposed target and off-target sequences with flanking regions, pairs of primers were designed to PCR and sequence the F0 rats. In comparison, off-targets predicted according to the number of mismatches by a previously published online tool, Cas-OFFinder [20], were manually retrieved from Ensembl, and genotyped by PCR and sequencing. Successful Cas9 induced mutations were confirmed by three criteria: 1) novel mutations confirmed by sequencing their PCR products and TA-clones; 2) mutations within +/-100 bp flankings of the expected target or off-target, and at least 50 bp away from the primers; 3) excluding any published SNP or other variations in wild type. TA-clones of the PCR products were constructed using Thermo Scientific CloneJET PCR Cloning Kit following its manual. Sanger sequencings were done by GenScript Inc using corresponding PCR primers. Construction and transcription of gRNA, mRNA and microinjection into rat embryos of SD background were done by Syncbiotech (http://www.syncbiotech.com/) in collaboration with China Medical University, Shenyang, China.

2.5 High-throughput gene knockout designs of rat P450 genes

To demonstrate the accuracy, speed and flexibility of the COD gene knockout design pipeline, all rat cytochrome P450 genes were automatically processed in batch through COD. The Entrez gene IDs of rat P450 genes were filled in the page “Gene Knockout Pipeline” on the COD website using default parameters. Output files were genbank files of each gene, prioritized coding exons, and labelled Cas9 targets.

2.6 Merging off-target analysis into Cas9 designer (COD2)

For user’s operational convenience, the off-target analysis was integrated into Cas9 designer to build COD2. User’s input and choices included a query sequence, a background species, the length of Cas9 target (17 to 20), and a threshold of minimum off-target score. Both NGG and NAG were considered as PAM when estimating off-targets. Two output files were created consisting of a genbank file with Cas9 targets labelled as misc_feature, and a .csv table of all off-targets for each designed Cas9 target. Each misc_feature contained a sum of identical scores (SIS) from off-targets identical to the candidate Cas9 target, a sum of non-identical scores (SNS) from off-targets similar but not identical to the candidate Cas9 target, a position of cut-site, and the direction of the Cas9 target. For example, a misc_feature of “Cas9: SIS=1: SNS=8.15: Cut=17 <” meant 1) there was only one copy in the genome identical to the Cas9 target which was itself; 2) the non-identical off-targets scored 8.15 in total in the genome for the Cas9 target; 3) the Cas9 target should cut at position 17 bp; and 4) the Cas9 target was in reversed direction. The output .csv table of off-targets included off-target score, alignment, sequence and map view links of off-targets for each designed Cas9 targets. Detailed instructions were accessible online by clicking “COD2” on the website (http://cas9.wicp.net).

The integration of off-target analysis into Cas9 designer may decelerate the computational speed. Therefore, a detailed comparison among the basic COD, COD2, and a popular MIT

CRISPR designer (http://crispr.mit.edu) [8] was conducted upon their computational time lapses using the same two input sequences from the demos of MIT CRISPR website. Top-ranked (top 6 out of totally 23 or 22) Cas9 targets designed by MIT CRISPR for input 1 and 2 were selected and verified whether they were reproduced by the basic COD or COD2 in their top 6 designs.

Off-targets estimated by MIT CRISPR or COD2 were also compared. Two of the best targets designed by MIT CRISPR were used as the seeds to estimate off-targets. For each target, the three most likely off-targets estimated by MIT CRISPR were compared to the three most likely off-targets estimated by COD2. The basic COD did not analyze any off-target, therefore, was excluded in the off-target comparison.

3 Results

The COD system automatically accorded the whole assembly line from the beginning of genomic sequence preparation, coding exon prioritization, Cas9 target screening and off-target prediction, to the finish line of genotyping strategies. To present, genomes of 21 species have been included into Cas9 target and off-target databases requested by users from Europe, Asia, North and South America. For each month, nearly a thousand Cas9 designs have been generated by COD to serve global users.

Results from several assignments performed by COD are listed below. To practice the workflow of COD gene knockout platform, a *Tnfrsf1a* knockout rat model was created successfully following instructions from COD. Also, a batch of rat P450 gene knockouts was designed successfully through the knockout pipeline to demonstrate the accuracy and speed of the automation. At last, an example of customized off-target scoring matrix was shown. For user's convenience, the off-target analysis was integrated into Cas9 designer successfully, and tested with two sample inputs.

3.1 *Tnfrsf1a* knockout rat

Entrez gene ID of 25625 was used as input on the page of "Gene Knockout pipeline" to design rat *Tnfrsf1a* knockout allele keeping all other parameters as default. The process took 40 seconds to produce results of a featured genbank sequence of *Tnfrsf1a* "25625.gb", genbank files of the prioritized exons containing designed Cas9 candidates "OUT-rat25625CDS0.gb", "OUT-rat25625CDS1.gb", "OUT-rat25625CDS2.gb", and a detailed pdf log file "25625 KO log.pdf" recording the whole process. Among the designed Cas9, one of the unique targets in rat genome, 5'-cagcagatggaattattctt-3', was picked from the exon 2 of "OUT-rat25625CDS0.gb" for further *Tnfrsf1a* knockout steps. All of these COD produced gene knockout design files were compacted into one zip file of "*Tnfrsf1a* KO Design.zip" as additional file 1.

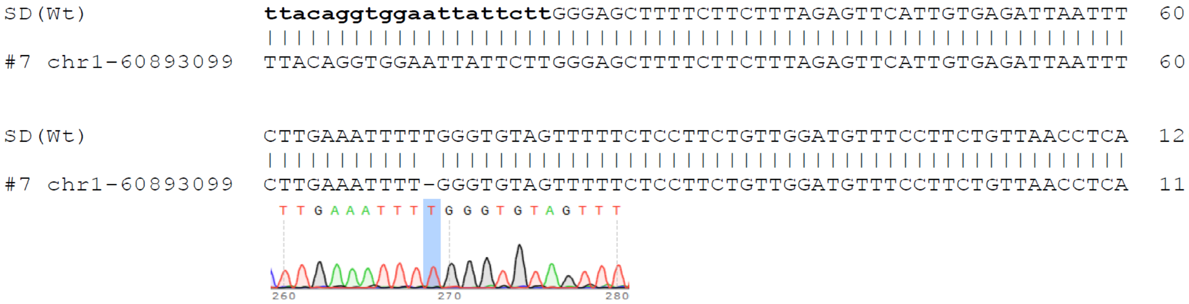
After transcription and microinjection of the selected gRNA targeting 5'-cagcagatggaattattctt-3' and Cas9 mRNA into 139 rat embryos, 122 survived embryos were embedded into 6 female surrogate rats. Thirty one F0 rats were born and genotyped by PCR and sequencing on the expected target of *Tnfrsf1a* exon 2. Five (#6, #7, #10, #23 and #31) showed mutations within exon 2 near the Cas9 target of 5'-cagcagatggaattattctt-3' as shown in Figure 2 and additional file 2 "Rat *Tnfrsf1a* Seq.zip". To evaluate the reality of computational predictions, the top seven off-targets predicted by COD (based on the highest scores) or Cas-OFFinder (based on the fewest mismatches) were sequenced taking rat #7 as an example. There was no overlapped subset shared between the off-targets predicted by COD vs. the off-targets predicted by Cas-OFFinder. A novel mutation of 1 bp deletion was detected near the most-likely off-target site predicted by COD in rat #7 at Chr1:60893099 (Figure 3). No novel mutation was detected among the other six COD predicted potential off-target sites. In contrast, there was no novel mutation detected in each of the seven off-targets predicted by Cas-OFFinder based on the number of mismatches. Detailed off-target prediction and sequencing files are listed in additional file 3 “Rat off targetsts.zip”. Primers were also stated.

```
Wt              GTGCTCCTGGCTCTGCTGATGGGGATACACCCGTCAGGGGTCACCGGACTGGTTCCTTCT 60
#1              GTGCTCCTGGCTCTGCTGATGGGGATACACCCGTCAGGGGTCACCGGACTGGTTCCTTCT 60
#6              GTGCCCCTGGCTCTGCTGATGGGGATACACCCGTCAGGGGTCACCGGACTGGTTCCTTCT 60
#7              GTGCTCCTGGCTCTGCTGATGGGGATACACCCGTCAGGGGTCACCGGACTGGTTCCTTCT 60
#10             GTGCTCCTGGCTCTGCTGATGGGGATACACCCATCAGGGGTCACCGGACTGGTTCCTTCT 60
#23             GTGCTCCTGGCTCTGCTGATGGGGATACACCCGTCAGGGGTCACCGGACTGGTTCCTTCT 60
#31             GTGCTCCTGGCTCTGCTGATGGGGATATACCCGTCAGGGGTCACCGGACTGGTTCCTTCT 60
                **** ********************** **** ***************************

Wt              CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCCCATCCAaag 120
#1              CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCCCATCCAAAG 120
#6              CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCCCAT------ 114
#7              CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCCCATCCAAAG 120
#10             CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCCCATC----- 115
#23             CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCGCATCCAAAG 120
#31             CTTGGTGACCGGGAGAAGAGGGATAATTTGTGTCCCCAGGGAAAGTATGCCCATCCAAAG 120
                ************************************************** ***

Wt              aataattccatctgctgCACCAAGTGCCACAAAGGTAGGAGACA 164
#1              AATAATTCCATCTGCTGCACCAAGTGCCACAAAGGTAGGAGACA 164
#6              ------TCCATCTGCTGCACCAAGTGCCACAAAGGTAGGAGACA 152
#7              AATAATTCCATCTGCTGCACCAAGTGCCACAAAGGTAAGAGACA 164
#10             -----------------------------CAAAGGTAGGAGACA 130
#23             AATAATTCCATCTGCTGCACCAAGTGCCACAAAGGTAGGAGACA 164
#31             AATAATTCCATCTGCTGCACCAAGTGCCACAAAGGTAGGAGACA 164
                                             ******** ******
```

Figure 2: Cas9 induced mutations in rat *Tnfrsf1a* exon 2. Five out of thirty one F0 rats showed mutations in exon 2 near the designed target after microinjection of Cas9 against *Tnfrsf1a*. In this CLUSTALW2 alignment, “wt” was the theoretical wild type genomic sequence with Cas9 target in lower-cased letters. #6, #7, #10, #23 and #31 showed point mutations marked with underlines. #10 and #23 also had deletions as indicated by “-”. Identical positions were marked by “*”. The sequences of other rats showed no change from the wild type, taking #1 for instance. Detailed sequencing files are included in additional file 2.



```
SD(Wt)            ttacaggtggaattattcttGGGAGCTTTTCTTCTTTAGAGTTCATTGTGAGATTAATTT   60
                  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
#7 chr1-60893099  TTACAGGTGGAATTATTCTTGGGAGCTTTTCTTCTTTAGAGTTCATTGTGAGATTAATTT   60

SD(Wt)            CTTGAAATTTTTGGGTGTAGTTTTTCTCCTTCTGTTGGATGTTTCCTTCTGTTAACCTCA   120
                  ||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
#7 chr1-60893099  CTTGAAATTTT-GGGTGTAGTTTTTCTCCTTCTGTTGGATGTTTCCTTCTGTTAACCTCA   119
```

Figure 3: A novel mutation was detected near the COD predicted off-target site at rat Chr1:60893099. Sequencing revealed a novel mutation near the most-likely off-target predicted by COD at rat chr1:60893099. The 1 bp deletion was labelled as “-” in the alignment between the TA-clone of rat #7 and wild type SD rat, and highlighted in the chromatogram. The predicted off-target was marked in lower-cased letters. Detailed sequencing files are included in additional file 3.

3.2 Design rat P450 knockouts

A list of rat P450 Entrez gene IDs was processed by COD on page “Gene knockout pipeline” to design gene knockouts in batch mode automatically. Out of 79 P450 genes, 72 knockouts were designed successfully with at least 3 unique Cas9 targets in their top three prioritized coding exons. Seven P450 genes failed to find unique Cas9 target because their exons were identical to their kindred genes. Cas9 designs for all genes were summarized in “rat P450 genes.xls”. For each of the succeeded 72 gene IDs, there were a genbank file named after its

Entrez ID (such as “286953.gb” for Cyp2b3), several genbank and png files of prioritized exons named as OUT-rat-EntrezID-cExon (such as “OUT-rat286953CDS0.gb” and its png graph file of the same name). Cas9 targets were labelled as misc_feature in the genbank files of prioritized exons as mentioned in methods. All output files were packed in “rat P450 KO Designs.zip” as additional file 4. The processing time to design one gene knockout was 30 to 90 seconds on average if three prioritized exons were processes for each gene.

3.3 Customized off-target scoring matrix

To demonstrate the customized off-target scoring matrix, an assumptive endonuclease recognizing a DNA sequence pattern of 5’ TggAAAATatAATCTGATGA 3’ was used as an input, with lower-cased letters on exchangeable positions where mutations were allowed. Suppose the mutations on position 9 and 10 would not disturb the catalytic activity, at all; though the mismatches on position 2 or 3 would reduce the enzyme activity to half. Accordingly, its off-target scoring matrix was assigned as (0 0.5 0.5 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0). After submission, two potential off-targets were predicted in human genome by COD. A more likely target was TGGAAAATtTAATCTGATGA on chr 5, 115148925 to 115148906 bp with a score of 1. A less likely one with score of 0.5 was TGaAAAATccAATCTGATGA on chr18, 46243899 to 46243880 bp. Mismatched positions are labelled in lower case.

Any user defined sequence-specific endonuclease can be analyzed in a similar approach to predict its targets and potential off-targets in a genome.

3.4 Integrated Cas9 designer (COD2)

COD2 successfully combined Cas9 designer with off-target analysis. Its computational speed was compared to the basic COD, and the MIT CRISPR designer (Table 1) using the same two input sequences. The basic Cas9 designer showed the highest speed because no off-target analysis was performed. It took much less time to complete the computation of Cas9 design and off-target analysis by COD2 in comparison to MIT CRISPR designer although the detailed hardware configuration was not inspected for the MIT CRISPR server.

Table 1: Computational time lapses of MIT CRISPR, basic COD and COD2

Input	MIT CRISPR	Basic COD	COD2
1	27 min	1 min	13 min
2	50 min	0.75 min	12 min

Note: Details of input files, parameters and output files are listed in “readme.txt” of additional file 5.

MIT CRISPR generated 23 Cas9 targets for input 1, and 22 Cas9 targets for input 2. Following a standard practice, the top-ranked 6 targets from MIT CRISPR were selected and listed in Table 2 for each of the two inputs. In comparison, COD2 picked the same top-ranked set of 6 targets as what MIT CRISPR picked for input 1. For input 2, 3 out of the 6 MIT CRISPR selected targets were picked by COD2 in its top 6. Five out of the 6 MIT CRISPR top-ranked targets for input 1, and 4 out of the 6 MIT CRISPR top-ranked for input 2, were reproduced by the basic COD. Overall, both the basic COD and COD2 shared 75% (9 out of 12) of their top 6 Cas9 targets with the top 6 targets designed by MIT CRISPR.

In Table 3, for the same two targets (target 1 and 2), the off-targets estimated by MIT CRISPR (off-target 1-1, 1-2, 1-3 and 2-1, 2-2, 2-3) were quite different from the ones estimated by COD2 (off-target 1-4, 1-5 and 2-4, 2-5, 2-6). Only 1 (off-target 2-2) out of 11 off-targets was reproduced by both MIT CRISPR and COD2.

The supplementary files are available at: http://dx.doi.org/10.6084/m9.figshare.1609664.

Table 2: MIT CRISPR, basic COD and COD2 designed similar sets of top-ranked Cas9 targets

Input	Top-ranked Cas9 Target	Rank according to score		
		MIT	Basic COD	COD2
1	GCTGCGTCGTCGTAGTTTTTTGG	1st	2nd	1st
	AGTCCAGCACTCGCTCGCGCCGG	2nd	1st	4th
	CTGCGTCGTCGTAGTTTTTTGGG	3rd	4th	3rd
	GGACCGGCGCGAGCGAGTGCTGG	4th	3rd	2nd
	GCGTCGTCGTAGTTTTTTGGGGG	5th	not included	6th
	TGCGTCGTCGTAGTTTTTTGGGG	6th	5th	5th
2	CTGTTTGTGCAGGGCTCCGAGGG	1st	2nd	3rd
	ACTGTTTGTGCAGGGCTCCGAGG	2nd	3rd	1st
	CGAGGGGACCCATGTGGCTCAGG	3rd	not included	8th
	TTAGCCACCCTGAGCCACATGGG	4th	not included	18th
	TGTCCTGGGACTGTTTGTGCAGG	5th	1st	4th
	GTCCTGGGACTGTTTGTGCAGGG	6th	6th	12th

Note: More details are in “CompareTop6RankTargets.xls” of additional file 5.

4 Discussion

Unlike in other industries, the production pipelines in biotech are usually non-standard because of the versatility of life itself. It severely hampered the productivity and efficiency of biotech industry. The automatic COD pipeline initiated standardized automatic designs for genome editions by a sequence-specific endonuclease (e.g. Cas9). Overall, it greatly enhanced the productivity with high accuracy, maintained high speed with accountability, and simplified personnel’s operations with modulated protocols.

Off-targeting has been a major concern ever since the origin of genome editing. Prediction and validation of off-targets in cells and animal models have been urgently demanded in the augmenting applications of Cas9 induced mutagenesis. The COD system comprehensively designed Cas9 targets, and predicted potential off-targets which were confirmed in a mutant rat indeed. The entire workflow was automated as a standard operational protocol which can be followed by a user with accuracy, efficiency and accountability. The result indicated that a real off-target may be more likely to be found using the COD system than merely counting the number of mismatches though further conclusive experiment with increased sample size is needed to be statistically significant.

In comparison to the MIT CRISPR designer, the superior computational efficiency of integrated COD2 may result from the benefit of optimized BLAST engine. Noticeably, the basic COD is still the best engine for batched high-throughput Cas9 pipelines because it spent only less than 1/10 of the computational time compared to others. The basic COD, COD2 and MIT CRISPR selected a similar subset of top-ranked Cas9 targets upon the same inputs despite their different screening algorithms and various computational time lapses. However, off-target estimations were largely discrepant between COD2 and MIT CRISPR. Based on the off-targets estimated for the same seed targets by COD2 or MIT CRISPR, we believe that COD2 predicted more reasonable off-targets than MIT CRISPR did. Because detailed inspection revealed that the COD2 selected off-targets with 4~5 mismatches near the 5' side of the 20 nt guide, which is consistent with other publications [7,9]. However, the mismatches on MIT CRISPR selected off-targets were more randomly distributed on the guide. The MIT CRISPR designer also selected more off-targets containing NAG as PAM, which is irrational because it supposed to have only lower than half of the activity of NGG [11, 22]. For instance, in Table 3, the off-target 1-4 and 1-5 are much better candidates in comparison to off-target 1-1, 1-2 and 1-3 considering the locations of mismatches and the inferior of NAG. Please note the observation is limited within the two sets of off-targets. On the other hand, COD and COD2 are also limited by the BLAST engine. If a potential off-target is discarded by the algorithm and thresholds of BLAST, it will not show up in COD or COD2.

Table 3: MIT CRISPR and COD2 generated different sets of off-targets for the same two targets of the best choices

Target 1	**GCTGCGTCGTCGTAGTTTTTTGG**	**MIT rank**	**COD2 rank**
Off-target 1-1	GCTGgGagGTCtTAGTTTTTGaG	1st	not included
Off-target 1-2	aCTGCagCGTCaTAGTTTTTGaG	2nd	not included
Off-target 1-3	GCTGaGTCGgCaTAGTTTTgGGG	3rd	not included
Off-target 1-4	cCTGCtTCcTCGTAGTTTTTTGG	not included	1st
Off-target 1-5	tagagGTaGTCGTAGTTTTTGGG	not included	2nd
Target 2	**CTGTTTGTGCAGGGCTCCGAGGG**	**MIT rank**	**COD2 rank**
Off-Target 2-1	CTGgaTGgGCAGGGCTCCGAGaG	1st	not included
Off-Target 2-2	CgGaaTtTGCAGGGCTCCGATGG	2nd	9th
Off-Target 2-3	CgGTTTGaaaAGGGCTCCGAGaG	3rd	not included
Off-Target 2-4	gcccTTGTcCAGGGCTCCGAAGG	not included	1st
Off-Target 2-5	tccTTTGaGgAGGGCTCCGATGG	not included	2nd
Off-Target 2-6	aagagTGTGgAGGGCTCCGAAGG	not included	3rd

Note: Lower-cased letters in sequences mark mismatches between an off-target and its corresponding target. More details are in "Compare3MostlikelyOffTargets.xls" of additional file 5. COD2 only estimated two off-targets for target 1.

In the future, several directions could be considered for further development. Recent publications indicated that the type of mismatches (e.g. A to G, C to T, etc.) could also affect the off-target score in a scale lower than the position of mismatches [8]. Therefore, considering a two dimensional scoring matrix might slightly improve the accuracy of off-target prediction in the future. Taking advantage of the high-throughput capability, all genes of more species could be automatically processed, and the results could be transformed into annotation files labelling any existing genome database as a public resource using the customized Distributed Annotation System (DAS) [46]. Genes sharing nearly identical exons are hard to design individual knockouts induced by unique Cas9 targets. An alternative approach could be using two unique Cas9 targets in introns flanking a critical exon as in previous publication [13] since introns are usually much less conserved than exons. Similar flanked design also applies to the floxed conditional knockout alleles, which are widely practiced among academic researches. On the other hand, it is possible to knockout multiple homologous genes using a commonly shared Cas9 target indicated by multiple hits in the off-target spreadsheet of COD output. Furthermore, it could be practically useful to label the functional domains of proteins near a selected Cas9 target by querying the database of CCDS [47] or Pfam [48].

5 Conclusions

The COD system has established a speedy, accurate, flexible and high through-put computational gene knockout pipeline supporting the Cas9 induced mutagenesis. Besides searching Cas9 candidates for a given DNA sequence, it can analyze off-targets for any sequence-specific endonuclease using a scoring matrix. Cas9 induced mutant animal models can be precisely designed, generated and genotyped following the instruction of COD. The automatic gene knockout pipeline can also be processed in a batch mode with efficiency. COD2 integrated Cas9 target design with off-target analysis, which showed superior computational efficiency in comparison to MIT CRISPR. It also demonstrated better accuracy than the MIT CRISPR designer when estimating off-targets in at least some cases.

The system is freely available at http://cas9.wicp.net.

Acknowledgements

COD was made to acknowledge the sacrificed research animals and plants. The Cas9 expression plasmid vector used to create rat mutation was kindly provided by Dr. Xingxu Huang, Nanjing University, China. Lots of implementations were inspired by interactions with COD users around the world.

References

[1] R. Sapranauskas, G. Gasiunas, C. Fremaux, et al. The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res*, 39(21):9275-9282, 2011.

[2] L. Cong, F. A. Ran, D. Cox, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*, 339(6121):819-823, 2013.

[3] P. Mali, L. Yang, K. M. Esvelt, et al. RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823-826, 2013.

[4] H. Wang, H. Yang, C. S. Shivalila, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*, 153(4):910-918, 2013.

[5] S. Waaijers and M. Boxem. Engineering the Caenorhabditis elegans genome with CRISPR/Cas9. *Methods*, 2014.

[6] Y. Niu, B. Shen, Y. Cui, et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*, 156(4):836-843, 2014.

[7] V. Pattanayak, S. Lin, J. P. Guilinger, et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol*, 31(9):839-843, 2013.

[8] P. D. Hsu, D. A. Scott, J. A. Weinstein, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol*, 31(9):827-832, 2013.

[9] Y. Fu, J. A. Foden, C. Khayter, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol*, 31(9):822-826, 2013.

[10] S. W. Cho, S. Kim, Y. Kim, et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res*, 24(1):132-141, 2014.

[11] P. Mali, J. Aach, P. B. Stranges, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*, 31(9):833-838, 2013.

[12] F. A. Ran, P. D. Hsu, C. Y. Lin, et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, 154(6):1380-1389, 2013.

[13] J. Zhou, J. Wang, B. Shen, et al. Dual sgRNAs facilitate CRISPR/Cas9 mediated mouse genome targeting. *FEBS J*, 2014.

[14] Z. Rong, S. Zhu, Y. Xu, et al. Homologous recombination in human embryonic stem cells using CRISPR/Cas9 nickase and a long DNA donor template. *Protein Cell*, 2014.

[15] B. Gennequin, D. M. Otte and A. Zimmer. CRISPR/Cas-induced double-strand breaks boost the frequency of gene replacements for humanizing the mouse Cnr2 gene. *Biochem Biophys Res Commun*, 441(4):815-819, 2013.

[16] L. Yang, P. Mali, C. Kim-Kiselak, et al. CRISPR-Cas-mediated targeted genome editing in human cells. *Methods Mol Biol*, 1114:245-267, 2014.

[17] L. Davis and N. Maizels. Homology-directed repair of DNA nicks via pathways distinct from canonical double-strand break repair. *Proc Natl Acad Sci U S A*, 111(10):E924-932, 2014.

[18] J. D. Sander, M. L. Maeder, D. Reyon, et al. ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Res*, 38(Web Server issue):W462-468, 2010.

[19] M. Ma, A. Y. Ye, W. Zheng, et al. A guide RNA sequence design platform for the CRISPR/Cas9 system for model organism genomes. *Biomed Res Int*, 2013:270805, 2013.

[20] S. Bae, J. Park and J. S. Kim. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics*, 2014.

[21] A. Xiao, Z. Cheng, L. Kong, et al. CasOT: a genome-wide Cas9/gRNA off-target searching tool. *Bioinformatics*, 2014.

[22] F. A. Ran, P. D. Hsu, J. Wright, et al. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*, 8(11):2281-2308, 2013.

[23] D. C. Swarts, M. M. Jore, E. R. Westra, et al. DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*, 507(7491):258-261, 2014.

[24] G. Sheng, H. Zhao, J. Wang, et al. Structure-based cleavage mechanism of Thermus thermophilus Argonaute DNA guide strand-mediated DNA target cleavage. *Proc Natl Acad Sci U S A*, 111(2):652-657, 2014.

[25] W. Kameshima, T. Ishizuka, M. Minoshima, et al. Conjugation of peptide nucleic acid with a pyrrole/imidazole polyamide to specifically recognize and cleave DNA. *Angew Chem Int Ed Engl*, 52(51):13681-13684, 2013.

[26] H. Katada, T. Harumoto, N. Shigi, et al. Chemical and biological approaches to improve the efficiency of homologous recombination in human cells mediated by artificial restriction DNA cutter. *Nucleic Acids Res*, 40(11):e81, 2012.

[27] A. M. Geurts, G. J. Cost, Y. Freyvert, et al. Knockout rats via embryo microinjection of zinc-finger nucleases. *Science*, 325(5939):433, 2009.

[28] T. Cermak, E. L. Doyle, M. Christian, et al. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res*, 39(12):e82, 2011.

[29] J. P. Guilinger, D. B. Thompson and D. R. Liu. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol*, 2014.

[30] W. C. Skarnes, B. Rosen, A. P. West, et al. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337-342, 2011.

[31] C. P. Austin, J. F. Battey, A. Bradley, et al. The knockout mouse project. *Nat Genet*, 36(9):921-924, 2004.

[32] J. B. Knaak, C. C. Dary, X. Zhang, et al. Parameters for pyrethroid insecticide QSAR and PBPK/PD models for human risk assessment. *Rev Environ Contam Toxicol*, 219:1-114, 2012.

[33] S. Miksys and R. F. Tyndale. Cytochrome P450-mediated drug metabolism in the brain. *J Psychiatry Neurosci*, 38(3):152-163, 2013.

[34] K. M. Gamber. 10 reasons your next animal model should be a rat: SAGE Labs; 2013.

[35] P. J. Kersey, J. E. Allen, M. Christensen, et al. Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res*, 42(Database issue):D546-552, 2014.

[36] D. Karolchik, R. Baertsch, M. Diekhans, et al. The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51-54, 2003.

[37] C. Camacho, G. Coulouris, V. Avagyan, et al. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009.

[38] D. Karolchik, A. S. Hinrichs, T. S. Furey, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, 32(Database issue):D493-496, 2004.

[39] A. Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database (Oxford)*, 2011:bar049, 2011.

[40] J. Ye, G. Coulouris, I. Zaretskaya, et al. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13:134, 2012.

[41] M. A. Larkin, G. Blackshields, N. P. Brown, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947-2948, 2007.

[42] Y. Fu, J. D. Sander, D. Reyon, et al. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol*, 2014.

[43] B. Shen, J. Zhang, H. Wu, et al. Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell Res*, 23(5):720-723, 2013.

[44] Y. Ma, B. Shen, X. Zhang, et al. Heritable Multiplex Genetic Engineering in Rats Using CRISPR/Cas9. *PLoS One*, 9(3):e89413, 2014.

[45] Y. Ma, X. Zhang, B. Shen, et al. Generating rats with conditional alleles using CRISPR/Cas9. *Cell Res*, 24(1):122-125, 2014.

[46] A. M. Jenkinson, M. Albrecht, E. Birney, et al. Integrating biological data--the Distributed Annotation System. *BMC Bioinformatics*, 9 Suppl 8:S3, 2008.

[47] K. D. Pruitt, J. Harrow, R. A. Harte, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*, 19(7):1316-1323, 2009.

[48] R. D. Finn, J. Mistry, B. Schuster-Bockler, et al. Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247-251, 2006.