Comparative Analysis of gRNA Design Software Tools for CRISPR Application on *orco* in Honey Bees

Sandy Zhang
Adviser: Yuri Pritykin

Abstract

Improvements in genome editing technologies, particularly CRISPR-Cas9, have resulted in the need for reliable tools that evaluate guide RNA (gRNA) efficiency and specificity. While most gRNA scoring tools have been extensively tested and optimized for the human genome, their applications to non-model organisms, such as the honey bee (Apis mellifera), are comparatively less explored. This study compares and analyzes the gRNA specificity scores generated by three computational tools — Benchling, GuideScan2, and Geneious — with a focus on the orco gene in honey bees. The results demonstrated a significant correlation between the scores, meaning these tools share overlapping evaluations since each of them uses some variation of the same algorithms. An estimated unified score was derived to further explore the results, suggesting the potential of creating a unified scoring system between software. However, due to the absence of experimental validation, evaluating the reliability of each tool remains difficult. This work highlights the need to validate CRISPR software tools in Apis mellifera and other genomes and offers insights into computational tools across less explored genomic contexts.

1. Introduction

CRISPR-Cas9 is a widely used gene-editing tool frequently utilized for research in the scientific community [9]. CRISPR-Cas9 technology offers a gateway for researchers to explore gene functions, as it allows researchers to edit specific parts of a living organism's genome. It uses a guide RNA (gRNA) to direct the Cas9 enzyme to a specific DNA sequence, where the enzyme then causes a DNA break to allow for modifications to the genome [11]. gRNAs are typically 20 base pairs long and their effectiveness relies on the presence of the protospacer adjacent motif (PAM), a short DNA sequence that's usually a sequence of 2-5 base pairs (bp) that marks proper target sites. The most

common one is “NGG” where “N” can be any nucleotide A C G T [5]. 91% of activate gRNAs contain an "NGG" PAM [16]. While alternative PAM sequences exist, they lead to notable but smaller gRNA activity [16]. For this paper, only the canonical PAM sequence ("NGG") will be observed.

When gRNA designs are generated, there are two scores that researchers focus on. One is the on-target score, which measures how well a given gRNA guides the Cas9 protein to the correct location [14]. Another is the specificity score, which measures the likelihood that a given gRNA only splits at the right target and not at similar off-target sites elsewhere in the organism’s genome [14]. The higher the scores, the better the efficiency and specificity respectively.

Advancements in CRISPR-Cas9 technology have led to the development of numerous computational tools that aid in the design of gRNAs. However, these tools lack standardization [13], with different tools often generating varying scores for the same gRNA. This compromises the ability of researchers to evaluate the effectiveness of a gRNA design. Furthermore, little research has been conducted to compare these computational tools systematically. This paper addresses this gap by evaluating three gRNA design tools — Benchling, GuideScan2, and Geneious — on their applicability to a non-model organism. The organism of interest for this paper is the honey bee, *Apis mellifera*. In particular, this paper is interested in the gene *orco* because this gene is crucial for the honey bee’s olfactory system (sense of smell), which governs many of their essential behaviors. Additionally, this study takes a step towards establishing a standardized system of gRNA evaluation by combining the outputs of the three analyzed tools into a unified scoring metric, potentially allowing for a more consistent and reliable gRNA design score.

2. Problem Background and Related Work

Honey bees play an essential role in agriculture and ecosystems. Understanding honey bees’ social and behavioral traits provides valuable insights into how these insects interact with their environment, adapt to challenges, and contribute to ecological balance.

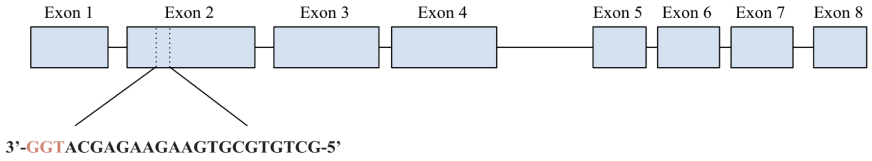

Figure 1: Illustrates the honey bee *orco* gene where the target is in Exon 2. The gRNA is bolded in black while the PAM sequence is bolded in light orange [2]

orco plays an important role in honey bee survival as they need it for locating food sources, identifying mates, navigating back to the hive, as well as other vital tasks. The *orco* protein localizes specific odorant receptors (ORs) to membranes of olfactory sensory neurons (OSNs), mostly located in the antennae, and forms channels with these ORs to respond to different odors [2]. Modification of such genes can have significant implications, as they strongly impact antennal gene expression.

Prior research on honey bees has used the CRISPR Guide RNA Design tool in Benchling (https://www.benchling.com) to design gRNAs targeting specific genes, such as the *orco* gene, to minimize off-target effects [2]. In the study, the gRNA targeted Exon 2 of the *orco* gene, as it had excellent on-target scores and minimal off-targets near genes in the genome [2]. This can be seen in Figure 1.

However, even for bees with complete knockouts (KO) of the *orco* gene, downregulated (i.e., decreased activity) *orco* expression was still observed, raising concerns about the precision and effectiveness of the gRNA design selected. The relationship between *orco* expression and the nature of the *orco* mutation still needs to be studied, but it remains possible that researchers failed to get a complete KO of the specific gene.

Although there are two scores of interest, this paper only focuses on the specificity score as it remains a major issue in CRISPR-Cas9 technology [6, 13, 15]. Web-based algorithms have limitations, as the software assumes that off-target sequences are closely related to the on-target site and hence may miss off-target sites with less sequence similarity [16]. This is why it's important to incorporate multiple tools during the design process.

To better understand how the effectiveness of gRNA designs is calculated, it's important to first

examine the computational algorithms used by the tools studied in this paper. Two prominent algorithms, the MIT-Broad Algorithm [7] and the CFD (Cutting Frequency Determination) Score [3], are the most popular ones. Both of these scoring algorithms use "synthetic" data — data constructed for research purposes — where a series of gRNAs targeting a specific dataset were mutated [14]. Measurements of the gRNA's ability to cleave the target site were taken and the results were used to make a linear regression algorithm to score the off-target sites [14]. The methods used by the two algorithms are similar, but they differ in their final construction [14]. The MIT-Broad Algorithm evaluates off-target activity using only the 20 bp target sequence (gRNA), whereas the CFD Score provides a more comprehensive analysis by incorporating both the target sequence and the PAM [7, 3]. Understanding the base algorithms allow for a better grasp of the computational strategies employed by tools like Benchling, GuideScan2, and Geneious in assessing CRISPR-Cas9 accuracy.

Guidescan2 (https://guidescan.com) was demonstrated to be more accurate than other tools when it came to enumerating potential off-targets and estimating specificity scores [12]. It also showed flexibility in preprocessing and analyzing new genomes, illustrating its power in custom gRNA designs and making it accessible in a wide range of applications [12]. The formula that Guidescan2 uses is based on the MIT-Broad algorithm as well as the CFD score [12].

$$\text{Specificity}(g) = \frac{1}{1 + \sum_{o \in \text{OffTargets}(g)} \text{CFD}(o, g)}$$

Guidescan2 calculates using a combination of the two popular algorithms with a specificity value that ranges from 0 to 1. This paper will convert that score to 0 to 100 to make it consistent with the other softwares' specificity scores.

Geneious (https://www.geneious.com) uses a similar method to calculate its specificity scores. The scoring algorithm for CRISPR sites uses the MIT-Broad algorithm [1]. Different weightings are used for mismatched bases between the CRISPR site and off-target sites in the final score, with the weights determined experimentally and based on their position [1].

Benchling specified on its website that it can use both MIT-Broad scoring and the CFD method, though CFD is only available for "NGG" PAM sites and gRNAs that are a length of 20 bp. This means Benchling can generate two different specificity scores.

All three software in this paper base their specificity score on either the algorithms CFD & MIT-Broad or a combination of the two algorithms. The exact formulas used by Benchling and Geneious weren't accessible as they weren't explicitly provided by the tools themselves. For the purposes of this paper, the scores will be between the ranges of 0 and 100, with a higher score denoting better specificity and less off-target activity.

Prior research has systematically evaluated available on-target design tools, genome-wide off-target cleavage site (OTS) detection techniques, and in silico genome-wide OTS prediction tools [8]. This led to the development of the integrated Genome-Wide Off-target cleavage Search platform (iGWOS), which integrated the available OTS prediction algorithms and datasets [8]. Yet little has been done to do the same evaluation for web-based tools.

One paper that worked on the computational design of guide RNAs to knockout the LasR gene of *Pseudomonas aeruginosa* did select several computational tools and web servers (CHOPCHOP, CRISPOR, Cas-Designer, and Benchling) to aid in the design of potential gRNAs [10]. However, instead of cross-comparing each tool, it selected the best gRNA hits from each tool, and, as long as the gRNA met all the parameters in two or more tools, the gRNA was selected for further analysis [10]. This work failed to fully compare the gRNAs generated by the tools, meaning the issue of potential discrepancies between computational tools still persists. These discrepancies are likely due to each tool using a slightly different variation of the existing algorithms to predict gRNA efficiency and specificity, which would lead to different gRNA design scores. Additionally, the way the parameters in the research were met differed between each tool. This means that gRNA designs that perform well in two tools may not necessarily be ranked similarly for the other tools.

While this paper focuses on three tools, they are a specific selection out of many gRNA design websites, such as CHOPCHOP, CRISPOR, CRISPR RGEN Tools, and E-CRISP [8]. Each tool has

its own distinct functionalities. Some tools work with custom genomes, while others are restricted to a certain type of genome background [8]. The three tools in this paper can design gRNAs from a multitude of genomes, making them decently versatile and representative of existing computational tools.

By designing and comparing gRNAs using tools like GuideScan2, Benchling, and Geneious, this study investigates the consistency of these computational tools while contributing to the potential testing of genetically edited bees.

3. Approach

The core approach of this paper is to design gRNAs for *orco* in honey bees using the three tools and compare them with one another. Through this process, similarities and differences between the different software can be observed.

The underlying questions that drive this research are if there are potential discrepancies between software and if so, if the score from one software could be more accurate than another. If the accuracy of these tools cannot be directly tested, the focus will transition to comparing the outputs from each tool to identify potential flaws or inconsistencies in their results. Given the time constraint, no experimental testing was completed so all results provided remain theoretical.

The three software were selected accordingly.

- Guidescan2 was selected because although it was never tested on the honey bee genome, it was tested extensively on human and mice genomes and it proved to be accurate in estimating specificity scores of gRNA designs [12]. To further test the capabilities of Guidescan2, it was compared with the other two software.
- Benchling was selected as it was used in a previous experiment on *orco* in honey bees [2]. Since it was used in a related work, it serves as a good point of comparison with the other two software. Additionally, this tool has been cited in multiple papers that have studied computational tools for CRISPR-Cas9 gene editing [10, 13].

- Geneious was selected because some researchers who have previously worked on *orco* [2] are currently using this software for their ongoing research on bumblebees. It was also selected because it allows for the design of custom gRNAs for any species, as it is able to provide custom sequences to check for off-target sites. A significant amount of software available online is suitable for a few provided species with integrated genomes in their databases, but not for custom sequences. In physical experiments, there's also the need to analyze the sequencing data every day to check if CRISPR worked in the injected samples (Experimentation isn't elaborated in this paper but is part of the process in most related research). Geneious provides a nice interface to edit the generated sequences and visually inspect for inference. Finally, Geneious serves as a one-stop solution to edit, analyze, and make publication-ready figures for sequence data. (Sarthok Rasique Rahman, private communication, Dec. 8, 2024)

What's unique about this project is that it directly compares the outputs from different tools to closely analyze the specificity scores. Although research in the past has looked into different computational tools for designing gRNAs, none ever went in depth to explore the differences between software for a specific gene of interest.

After comparing the specified software tools, a unified score was developed by combining the individual scores outputted by each software. This unified score acts like a standardized metric, helping address potential discrepancies between the tools while also presenting a single score for gRNA effectiveness. Using this unified score as a benchmark, several regression algorithms were tested to predict the unified score based on the individual software outputs. This approach allows for the input of scores from the different computational tools to obtain a predicted unified score. The idea is to help standardize the computational tools, making it easier for researchers working in this field to analyze gRNAs.

Additionally, while most CRISPR design algorithms have been extensively tested on human genomes, this research explores their applicability to a less-studied but equally important genome.

This has the potential to uncover biases or limitations in current software. Ultimately this paper hopes to enhance researchers' understanding of gene functions and their implications for future CRISPR study.

4. Implementation

The honey bee genome was retrieved through NCBI using the RefSeq assembly (GCF_003254395.2). It was downloaded into a FASTA file, which is a text-based format that represents nucleotide or amino acid sequences. Similarly, the *orco* gene was retrieved through NCBI and downloaded into a FASTA file.

4.1. Guidescan2

Guidescan2 doesn't include the honey bee genome on their website, so the genome was manually processed through the terminal. This was done using the open source software package (https://github.com/schmidt73/guidescan-cli).

The two sub-commands important to this study are index and enumerate. The index command constructs a genomic index from the honey bee FASTA file. The compressed genomic index is used to search for off-targets. The enumerate command enumerates off-targets against the genomic index for a set of kmers, a collection of short nucleotide sequences. Kmers can be generated using the python script "*generate_kmers.py*". The kmer length of interest in this paper is 20 with a PAM sequence of "NGG". The enumerate command can take several options, but this paper only uses the option -m/–mismatches, which is the mismatch radius to search for kmer matches.

```
./guidescan index
../bin/ncbi_dataset_honey/ncbi_dataset/data/GCF_003254395.2/
GCF_003254395.2_Amel_HAv3.1_genomic.fna
```

This line indexed the honey bee FASTA file to make it easier to search for off-targets across the honey bee genome. It's important to index the entire genome and not the specific gene to ensure the outputted specificity scores are accurate.

```
python generate_kmers.py ../bin/Orco_datasets/ncbi_dataset/data/
gene.fna > orco_gene_honeybee.txt
```

This line generated a set of kmers for the *orco* gene. With no parameters set, the kmers would be of length 20 with the canonical PAM sequence. It's important to generate kmers for the gene of interest and not the entire genome.

```
./guidescan2 enumerate ../bin/ncbi_dataset_honey/ncbi_dataset/data/
GCF_003254395.2/GCF_003254395.2_Amel_HAv3.1_genomic.fna.index
-kmers-file
../scripts/orco_gene_honey.txt -mismatches 4 -o orco_gene_honeybee.txt
```

This line resulted in the enumeration of off-targets for the *orco* gene against the entire honey bee genome using the index generated previously. The mismatch was set to 4, meaning Guidescan2 will output any off-targets that have a maximum of 4 mismatches from the original gRNA sequence. The radius is set to 4 in this paper, since the search complexity grows exponentially with the increase in parameters.

GuideScan2 outputted a list of 511,826 possible gRNA designs. When no parameters were inputted — meaning the mismatches were set to the default value of 3 — only about 10,000 gRNA were generated.

Attempts to upload a version of the honey bee gRNA database onto the main website were made but remained unsuccessful. Enumerations were initially run on a personal device, but over 95 GB of files were downloaded, resulting in insufficient storage. Several enumerations that took about 50 hours each were also run on Adroit. However, there were issues with getting Adroit to work properly with the data when trying to decode the Guidescan2 database into a human-readable CSV file. Several attempts to run the python script "*decode_database.py*" using the SAM file, a file format that's commonly used in bioinformatics, and using the honey bee FASTA file all gave the same error message stating

<does not contain alignment data>

While the database is tangent to the study and doesn't pose an issue in the context of this paper, it can be addressed as a future work.

4.2. Benchling

Benchling already had the honey bee genome in its system. The honey bee genome was not the most up-to-date version as it used the older honey bee genome, assembly Amel_4.5. The existing genome was then used to design and analyze guides. gRNA guides were only generated for each exon, as the software does not allow for more than 5,000 base pairs at a time.

To design gRNA guides, Benchling has an option called "Design and analyze guides." There, the inputs were specified to be a single guide of length 20 with the genome being the assembly Amel_4.5 and the PAM being the generic "NGG", since only "NGG" PAMs can be used for the CFD off-target scores. The advanced setting allows for the change of off-target score to use either the MIT-Broad scoring method or the CFD score. To better understand the specificity scores, both calculations of specificity scores from Benchling were analyzed. Using those settings, each exon was selected for a list of generated gRNAs. The total gRNA designs generated across all 8 exons was 137, one of which was the gRNA the past research on *orco* in honey bees worked with [2].

4.3. Geneious

For Geneious, similar to Benchling, only the exons of the gene were analyzed. The PAM sequence was set to "NGG", each gRNA sequence was set to 20bp, and the maximum number of mismatches was set to 4. The gene was scored against the whole honey bee genome to get more accurate results, as the software looks at off-target sites across the whole honey bee genome. It should be mentioned that Geneious skips any sequences in the off-target database that are exact duplicates of the target sequence. The total gRNA designs generated across all 8 exons was 144. The generation of the scores and gRNA was done by Postdoctoral Research Associate Sarthok Rasique Rahman in the Kocher Lab.

4.4. Comparative Analysis of Tools

After all the gRNA was generated, the gRNA were matched by exons, which were then organized into sheets. Although the assembly used in Benchling was an earlier version than that of Geneious and Guidescan2, most of the gRNA ended up matching, with the exception of a select few in the last exon. To find gRNA matches, a short code was written that would search through the results of the around 500,000 generated designs by Guidescan2. This made it easier to find matches and potential off-target sites for the specific gRNA design. Over the 8 exons of the *orco* gene, there were 134 overlapping gRNA designs. Exon 1 had seven, exon 2 had forty-one, exon 3 had twenty-eight, exon 4 had thirty-nine, exon 5 had two, exon 6 had three, exon 7 had ten, and exon 8 had four gRNA designs. Since factors such as sequence length, PAM sequence, and number of mismatches are controlled, the evaluation and comparison of the specificity scores more accurately reflects the algorithm each software uses to calculate such scores.

Five comparisons were analyzed.

- Geneious vs. Guidescan2
- Benchling (CFD) vs. Geneious
- Benchling (CFD) vs. Guidecan2
- Benchling (MIT-Broad) vs. Geneious
- Benchling (MIT-Broad) vs. Guidescan2

For each of these five comparisons, the following factors were analyzed:

- Average
- Standard Deviation
- Wilcoxon p-value (Paired)
- Spearman correlation
- Spearman p-value
- Ranking of gRNA designs

The average and standard deviation are preliminary analyses that were done similar to the Wilcoxon p-value. The Wilcoxon paired test instead of the unpaired test was used, as each software measured the score for the same gRNA design in the same exon in the same gene. The Spearman correlation and p-value were calculated to see the correlations between the software. Since the methods of calculation for each software may be different, it is hard to compare by purely looking at the measures of central tendency.

4.5. Unified Scoring Method

After the analysis of the individual software scores, it was of interest to consider a unified score. A preliminary model was developed that calculates and predicts estimated unified scores based on the scores from the three software. Although the model still isn't fully accurate, as the estimated score is just a benchmark, it remains a good stepping stone to a future scoring system between the different software.

PCA was used to calculate the estimated scores, because there is clear correlation between the scores produced by the software so combining the scores into a single dimension could avoid multicollinearity. Additionally, reducing the scores to a single principal component allows for the creation of a unified score for the tools this paper works with. StandardScalar was first run to tackle the issue of the means and variance of the different software varying too much. After doing so, PCA was performed and the output scores were scaled to the 0–100 range that was also used for the three software.

The model was then trained and the algorithms below were tested to predict unified scores.

- Linear
- Lasso
- Random Forest

Linear was used first because it was the simplest and the correlations between the software seemed

linear. The result had an almost perfect R^2 score of 1 while the MRE (Mean Squared Error) was extremely small — 1.25236e-27. This led to suspicion of potential overfitting.

Lasso regression was then used, since it helps prevent overfitting. Additionally, since the software was all somewhat correlated, lasso helped with multicollinearity. With an $\alpha = 1.0$, the MRE remained relatively small at 0.6875159, while the R^2 score was 0.997. The α is a regularization parameter to prevent overfitting as it controls how much the model shrinks its coefficients. However, as shown in Figure 3, there were a few major outliers. Both linear and lasso regression don't handle outliers particularly well.

This led to the use of Random Forest Regression. Random Forest learns based on decision trees; it essentially combines multiple trees to make predictions. It works well with nonlinear relationships, giving it more flexibility. It is also less affected by outliers. Setting the number of estimators to 3500 resulted in the lowest MRE and highest R^2 score, with a MRE of 24.59386 and a R^2 score of 0.8942.

This is not surprising. A paper that worked on predictive Modeling of CRISPR-Cas9 guide efficiency did a systematic comparison of different predictive models and found that Random Forest always did worse than linear and lasso regression [4]. In the research, lasso did better than linear regression [4] but the comparison was done with Spearman correlation, not MRE or R^2 scores, which could explain the discrepancy. Additionally, it only investigated for on-target predictions and not off-target predictions [4].

Access to my work and code can be found here (https://github.com/sandyzyn/IW-Project).

5. Evaluation

For the analysis, a general overview of the comparison results is observed, then it'll delve deeper into individual gRNA analysis as well as rankings. Lastly, it will cover the results from the unified scoring method. Exon 5, 6, and 8 won't be discussed since these exons have too few gRNA designs to make accurate comparisons.

5.1. Overview of Comparitive Analysis

Below, a detailed analysis is conducted to interpret the meaning of the values presented in the 5 tables. Since Benchling calculates scores in two different ways, both of its methods were compared with the other software.

Table 1: Analysis of Geneious vs. Guidescan2 Specificity Scores

Exon	Avg Guidescan2	Avg Geneious	Guidescan2 Std	Geneious Std	Wilcoxon p-value	Spearman Corr	Spearman p-value
Exon 1	44.70075714	97.30714286	24.91866208	1.916722152	0.015625	0.785714286	0.036238463
Exon 2	58.28923902	98.44146341	20.32650671	1.636313786	3.56939E-08	0.883467675	2.05065E-14
Exon 3	73.80683214	98.59964286	22.97010879	1.801080914	8.2981E-06	0.911459189	1.57777E-11
Exon 4	69.02059231	98.91538462	17.95410643	1.005552601	7.73973E-08	0.653499607	6.46802E-06
Exon 5	90.59365	99.645	5.553545949	0.233345238	0.5	-1	
Exon 6	48.4531	98.48	16.55010999	1.422638394	0.25	0.5	0.666666667
Exon 7	48.49209	97.766	22.569312	1.812004415	0.007685794	0.733333333	0.015800596
Exon 8	81.40755	97.98	30.61186643	3.768085279	0.125	0.8	0.2

Table 1 illustrates the scores for the software Geneious and Guidescan2. Geneious's standard deviation is small while Guidescan2 is quite substantial. From the Wilcoxon p-value in Exons 2–4, it is clear that the value is significant as the p-values are much smaller than 0.05. indicating there is a difference in score between the two software. Exons 1 and 7 also illustrated a significant p-value. This could be due to many things, depending on how the software tweaks the scoring algorithms. Looking at the Spearman correlation we see that Exons 1–3 and 7 have a pretty high correlation while Exon 4 has a moderate correlation. The Spearman p-value for the five Exons 1–4 and 7 illustrate that the correlation is indeed significant. This means that although there is a score discrepancy, the scores between Geneious and Guidescan2 are quite consistent.

Table 2: Analysis of Benchling (MIT-Broad) vs. Guidescan2 Specificity Scores

Exon	Avg Benchling	Avg Guidescan2	Benchling Std	Guidescan2 Std	Wilcoxon p-value	Spearman Corr	Spearman p-value
Exon 1	47.78025791	44.70075714	1.370756942	24.91866208	0.6875	0.857142857	0.013697327
Exon 2	49.03244498	58.28923902	0.98665663	20.32650671	0.006653709	0.68554007	7.59058E-07
Exon 3	49.01683393	73.80683214	1.243571455	22.97010879	1.29491E-05	0.854523068	7.18343E-09
Exon 4	51.76823384	69.02059231	10.64903949	17.95410643	9.68148E-06	0.550607287	0.000281315
Exon 5	49.87818735	90.59365	0.041880308	5.553545949	0.5	-1	
Exon 6	48.15213453	48.4531	2.188917407	16.55010999	1	0.5	0.666666667
Exon 7	48.4714193	48.49209	1.542989282	22.569312	0.6953125	0.854545455	0.001636803
Exon 8	48.9154969	81.40755	1.561028173	30.61186643	0.25	0.8	0.2

Table 2 illustrates the scores for the software Benchling (MIT-Broad) and Guidescan2. Exons 2–4 had significant Wilcoxon p-value meaning there is a difference between the specificity scores

between the two software. For the other exons, there does not seem to be a significant difference. This could be due to the smaller amounts of gRNA design. However, it is important to note that Exons 1 and 7 have a moderate amount of gRNA so the fact that the p-value is not significant in these two exons indicates that the score discrepancy is less than that of Table 1, which compares Geneious and Guidescan2. Exons 1, 3, and 7 have a high correlation while Exons 2 and 4 have a moderate correlation. The Spearman p-value for the Exons 1–4 and 7 indicate significance, meaning there does exist a correlation between the two software.

Table 3: Analysis of Benchling (MIT-Broad) vs. Geneious Specificity Scores

Exon	Avg Benchling	Avg Geneious	Benchling Std	Geneious Std	Wilcoxon p-value	Spearman Corr	Spearman p-value
Exon 1	47.78025791	97.30714286	1.370756942	1.916722152	0.015625	0.928571429	0.002519472
Exon 2	49.03244498	98.44146341	0.98665663	1.636313786	9.09495E-13	0.748595095	1.8245E-08
Exon 3	49.01683393	98.59964286	1.243571455	1.801080914	7.45058E-09	0.83431642	3.45304E-08
Exon 4	51.76823384	98.91538462	10.64903949	1.005552601	3.63798E-12	0.570039645	0.000151711
Exon 5	49.87818735	99.645	0.041880308	0.233345238	0.5	1	
Exon 6	48.15213453	98.48	2.188917407	1.422638394	0.25	1	0
Exon 7	48.4714193	97.766	1.542989282	1.812004415	0.001953125	0.83030303	0.002940227
Exon 8	48.9154969	97.98	1.561028173	3.768085279	0.125	1	0

Table 3 illustrates the scores for the software Benchling (MIT-Broad) and Geneious. Exons 1–4 and 7 all illustrate a significant p-value for the Wilcoxon paired test. This means that there is a clear discrepancy between the scores. This makes sense since both scores have small standard deviation with scores from Geneious having scores no lower than 90 while the scores from Benchling stay close to 50. The Spearman correlation however illustrates that there is a strong correlation in Exons 1–3 and 7. There is a moderate correlation in Exon 4. The Spearman p-value illustrated significance for Exons 1–4 and 7, meaning there is a correlation between the scores of the two software.

Table 4: Analysis of Benchling (CFD) vs. Guidescan2 Specificity Scores

Exon	Avg Benchling	Avg Guidescan2	Benchling Std	Guidescan2 Std	Wilcoxon p-value	Spearman Corr	Spearman p-value
Exon 1	23.40279056	44.70075714	10.15513946	24.91866208	0.015625	1	0
Exon 2	33.74270803	58.28923902	8.9128795	20.32650671	1.81899E-12	0.921254355	1.39219E-17
Exon 3	38.56320025	73.80683214	9.586712569	22.97010879	7.45058E-09	0.936909821	2.23034E-13
Exon 4	39.45060321	69.02059231	9.402787585	17.95410643	3.63798E-12	0.827530364	8.23778E-11
Exon 5	47.48530965	90.59365	1.564459116	5.553545949	0.5	1	
Exon 6	28.06815337	48.4531	7.548612161	16.55010999	0.25	0.5	0.666666667
Exon 7	29.00734724	48.49209	10.02568701	22.569312	0.001953125	0.927272727	0.000112035
Exon 8	41.23124278	81.40755	10.86027636	30.61186643	0.125	0.2	0.8

Table 4 illustrates the scores for the software Benchling (CFD) and Guidescan2. Exons 1–4 and 7 illustrated significance in the Wilcoxon paired test indicating that there is discrepancy between

the scores. However, the correlation between the two software for Exons 1–4 and 7 is really high. In fact, the correlation in Exon 1 is 1, indicating perfect correlation. The Spearman p-value also validates this correlation as Exons 1–4 and 7 have a p-value less than 0.05, indicating significance. Of all the comparisons, this table showed the highest correlation values. This makes sense since, as mentioned before, Guidescan2 is based on the MIT-Broad algorithm with the addition of CFD.

Table 5: Analysis of Benchling (CFD) vs. Geneious Specificity Scores

Exon	Avg Benchling	Avg Geneious	Benchling Std	Geneious Std	Wilcoxon p-value	Spearman Corr	Spearman p-value
Exon 1	23.40279056	97.30714286	10.15513946	1.916722152	0.015625	0.785714286	0.036238463
Exon 2	33.74270803	98.44146341	8.9128795	1.636313786	9.09495E-13	0.827532148	2.53084E-11
Exon 3	38.56320025	98.59964286	9.586712569	1.801080914	7.45058E-09	0.865836259	2.68171E-09
Exon 4	39.45060321	98.91538462	9.402787585	1.005552601	3.63798E-12	0.478780633	0.002046799
Exon 5	47.48530965	99.645	1.564459116	0.233345238	0.5	-1	
Exon 6	28.06815337	98.48	7.548612161	1.422638394	0.25	1	0
Exon 7	29.00734724	97.766	10.02568701	1.812004415	0.001953125	0.6	0.066688
Exon 8	41.23124278	97.98	10.86027636	3.768085279	0.125	0.4	0.6

Table 5 illustrates the scores for the software Benchling (CFD) and Geneious. Exons 1-4 and 7 illustrated significance in the Wilcoxon paired test indicating that there is a discrepancy between the scores. For Exons 1-3, there is a strong correlation while Exon 7 has a moderate correlation. Exon 4 showed a relatively weaker correlation with a value of less than 0.5. The Spearman p-value indicates that the correlation is indeed significant, other than that of Exon 7.

The comparison of central tendency can be seen in Figure 2.

A thing to note is that for Geneious, each specificity score is above 90%, which is somewhat suspicious as it indicates that all the gRNA designs generated have very little off-target effects. The consistency in high scores could be indication that the tool is potentially overlooking off-target effects. While this paper won't dive deeper into this issue, it could serve as an interesting study for future research.

5.2. Individual gRNA Designs

Of course, just looking at the correlations would not be enough. The individual rankings of each gRNA design were also analyzed, which led to some interesting results. For the interest of

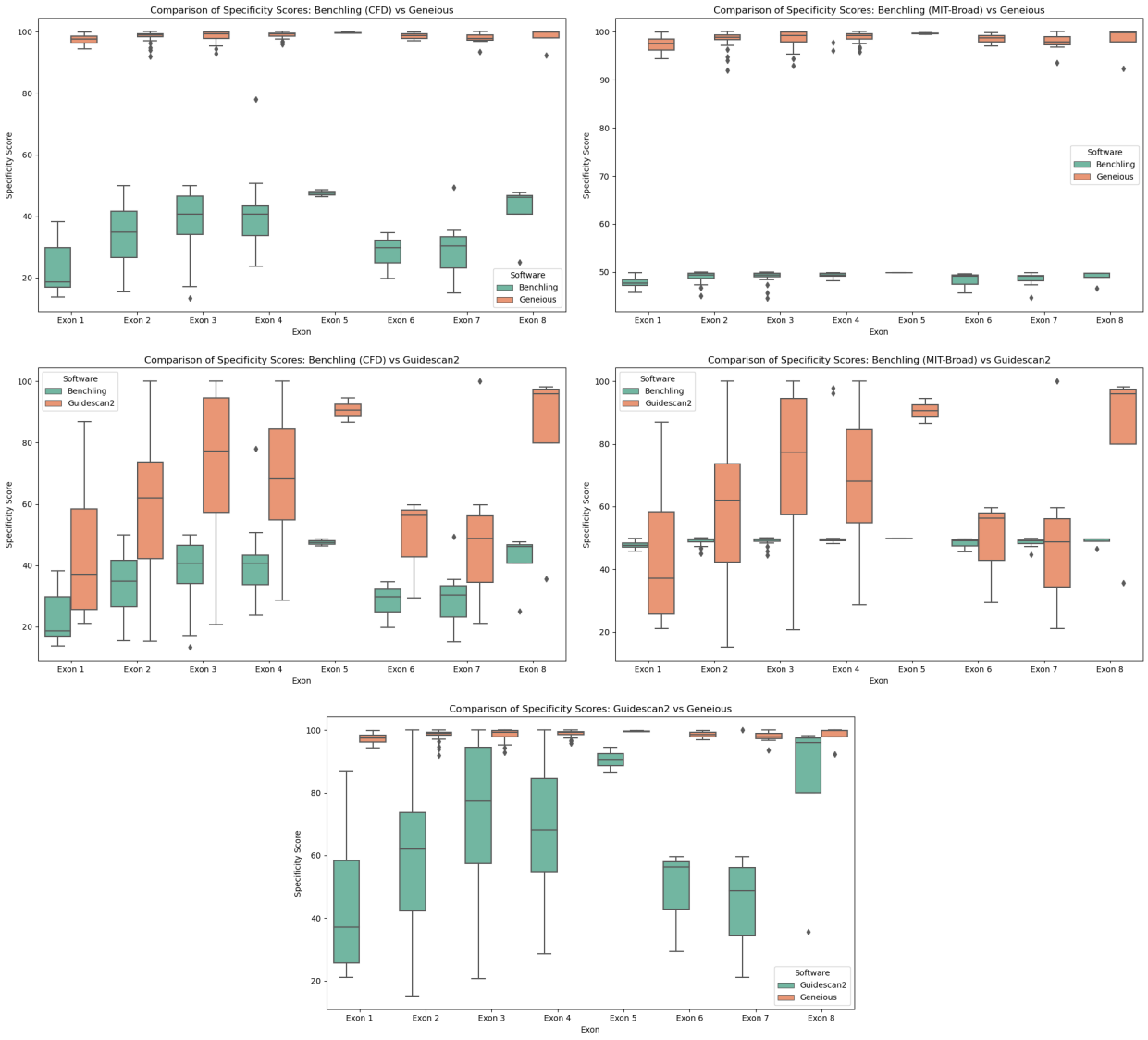

Figure 2: Specificity score comparison across 8 exons

discussion, only Exons 2-4 will be discussed since they contain the most gRNA designs, resulting in more accurate results.

There was only one gRNA which all software ranked the same. This gRNA was found in Exon 2, which will be talked more about below. Every other gRNA had slight or major deviations in ranking. Geneious and Benchling (MIT-Broad) had the worst correspondences in rankings, with 15 gRNA matches across the three exons and 24 gRNA that were more than 10 rankings off. Benchling (CFD) and Guidescan2 had the best correspondences in rankings, with 24 exact gRNA matches and only 8 gRNA that were more than 10 rankings apart. Exon 4 consistently did the worst in individual

ranking across all 5 comparisons. This makes sense since — looking back at all the tables, Exon 4 also seemed to have the lowest correlation score among the comparisons. For the Geneious vs. Benchling (CFD) comparison, although it had more matches throughout the three exons, there were no gRNA that were ranked the same in Exon 4 while 16 out of 39 gRNAs had a difference of 10 or more rankings. Looking back at Table 5, it is also clear that Exon 4 had the lowest correlation value. In fact, it is the lowest compared to all the other tables. This comes as a surprise if one just looks at Figure 2's correlation graph since the graph for the comparison of Geneious vs. Benchling (MIT-Broad) seemed to have a high correlation with the exception of a few outliers, which will be addressed later in this section.

Looking more closely at the selected gRNA (GCTGTGCGTGAAGAAGAGCA) from the previous experiment on *orco* [2], it is shown that the specificity score is not ranked as high. In fact, out of the 41 gRNA compared, it was ranked 27 in Benchling (MIT-Broad), 31 in Benchling (CFD), 34 in Guidescan2, and 36 in Geneious. The gRNA design that all three software collectively agree to be the best in terms of specificity score is TCTACAGAACGCTTGGCATA, which received a score of 50 from both of Benchling's score calculations and 100 from both Guidescan2 and Geneious. This gRNA could be of particular interest to future researchers who decide to work on *orco* in honey bees in the future.

In Figure 3, there seemed to be a consistent outlier when comparing Benchling to Guidescan2 and Geneious. These outliers are all from Exon 4. This is because in Exon 4, there are two gRNAs that Benchling ranked really high, with scores of 96.0949503 & 97.782487 for the MIT-Broad score and 50.678589 & 78.053705 for the CFD score. Comparing the two specific gRNAs (TGGCAGTGCAGAGAGCCGA & TGCAGAGCAGCCGAAGGAAC) against the Guidescan2 results, it was found that both tools generated two gRNA sequences with 4 mismatches for the first gRNA and both generated eleven gRNA sequences with 4 mismatches for the second gRNA. Guidescan2 did, however, generate another gRNA (CGTAGAGCAGTCGAAGGAAA) with PAM sequence "GGG" and mismatch of 4 for the second gRNA. After analyzing the individual off-target

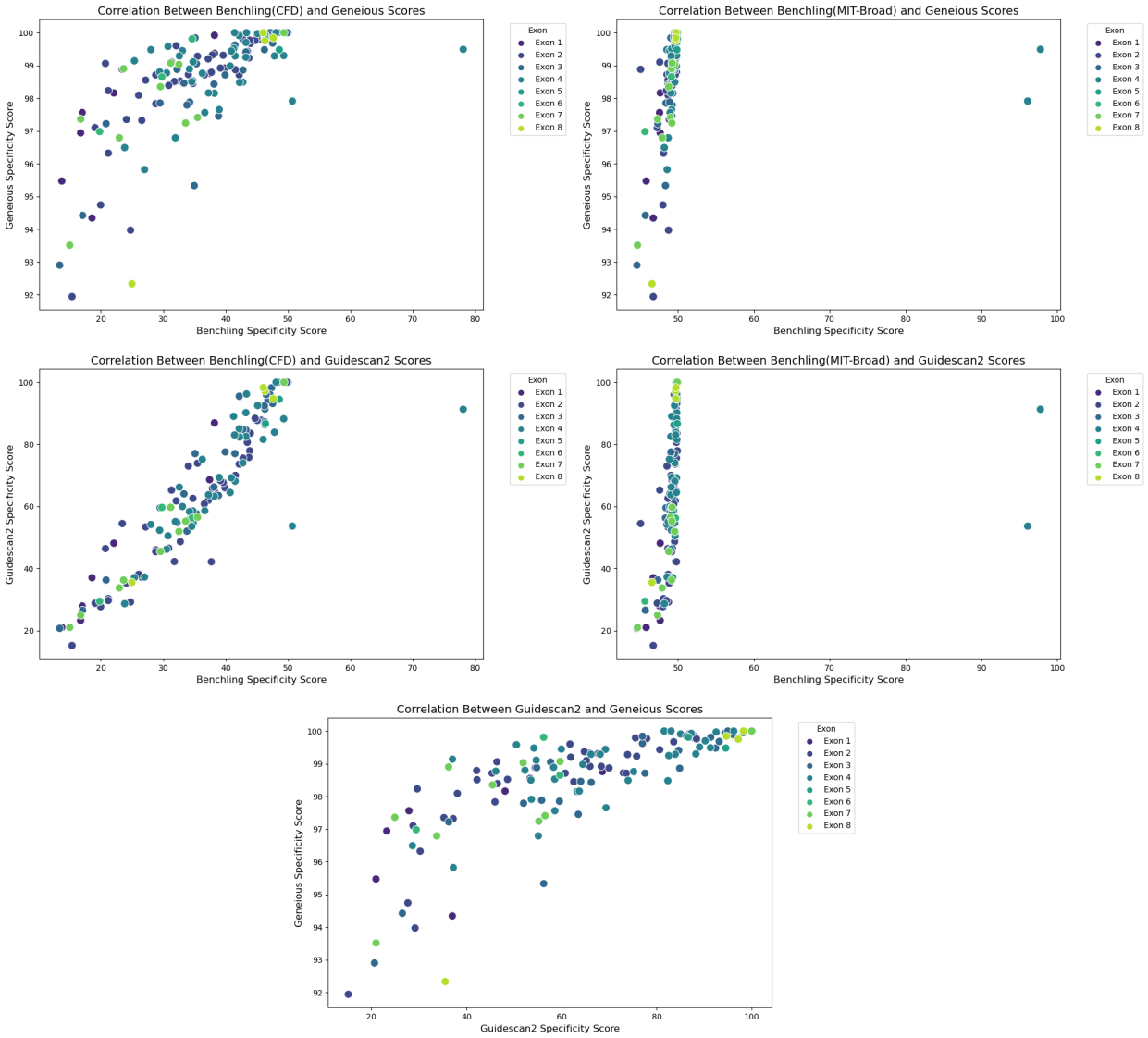

Figure 3: Specificity score correlation of the 8 exons

sites for the two gRNA designs, there seemed to be a slight inconsistency in Benchling. Every other gRNA in Exon 4 included the exact sequence with a score of 100 while the two gRNA mentioned above did not include the exact sequence. This seemed to have heavily affected the way the scoring was generated, which resulted in the major outliers as shown in the figure. A similar issue may have occurred in the past research on the knockout of the *LasR* gene in *Pseudomonas aeruginosa*. It is possible that the high-hit gRNA the researchers selected could also potentially be from this inconsistency. 15 best-performing hits were selected from a list of 115 possible gRNAs with all 15 hits having a specificity score of above 90 [10]. No comparison can be made with Geneious since

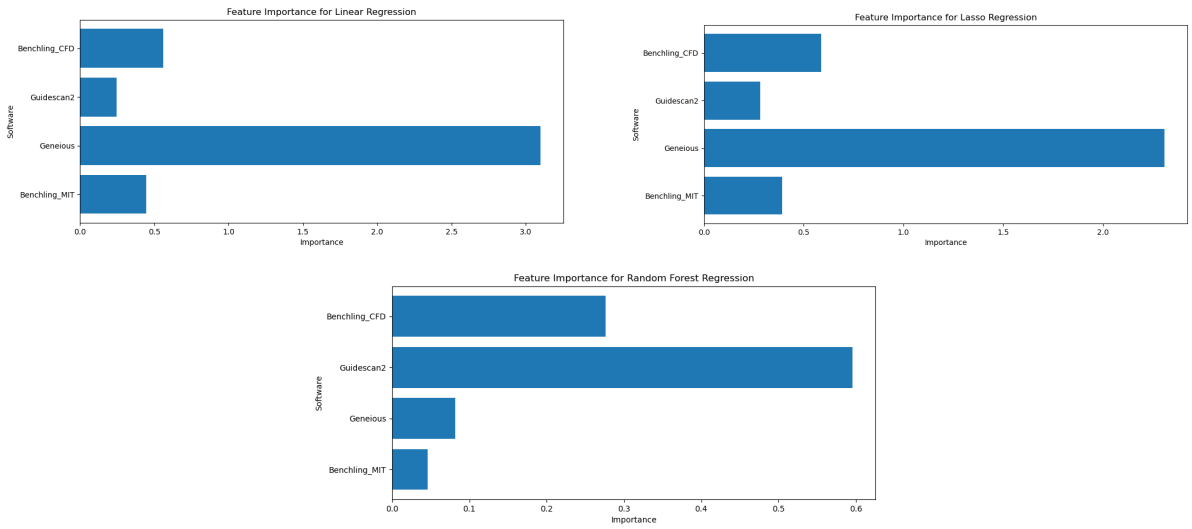

Figure 4: Feature Importance for Linear vs. Lasso vs. Random Forest Regression

access to the data was limited and only the gRNA sequences and scores were analyzed.

5.3. Unified Scoring System

With the given correlation and individual rankings, the development of a unified score also resulted in some interesting findings. In Figure 4, lasso and linear resulted in similar feature importance with the software Geneious placed at the highest importance. Random Forest selected Guidescan2 as the most important. From the MRE and R^2 scores, linear regression seemed to be the best model, but without the ability to validate the accuracy, it is hard to say which model works the best in reality. It is to note, as mentioned above, that the values of Geneious provided interesting specificity scores that require deeper investigation.

Additionally, the heavy lean towards Benchling (CFD) and Guidescan2 for Random Forest Regression connects back to the correlation found between the two software. With these two software specificity score correlation being the highest among the other comparisons, it makes sense that if Guidescan2's importance is high, then that of Benchling (CFD) would be relatively high as well.

6. Conclusions and Future works

6.1. Conclusion

It cannot be concluded which software is more accurate than another since no experimental testing was done. While the correlations between the different software illustrated some consistency, researchers should still be careful when using such tools as some variability still remain. This can be seen with Benchling's occasional scoring inconsistency and the discrepancies observed in Exon 4 between the tools. Exon 2 and 3, on the other hand, had the least discrepancies, making these two exons the preferable choices when selecting gRNAs for future experiments with *orco*. These findings support the idea that the implementation of scoring algorithms should be improved to reduce variability and improve reliability, which emphasizes the need for a standardized scoring system. Each software attempts to improve the scoring system in order to improve off-target effects but if the scoring system can be implemented across all software or mainstream software, it makes it significantly simpler for researchers. Research in the past has already mentioned this possibility. If the research community started reporting gRNAs on- and off-target activity to shared databases, such information could greatly improve guide selection, and enable fair comparison and benchmarking across publications while enabling more precise genome editing [13].

It is to be noted that the unified score in this paper is still very preliminary as there was not any accurate data to compare the predicted score against. However, that is not to say it cannot be expanded on in the future. The usage of similar methods to calculate the specificity scores makes it easier for software to adapt to a unified scoring system to make the selection of gRNA for experiments an easier process for researchers. Further research can expand on the unified score by incorporating more computational tools and potentially standardizing the methods for measuring and reporting off-target scores.

Although the computational tools discussed in this paper could help in designing gRNAs that are different from other off-target sites in the genome, there's still many other features that are

important to gRNA specificity. This includes the impact of seed sequence on gRNA abundance, seed abundance in the genome, and epigenetic feature [15].

6.2. Future Works

The findings suggest that some tools, such as Geneious, may be lenient in its specificity scoring as it consistently had scores above 90%. This raises the issue of the software's ability to detect off-target effects, suggesting a potential need for improvement in the software's implementation of the MIT-Broad algorithm. Future work could entail diving deeper into the methodology that Geneious uses and making sure it catches the off-targets sites.

Additionally, further work can be done to upload a version of the honey bee gRNA database onto the Guidescan2 web interface. This will be of good help for future researchers working on honey bees as it provides them with the information necessary without the need to run everything manually.

References

[1] "CRISPR - Geneious Prime User Manual — manual.geneious.com," https://manual.geneious.com/en/latest/CRISPR.html#specificity-scoring, [Accessed 07-01-2025].

[2] Z. Chen, I. M. Traniello, S. Rana, A. C. Cash-Ahmed, A. L. Sankey, C. Yang, and G. E. Robinson, "Neurodevelopmental and transcriptomic effects of crispr/cas9-induced somatic orco mutation in honey bees," *Journal of Neurogenetics*, vol. 35, no. 3, pp. 320–332, 07 2021. Available: https://doi.org/10.1080/01677063.2021.1887173

[3] J. G. Doench *et al.*, "Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9," *Nature Biotechnology*, vol. 34, no. 2, pp. 184–191, 2016.

[4] N. Fusi, I. Smith, J. Doench, and J. Listgarten, "In silico predictive modeling of crispr/cas9 guide efficiency," *bioRxiv*, 2015. Available: https://www.biorxiv.org/content/early/2015/06/26/021568

[5] D. Gleditzsch, P. Pausch, H. Müller-Esparza, A. Özcan, X. Guo, G. Bange, and L. Randau, "Pam identification by crispr-cas effector complexes: diversified mechanisms and structures," *RNA Biology*, vol. 16, no. 4, pp. 504–517, 04 2019. Available: https://doi.org/10.1080/15476286.2018.1504546

[6] M. Haeussler *et al.*, "Evaluation of off-target and on-target scoring algorithms and integration into the guide rna selection tool crispor," *Genome Biology*, vol. 17, no. 1, p. 148, 2016. Available: https://doi.org/10.1186/s13059-016-1012-2

[7] P. D. Hsu *et al.*, "Dna targeting specificity of rna-guided cas9 nucleases," *Nature Biotechnology*, vol. 31, no. 9, pp. 827–832, 2013. Available: https://doi.org/10.1038/nbt.2647

[8] C. Li, W. Chu, R. A. Gill, S. Sang, Y. Shi, X. Hu, Y. Yang, Q. U. Zaman, and B. Zhang, "Computational tools and resources for crispr/cas genome editing," *Genomics, Proteomics & Bioinformatics*, vol. 21, no. 1, pp. 108–126, 2023. Available: https://www.sciencedirect.com/science/article/pii/S1672022922000274

[9] T. Li, Y. Yang, H. Qi, W. Cui, L. Zhang, X. Fu, X. He, M. Liu, P.-f. Li, and T. Yu, "Crispr/cas9 therapeutics: progress and prospects," *Signal Transduction and Targeted Therapy*, vol. 8, no. 1, p. 36, 2023. Available: https://doi.org/10.1038/s41392-023-01309-7

[10] L. Radha KesavanNair, "Computational design of guide rnas and vector to knockout lasr gene of pseudomonas aeruginosa," *Gene and Genome Editing*, vol. 6, p. 100028, 2023. Available: https://www.sciencedirect.com/science/article/pii/S2666388023000047

[11] M. Redman, A. King, C. Watson, and D. King, "What is crispr/cas9?" *Archives of Disease in Childhood - Education and Practice*, vol. 101, no. 4, pp. 213–215, 2016. Available: https://ep.bmj.com/content/101/4/213

[12] H. Schmidt, M. Zhang, H. Mourelatos, F. J. Sánchez-Rivera, S. W. Lowe, A. Ventura, C. S. Leslie, and Y. Pritykin, "Genome-wide crispr guide rna design and specificity analysis with guidescan2," *bioRxiv*, 2022. Available: https://www.biorxiv.org/content/early/2022/05/03/2022.05.02.490368

[13] J. Tycko, V. E. Myer, and P. D. Hsu, "Methods for optimizing crispr-cas9 genome editing specificity," *Molecular Cell*, vol. 63, no. 3, pp. 355–370, 2025/01/07 2016. Available: https://doi.org/10.1016/j.molcel.2016.07.004

[14] L. O. W. Wilson, A. R. O'Brien, and D. C. Bauer, "The current state and future of crispr-cas9 grna design tools," *Frontiers in Pharmacology*, vol. 9, 2018. Available: https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2018.00749

[15] X. Wu, A. J. Kriz, and P. A. Sharp, "Target specificity of the crispr-cas9 system," *Quantitative Biology*, vol. 2, no. 2, pp. 59–70, 2025/01/07 2014. Available: https://doi.org/10.1007/s40484-014-0030-x

[16] X.-H. Zhang, L. Y. Tee, X.-G. Wang, Q.-S. Huang, and S.-H. Yang, "Off-target effects in crispr/cas9-mediated genome engineering," *Molecular Therapy Nucleic Acids*, vol. 4, 2025/01/07 2015. Available: https://doi.org/10.1038/mtna.2015.37