A peer-reviewed version of this preprint was published in PeerJ on 18 November 2015.

Roberson ED. 2015. Identification of high-efficiency 3′GG gRNA motifs in indexed FASTA files with ngg2. PeerJ Computer Science 1:e33 https://doi.org/10.7717/peerj-cs.33

Identification of high-efficiency 3'GG gRNA motifs in indexed FASTA files with ngg2

Elisha D.O. Roberson[1,2,*]

[1]Department of Internal Medicine, Division of Rheumatology, Washington University, St. Louis, MO, USA.

[2]Department of Genetics, Washington University, St. Louis, MO, USA.

[*]Elisha D.O. Roberson, Ph.D.
Washington University
Depts. of Internal Medicine and Genetics,
Division of Rheumatology
660 South Euclid Ave.
Campus Box 8045
St. Louis, MO 63110
eroberso@dom.wustl.edu

Abstract

CRISPR/Cas9 is emerging as one of the most used methods of genome modification in organisms ranging from bacteria to human cells. However, the efficiency of editing varies tremendously site-to-site. A recent report identified a novel motif, called the 3'GG motif, which substantially increases the efficiency of editing at all sites tested. Furthermore, they highlighted that previously published gRNAs with high editing efficiency also had this motif. I designed a python command-line tool, ngg2, to identify 3'GG gRNA sites from indexed FASTA files. As a proof-of-concept, I screened for these motifs in six genomes: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, and *Homo sapiens.* I identified more than 24 million single match 3'GG motifs in these reference genomes. Greater than 87% of all protein coding genes in the six reference genomes had at least one overlapping unique 3'GG gRNA site. In particular, more than 96% of mouse and 99% of human protein coding genes have at least one unique, overlapping 3'GG gRNA. These identified sites can be used as a starting point in gRNA design, and the ngg2 tool provides an important ability to identify high-efficiency editing sites in non-model species.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Introduction

Genome engineering allows for the targeted deletion or modification by homology directed repair of a target locus. Currently, one of the most popular methods for genome manipulation is the clustered regularly interspaced short palindromic repeat (CRISPR) / CRISPR associated protein 9 (Cas9) system adapted from *Streptococcus pyogenes*. The CRISPR/Cas system was initially thought to represent a novel DNA repair mechanism, but was eventually found to provide heritable bacterial immunity to invading exogenous DNA from sources, such as plasmids and bacteriophages (Barrangou et al. 2007; Makarova et al. 2006). During endogenous CRISPR/Cas9 function, foreign DNA integrates into the CRISPR locus. The bacterial cell then expresses the pre-CRISPR RNA (crRNA) and a trans-activating crRNA (tracrRNA) that pair to form a complex that is cleaved by RNAse III (Deltcheva et al. 2011). The resulting RNA is a hybrid of the pre-crRNA and the tracrRNA, and includes a 20 bp guide RNA (gRNA) sequence. The gRNA is incorporated into Cas9 and can then guide the cleavage of a complementary DNA sequence by the nuclease activity of the Cas9 protein. The topic of CRISPR-Cas genome editing has been reviewed extensively elsewhere (Doudna & Charpentier 2014; Hsu et al. 2014; Jiang & Doudna 2015; Mali et al. 2013).

Codon-optimized versions of Cas9 are available for a wide range of organisms, and can easily be synthesized if it is not already available. Transfecting cells with Cas9 plasmid along with a fused crRNA-tracrRNA hybrid construct called a single-guide RNA (sgRNA) allows for temporary activity of Cas9. Keeping a stock of plasmids with a sgRNA backbone minus the gRNA site makes it easy to quickly generate new sgRNA plasmids by site-directed mutagenesis. The Cas9 protein loaded with the sgRNA will bind to sites complementary genomic loci, but will only cut it if a protospacer adjacent motif (PAM) site immediately follows the complementary sequence (Mojica et al. 2009). The PAM site for type-I CRISPR is an NGG. Therefore, a gRNA site can be defined as N_{20}NGG. It is important to note that constitutively expressed sgRNAs typically use a U6 snRNA promoter that strongly prefers a G starting base. For U6 compatibility, sequences starting with A, C, or T may be used if they are cloned into a sgRNA vector with an appended G base, resulting in a 21 bp gRNA (Farboud & Meyer 2015; Ran et al. 2013b). The subset gRNA sites contain a starting G base (GN_{19}NGG), and can be cloned as a 20 bp gRNA, which I will refer to as canonical gRNA sites.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

The rate of editing using the CRISPR/Cas9 system is far higher than homologous recombination, but higher efficiency is still desirable. The introduction of a longer stem in part the sgRNA stem-loop structure and the flip of a single A in a polyA track of a separate sgRNA stem-loop, called the flip + extension (F+E) sgRNA design, resulted in increased Cas9 editing efficiency (Chen et al. 2013). Recently, another improvement was reported that increases efficiency. gRNA sites with a GG motif adjacent to the PAM site, called 3'GG gRNAs, have far higher activity than equivalent gRNA sites in the same region (Farboud & Meyer 2015). These sites take the form of N_{18}GGNGG.

In this manuscript, I report a python command-line tool, ngg2, for identification of high-efficiency 3'GG gRNA motifs from indexed FASTA files. Tools already exist to identify Cas9 gRNA targets in common model organisms. However, support for less common and non-model organisms is limited. This tool will enable the easy identification of high-efficiency gRNA sites in any genome. As a proof of concept, I report all 3'GG gRNA motifs in 6 model species, identifying more than 35 million sites, of which more than 24 million are unique matches within the reference genome for that species. More than 90% of all protein coding genes in 5/6 species have at least one unique 3'GG gRNA overlapping it for potential editing.

Materials & Methods

ngg2 Motif identification

I designed ngg2 using python with compiled regular expressions for the 3'GG gRNA plus PAM motif. ngg2 identifies these sites on both the sense and antisense strands, and optionally can be restricted to only canonical sites starting with a G for 20 bp gRNAs. Sites can be identified for a specified region, a whole contig, or all contigs in the input FASTA file. ngg2 uses the FASTA index to directly seek the genomic target without reading the entire file. This tool only identifies potential editing sites, and does not report uniqueness of the gRNA. As part of this manuscript, I generated files listing the uniqueness of each gRNA for each species, including gene overlaps. I recommend checking any identified site with blast / blat to ensure the minimum number of potential off target sites. ngg2 output includes the contig name, start and end positions, the gRNA sequence, the PAM sequence, and whether the site starts with a G.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Multi-species site identification

I used ngg2 to identify all 3' GG gRNA motifs 6 commonly studied organisms: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Mus musculus*, and *Homo sapiens*. I used a GNU Make script for genome downloads and site identification to enable reproducibility. The Makefile downloads the top-level (*Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*) or primary assembly (*Homo sapiens*, *Mus musculus*) genomes from Ensembl Release 79, indexes the FASTA files with samtools, downloads gene annotations for later intersection, runs ngg2 on all contigs for each FASTA file, and calculates GC content for each genome. I did not allow for 'N' bases in the gRNA sequence or PAM site. I based the GC content on only non-N base content in each genome. I counted occurrences of each gRNA sequence within each species, annotated the overlapping genes for each gRNA, and determined the number of genes cut by any 3'GG gRNAs as well as by unique gRNAs using R (v3.1.2) with the GenomicRanges package (v1.18.4) (Lawrence et al. 2013; R Core Team 2014). I calculated all summary statistics, counts, and figures using RStudio (v0.98.1102) Markdown with knitr (Xie 2013). I used many accessory functions from plyr, dplyr, tidry, stringr, and magrittr in the initial analysis, and I generated all plots with ggplot2.

Results

3'GG gRNA sites are common in each species

Overall, I identified greater than 35 million 3'GG gRNA sites in the six reference genomes (Table 1). Some of these gRNA sequences were not unique in a genome, leaving more than 24 million unique 3'GG sites. Approximately 6 million of the 24 million unique sites were canonical G starting motifs. The sites identified in each species with the gRNA sequence, PAM sequence, genome coordinates, annotated overlapping genes, and number of perfect genome matches are available for download (Roberson 2015). The R scripts, python files, and Make files are also available in a public repository for reproducibility.

The genomes I analyzed had vastly different sizes, ranging from approximately 12 Mb for yeast to greater than 3 Gb for humans, and as a result had dramatically different numbers of 3'GG gRNA sites per genome. Therefore, I also assessed the site density per megabase of reference genome size (Table 2). Unique sites averaged a density of 3,091

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

sites / Mb, or 1 unique site per 324 bp. *D. rerio* had the lowest density at 1,733 unique sites / Mb, while *D. melanogaster* had the highest density at 4,062 unique sites / Mb. The low density of unique sites in zebrafish may be due to genome complexity from previous duplication events.

Little strand bias observed for 3'GG gRNA sites

The strand of each gRNA site with respect to the reference was included in the ngg2 output files. I plotted the split of sense / antisense sites in each genome to visualize strong deviations from an expection of no strand bias (Fig. 1). For each organism, I considered every gRNA site as an independent Bernoulli trial with a 50% probability of success, and considered a "Sense" strand designation as a trial successful outcome (Supp. Table 1). *C. elegans, D. melanogaster,* and *H. sapiens* all demonstrated a strand bias significantly different from the expected ratio for all 3'GG sites. However, while the difference is significant, it may be unimportant. Wildtype Cas9 cleaves both DNA strands simultaneously, and therefore the strand of the target sequence doesn't matter. Strategies that employ dual nickases to reduce off target effects could be affected by such bias, as they require two separate gRNA sites on opposite strands (Ran et al. 2013a). The difference observed is only less than 0.4% different from expected 50% ratio, and whether this functionally affects the ability to choose paired 3'GG gRNAs remains to be seen.

CGG & GGG PAM sites are underrepresented

I visualized the distribution of the four PAM sites (AGG, CGG, GGG, TGG) as a stacked bar chart of each sites proportion of the total identified sites in each species (Fig. 2). In general, the AGG and TGG sites represented the majority of 3'GG gRNA sites in all species. I tested whether PAM site distribution differed from chance based on the GC content of the reference genome. For each species, I considered each PAM site a Bernoulli trial, and defined success as either CGG or GGG site identity. The probability of success was set equal to the estimated genome-wide GC content calculated from the reference genome, excluding N (Supp. Table 2). Only *D. melanogaster* met the expected GC success rate for 3'GG gRNA sites. The rate of picking a CGG or GGG PAM was less than the genome GC content in *S. cerevisiae, D. rerio, M. musculus,* and *H. sapiens.* In particular, the estimate for both *M. musculus* and *H. sapiens* is approximately 20% lower than the average genome GC content. This is not necessarily unexpected. The CGG PAM site includes a 5' CpG dinucleotide that is generally underrepresented due to the relatively high frequency of methyl-cytosine deamination to thymine in this context. *C.*

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

elegans was the exception, with CGG and GGG PAM selection greater than the expected frequency. However, this species lacks DNA methylation and would not necessarily be at an advantage to limit CpG dinucleotides.

Most protein coding genes overlap at least one unique 3'GG gRNA

A common use of genome engineering is to knock out or otherwise modify the function of a protein coding gene. The efficiency of such edits is critical, as just introducing frame-shifting mutations can require screening a large number single-cell clones or derived animals to identify a successful edit. As part of this study, I annotated for each gRNA in the 6 species if there was any overlap with a gene. Conversely, I also annotate a count of how many of each of the four classes (all sites, all unique sites, canonical sites, and unique canonical sites) overlap every gene. No less than 87% of any species' genes overlap at least one unique 3'GG gRNA (Fig. 3, Supp. Table 3). This catalog of potential sites demonstrates that most protein coding genes can be targeted by at least one 3'GG gRNA site to achieve high editing efficiency.

Discussion

In this manuscript, I have described a new tool for identifying 3'GG gRNA sites and presented a catalog of potential editing sites in 6 species. Importantly, many genomic loci can be targeted by unique 3'GG gRNA sites for efficient genome modification. Blast / blat searching of gRNA sequences is critical to ensure that there is not an abundance of exact or near matches in the target genome. It is also important to consider the target genome's specific genotypes when designing a gRNA. In particular, variants that alter PAM sites away from NGG will not be cleaved by Cas9 even if the gRNA is an exact match. Identification of potential gRNA sites from reference genomes is a starting point for genome editing, but careful study of target sites is required before a final design is selected.

The accuracy of editing can be improved by using two gRNAs and a mutant Cas9 nickase. I observed some significant, but low-effect strand bias in these genomes. This may lead to some loci not being compatible with paired 3'GG gRNA sites. When possible, choosing paired 3'GG gRNA sites should be strongly considered. Efficiencies of less than 10% were increased to 50% efficiency or greater by using the 3'GG strategy (Farboud & Meyer 2015). As such, using paired 3'GG gRNAs with a nickase may give the best of both worlds with both high accuracy and high efficiency.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

It is important to note that ngg2 will operate on any indexed FASTA file. Many gRNA site finding tools are limited to catalogs of gRNA sites in model organisms. This tool fills an important gap for individuals working outside of commonly used species, allowing for rapid and accurate identification of high efficiency 3'GG gRNA sites. The provided gRNA site survey and associated tool, ngg2, represent a valuable resource for designing genomic modification strategies.

Acknowledgments

This work was performed in the Human Genomics and Bioinformatics Facility of the Rheumatic Disease Core Center at Washington University (P30 AR048335). I wish to thank Dr. Li Cao for her helpful comments during the preparation of this manuscript.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Figures

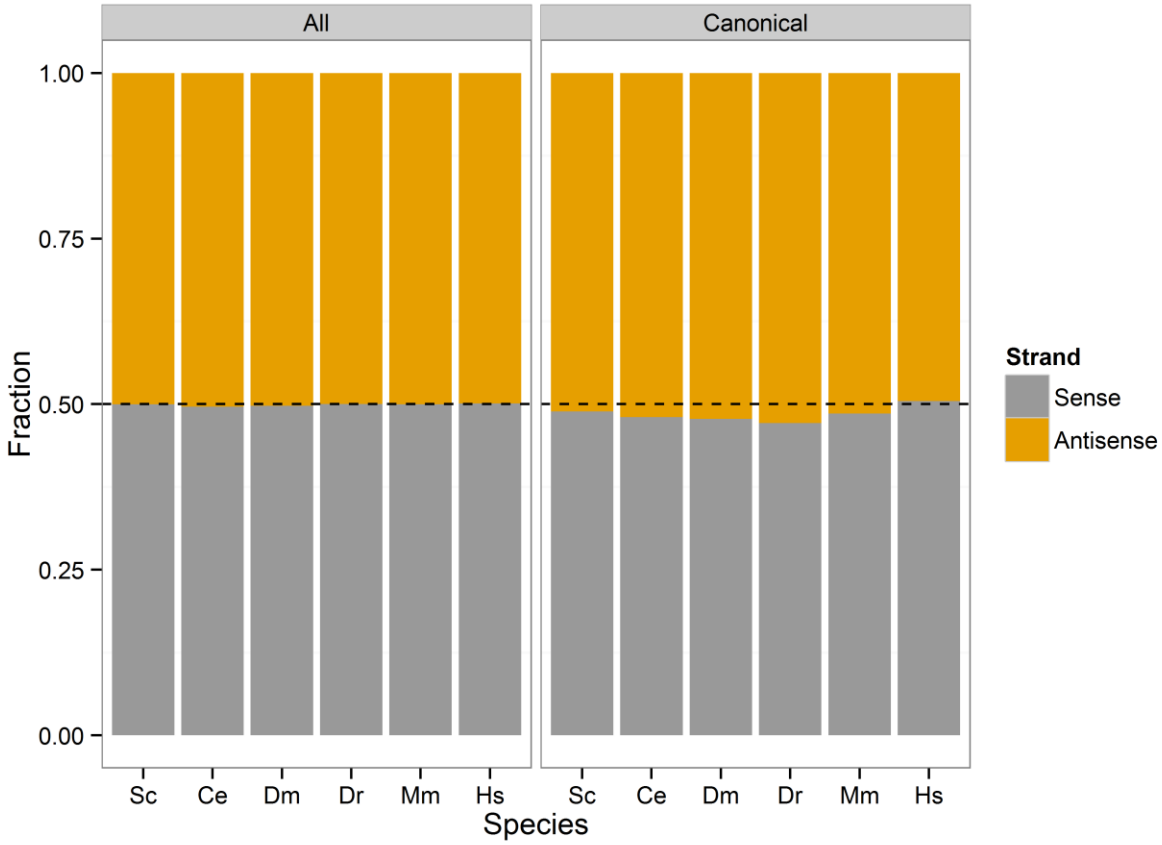

Fig. 1 - gRNA strand compared to reference genome

The fraction of gRNA sites sense and antisense to the reference are shown for each species. The x-axis labels are the two-letter species abbreviations. When considering all gRNA sites, there is not dramatic bias from a 50/50 strand ratio. The canonical sites are more skewed toward being antisense to the reference.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

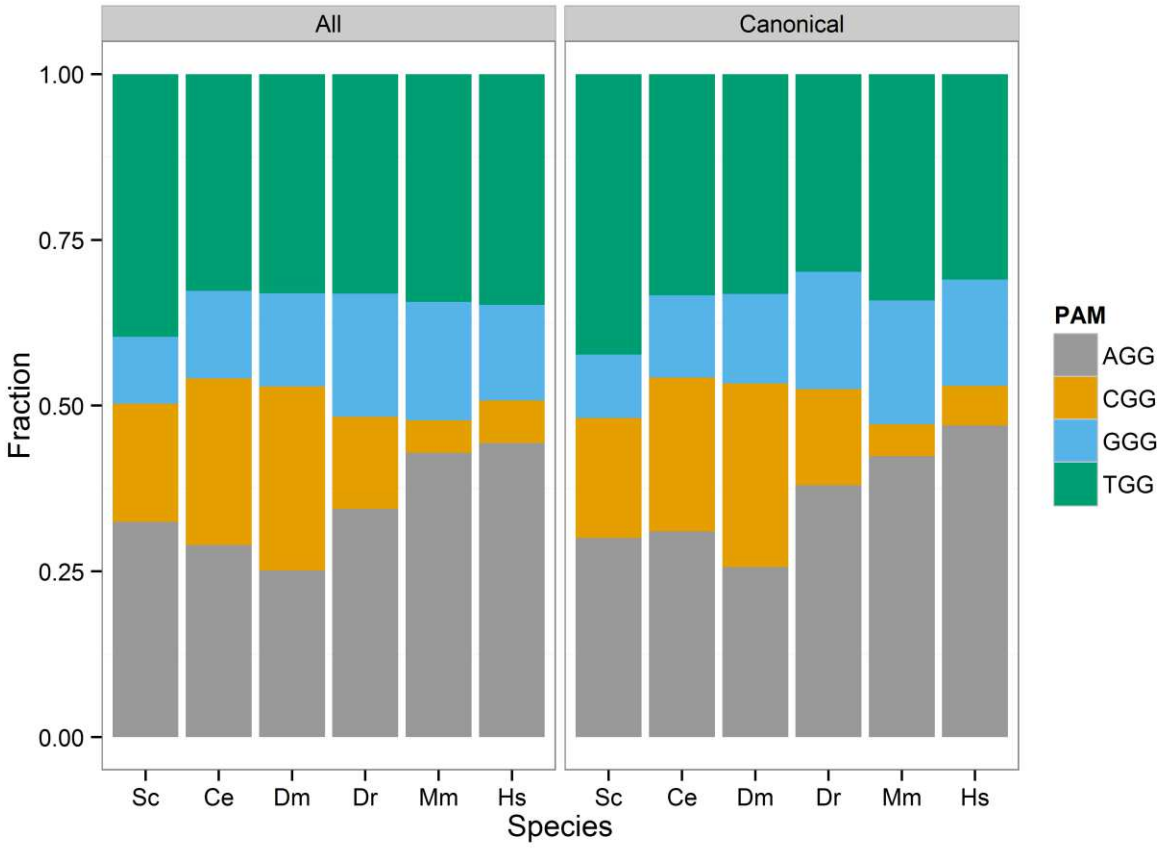

Fig. 2 - PAM site usage

Each species has four potential protospacer adjacent motifs (PAM) possible for identified gRNA sites. The stacked bar chart shows the fraction of all PAM sites each motif occupies. The CGG motif, that includes a CpG dinucleotide, is the least prevalent motif in the zebrafish, mouse, and human genomes.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

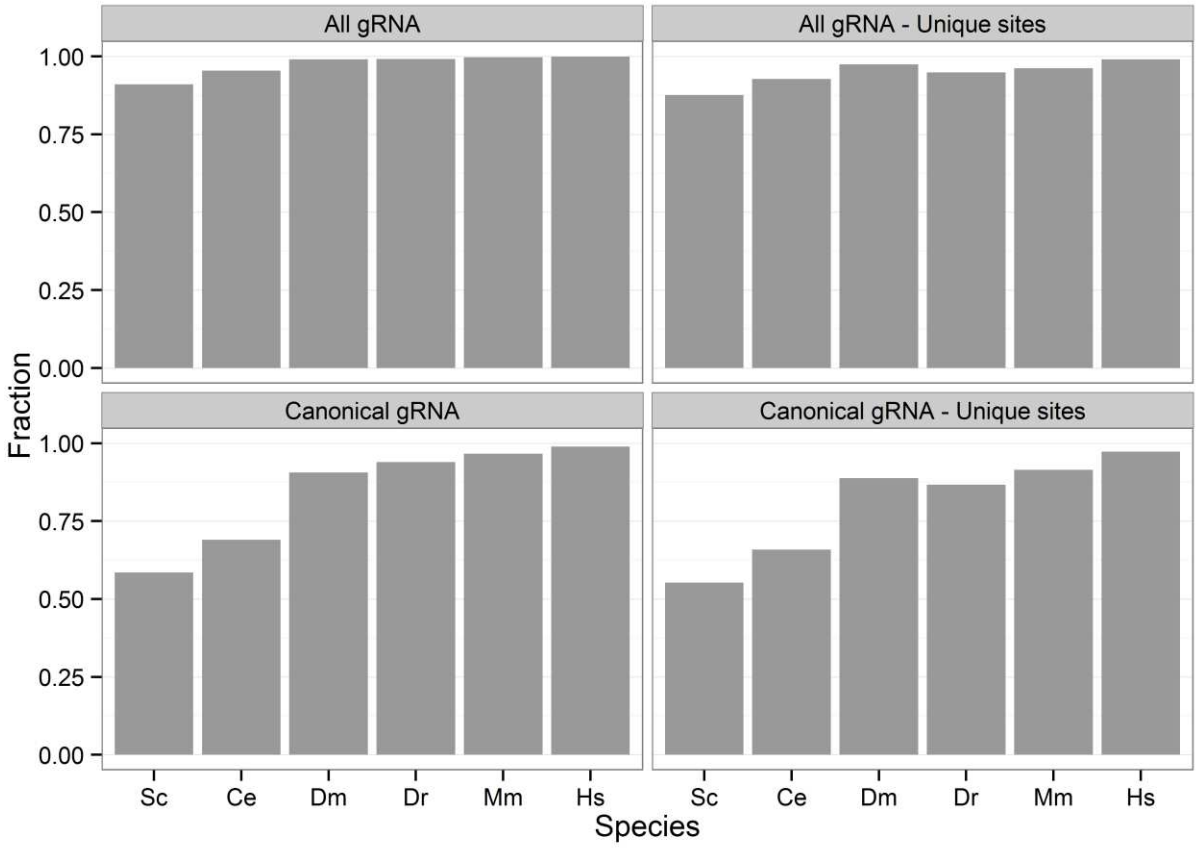

Fig. 3 - Overlap of protein coding genes and 3′GG gRNA sites

The fraction of protein coding genes overlapped by at least one gRNA of each class are shown for each species. The majority of protein coding genes have at least one unique overlapping 3′GG gRNA site.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Tables

	All gRNAs		Canonical gRNAs	
	All	**Unique**	**All**	**Unique**
Saccharomyces cerevisiae	38,430	35,733	8,328	8,148
Caenorhabditis elegans	293,597	257,317	62,874	60,669
Drosophila melanogaster	672,135	583,799	162,986	159,412
Danio rerio	4,422,730	2,448,177	626,202	561,214
Mus musculus	13,436,734	9,965,896	2,591,654	2,482,936
Homo sapiens	16,454,683	11,145,670	2,934,283	2,797,025
Total	35,318,309	24,436,592	6,386,327	6,069,404

Table 1 - Count of gRNA classes in each species

All N_{18}GGNGG motifs are included in the All category for All gRNAs. However, multiple edit sites in the genome for one gRNA are typically disadvantageous. Unique gRNAs were only observed in a gRNA plus PAM context once in a given genome.

	All gRNAs		Canonical gRNAs	
Species	**All**	**Unique**	**All**	**Unique**
Saccharomyces cerevisiae	3,161.11	2,939.27	685.03	670.23
Caenorhabditis elegans	2,927.59	2,565.82	626.94	604.96
Drosophila melanogaster	4,676.50	4,061.89	1,134.01	1,109.14
Danio rerio	3,131.21	1,733.27	443.34	397.33
Mus musculus	4,920.31	3,649.35	949.02	909.21
Homo sapiens	5,308.39	3,595.67	946.62	902.34

Table 2 - 3′GG gRNA Sites per Megabase Genome Size

Reference genome size was determined from the FASTA index. The number of unique 3′GG gRNA sites in the genome is encouraging, with an average across all species of one site per 324 bp.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Supplementary

	All gRNAs			Canonical gRNAs		
Species	**estimate**	**p.value**	**p.adj**	**estimate**	**p.value**	**p.adj**
Saccharomyces cerevisiae	0.500	9.55E-01	1.00E+00	0.489	3.90E-02	1.56E-01
Caenorhabditis elegans	*0.496*	*1.44E-05*	*8.66E-05*	*0.481*	*2.29E-24*	*1.60E-23*
Drosophila melanogaster	*0.498*	*1.82E-04*	*9.12E-04*	0.478	2.78E-79	2.78E-78
Danio rerio	0.500	1.41E-01	4.19E-01	0.471	3.95E-323	4.74E-322
Mus musculus	0.500	8.50E-01	1.00E+00	0.486	6.92E-323	7.61E-322
Homo sapiens	*0.501*	*3.59E-25*	*2.87E-24*	0.505	6.85E-70	6.16E-69

Supp. Table 1 - Strand preference for gRNA sites

The estimate is the binomial estimate of choosing the sense strand for a given gRNA, and the p-values are the significance of the strand bias. Half of the examined genomes show statistically significant, but low effect strand bias.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

		All gRNAs			Canonical gRNAs		
Species	**GC**	**estimate**	**p.value**	**p.adj**	**estimate**	**p.value**	**p.adj**
Saccharomyces cerevisiae	0.382	*0.279*	*4.94E-324*	*5.93E-323*	*0.277*	*9.77E-96*	*3.91E-95*
Caenorhabditis elegans	0.354	*0.384*	*7.99E-236*	*4.00E-235*	0.356	3.40E-01	3.40E-01
Drosophila melanogaster	0.417	0.418	1.16E-01	2.33E-01	*0.412*	*1.56E-05*	*4.69E-05*
Danio rerio	0.366	*0.325*	*7.41E-323*	*6.92E-322*	*0.322*	*3.46E-323*	*3.80E-322*
Mus musculus	0.416	*0.228*	*1.33E-322*	*9.34E-322*	*0.235*	*6.92E-323*	*6.92E-322*
Homo sapiens	0.407	*0.209*	*1.48E-322*	*9.34E-322*	*0.220*	*6.92E-323*	*6.92E-322*

Supp. Table 2 - PAM site frequency compared to genome GC content

The average genome GC content and the estimated chance of picking a GC PAM site (CGG or GGG) is shown for each species. GC content was calculated from the downloaded reference files. For all 3'GG gRNA sites, only *D. melanogaster* had no strand bias. Most genomes show significantly fewer CGG and GGG PAM sites than expected based on genome GC content.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

	All gRNAs		Canonical gRNAs	
Species	**All**	**Unique**	**All**	**Unique**
Saccharomyces cerevisiae	0.911	0.876	0.585	0.552
Caenorhabditis elegans	0.954	0.928	0.690	0.658
Drosophila melanogaster	0.991	0.975	0.906	0.888
Danio rerio	0.992	0.949	0.939	0.867
Mus musculus	0.998	0.962	0.967	0.915
Homo sapiens	0.999	0.990	0.990	0.973

Supp. Table 3 - Fraction protein coding genes with overlapping gRNAs

The proportion of protein-coding genes with at least one gRNA of the given type overlapping the annotated transcript start and end are shown for each species. Both *C. elegans* and *S. cerevisiase* have a lower rate of overlap with unique gRNAs. However, even the lowest fraction of unique overlap (yeast) still has >87% of protein-coding genes with at least one unique 3'GG gRNA site.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

References

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, and Horvath P. 2007. CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* 315:1709-1712.

Chen B, Gilbert Luke A, Cimini Beth A, Schnitzbauer J, Zhang W, Li G-W, Park J, Blackburn Elizabeth H, Weissman Jonathan S, Qi Lei S, and Huang B. 2013. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* 155:1479-1491.

Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, and Charpentier E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471:602-607.

Doudna JA, and Charpentier E. 2014. The new frontier of genome engineering with CRISPR-Cas9. *Science* 346:1258096.

Farboud B, and Meyer BJ. 2015. Dramatic Enhancement of Genome Editing by CRISPR/Cas9 Through Improved Guide RNA Design. *Genetics* 10.1534/genetics.115.175166.

Hsu PD, Lander ES, and Zhang F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157:1262-1278.

Jiang F, and Doudna JA. 2015. The structural biology of CRISPR-Cas systems. *Current Opinion in Structural Biology* 30:100-111.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, and Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol* 9:e1003118.

Makarova K, Grishin N, Shabalina S, Wolf Y, and Koonin E. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* 1:7.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015

Mali P, Esvelt KM, and Church GM. 2013. Cas9 as a versatile tool for engineering biology. *Nat Meth* 10:957-963.

Mojica FJM, Díez-Villaseñor C, García-Martínez J, and Almendros C. 2009. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155:733-740.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. 3.1.2 ed: R Foundation for Statistical Computing.

Ran FA, Hsu Patrick D, Lin C-Y, Gootenberg Jonathan S, Konermann S, Trevino AE, Scott David A, Inoue A, Matoba S, Zhang Y, and Zhang F. 2013a. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* 154:1380-1389.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, and Zhang F. 2013b. Genome engineering using the CRISPR-Cas9 system. *Nat Protocols* 8:2281-2308.

Roberson E. 2015. Survey of 3'GG gRNA sites in 6 genomes http://dx.doi.org/10.6084/m9.figshare.1371077.

Xie Y. 2013. *Dynamic Documents with R and knitr*: Chapman and Hall/CRC.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.969v1 | CC-BY 4.0 Open Access | rec: 8 Apr 2015, publ: 8 Apr 2015