Databases and ontologies

WGE: a CRISPR database for genome engineering

Alex Hodgkins, Anna Farne, Sajith Perera, Tiago Grego, David J. Parry-Smith, William C. Skarnes* and Vivek Iyer*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on January 21, 2015; revised on April 25, 2015; accepted on May 12, 2015

Abstract

Summary: The rapid development of CRISPR-Cas9 mediated genome editing techniques has given rise to a number of online and stand-alone tools to find and score CRISPR sites for whole genomes. Here we describe the Wellcome Trust Sanger Institute Genome Editing database (WGE), which uses novel methods to compute, visualize and select optimal CRISPR sites in a genome browser environment. The WGE database currently stores single and paired CRISPR sites and pre-calculated off-target information for CRISPRs located in the mouse and human exomes. Scoring and display of off-target sites is simple, and intuitive, and filters can be applied to identify high-quality CRISPR sites rapidly. WGE also provides a tool for the design and display of gene targeting vectors in the same genome browser, along with gene models, protein translation and variation tracks. WGE is open, extensible and can be set up to compute and present CRISPR sites for any genome.

Availability and implementation: The WGE database is freely available at www.sanger.ac.uk/htgt/wge

Contact: vvi@sanger.ac.uk or skarnes@sanger.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

CRISPR-Cas technology is a powerful tool for genome editing that can be applied to virtually any species, from viruses to plants to mammals (Hsu et al., 2014; Mali et al., 2013). The CRISPR-Cas system, exemplified by Cas9 from Streptococcus pyogenes, is an RNA-guided endonuclease that can be targeted to specific sequences by Watson-Crick base pairing between a guide RNA (gRNA) molecule and a 20 bp target sequence adjacent to an obligate NGG protospacer adjacent motif (PAM). By replacing the first 20 bp of the gRNA with the desired target sequence Cas9 can be re-programmed to induce a double-stranded break at any N(20)NGG site in the genome. The Cas9 endonuclease will tolerate some mismatches in the alignment between the gRNA and target DNA sequence and off-target damage can occur at other sites in the genome with high sequence similarity to the CRISPR site (Hsu et al., 2014). Therefore, when designing gRNAs for genome editing, it is important to consider potential off-target sites as well as the position of the gRNA site within a gene.

Many web-delivered software solutions are available for choosing highly specific CRISPR sites in vertebrate genomes (Supplementary Table S1). Some require the input of target or gRNA sequences (Bae et al., 2014; Hsu et al., 2014; http://www.benchling.com). Others provide a basic genome browser (http://horizon.deskgen.com; Montague et al., 2014). One can be run locally (Bae et al., 2014), and a few have more involved scoring schemes, both for gRNA’s and off-targets (e.g. Heigwer et al., 2014; http://www.benchling.com). Some use a heuristic short-read aligner such as Bowtie (Langmead and Salzberg, 2012) to compute gRNA specificity (Ma et al., 2013; Montague et al., 2014; Sander et al., 2010).

Here we describe WGE (http://www.sanger.ac.uk/htgt/wge). In contrast to existing websites, WGE guides the genome editing
2 Methods

2.1 CRISPR-analyser: CRISPR display and scoring

The WGE CRISPR-Analyser package identifies CRISPR sites by scanning every 23 bases of the reference genome, searching for a CC as the first two bases (indicating a PAM site on the reverse strand) or a GG as the final two bases (PAM site on forward strand). The CRISPR-Analyser software also provides very fast, genome-wide off-target scoring (approximately 3 seconds per CRISPR site). Off-target potential is found by directly comparing the CRISPR sequence to all other possible matches in the genome, with up to 4 bp of mismatch. This is done very rapidly by building an in-memory index of all CRISPR sequences (Supplementary Information Section 2 and Supplementary Fig. S2). In contrast to other packages (e.g. Naito et al. 2014), no other alignment software is used. Using 80 parallel processes, we computed genome-wide off-target scores for all CRISPR sites (PAM = NGG) in the mouse and human exomes—including 200 bp of sequence. These results are stored in the WGE database. The storage of CRISPR sites and their off-target scores allows users to rapidly browse the genome and alter filter criteria to select CRISPRs. Website users can also initiate real-time off-target scoring for previously un-scored CRISPRs: the resulting scores are stored and made available to all users. All data on the website is accessible via an API.

Our focus is on mouse and human genomes, but the components have been written to make the process of installing and extending to other genomes easy. See, for example (https://github.com/coronin/CRISPR-Analyser) which extends the CRISPR-Analyser software to NGG and NAG-pam sites in the dog genome.

2.2 WGE database and WGE website

All CRISPR sites, pre-computed and user-requested off-target scores, as well as user-generated vector designs are stored in the WGE Database (Supplementary Information Section 3). Bulk downloads of the stored CRISPR locations and off-target information are available via the WGE website or by directly querying the WGE database with a REST API (Supplementary Information Section 1). The WGE website is built using Genoverse (www.genoverse.org) and standard web development components (Supplementary Information Section 4).

3 Results and discussion

Users enter the CRISPR-finding part of the WGE website by selecting a species—currently human or mouse—and the marker symbol of the gene to inspect. They are then prompted to select a target Ensembl exon. All possible CRISPR sites and paired sites (Supplementary Fig. S3) are shown on a scrollable Genoverse genome browser view, which incorporates the current gene models from Ensembl (Flicek et al. 2014), protein translation, available variation and any user-generated targeting vector designs (see below). ‘Paired’ CRISPR sites (Shen et al. 2014) are shown in WGE when CRISPR sites on opposite strands have a separation of up to 30 bp, or an overlap of up to 10 bp.

Our scoring system reports the number of similar sequences in the genome with up to four mismatches (excluding the PAM region), summarized in a simple output string. For example, a score ‘0:1, 1:0, 2:0, 3: 4, 4:5’ indicates there is 1 genomic site with 0 mismatches (the CRISPR site itself), no off-target sites with 1 or 2 mismatches, and an increasing number of potential off-target sites with 3 and 4 mismatches. By clicking on a CRISPR site, off-target information is displayed with a link to a summary report which highlights the bases that differ within the off-target sequences (Fig. 1) and reports their genomic coordinates and genomic context (intergenic, intron, exon). In this way, users can immediately grasp the off-target potential for each CRISPR site.

Users can also filter CRISPR sites based on their stored off-target characteristics (Supplementary Fig. S4). Using this interface, hundreds of possible CRISPR sites can be narrowed down and evaluated to select the optimal site(s) for an editing task. WGE also mimics other CRISPR-finding websites by allowing a user to directly paste in genomic sequence, which is analysed rapidly to show CRISPR sites and their pre-computed off-targets (Supplementary Fig. S5).

WGE can be used to design PCR primers for the assembly of gene targeting vectors by Gibson assembly (Gibson et al. 2009) or other similar PCR-based methods. This involves first choosing a...
target exon, and then adjusting design parameters via a web interface to allow the primer calculations to be run (Supplementary Fig. S6). The resulting targeting vector designs can be bookmarked, edited, and are displayed alongside CRISPR sites in the genome browser (Supplementary Fig. S7).

WGE provides the user with a highly visual method of rapidly designing genome edits using CRISPRs and targeting vectors. We plan to exploit this open and extensible platform to incorporate more genomes as needed, efficiency-based CRISPR scoring strategies and other useful browser tracks.

Acknowledgements
The authors thank the Database Group and Informatics Support Group at the Wellcome Trust Sanger Institute for help in setting up and maintaining the servers for WGE.

Funding
The National Institutes of Health (NIH) [1 U54 HG006370-01] and Wellcome Trust Core Funding.

Conflict of interest: none declared.

References