THESIS

ESTIMATING POTENTIAL BIOSAFETY RISK OF A CRISPR-CAS9 SYSTEM FOR TARGETED KILLING OF CERTAIN PATHOGENS IN BEEF CATTLE PRODUCTION USING OMIC-BASED ANALYSIS METHODOLOGIES

Submitted by

Jixin Dong

Department of Animal Sciences

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2021

Master's Committee:

Advisor: Hua Yang

Keith E. Belk
Jessica L. Metcalf
Tiffany L. Weir

ABSTRACT

ESTIMATE POTENTIAL BIOSAFETY RISK OF A CRISPR-CAS9 SYSTEM FOR TARGETED KILLING OF CERTAIN PATHOGENS IN BEEF CATTLE PRODUCTION USING OMIC-BASED ANALYSIS METHODOLOGIES

The CRISPR-Cas9 system has emerged as a programmable and versatile tool for precise gene editing purposes. In addition to gene editing, the CRISPR-Cas9 system can be developed to kill targeted bacteria. In our previous studies, we developed a CRISPR-Cas9 targeted killing system with a guide RNA designed to specifically recognize the Shiga-toxic genes (*stx1* and *stx2*). Delivery of this system into *E. coli* cells could effectively kill Shiga-toxin producing *E. coli* (STEC) cells. This current study was conducted to estimate potential biosafety risk associated with our CRISPR-Cas9-based targeted killing system when applied to kill STEC cells in a bovine cell line model system using next generation sequencing (NGS) analysis.

A bovine cell line CPA 47 (ATCC® CRL-1733™) was cultured to reach 90% confluence. Then the bovine cells were subjected to one of four treatments: 1) bovine cell control: without any CRISPR treatment; 2) CRISPR/gRNA: treated with phages that carry the CRISPR system with the guide RNA targeting *stx* genes (10^6 PFU/flask); 3) CRISPR+O157: treated with phages that carry the CRISPR system but without the guide RNA targeting *stx* genes (10^6 PFU/flask) and *E. coli* O157:H7 strain Sakai cells (10^5 CFU/flask); and 4) CRISPR/gRNA+O157: treated with phages that carry the CRISPR system with the guide RNA targeting *stx* genes (10^6 PFU/flask) and *E. coli* O157:H7 strain Sakai cells (10^5 CFU/flask). Each treatment was conducted in four replicates for a total of 16 samples. After application of treatments, bovine

cells from each sample were collected and divided into two portions: half for whole genome sequencing (WGS) and half for protein analysis. Whole genome DNA from each sample was extracted, purified, and sent to Novogene Bioinformatics Technology (Beijing, China) for library construction and WGS. Raw reads were subjected to quality control (QC) procedures to remove unusable reads. Clean reads after QC were aligned to the bovine reference genome (NCBI access number: ARS-UCD 1.2 USDA ARS) using Burrows-Wheeler Aligner (BWA) with default parameter. Based on the mapping results, SAMtools was used to detect individual SNP/InDel variants, and ANNOVAR was used for functional annotation of the detected variants.

A total of 1078 Gb of output with 3593.12 million paired end reads (150 bp) were obtained for all 16 samples after NGS. After QC, a total of 1073 Gb of clean data output were obtained for all 16 samples. The average output for the four replicates within the control, CRISPR/gRNA, CRISPR+O157, and CRISPR/gRNA+O157 treatments was 59.1, 72.3, 72.3, and 65.9 Gb, respectively. The mapping rate of each sample ranged from 99.48% to 99.75%, and the 4X coverage ranged from 89.45% to 98.23%. For SNP detection, a total of 94,796,157 SNPs were identified in all 16 samples when compared with the reference genome. The number of SNPs with each sample ranged from 5,225,269 to 6,192,930. Of the total number of SNPs, 60.01% were located in intergenic regions (regions between genes), 36.40% in intronic regions (non-coding sequences of genes), and 0.79% in exonic regions (coding sequences of genes). Further analysis of exonic regions showed that the average SNPs for each treatment of control, CRISPR/gRNA, CRISPR+O157, and CRISPR/gRNA+O157 was 0.1964%, 0.2002%, 0.2002%, and 0.1948%, respectively. SNPs within each functional class, for example, stop loss, stop gain, synonymous, non-synonymous, at slicing sites, and upstream or downstream from transcription termination sites, the number of SNPs all showed no significant differences *($P > 0.05$)* among

the control and the three CRISPR treatments. For InDel detection, a total of 11,949,421 InDels were identified in all 16 samples, with each sample having between 604,387 and 817,716 InDels. Of the total number of InDels, 62.78% were located in intergenic regions, 38.54% in intronic regions, and 0.15% in exonic regions. The average exonic InDels in the control, and CRISPR/gRNA, CRISPR+O157, and CRISPR/gRNA+O157 treatments was 0.0371%, 0.0372%, 0.0375%, and 0.0372%, respectively. InDels within each functional class, for example, stop loss, stop gain, frameshift insertion/deletion, non-frameshift insertion/deletion, at slicing sites, and upstream or downstream from transcription termination sites, the number of InDels all showed no significant differences *($P > 0.05$)* among the control and the three CRISPR treatments.

Neither the SNP nor InDel data showed significant differences *($P > 0.05$)* in the number of SNPs/InDels between the control and the CRISPR treatments when their WGS data were compared with the bovine reference genome. These results go along with our initial prediction that the biosafety concern of our CRISPR-Cas9 system should be low because our CRISPR system was designed to make a cleavage on target bacterial genomes but not on cattle genomes. In addition to coding regions, we are continuing with our analysis of the variants in non-coding regions. Results from this study will provide insights on how to further improve approaches or develop criteria on biosafety evaluation of the CRISPR-Cas9 system. Completion of this project will provide the beef industry with biosafety information regarding application of CRISPR as an alternative to antibiotics in cattle production.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank several people who have offered invaluable assistance in the preparation of the thesis. First of all, I would like to thank my fellow students who have been with me for two years for all your help, friendship. Thanks for the two happy years we have spent together.

Then, I would like to thank the committee members for their patience, guidance, and support during my Master's Program. My deepest gratitude goes foremost to my advisor Dr. Yang Hua. Thank you for the encouragement and patience in supporting me through the complex experiments. I have great respect for your intelligence and rich experience. Every idea you suggested helped me to advance the project smoothly. Every piece of advice you have given benefited me in the project. You have been a great mentor to me both in my studies and in my life. I feel very fortunate to have met such an incredible advisor like you. I must express my most sincere thanks to you, thank you very much. I really appreciated all your hard work and patience.

Thank you to the Department Head and my committee member, Dr. Keith Belk, for the excellent opportunity to learn and work in the meat science group. Thank you for your encouragement and care and for your tireless work in positively leading the group.

Thanks to two friends: Yao Zhu, a graduate student majoring in Statistics, helping me solve some coding problems with data analysis; and Cheris, for the accompany while I was working on my thesis.

Finally, I would like to thank all my professors, family, and friends who have always encouraged and supported me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

CRISPR-Cas9 Technology

Since the beginning of the 21st century, gene-editing technology has made incredible progress in biotechnology. Gene editing technology allows scientists to act as "creators" and modify plant and animal genes for different purposes. Clustered Regularly Interspaced Short Palindromic Repeat-Cas9 (CRISPR-Cas9) is currently the most popular gene-editing tool that allows targeted editing of specific sites in the genome, including insertion, repair, and replacement. Therefore, the CRISPR-Cas9 technology is being studied in a wide range of fields. It has been used to research cancer, genetic diseases, pathogen detection, killing of bacteria, and drug-resistant bacteria strains. The CRISPR-Cas9 system has been successfully used to disrupt *HIV-1* provirus (Ebina et al., 2013), *human papillomavirus* (HPV) (Kennedy et al., 2014), and *hepatitis B* virus (Kennedy et al., 2015). In addition, the CRISPR-Cas9 system has been used to target human hereditary liver diseases (Yang et al., 2016; Yin et al., 2016); and in cancer (Chen et al., 2019) and Hutchinson-Gilford premature aging syndrome (Beyret et al. 2019) CRISPR_Cas9 has shown great promise as a treatment. In addition, the CRISPR-Cas9 system is used for nucleic acid detection associated with pathogens or diseases (Gootenberg et al., 2017), and Chen et al. (2018) detected two different genotypes of HPV in patient samples accurately. With the CRISPR-Cas9 technology, more research directions continue to develop in the fields of medical, natural and agricultural sciences.

Natural CRISPR system in bacterial cells and their functions

The CRISPR-Cas9 system is an RNA-based adaptive immune system in bacteria and archaea (Barrangou et al., 2007). The system was first discovered in *Streptococcus thermophilus*,

including the CRISPR-associated endonuclease Cas9 (Barrangou et al., 2007). The native CRISPR-Cas9 system is used to defend the exogenous DNA from virus infection or plasmids. When the virus first invades, the bacteria confer resistance to the virus by integrating short repetitive sequences (spacer) of viral DNA into the bacterial genome (Barrangou et al., 2007). When bacteria are infected secondly by the same virus, transcripts of these short repeats with Cas9 endonucleases bind to the invading viral DNA, activating endonuclease activity, and thereby protecting the bacteria from destroying the viral DNA (Barrangou et al., 2007).

The mechanism of CRISPR-Cas9 system

The natural CRISPR-Cas9 system has been adapted as a gene editing tool because scientists found that it could be reconstituted by using three main components: trans-activating CRISPR RNA (tracrRNA), CRISPR RNA (crRNA) and Cas9 protein. Single-guide RNA (sg RNA) is composed of tracrRNA and crRNA (Doundna & Charpentier, 2014). Therefore, CRISPR-Cas9 system is a functional complex consisting of single-guide RNA and endonuclease protein Cas9. The Cas9-sgRNA recognizes the proto-spacer-adjacent motif (PAM) sequence of the target DNA, unlocking the double strand of the target DNA. Then, sgRNA is paired with the target DNA, formed an R ring, and Cas9 cleaves the DNA double-strand.

When the guide RNA in CRISPR-Cas9 system is designed to target Eukaryotes' genome, it works as a gene editing tool. However, when the guide RNA in CRISPR-Cas9 system is designed to target bacterial genomes, it works as an antimicrobial. The reason for this difference is that the double-strand break (DSB) can be repaired by eukaryotes, but it is lethal to bacteria. Eukaryotes have self-repair mechanisms through non-homologous end joining (NHEJ) and homologous repair (HR) (El-Mounadi et al. 2020). While multiple studies have indicated that cleavage of the chromosome leads to cell death in many bacteria (Jiang et al., 2013). Therefore,

it is also reasonable to use the CRISPR-Cas9 system to kill bacteria by deleting specific genes and not reinserting others.

Other study of using CRISPR-Cas9 system to control bacteria

An example of using CRISPR-Cas9 antimicrobials was to control bacterial drug resistance. In some studies, scientists designed the guide RNA to target the resistant gene, which allowed the bacteria to resist antibiotics such as penicillin and methicillin (Shabbir et al., 2019; Van Der Oost et al., 2014). With help from the programmed CRISPR-Cas9 system, they transformed the antimicrobial resistance (AMR) pathogens to antibiotics-sensitive cells by removing resistance genes. The antibiotic resistance gene (like the *mecA* gene) was precisely degraded in methicillin-resistant *Staphylococcus aureus* (MRSA). Their results showed that MRSA without *mecA* gene could be easily killed by penicillin and other analogues, with killing efficiency ranging between 10^3 and 10^5 compared to the control (MRSA with *mecA* gene).

Previous studies of the CRISPR-Cas9 system to control STEC at CSU

Previous studies used the CRISPR-Cas9 system as an antimicrobial to selectively kill *E. coli* O157:H7 (Yang et al. 2017; Yang et al. 2018). A designed guide RNA was created to target the Shiga-toxin genes (*stx1* and *stx2*) in *E. coli* O157:H7, and when combined with endonuclease Cas9, the CRISPR-Cas9 system had the ability to recognize *stx1* and *stx2* genes and cleave bacterial chromosomes at the loci of those genes. To deliver the CRISPR-Cas9 systems into cells and to target *stx1* and *stx2* genes, the system was packaged into a phagemid. This phage-mediated system was constructed to deliver the CRISPR-Cas9 based targeted killing system into *E. coli* O157:H7 in both pure culture and cattle rumen fluid. Results demonstrated that the CRISPR-Cas9 systems were effectively delivered and then targeted on *E. coli* O157:H7. Compared to the group with only phagemids, a significant reduction of *E. coli* O157:H7 was

observed in the group with CRISPR-Cas9 targeted killing system in both environments (pure culture and cattle rumen fluid).

In conclusion, CRISPR-Cas9 technology has been popular with great diversity of research in different fields because the CRISPR-Cas9 system could be programmable and versatile by changing the DNA target-binding sequence in the guide RNA. The endonuclease Cas9 with programmed guide RNA could efficiently recognize, bind, and cleave any double-strand DNA sequence of target genes that we interest.

Biosafety Concern

After demonstrating that the CRISPR-Cas9 system is useful and effective in selectively killing certain strains of pathogens, it is essential to assess the biosafety risk of the CRISPR-Cas9 system before it is used in beef production. It is not worth being commercialized and applied in real production if biosafety issues outweigh benefits. The CRISPR-Cas9 technology as biotechnology allows specific modification of genes, causing mutations or genetic damage directly or indirectly. Although there are substantial individual and societal benefits from applying CRISPR-Cas9 technology, its biosafety risks create significant concerns for individuals, society, and the environment. The biosafety aspect was highlighted very early as a critical limitation that would need to be resolved before any application of gene editing on humans or release into the environment could occur (Akbari et al., 2015). On the other hand, the biosafety has received more attention due to the identification of gene editing tools as one of the national security threats by relevant government officials (Oye et al., 2014).

The potential biosafety risk caused by applying CRISPR-Cas9 technology are not adequately corroborated. Also, there is a debate regarding the methodologies that should be applied to evaluate the biosafety risks of CRISPR-Cas9 technologies. Therefore, using CRISPR-

Cas9 technology in the food industry for production and safety purposes in vegetables and meats (Sovová et al. 2016) creates unique and novel challenges to biosafety. Additionally, the application of CRISPR-Cas9 technology in agricultural plants and animals for breeding brings plenty of biosafety concerns. When studies involve treating various genetic diseases, infectious diseases, and cancers, the safety of CRISPR-Cas9 technology should also be considered.

Main biosafety issues that could be attributed to gene editing tools are the number of off-target effects and the potential epigenetic effects (Rath et al., 2017). These biosafety issues not only are possible when using the CRISPR-Cas9 system, but also when other gene-editing tools are used. Therefore, it is reasonable to consider that approaches used to study the biosafety risk of other gene editing tools could be used to assess the biosafety risk of the CRISPR-Cas9 system. Moreover, early identification, assessment, and governance of the risks of the CRISPR-Cas9 system will contribute to long-term development of this system. Currently, there are several examples of biosafety testing, such as using somatic gene therapy modified by immune cells to treat cancer (Rath et al., 2017), modifying HIV with CRISPR-CAS-based technology (Rath et al., 2017). The gene editing techniques are safe and effective in their clinical trials. Results of a few studies are not sufficient to illustrate whether the biosafety of the CRISPR-Cas9 system is low or high. However, as biosafety testing and analysis results accumulate, we will have a more comprehensive understanding of the biosafety of the CRISPR-Cas9 system.

Off-target effects of the CRISPR-Cas9 systems as gene editing tools against Eukaryotes

The main biosafety issue limiting development and application of CRISPR-Cas9 technology is the off-target effect. As a gene editing tool, CRISPR-Cas9 system works directly on the genome of animals and plants. The off-target impact would occur due to some mismatch between CRISPR-guide RNA and target RNA sequence or genes that are very similar in the

genome. In a study of Zhang et al. (2015), RGEN (RNA-guided endonuclease) induced mutations at sites other than the intended on-target site, leading to the high frequency of off-target activity (over 50%).

Specificity of the CRISPR/Cas9 gene-editing system is determined by the recognition sequence (~20nt) on guide RNA. However, in complex biological genomes, the recognition sequence of guide RNA may be a partial match with non-target DNA (Doench et al., 2016). These local matches can be divided into two categories (Doench et al., 2016): (1) guide RNA and off-target DNA sequences have the same length, but there is a base mismatch; (2) The length of the off-target DNA sequence is several bases longer than that of guide RNA, and the correct pairing of other bases can be achieved by the formation of DNA or RNA bulge. Furthermore, previous studies have shown that different guide RNA structures affect the cleavage of on-target and off-target sites (Zhang et al., 2015). Therefore, designing guide RNAs that could reduce off-target effects is an appropriate way to decrease the biosafety risk of CRISPR-Cas9 system.

Biosafety concerns over CRISPR-Cas9 targeted killing system against bacteria

In previous studies, we designed a guide RNA that targeted the Shiga-toxin gene (*stx1 & stx2*) in *E. coli* O157:H7. Then, a phage-mediated system was constructed for delivery of the CRISPR-Cas9 system into *E. coli* O157:H7. These studies demonstrated that the CRISPR-Cas9 system can effectively kill *E. coli* O157:H7 cells inoculated into cattle rumen fluid.

When applying the CRISPR-Cas9 targeted killing system on cattle, the biosafety concern is that the CRISPR-Cas9 targeted killing system would impact endogenous cells of cattle. To be specific, it is necessary to assess whether the CRISPR-Cas9 targeted killing system causes nucleic acid and protein changes in bovine cells.

Theoretically, the CRISPR-Cas9-based targeted killing system should have a low biosafety risk. This is primarily because the CRISPR-Cas9 targeted killing system used in this project was designed to target virulence genes in *E. coli* O157:H7 (bacterial cells), but not cattle cells. So, in theory, the CRISPR-Cas9 system should not impact the genome of cattle directly. Therefore, even if the CRISPR-Cas9 targeted killing system affects any off-target cells, it would occur in bacterial cells only, not affecting cattle cells. In addition, a phage-mediated system is used to deliver the CRISPR-Cas9 system into *E. coli* O157:H7. The phage, also known as a bacteriophage, is very species-specific regarding host and usually only infects a single bacterial species or even specific strains within a species. The bacteriophages used to deliver the CRISPR-Cas9 system were specifically chosen to infect *E. coli* O157:H7, thus the CRISPR-Cas9 system should not be delivered into cattle cells. That further reduces potential risks of bovine cells being affected by the Shiga toxin targeted CRISPR-Cas9 system.

Another biosafety concern is that the CRISPR-Cas9 Shiga toxin targeted *E. coli* O157:H7 killing system may mistakenly kill other types of microbiological cells in the rumen and intestine of cattle that are beneficial to the cattle. Although infrequent, such a scenario could have the undesirable consequence of upsetting the balance of the gut flora. Though this concern is not our goal in the project, it is worthy of attention and exploration in the future.

So far, the biosafety concerns associated with use of the CRISPR-Cas9 Shiga-toxin gene targeted killing system on non-target organisms has not been studied. And no other similar studies have been done in estimating the biosafety risk of CRISPR-Cas9 targeted killing system.

Omics-based methodologies

Previous studies commonly use omics-based methodologies to assess macro/micro molecules (such as genes, transcripts, proteins, and metabolites) in cells. Omics-based

methodologies contain applications of high-throughput genomics, transcriptomics, proteomics, metagenomics, epigenomics, etc. (Gomez-Chiarri et al., 2015). The high-throughput sequencing technologies and CRISPR-Cas technologies benefit each other, and both have made a great progress. The development of fast, low-cost, high-throughput technologies has led to an increase in the number of CRISPR-Cas technology studies year by year (Gomez-Chiarri et al., 2015) and also provides an important approach to detect biosafety risks.

Omics-based disciplines focus on analysis of the structure and function of the genetic material (genomics), expressed genes (transcriptomics), proteins (proteomics), and low molecular weight metabolites (metabolomics) in an organism (Gomez-Chiarri et al., 2015). These technologies allow researchers to investigate the complex relationships between genotypes, phenotypes, and the environment by simultaneously analyzing large numbers of individuals, genes, proteins, or metabolites in samples directly collected from the environment (Carvalho & Creer et al., 2009). Therefore, omics-based analysis is particularly suited to study the complex interactions between genome editing tools, individuals, and the environment that may lead to cellular variants. All these omics tools can be applied to either single organism or complex mixtures of organisms.

The advantages of next generation sequencing

Next generation sequencing (NGS) is commonly used to identify mutations and variants in cells. The biggest advantage of NGS technology is its ability to perform massively parallel sequencing of genomes. In the same conditions, NGS technology can sequence and report DNA samples in a much shorter time than traditional sequencing technologies (Luthra et al., 2015).

Moreover, traditional sequencing technologies require a large quantity of multiple input of nucleic acids, whereas NGS technology requires a relatively small amount of DNA or RNA in

a single input (Luthra et al., 2015). This advantage reduces the total cost of using NGS technology. Additionally, NGS is also appropriate for detecting various genomic variants in experimental samples (Luthra et al., 2015), such as single nucleotide variants (SNVs), small-size insertions and deletions (InDels), copy number variation (CNVs). NGS technology can also be used to repeat sequencing of target fragments of interest's genes, allowing higher confidence and sensitivity of variation detection.

Studies demonstrated that NGS has developed dramatically in producing increased data output and increasing efficiencies (Luthra et al., 2015; Hu et al., 2021). These considerable advantages listed above make NGS technology desirable for finding variants in this project.

The challenges of next generation sequencing

Next-generation sequencing technologies also face several challenges. One challenge is that NGS processes are associated with sequence errors originating from library preparation, the sequencing process itself, or data analysis (e.g., read maps, variant calling), resulting in incorrect calls to DNA bases or variants (Buermans & Den Dunnen, 2014). In addition, incorrect sequence variants may also originate from the process of data analysis and require further computational correction of errors (Bacher, 2018). Besides, NGS platforms commonly generate short reads between 100 ~ 500 bp. The disadvantage of short-read sequences is the tendency to miss structural variants (Bacher, 2018), such as longer insertions and deletions. Although paired-end sequencing can alleviate this problem by allowing the same reads from both ends to provide additional localization information, this error correction approach requires more complicated library preparation and data analysis, more sequencing time, and larger data volumes (Buermans & Den Dunnen, 2014). Finally, post-sequencing steps such as filtering variant numbers, comparing data with reference genome or databases, and interpretation of gene variants have

become a challenge (Bacher, 2018). The output data of NGS is huge, and its analysis is usually time-consuming. People with in-depth bioinformatic knowledge and experience may be able to avoid repeated analysis, and thereby saving some time.

Small variants - single nucleotide polymorphism (SNP)

Single nucleotide polymorphism (SNP) is a DNA sequence polymorphism caused by a variation in a single nucleotide at the genomic level (Vignal et al., 2002). Since only one single deoxynucleotide is altered, SNP belongs to single-nucleotides variants (SNV) which is a small variant being detected by next-generation sequencing. Single nucleotide polymorphisms can be classified depending on their location in the gene (Figure 1), such as exonic regions (protein-coding regions), intronic regions (non-coding regions of protein), and intergenic regions (regions between genes). Any base can be mutated in genomic DNA so that SNPs can be both within the gene sequence and on non-coding sequences outside the gene.

In general, SNPs located within coding regions (cSNPs) could receive more attention because of their ability to directly influence the amino acid sequence of the protein, altering the phenotype of the organism. Additionally, the number of coding regions are very small in the genome. For example, coding regions account for approximately 1 precent of the human genome (Zhao, 2012), but they are pivotal. And far more studies have been done on cSNPs than on variants located in non-coding regions.

There are four types of single nucleotide polymorphisms (SNPs) in the exonic region: synonymous, non-synonymous, stop gain and stop loss. Due to the degeneracy of the gene codon, single nucleotide polymorphisms (SNPs) containing gene-coding sequences may not necessarily change the protein's amino acid sequence. A synonymous single nucleotide polymorphism (synonymous SNP) does not affect the protein sequence, while a non-

synonymous single nucleotide polymorphism (non-synonymous SNP) changes the amino acid sequence of proteins. Both stop gain and stop loss belong to the non-synonymous SNP type. Stop gain SNP changes one base of a codon, leading to a premature stop codon (so a stop codon is gained), which would likely lead to a shortening of the protein sequence. On the contrary, stop loss SNP changes one nucleotide acid of a codon, resulting in a loss of a stop codon. It has the consequence that the protein sequence would be longer than the original one or may not be translated. Non-synonymous variants are likely to lead to changes in gene functionality, so we would focus more on exonic SNPs, especially non-synonymous types, in the result and further discussion.

SNPs in non-coding regions of the gene (intronic region) could also harbor crucial functional elements that affect gene expression and RNA splicing (Krawczak et al., 2007). SNP in intergenic regions usually brings no effects or unknown effects due to insufficient research, but recent studies (Schierding et al., 2016; Kumer et al., 2013) showed that the intergenic regions might be attributed to potential regulatory functions of DNA sequence. Besides, upstream and downstream of the gene (belongs to intergenic regions) are considered more important because of containing promoters and enhancers (Hua et al., 2018), which may regulate the expression of discrete genes over thousands of base-pair distances on the genome. Upstream SNP is defined as a SNP located within 1000 base pairs (bp) away from the gene's transcription start site, toward the 5' end of the coding strand. Similarly, downstream SNP serves as a SNP located within 1000 bp away from the transcription termination site of the gene region toward the 3' end.

Additionally, two types of DNA substitution caused by the SNP should be noticed. Transitions are interchanges of two-ring purines (A⇔G) or one-ring pyrimidines (C⇔T) (Vignal et al., 2002). Transversions are interchanges of purines for pyrimidine bases, which involve

exchange of one-ring and two-ring structures (G⇔C, G⇔T, A⇔C, A⇔T), or vice versa (Vignal et al., 2002). Due to the molecular mechanisms that allows them to generate, the number of transversions is twice as many possible than that of transitions (Vignal et al., 2002). If the mutation is random, then the transitions over transversions ratio should be 0.5 theoretically.

Small variants – small insertion or deletion (InDel)

InDels refer to the insertion or deletion of ⩽ 50 bp sequences in the double-stranded DNA. Single nucleotide polymorphisms (SNPs) are ubiquitous, and the relative maturity of detection technologies (Narzisi & Schatz, 2015), thus scientists have focused on single nucleotide polymorphisms (SNPs) in genomic studies in the last two decades. However, in recent years, an increasing number of studies have begun to analyze the role of insertion and deletion (InDels) variants (Narzisi & Schatz, 2015). That is because they are involved in protein production through small InDels that cause frameshift mutations as well as larger InDels that fundamentally alter genes, change splicing and binding sites, or disrupt other important genomic sequences (Narzisi & Schatz, 2015). Although InDels generally occur less frequently than SNPs, there is great diversity in the size of InDels (Montgomery et al., 2013). Either one single nucleotide insertion or deletion or multi-nucleotides insertion or deletion may lead to significant variants in the genome (Montgomery et al., 2013). In 2002, Webel et al. identified a total of 2000 InDels in human's genome and their study was one of the earliest genome-wide InDel discovery efforts in human. Most of these InDels (96%) were 2–16 bp in length, and the largest InDel was 55 bp in length. In 2010, a study (Mullaney et al., 2010) confirmed that InDel variation is the second most abundant form of genetic variation in humans after SNPs (79% of the variants detected were SNPs and 21% of the variants were InDels). It is valuable to analyze InDel variants, even if time-consuming.

The same as SNPs, InDels could also be identified depending on their location: exonic regions, intronic regions, and intergenic regions (Figure 1). InDels located in exonic regions may contribute to protein changes that may have diverse effects on the organism's phenotype, whereas located in intronic and intergenic regions may give rise to gene expression changes, RNA splicing, or regulatory functions.

The InDels in exonic regions consist of frameshift insertion/deletion, non-frameshift insertion/deletion, stop gain and stop loss. A frameshift InDel involves the insertion or deletion of nucleotides that cannot be divisible by three. Since the genetic codons are nucleotide triplets, and each group of three bases corresponds to one of the 20 different amino acids used to build a protein. If a mutation disrupts the reading frame of the messenger RNA (mRNA), then the entire DNA sequence following the mutation would be misread. In other words, the insertion or deletion of bases that are not multiples of three results in the displacement of the entire DNA strand. It would then encode a completely different peptide segment. The protein sequence may be changed entirely and likely to produce a premature stop codon, and even would result in the protein not being translated. A frameshift insertion or deletion probably generates a range of phenotypic and molecular effects because of changing protein expression and amino acid sequences.

Compared to frameshift insertion/deletion, non-frameshift insertion or deletion causes much less regrets because it leads to the gain or loss of several nucleotides divisible by three, such that the reading frame is not disrupted. Therefore, the protein sequence differs from the original (or wildtype) with the addition or deletion of one or more amino acids. Consequently, the protein still presents but slightly makes a difference in composition or structure due to the degeneracy of the codon.

The location information and effects of stop gain, stop loss, upstream and downstream in InDel variants are the same as those of SNP variants and will not be repeated in this section.

In conclusion, NGS is used to detect SNP and InDel variants (as a couple of purposes) which indicates the bovine cells' genome may have some changes. Analyzing these variants will help us to evaluate the potential biosafety risk of the CRISPR-Cas9 targeted killing system. Currently, there is no well-developed approaches or criteria to analyze the biosafety risk of CRISPR-Cas9 system, and NGS is a common method to identify variants that has been widely accepted. The advantage of this method is that variants can be found in a short time and the amount of required DNA samples is relatively low. However, it also has some limitations. For example, NGS data and variants data are tremendous and complex, requiring sufficient knowledge, experience, and enough time to do analysis and report results. Moreover, NGS technology and variants calling are not 100% accurate. Nevertheless, we believe that NGS is the best option to evaluate the biosafety risk of CRISPR-Cas9 system from a genomic level.

CHAPTER 2: ESTIMATE POTENTIAL BIOSAFETY RISK OF A CRISPR-CAS9 SYSTEM FOR TARGETED KILLING OF CERTAIN PATHOGENS IN BEEF CATTLE PRODUCTION USING OMIC-BASED ANALYSIS METHODOLOGIES

Introduction

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR associated (Cas) genes are naturally employed in bacteria as the adaptive immune system by protecting against invasion of mobile elements like viruses and plasmids (Wiedenheft et al., 2012). The CRISPR-Cas9 system has emerged as a programmable and versatile tool in a wide variety of organisms for precise gene editing and as an antimicrobial to target the killing of specific bacteria. Although other antimicrobials like antibiotics have been effectively used in beef cattle production, such treatment may contribute to antibiotic-resistance bacteria and decreases efficiency of killing pathogenic bacteria. This would prevent the beef industry from growing.

An essential feature of the CRISPR-Cas9-based targeted killing system is that it can be easily programmed to selectively remove any unwanted bacteria and/or genes without affecting non-targeted bacterial populations in beef cattle production systems. The use of a programmable and versatile CRISPR-Cas9 system as an alternative to antibiotics would alleviate the possibility of selection pressure associated with long-term use of conventional antibiotics, a pressure that may favor the development of antibiotic-resistant bacteria. Thus, in addition to controlling pathogenic bacteria, the development and spread of antibiotic resistance in beef cattle production could be prevented using a targeted killing system based on CRISPR-Cas9. The CRISPR-Cas9 system can also be used as a "vaccine" to susceptible commensals, protecting them from the further acquisition of antibiotic resistance genes. Moreover, the CRISPR-Cas9 system could be

developed as a novel and intelligent antimicrobial agent to manipulate the composition of heterogeneous gut flora, promoting a healthy microbiome through microbiome editing. The successful application of CRISPR-Cas9 systems in beef cattle production will eventually help achieve the sustainable goal of reducing antibiotics while still retaining the health and performance of beef cattle.

In part, antibiotic resistance will likely be addressed as a societal public health concern by reducing long-term agricultural antibiotic use in beef cattle production through the development of a novel CRISPR-Cas9-based targeted killing system. To achieve our goal, a CRISPR-Cas9 targeted killing system was designed to selectively kill *E. coli* O157:H7 with a guide RNA that targets Shiga-toxin genes (*stx1* and *stx2*) (Yang et al., 2017). In 2018's study (Yang et al., 2018), a phage-mediated system was constructed to efficiently deliver the CRISPR-Cas9-based targeted killing system into *E. coli* O157:H7 inoculated in cattle rumen fluid. After the CRISPR-Cas9-based targeted killing system proved to be efficient, a cautionary evaluation of our CRISPR system to demonstrate its biosafety becomes an imperative step to complete the development of this novel strategy before its potential implementation in beef production.

How to assess the biosafety of CRISPR-edited products is still being debated around the world. Scientists not only have concerns about the biosafety of using CRISPR technology, but also argue about the correct, accuracy methods to estimate the biosafety risk of CRISPR-based products. Scientific efforts towards understanding the extent of CRISPR's off target effects have not provided consistent results. In recent years, there have been increasing topics investigating the off-target effects due to CRISPR. At the same time, there is active research on how to reduce the off-target effects of CRISPR. More scientific studies are needed to determine the conditions

under which such off-target effects occur *in vitro* and how to avoid off-target effects of CRISPR by a good design of guide RNA or Cas9 protein.

This project used omics-based analysis methods to evaluate the potential biosafety risk of CRISPR-Cas9 Shiga toxin targeting system. Next-generation sequencing (NGS) analysis was used to identify variants and mutations in bovine cell lines after treatments. This project will provide the beef industry with biosafety information regarding the application of CRISPR as an alternative to antibiotics in cattle production. Furthermore, the biosafety evaluation model generated from this study will also be helpful to the beef industry for making decisions over choosing the appropriate CRISPR technique forms with the fewest biosafety concerns, so that a variety of CRISPR-related techniques/products become available in the future.

Materials and Methods

Bovine cell culture

A bovine cell line CPA 47 (ATCC® CRL-1733™) was purchased from the American Type Culture Collection (ATCC). The bovine cells were seeded into 25 cm^2 flasks in 10 mL of Ham's F12K medium containing 1.5 g/L sodium bicarbonate (ATCC® 30-2004™) with 10% horse serum (ATCC® 30-2041™). The flasks were incubated at 37°C for approximately 7 days in a CO_2 incubator to reach 90% confluence.

Exposure of bovine cells to CRISPR treatments

Bovine cells were washed with an ATCC-formulated Dulbecco's Phosphate Buffered Saline (D-PBS, 1X) solution (ATCC® 30-2200™), and then subjected to one of the following four treatments:

1) Bovine cell control (i.e., bovine cells without any CRISPR treatment).

2) CRISPR/gRNA treatment: phages that carry the CRISPR system with the guide RNA targeting *stx* genes were added to the bovine cells (10^6 PFU/flask). Even though the CRISPR system with guide RNA can kill STEC cells, without the presence of STEC cells, the CRISPR system will not actually execute its function. Therefore, this treatment was to simulate a scenario of bovine cells being exposed to the CRISPR system that is not actively functioning.

3) CRISPR+O157 treatment: phages that carry the CRISPR system but without the guide RNA targeting *stx* genes (10^6 PFU/flask) and *E. coli* O157:H7 strain Sakai cells (10^5 CFU/flask) were added to the bovine cells. Without the gRNA that targets *stx* genes, the CRISPR system will not be able to recognize and kill any *E. coli* O157:H7 cells present in the environment. Therefore, this treatment also was to simulate a scenario of bovine cells being exposed to the CRISPR system that is not actively functioning.

4) CRISPR/gRNA+O157 treatment: phages that carry the CRISPR system with the guide RNA targeting *stx* genes (10^6 PFU/flask) and *E. coli* O157:H7 strain Sakai cells (10^5 CFU/flask) were added to the bovine cells. The CRISPR system carrying the gRNA targeting *stx* genes will be able to recognize *stx* genes and cleave the chromosomes of *E. coli* O157:H7 cells that are present in the environment. This treatment was to simulate a scenario of bovine cells being exposed to the CRISPR system that is actively functioning.

Each treatment contained 4 replicates (n = 4). Following addition of one of the above treatments to the bovine cells, all samples were incubated at 37°C in a CO_2 incubator for 3 hours. After incubation, 1 mL of a penicillin-streptomycin solution (ATCC® 30-2300™) was added to each flask. Then, all samples were incubated at 37°C in a CO_2 incubator for 3 days. On the fourth day, the old culture medium was removed from each flask and replaced with fresh cell culture medium containing the same components. All samples were incubated at 37°C in a CO_2

incubator for another 4 days. On the seventh day, samples were washed three times using 0.1% PBS, detached using trypsin-EDTA solution (1X, ATCC® 30-2101™) and harvested by centrifugation at 130 × *g* for 10 min. The collected bovine cells for each sample were divided into two portions: half for whole genome sequencing and half for protein analysis.

DNA extraction and sequencing

Whole genome DNA from each sample was extracted using a QIAGEN Blood & Tissue kit (DNeasy®, Germantown, MD) and purified using a Monarch Genomic DNA Purification Kit (NEW ENGLAND Biolabs® Inc., Ipswich, MA). The purified DNA samples were sent to Novogene Bioinformatics Technology (Beijing, China) for library construction and whole genome sequencing. Briefly, the genomic DNA was first randomly fragmented to the size of 350 bp, then DNA fragments were end polished, A-tailed, ligated with Illumina sequencing adapters, and PCR enriched with P5 and P7 primer oligos. The PCR products to be used as the final constructs of libraries were purified (AMPure XP system) and subjected to a quality control testing that included a size distribution measurement using the Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and a molarity measurement using real-time PCR. The Illumina Novaseq 6000 platform (Illumina Inc., San Diego, CA) was utilized for genomic DNA sequencing to generate 150-bp paired-end reads with a minimum coverage of 10× for around 99% of the genome (mean coverage of 30×).

DNA sequencing data analysis

Original image data generated by the sequencer were converted into sequence data via base calling (Illumina pipeline CASAVA v1.8.2). The base calling was subjected to quality control (QC) procedures to remove unusable reads. Raw reads filtering criteria was conducted as follows to remove:

1) Paired reads with either of the two reads containing adapter contamination.

2) Paired reads with uncertain nucleotides (N) constituting more than 10 percent of either read.

3) Paired reads with low quality nucleotides ($Q \leq 5$, base quality less than 5) constituting more than 50 percent of either read.

Clean reads after quality control procedures were aligned to the bovine reference genome (NCBI access number: ARS-UCD 1.2 USDA ARS) using Burrows-Wheeler Aligner (https://github.com/lh3/bwa) with default parameters (Li and Durbin, 2009). The subsequent processing, including removal of duplicates was performed using SAMtools (http://www.htslib.org) and PICARD (http://picard.sourceforge.net) (Li et al., 2009).

The individual SNP/InDel variants were detected by SAMtools with the following criteria:

1) The number of supporting reads for each SNP should be more than 4.

2) the mapping quality of each SNP should be higher than 20.

ANNOVAR was used for functional annotation of the detected variants (https://www.openbioinformatics.org/annovar/annovar_download_form.php) (Wang and Hakonarson, 2010). The UCSC Known Genes datasets hosted at the University of California (Santa Cruz, CA) were used for gene and region annotations (https://genome.ucsc.edu).

Statistical analysis of variants

All statistical analyses were performed in the control and the three CRISPR treatments with four independent replicates for each group. Data were analyzed using One-way ANOVA, in Rstudio release 1.3.1073 (Rstudio Teams, Boston, MA). For significant effects *($P < 0.05$)*, SNP/InDel means were compared using F-test.

Result and Discussion

DNA sequencing quality

A total of 1,078 Gb of output with 3593.12 million paired-end reads (150 bp) was obtained for all the 16 samples after next generation sequencing (NGS). The average output for the 4 replicates within the control, CRISPR/gRNA, CRISPR+O157 and CRISPR/gRNA+O157 treatments was 59.1, 72.3, 72.3 and 65.9 Gb, respectively. Sequencing raw reads often contain low quality reads or reads with adaptors, which will affect the quality of downstream analysis. To avoid this, the raw reads were filtered based on the criteria described in the to obtain clean reads. Detailed statistics of sequencing data are listed in Table 1. Effective rates were ≥ 99.13% and the error rates were ≤ 0.03 for each sample. Illumina sequencing used the parameter of Q-score to indicate its sequencing quality. The relationship between the Illumina Q-quality score and Phred score in base calling, and its corresponding error and correct rates are shown in Table 2. As shown in Table 1, each sample had over 93.68% of the sequencing data with a Q30 score (error rate < 0.01% and correct rate > 99.9%). For NGS, the sequencing platform, chemical reactants, and sample quality can influence sequencing quality and base error rate. Paired-end sequencing data with a Q30 above 80% and a per base error rate below 6% are generally considered as acceptable for sequencing quality. The statistics showed that the quality of our DNA sequencing was much higher than this criterion, so it was reliable to utilize the clean data after quality control for further analysis.

Sequence alignment to a reference genome

Reads in clean data were aligned with the bovine reference sequence (ARS-UCS 1.2) using BWA software (Li and Durbin, 2009). Statistics associated with the reference genome are shown in Table 3. The mapping rates of samples reflect the similarity between the reference

genome and the genome of each sample. The mapping depth and coverage rates which indicate the evenness and homology of each sample with the reference genome were counted according to the alignment results (Table 4). For the current 2,715,853,792 bp reference genome, the mapping rate of each sample ranged from 99.48% to 99.75%, the average depths ranged from 15.76X to 25.28X, and the 4X coverages ranged from 89.45% to 98.23% (Table 4). These mapping results indicate the data are qualified to be used for subsequent variants detection and other related analysis.

SNP detection and annotation

Single nucleotide polymorphism (SNP) refers to a variation in a single nucleotide that may occur at some specific position in the genome, including transition and transversion of a single nucleotide. A total of 94,796,157 SNPs were identified in all 16 samples that were sequenced. The number of SNPs with each sample ranged from 5,225,269 to 6,192,930. The frequency of transition-to-transversion (Ti/Tv) is shown in Figure 2. The transition-to-transversion (Ti/Tv) ratio for each sample was in the range of 2.24 to 2.27, which is consistent with the approximate ratio of 2.2 in bovine, as reported by Swane et al. (2019). Our treatments did not cause a significant impact on the transition-to-transversion ratio, which indicated that the detected SNPs were more likely to occur randomly.

Tables 5 and 6 are summaries of SNP detection and annotation in coding and non-coding regions, respectively. SNPs found in the protein coding regions of the genome have significant impacts on bovine phenotypes and diseases. SNPs in the coding region are of two types: synonymous and non-synonymous SNPs. Non-synonymous SNPs usually result in amino acid changes in the protein product of genes, and cause protein malfunction or nonfunctional protein products. Synonymous SNPs do not result in amino acid changes in the protein, but they may

affect protein expression and function in other way. Non-coding DNA includes structural DNA (not transcribed), functional RNA (transcribed, but not translated), and introns (transcribed, but removed before translation). SNPs that are not in protein-coding regions may contribute to gene regulation, evolution, and variation by affecting gene splicing, transcription factor binding, mRNA degradation, DNA methylation, histone modification and so on.

Of the total number of SNPs, 60.01% were in intergenic regions (regions between genes), 36.40% in intronic regions (non-coding sequences of genes), and 0.79% were in exonic regions (coding sequences of genes) (Tables 5 and 6). The average SNPs in exonic regions for each treatment of control, CRISPR/gRNA, CRISPR+O157 and CRISPR/gRNA+O157 was 0.1964%, 0.2002%, 0.2002% and 0.1948%, respectively (Table 5). Furthermore, the average percentages of non-synonymous and synonymous SNPs were in the range of 0.0847 to 0.0873% and 0.1092 to 0.1120%, respectively, for all four treatments (i.e., the control and the three CRISPR treatments) (Table 5). The number of SNPs within each functional class in exonic regions, for example, stop loss, stop gain, synonymous, non-synonymous, all showed no significant differences *($P > 0.05$)* among the control and the three CRISPR treatments (Tables 7). And the number of SNPs within each functional class in intronic and intergenic regions, for example, slicing sites and upstream or downstream from transcription termination sites, all showed no significant differences *($P > 0.05$)* among the control and the three CRISPR treatments (Table 8). All these data together suggested that none of the CRISPR treatments resulted in significant differences ($P > 0.05$) in the number of SNPs between the control and the treatments when their NGS data were compared with a bovine reference genome.

All the SNPs in Tables 5 and 6 were called by comparison of the genome of the bovine cell line with that of reference genome. Because our bovine DNA was from an immortalized cell

line, there may be numerous SNPs within the genomes of these cells that are due to the natural differences in genomes between the bovine cell line used in this study and the reference. The SNPs identified by comparing with the reference genome include both the variants from the genome differences and the variants between the control and CRISPR treatments. Since the variants originating from the CRISPR treatments were of most interest to us, an appropriate way to differentiate these two types of variants will facilitate our further analysis. We considered that any SNPs common across all the 16 samples were most likely the variants from the genome differences and not from the CRISPR treatments. Each sample contained 35,669 SNPs that were common across all 16 samples in exonic regions, with the detected SNPs from each sample ranged from 46,027 to 49,448 SNPs (Figure 3). Therefore, we excluded those SNPs shared by all the 16 samples (natural difference type SNPs) from our further analysis and used the remaining SNPs in the control group as an initial baseline to make the comparison.

Table 9 shows the differences in SNPs in exonic regions between the control and the three CRISPR treatments. Some of these exonic SNPs were found in all four replicates of each treatment, while some were only found in 1 of the 4 replicates. It is reasonable to assume that the SNPs present in all 4 replicates should be more likely to be associated with the CRISPR treatment that was applied to the 4 replicates. When compared to the control, the CRISPR/gRNA, CRISPR+O157, and CRISPR/gRNA+O157 treatments had 27, 30, and 26 SNPs, respectively, that were present in all four replicates (Table 9). The details of these SNPs are shown in Table 10.

Further investigation of the locations of these exonic SNPs showed that the CRISPR/gRNA and CRISPR+O157 treatments shared 3 SNPs, CRISPR/gRNA and CRISPR/gRNA+O157 shared 3 SNPs, CRISPR+O157 and CRISPR/gRNA+O157 shared 1 SNP, and the others were

individual SNPs with no overlap with any other the CRISPR treatments. Although the bovine cells were exposed to the same CRISPR system (with or without a gRNA targeting *stx* genes), only a few overlapping variants were found across the CRISPR treatments. With these exonic SNP results, we were not able to identify any obvious SNP patterns that may have originated from the three CRISPR treatments. Therefore, we considered that these SNPs occurred randomly and may not be attributed to the consequence of CRISPR exposure or function.

Findings coincided with a hypothesis that any biosafety risk related to our phage-delivered CRISPR-Cas9 system should be less than those associated with the CRISPR systems that are used as gene editing tools to make genome alterations in cattle. When used as a gene editing tool, it needs to be delivered into bovine cells that then make direct cleavages on their genomes. The CRISPR-Cas9 system we developed for the control of STECs uses bacteriophages to deliver the CRISPR-Cas9 system into bacterial cells, a specific gRNA to recognize *stx* genes, and the CRISPR-Cas9 system to make a cleavage on target bacterial genomes. Therefore, exposure of bovine cells to our CRISPR-Cas9 system should be considered as "indirect exposure" and our CRISPR system was designed to keep such exposure outside of bovine cells. On the other hand, exposure to gene editing CRISPR tools should be considered as "direct exposure" because these CRISPR tools are designed to get into bovine cells and have a direct function on their genomes. These essential differences between our CRISPR system and gene editing CRISPR tools should result in different levels of biosafety concerns when applying to cattle. It is reasonable to deduce that the "indirect exposure" would pose a relatively low risk to cattle.

InDel detection and annotation

InDel refers to the insertion or deletion of ≤ 50 bp sequences in the DNA. InDels are the

second most abundant type of genetic variant following SNPs in cattle. InDels in coding regions can significantly impact gene expression, particularly through frameshifts resulting in prematurely terminated protein products or changed splice variants.

A total of 11,949,421 InDels were identified in all 16 samples with each sample having between 604,387 and 817,716 InDels. The length distribution of InDels is shown in Figure 4. The majority of the identified InDels had lengths less than 6 bp (83.18%). Statistics tied to InDel detection and annotation are shown in Tables 11 and 12.

Of the total number of InDels, 62.78% were in intergenic regions, 38.54% in intronic regions, and 0.15% in exonic regions. The average exonic InDels in the control, CRISPR/gRNA, CRISPR+O157 and CRISPR/gRNA+O157 treatment groups were 0.0371%, 0.0372%, 0.0375%, and 0.0372%, respectively (Table 11). The frameshift InDel commonly results in the displacement of the entire DNA strand and will then encode a completely different peptide segment. The average percentages of frameshift deletion and frameshift insertion were in the range of 0.0112% to 0.0115% and 0.0091% to 0.0094%, respectively, for all four treatments (i.e., the control and three CRISPR treatments) (Table 11). The number of InDels within each functional class in exonic regions, for example, stop loss, stop gain, frameshift insertion/deletion, non-frameshift insertion/deletion, all showed no significant differences ($P > 0.05$) among the control and the three CRISPR treatments (Tables 13). And the number of InDels within each functional class in intronic and intergenic regions, for example, slicing sites and upstream or downstream from transcription termination sites, all showed no significant differences ($P > 0.05$) among the control and the three CRISPR treatments (Table 14). As noted for the SNPs data, all these InDel data also did not show any significant differences ($P > 0.05$) in the number of InDels between the control and the CRISPR treatments when their NGS data were compared with the

bovine reference genome.

As we described above for the SNPs, it is also ideal to differentiate the detected InDels existing between our cell line and the reference genome, and the variants originated from the CRISPR treatments for our further analysis. However, the identification of InDels that are common across all the samples is far more complicated than it is for SNPs. We are still working on developing appropriate methods to achieve this purpose.

Conclusion

In the thesis, we presented a genome analysis of the bovine cells after indirect exposure to the CRISPR system. The identification and annotation of SNPs and InDels in the genome is essential for evaluating CRISPR biosafety. Our primary analysis focused on the variants located in exonic regions of SNPs. From the results of variant SNPs, no obvious SNP patterns that may have originated from our CRISPR treatments could be detected. This matched our initial hypothesis that the potential biosafety risk of our Shiga toxin targeted CRISPR-Cas9 system should be low because of the nature of its indirect exposure to bovine cells. However, in addition to coding regions, millions of variants were also identified in non-coding regions. These SNPs and/or InDels may potentially play a significant role in gene regulation and affect protein function in other ways. We are continuing with our analysis of the variants in non-coding regions.

Table 1. Statistics of the sequencing data.

Treatment	Sample	Raw data[a] (G)	Clean data[b] (G)	Effective[c] (%)	Error[d] (%)	Q20[e] (%)	Q30[f] (%)	GC[g] (%)
Control[h]	S1	54.5	54.1	99.23	0.025	97.89	94.45	47.25
	S2	55.0	54.7	99.26	0.023	98.11	94.87	46.21
	S3	61.0	60.8	99.58	0.02	98.44	95.45	44.47
	S4	65.9	65.7	99.62	0.02	98.43	95.38	44
CRISPR/gRNA[i]	S1	61.5	61.2	99.58	0.03	97.61	93.68	45.77
	S2	68.2	67.9	99.64	0.02	97.91	94.32	45.39
	S3	71.7	71.4	99.63	0.02	98.28	94.96	43.56
	S4	87.7	87.4	99.61	0.02	98.3	95.05	43.79
CRISPR+O157[j]	S1	64.8	64.4	99.61	0.03	97.84	94.13	45.58
	S2	68.3	68.0	99.61	0.02	98.38	95.23	43.88
	S3	78.0	77.7	99.63	0.02	98.41	95.32	43.63
	S4	78.1	77.8	99.58	0.02	98.27	95.02	44.71
CRISPR/gRNA+O157[k]	S1	60.2	59.7	99.13	0.03	97.6	93.90	48.02
	S2	61.5	60.8	98.93	0.03	97.58	93.92	48.91
	S3	68.8	68.5	99.58	0.02	98.36	95.21	44.46
	S4	72.9	72.6	99.59	0.02	98.4	95.35	44.54

[a]Raw data (G): The original sequencing data calculated in G (Gigabits).
[b]Clean data (G): The sequencing data after quality control.
[c]Effective (%): The ratio of clean data to raw data.
[d]Error (%): Base error rates.
[e]Q20 (%): The percentage of bases with Phred score ≥20.
[f]Q30 (%): The percentage of bases with Phred score ≥30.
[g]GC: The percentage of G and C contents in the clean data.
[h]Control: bovine cells only
[i]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[j]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[k]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7

Table 2. Sequencing error rate and corresponding base quality value.

Phred score	Error rate	Correct rate	Q-score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

Table 3. Statistics of the reference genome.

Ref Name	Seq number[a]	Total length[b] (bp)	GC content[c] (%)	Gap rate[d] (%)	N50 length[e] (bp)	N90 length[f] (bp)
ARS-UCD1.2	2,211	2,715,853,792	41.93	0.00	103,308,737	51,098,607

[a]Seq number: the total number of the assembled genomic sequences.
[b]Total length: the total length of the assembled genomic sequence.
[c]GC content: the GC content of the reference genome.
[d]Gap rate: the proportion of unknown sequence (N) in the reference genome assembly.
[e]N50 length: the length of scaffold N50, of which 50% of the sequence is higher than this level.
[f]N90 length: the length of scaffold N90, of which 90% of sequence is higher than this level.

Table 4. Statistics of mapping rates, depths and coverages.

Treatment	Sample	Mapped reads[a]	Total reads[b]	Mapping rate[c] (%)	Average depth[d] (X)	Coverage at least 1X[e] (%)	Coverage at least 4X[f] (%)	Duplicate[g] (%)
Control[h]	S1	3.59×10^8	3.61×10^8	99.53	16.28	98.56	95.25	16.86
	S2	3.63×10^8	3.65×10^8	99.63	16.29	98.57	95.54	18.46
	S3	4.04×10^8	4.05×10^8	99.74	18.55	98.57	97.98	16.69
	S4	4.37×10^8	4.38×10^8	99.75	19.80	98.59	98.05	18.03
CRISPR/gRNA[i]	S1	4.07×10^8	4.08×10^8	99.60	18.47	98.61	97.51	16.74
	S2	4.51×10^8	4.53×10^8	99.65	20.47	98.59	97.89	17.13
	S3	4.75×10^8	4.76×10^8	99.73	21.35	98.63	98.13	18.69
	S4	5.81×10^8	5.83×10^8	99.73	25.28	98.67	98.23	20.96
CRISPR+O157[j]	S1	4.28×10^8	4.29×10^8	99.66	19.53	98.60	97.68	16.42
	S2	4.52×10^8	4.54×10^8	99.74	20.73	98.60	98.10	17.12
	S3	5.17×10^8	5.18×10^8	99.74	22.94	98.62	98.15	19.67
	S4	5.17×10^8	5.18×10^8	99.72	22.54	98.63	98.14	20.41
CRISPR/gRNA +O157[k]	S1	3.96×10^8	3.98×10^8	99.48	15.76	98.16	89.61	26.50
	S2	4.03×10^8	4.05×10^8	99.50	17.36	97.98	89.45	20.53
	S3	4.55×10^8	4.57×10^8	99.73	20.63	98.59	98.07	17.77
	S4	4.83×10^8	4.84×10^8	99.74	21.63	98.61	98.11	18.44

[a]Mapped reads: The number of clean reads mapped to the reference assembly, including both single-end reads and reads in pairs.
[b]Total reads: Total number of reads in clean data.
[c]Mapping rate: The ratio of reads mapped to the reference genome assembly to the total clean reads.
[d]Average depth (%): The average depth of the mapped reads at each site, calculated by dividing the size of the mapped reads by the size of the assembled genome.
[e]Coverage at least 1X (%): The percentage of the assembled genome with more than one read at each site.
[f]Coverage at least 4X (%): The percentage of the assembled genome with ≥4X coverage at each site.

[g]Duplicate (%): The duplication rate is the fraction of mapped reads where any 2 reads share the same 5′ and 3′ coordinates.
[h]Control: bovine cells only
[i]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[j]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7

Table 5. SNP detection and annotation in exonic regions when compared to the reference genome.

Treatment	Sample	Stop gain[a]	Stop loss[b]	Synonymous[c]	Non-synonymous[d]
Control[g]	S1	174	39	25880	20045
	S2	174	40	25776	19901
	S3	176	42	26376	20514
	S4	182	39	26365	20479
	Subtotal[e]	706	160	104397	80939
	Percentage[f]	0.0007%	0.0002%	0.1101%	0.0854%
CRISPR/gRNA[h]	S1	182	41	26367	20427
	S2	180	41	26563	20713
	S3	181	40	26491	20625
	S4	187	43	26778	20949
	Subtotal	730	165	106199	82714
	Percentage	0.0008%	0.0002%	0.1120%	0.0873%
CRISPR+O157[i]	S1	183	41	26361	20513
	S2	185	39	26553	20666
	S3	177	42	26565	20729
	S4	185	38	26683	20830
	Subtotal	730	160	106162	82738
	Percentage	0.0008%	0.0002%	0.1120%	0.0873%
CRISPR/gRNA +O157[j]	S1	169	37	25087	19371
	S2	181	37	25241	19511
	S3	191	41	26631	20745
	S4	180	41	26541	20704
	Subtotal	721	156	103500	80331
	Percentage	0.0008%	0.0002%	0.1092%	0.0847%

[a]Stop gain: A nonsynonymous SNP that leads to the introduction of stop codon at the variant site.
[b]Stop loss: A nonsynonymous SNP that leads to the removal of stop codon at the variant site.
[c]Synonymous: Synonymous SNPs result in different alleles but still encode for the same amino acid.
[d]Non-synonymous: Non-synonymous SNPs result in different alleles that encode for different amino acids.
[e]Subtotal: Sum of SNPs from the four replicates in one treatment.
[f]Percentage: Subtotal SNPs were divided by total SNPs.
[g]Control: bovine cells only
[h]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[i]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[j]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7

Table 6. SNP detection and annotation in intronic and intergenic regions when compared to the reference genome.

Treatment	Sample	Intronic[a]	Splicing[b]	Intergenic[c]	Upstream[d]	Downstream[e]	Upstream/ downstream[f]
Control[i]	S1	2067242	168	3415037	33463	33871	880
	S2	2070277	174	3435124	33127	33936	873
	S3	2194672	178	3663501	34911	36266	887
	S4	2207597	173	3692662	35046	36297	893
	Subtotal[g]	8539788	693	14206324	136547	140370	3533
	Percentage[h]	9.009%	0.001%	14.986%	0.144%	0.148%	0.004%
CRISPR/gRNA[j]	S1	2174630	177	3602803	34996	35824	907
	S2	2200945	178	3657017	35153	36230	899
	S3	2223044	169	3724185	35028	36654	871
	S4	2244065	177	3766178	35631	37092	894
	Subtotal	8842684	701	14750183	140808	145800	3571
	Percentage	9.328%	0.001%	15.560%	0.149%	0.154%	0.004%
CRISPR+O157[k]	S1	2181293	175	3618066	34878	35938	894
	S2	2216424	176	3709715	35102	36441	880
	S3	2230626	176	3736272	35389	36775	895
	S4	2227481	177	3728158	35593	36776	905
	Subtotal	8855824	704	14792211	140962	145930	3574
	Percentage	9.342%	0.001%	15.604%	0.149%	0.154%	0.004%
CRISPR/gRNA+O157[l]	S1	1915512	164	3148629	31331	31243	856
	S2	1917584	160	3148849	31718	31516	864
	S3	2214196	180	3701339	35254	36467	900
	S4	2219551	178	3711072	35441	36504	901
	Subtotal	8266843	682	13709889	133744	135730	3521
	Percentage	8.721%	0.001%	14.462%	0.141%	0.143%	0.004%

[a]Intronic: Intronic region is the non-coding region of an RNA transcript (encoding DNA region).
[b]Splicing: SNPs located at splicing sites.

[c]Intergenic: Intergenic region is a stretch of DNA sequence located between genes (non-coding DNA region).
[d]Upstream: SNPs located within 1 kb upstream (away from transcription start site) of the gene.
[e]Downstream: SNPs located within 1 kb downstream (away from transcription termination site) of the gene region.
[f]Upstream/downstream: SNPs located within 2 kb of intergenic regions, which is within 1 kb downstream or upstream of the genes.
[g]Subtotal: Sum of SNPs from the four replicates within one treatment group.
[h]Percentage: Subtotal SNPs were divided by total SNPs.
[i]Control: bovine cells only
[j]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[k]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[l]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7

Table 7. One-way ANOVA analysis of the number of SNPs in exonic regions in the control and each CRISPR treatment.

Treatment	SNP Mean ± SE			
	Stop gain[a]	Stop loss[b]	Synonymous[c]	Non-synonymous[d]
Control[e]	176±2.73A[i]	40±0.875A	26099±227A	20235±211A
CRISPR/gRNA[f]	182±2.73A	41±0.875A	26550±227A	20678±211A
CRISPR+O157[g]	182±2.73A	40±0.875A	26540±227A	20684±211A
CRISPR/gRNA+O157[h]	180±2.73A	39±0.875A	25875±227A	20083±211A

[a]Stop gain: A nonsynonymous SNP that leads to the introduction of stop codon at the variant site.
[b]Stop loss: A nonsynonymous SNP that leads to the removal of stop codon at the variant site.
[c]Synonymous: Synonymous SNPs result in different alleles but still encode for the same amino acid.
[d]Non-synonymous: Non-synonymous SNPs result in different alleles that encode for different amino acids.
[e]Control: bovine cells only
[f]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[g]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[h]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7
[i]A: Within a column, means followed by the same capital letter are not significantly different ($P > 0.05$).

Table 8. One-way ANOVA analysis of the number of SNPs in intronic and intergenic regions in the control and each CRISPR treatment.

Treatment	SNP Mean ± SE					
	Intronic[a]	Splicing[b]	Intergenic[c]	Upstream[d]	Downstream[e]	Upstream/ downstream[f]
Control[g]	2134947±48309A[k]	173±2.9A	3551581±91286A	34137±615A	35092±831A	883±7.81A
CRISPR/gRNA[h]	2210671±48309A	175±2.9A	3687546±91286A	35202±615A	36450±831A	893±7.81A
CRISPR+O157[i]	2213956±48309A	176±2.9A	3698053±91286A	35240±615A	36482±831A	894±7.81A
CRISPR/gRNA+O157[j]	2066711±48309A	170±2.9A	3427472±91286A	33436±615A	33932±831A	880±7.81A

[a]Intronic: Intronic region is the non-coding region of an RNA transcript (encoding DNA region).
[b]Splicing: SNPs located at splicing sites.
[c]Intergenic: Intergenic region is a stretch of DNA sequence located between genes (non-coding DNA region).
[d]Upstream: SNPs located within 1 kb upstream (away from transcription start site) of the gene.
[e]Downstream: SNPs located within 1 kb downstream (away from transcription termination site) of the gene region.
[f]Upstream/downstream: SNPs located within 2 kb of intergenic regions, which is within 1 kb downstream or upstream of the genes.
[g]Control: bovine cells only
[h]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[i]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[j]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7
[k]A: Within a column, means followed by the same capital letter are not significantly different ($P > 0.05$).

Table 9. Comparison of the CRISPR treatments against the control group for SNPs in exonic regions.

Occurrence time	Number of SNPs		
	CRISPR/gRNA[a]	CRISPR+O157[b]	CRISPR/gRNA+O157[c]
1[d]	8933	8810	10343
2[e]	1668	1732	2143
3[f]	323	312	286
4[g]	27	30	26

[a]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[b]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[c]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7
[d]1: One out of 4 replicates showed a variant from the control.
[e]2: Two out of 4 replicates showed the same variant from the control.
[f]3: Three out of replicates showed the same variant from the control.
[g]4: All four replicates showed the same variant from the control.

Table 10. Details of the SNPs that occurred in 4 replicates.

	Synonymous[d]	Non-synonymous[e]	Stop gain[f]	Unknown[g]
CRISPR/gRNA[a]	7	20	0	0
CRISPR+O157[b]	11	17	0	2
CRISPR/gRNA +O157[c]	10	12	1	3

[a]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[b]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[c]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7
[d]Synonymous: Synonymous SNPs result in different alleles but still encode for the same amino acid.
[e]Non-synonymous: Non-synonymous SNPs result in different alleles that encode different amino acids.
[f]Stop gain: A non-synonymous SNP that leads to the introduction of stop codon at the variant site.

Table 11. Statistics of exonic InDels detection and annotation when compared to the reference genome.

Treatment	Sample	Stop gain[a]	Stop loss[b]	Frameshift deletion[c]	Frameshift insertion[d]	Non-frameshift deletion[e]	Non-frameshift insertion[f]
Control[i]	S1	13	2	341	276	275	223
	S2	15	1	331	266	264	209
	S3	12	2	343	266	264	215
	S4	17	2	340	274	259	209
	Subtotal[g]	57	7	1355	1082	1062	856
	Percentage[h]	0.0005%	0.0001%	0.0113%	0.0091%	0.0089%	0.0072%
CRISPR/gRNA[j]	S1	15	2	335	280	273	213
	S2	14	2	337	288	271	214
	S3	17	1	339	275	246	195
	S4	16	2	341	284	273	213
	Subtotal	62	7	1352	1127	1063	835
	Percentage	0.0005%	0.0001%	0.0113%	0.0094%	0.0089%	0.0070%
CRISPR+O157[k]	S1	16	2	345	266	279	213
	S2	15	1	333	276	255	212
	S3	18	2	352	276	269	211
	S4	12	2	345	283	281	213
	Subtotal	61	7	1375	1101	1084	849
	Percentage	0.0005%	0.0001%	0.0115%	0.0092%	0.0091%	0.0071%
CRISPR/gRNA +O157[l]	S1	16	1	327	263	269	210
	S2	15	2	329	268	273	219
	S3	12	2	341	283	280	220
	S4	17	2	342	273	264	215
	Subtotal	60	7	1339	1087	1086	864
	Percentage	0.0005%	0.0001%	0.0112%	0.0091%	0.0091%	0.0072%

[a]Stop gain: An InDel that leads to the introduction of stop codon at the variant site.
[b]Stop loss: An InDel that leads to the removal of stop codon at the variant site.
[c]Frameshift deletion: InDel mutation changing the open reading frame with deletion.
[d]Frameshift insertion: InDel mutation changing the open reading frame with insertion.
[e]Non-frameshift deletion: InDel mutation without changing the open reading frame with deletion sequences of 3 or multiple of 3 bases.
[f]Non-frameshift insertion: InDel mutation without changing the open reading frame with insertion sequences of 3 or multiple of 3 bases.
[g]Subtotal: Sum of InDels from the four replicates within one treatment group.
[h]Percentage: Subtotal InDels were divided by total InDels.
[i]Control: bovine cells only
[j]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[k]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[l]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7

Table 12. Statistics of InDels (intronic and intergenic) detection and annotation when compared to the reference genome.

Treatment	Sample	Intronic[a]	Splicing[b]	Intergenic[c]	Upstream[d]	Downstream[e]	Upstream/ downstream[f]
Control[i]	S1	261419	111	420580	5316	4653	149
	S2	263779	102	428391	5094	4682	141
	S3	296595	114	484461	5599	5237	151
	S4	299285	108	490929	5597	5366	144
	Subtotal[g]	1121078	435	1824361	21606	19938	585
	Percentage[h]	9.382%	0.004%	15.267%	0.181%	0.167%	0.005%
CRISPR/gRNA[j]	S1	286455	106	460191	5706	5192	164
	S2	295328	110	479245	5813	5280	162
	S3	305743	102	503944	5602	5413	141
	S4	313998	112	517662	5867	5591	149
	Subtotal	1201524	430	1961042	22988	21476	616
	Percentage	10.055%	0.004%	16.411%	0.192%	0.180%	0.005%
CRISPR+O157[k]	S1	289810	102	467736	5673	5188	163
	S2	303584	111	499726	5644	5389	152
	S3	308655	109	507134	5734	5483	157
	S4	306535	106	501648	5890	5462	158
	Subtotal	1208584	428	1976244	22941	21522	630
	Percentage	10.114%	0.004%	16.538%	0.192%	0.180%	0.005%
CRISPR/gRNA +O157[l]	S1	232163	101	372228	4818	4138	142
	S2	235648	100	376276	4990	4153	151
	S3	301845	108	493884	5794	5365	161
	S4	303896	105	497582	5875	5412	158
	Subtotal	1073552	414	1739970	21477	19068	612
	Percentage	8.984%	0.003%	14.561%	0.180%	0.160%	0.005%

[a]Intronic: Intronic region is the non-coding region of an RNA transcript (encoding DNA region).
[b]Splicing: InDels located in the splicing site.
[c]Intergenic: Intergenic region is a stretch of DNA sequences located between genes (non-coding DNA region). InDels located within the > 2 kb intergenic region.
[d]Upstream: InDels located within 1 kb upstream (away from transcription start site) of the gene.
[e]Downstream: InDels located within 1 kb downstream (away from transcription termination site) of the gene region.
[f]Upstream/downstream: InDels located within 2 kb intergenic region, which is within 1 kb downstream or upstream of the genes.
[g]Subtotal: Sum of InDels from the four replicates within one treatment group.
[h]Percentage: Subtotal InDels were divided by total InDels.
[i]Control: bovine cells only
[j]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[k]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E.coli* O157:H7
[l]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E.coli* O157:H7

Table 13. One-way ANOVA analysis of the number of InDels in exonic regions in the control and each CRISPR treatment.

Treatment	InDel Mean ± SE					
	Stop gain[a]	Stop loss[b]	Frameshift deletion[c]	Frameshift insertion[d]	Non-frameshift deletion[e]	Non-frameshift insertion[f]
Control[g]	14±1.05A[k]	2±0.25A	339±3.15A	270±3.36A	262±5.04A	214±3.06A
CRISPR/gRNA[h]	16±1.05A	2±0.25A	338±3.15A	282±3.36A	266±5.04A	209±3.06A
CRISPR+O157[i]	15±1.05A	2±0.25A	344±3.15A	275±3.36A	271±5.04A	212±3.06A
CRISPR/gRNA+O157[j]	15±1.05A	2±0.25A	335±3.15A	272±3.36A	272±5.04A	216±3.06A

[a]Stop gain: An InDel that leads to the introduction of stop codon at the variant site.
[b]Stop loss: An InDel that leads to the removal of stop codon at the variant site.
[c]Frameshift deletion: InDel mutation changing the open reading frame with deletion.
[d]Frameshift insertion: InDel mutation changing the open reading frame with insertion.
[e]Non-frameshift deletion: InDel mutation without changing the open reading frame with deletion sequences of 3 or multiple of 3 bases.
[f]Non-frameshift insertion: InDel mutation without changing the open reading frame with insertion sequences of 3 or multiple of 3 bases.
[g]Control: bovine cells only
[h]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[i]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[j]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7
[k]A: Within a column, means followed by the same capital letter are not significantly different ($P > 0.05$).

Table 14. One-way ANOVA analysis of the number InDels in intronic and intergenic regions in the control and each CRISPR treatment.

Treatment	InDel Mean ± SE					
	Intronic[a]	Splicing[b]	Intergenic[c]	Upstream[d]	Downstream[e]	Upstream/ downstream[f]
Control[g]	280270±11787A[k]	109±2.16A	456090±21280A	5402±154A	4984±209A	146±3.81A
CRISPR/gRNA[h]	300381±11787A	108±2.16A	490260±21280A	5747±154A	5369±209A	154±3.81A
CRISPR+O157[i]	302146±11787A	107±2.16A	494061±21280A	5735±154A	5380±209A	158±3.81A
CRISPR/gRNA+O157[j]	268388±11787A	104±2.16A	434992±21280A	5369±154A	4767±209A	153±3.81A

[a]Intronic: Intronic region is the non-coding region of an RNA transcript (encoding DNA region).
[b]Splicing: SNPs located at splicing sites.
[c]Intergenic: Intergenic region is a stretch of DNA sequence located between genes (non-coding DNA region).
[d]Upstream: SNPs located within 1 kb upstream (away from transcription start site) of the gene.
[e]Downstream: SNPs located within 1 kb downstream (away from transcription termination site) of the gene region.
[f]Upstream/downstream: SNPs located within 2 kb of intergenic regions, which is within 1 kb downstream or upstream of the genes.
[g]Control: bovine cells only
[h]CRISPR/gRNA: bovine cells + CRISPR with the guide RNA targeting *stx* genes
[i]CRISPR+O157: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
[j]CRISPR/gRNA+O157: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7
[k]A: Within a column, means followed by the same capital letter are not significantly different ($P > 0.05$).

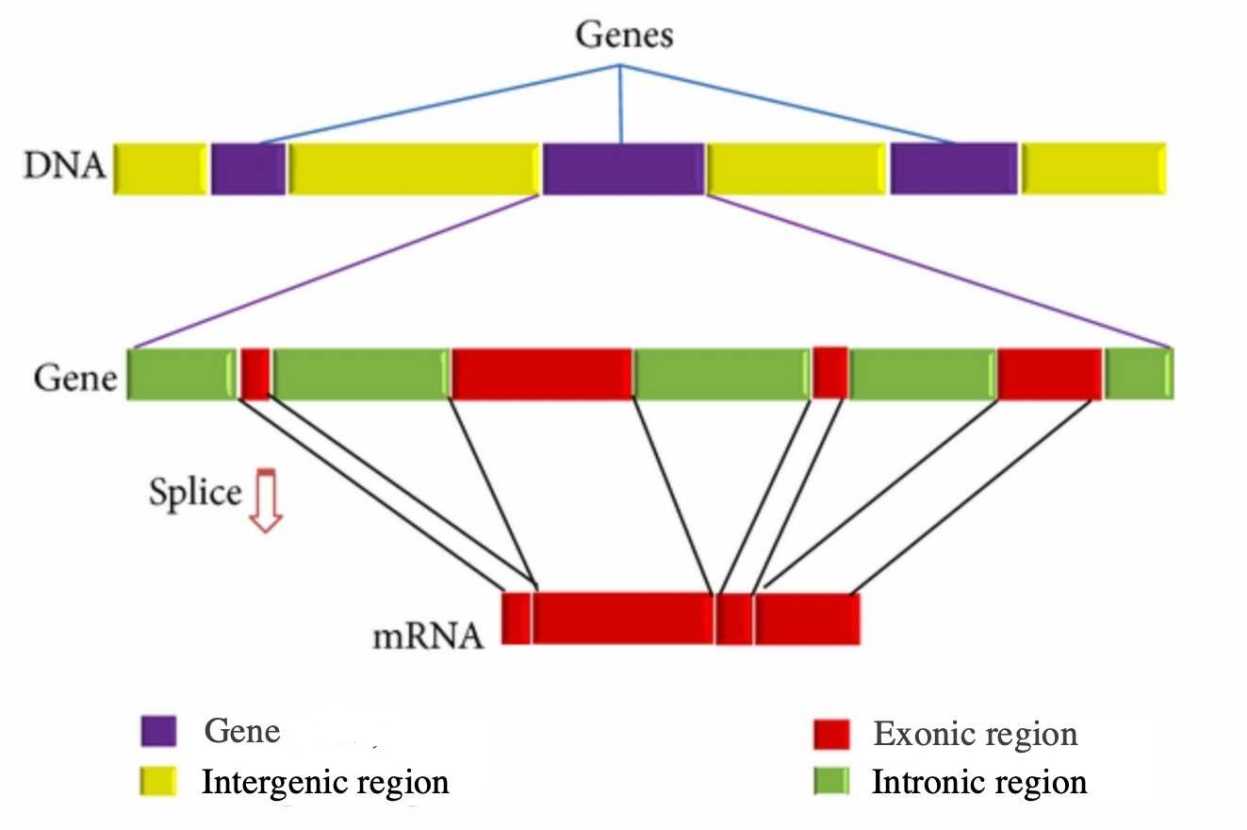

Figure 1: The DNA structure of eukaryotes and the splicing process.

This figure shows that eukaryotic DNA consists of exonic, intronic and intergenic regions, and the exonic regions are interrupted by introns in eukaryotic DNA. (Modified from Liu & Luan, 2014).

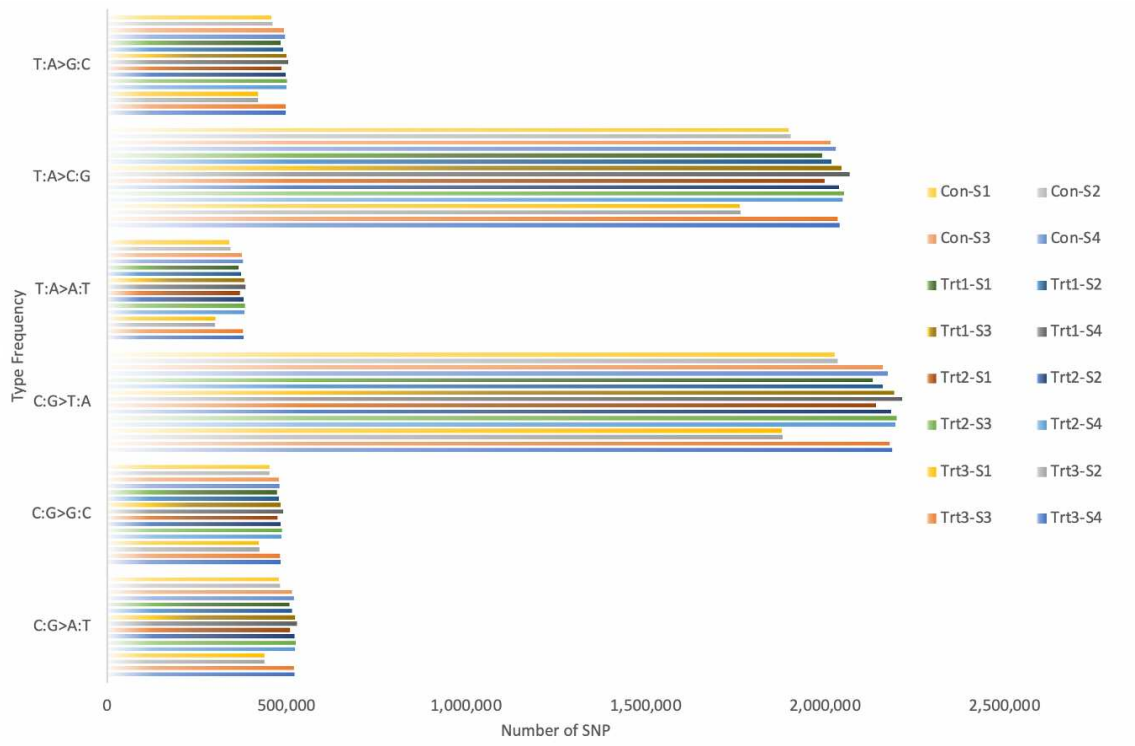

Figure 2: The frequency of each SNP type in each sample.

Con: bovine cells only
Trt1: bovine cells + CRISPR with the guide RNA targeting *stx* genes
Trt2: bovine cells + CRISPR without guide RNA+*E. coli* O157:H7
Trt3: bovine cells + CRISPR with the guide RNA targeting *stx* genes+*E. coli* O157:H7

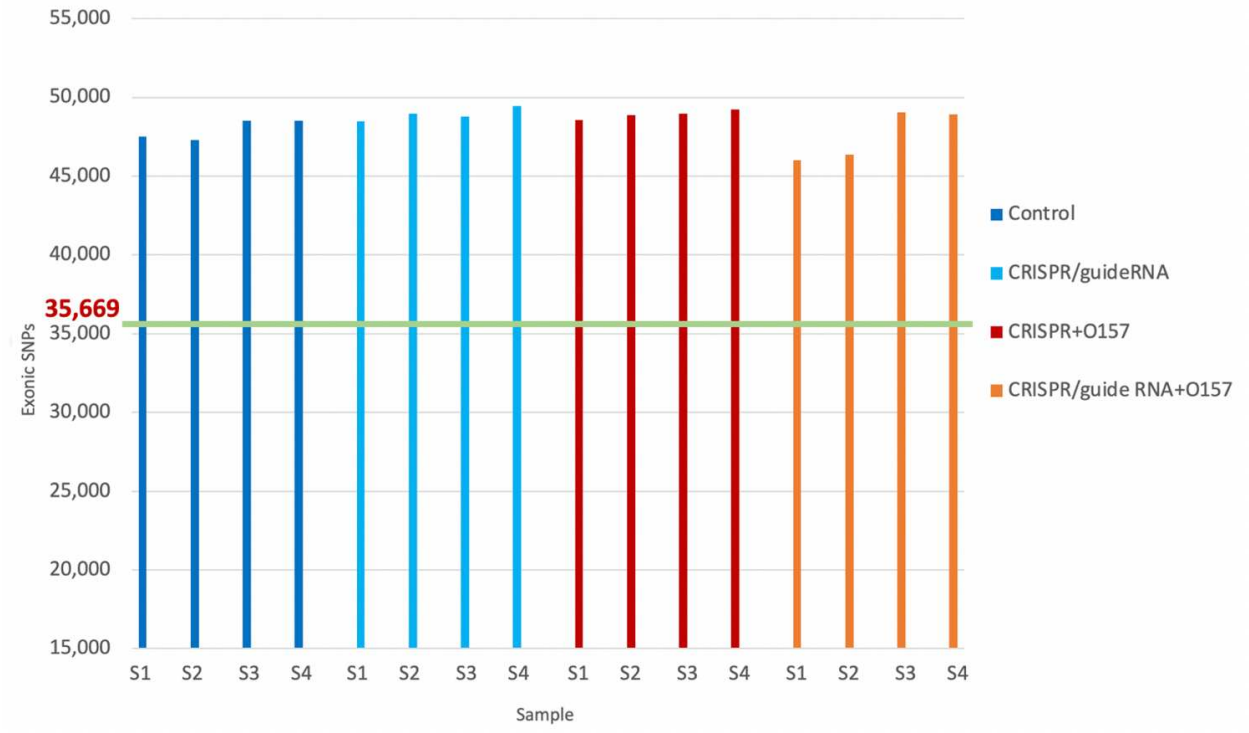

Figure 3: Counts of exonic SNPs in each sample.

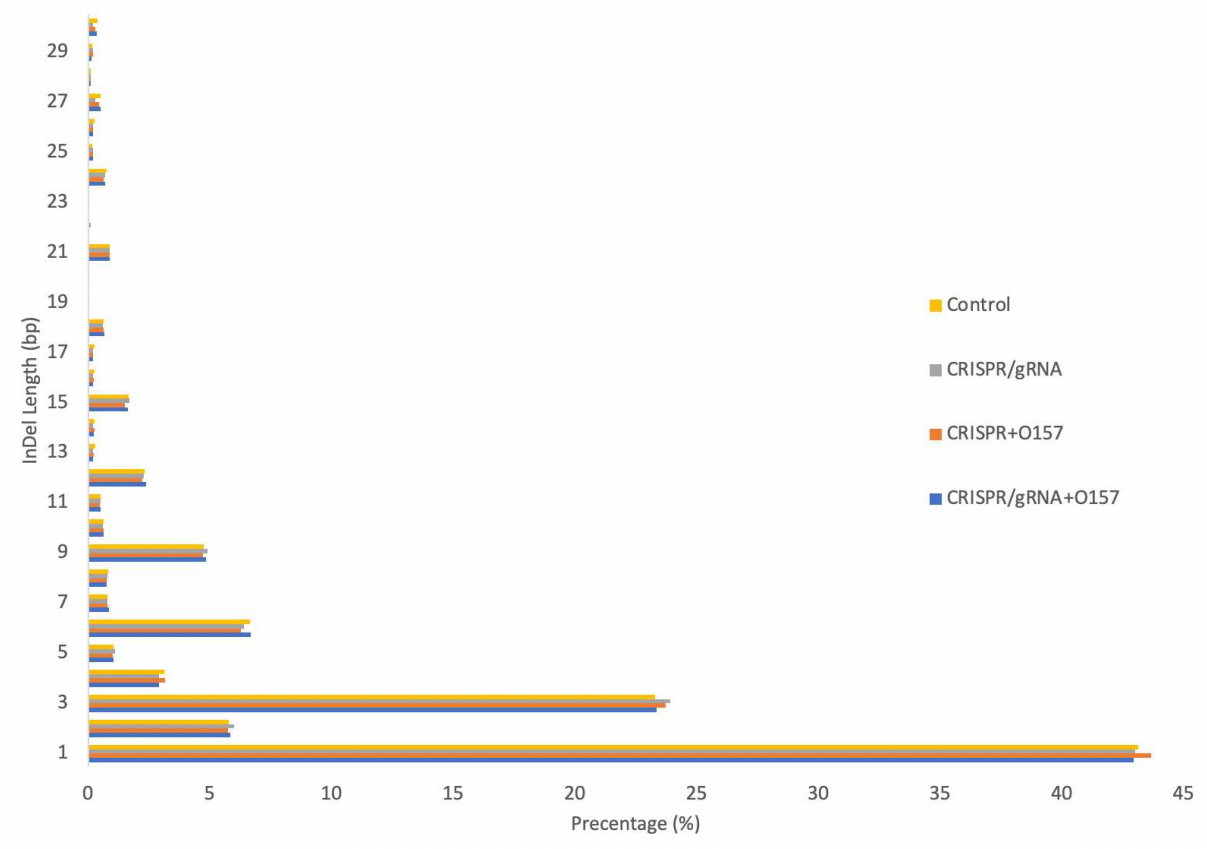

Figure 4: Length distribution of InDels in the control and CRISPR treatments (exonic regions).

REFERENCES

Akbari O, Bellen H, Bier E, Bullock SL, Burt A, Church GM, Cook KR, Duchek P, Edwards OR, Esvelt KM1, Gantz VM, Golic KG, Gratz SJ, Harrison MM, Hayes KR, James AA, Kaufman TC, Knoblich J, Malik HS, Matthews KA, O'Connor-Giles KM, Parks AL, Perrimon N, Port F, Russell S, Ueda R, Wildonger J. 2015. Safeguarding gene drive experiments in the laboratory. Science 349(6251):927–929. doi: 10.1126/science.aac7932.

Bacher U, Shumilov E, Flach J, Porret N, Joncourt R, Wiedemann G, ... & Pabst T. 2018. Challenges in the introduction of next-generation sequencing (NGS) for diagnostics of myeloid malignancies into clinical routine use. Blood cancer journal. 8(11):1-10. doi: 10.1038/s41408-018-0148-6.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, ... & Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 315(5819): 1709-1712. doi: 10.1126/science.1138140.

Buermans HPJ, & Den Dunnen JT. 2014. Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. 1842(10): 1932-1941. doi: 10.1016/j.bbadis.2014.06.015.

Beyret E, Liao HK, Yamamoto M, Hernandez-Benitez R, Fu Y, Erikson G, ... & Belmonte JCI. 2019. Single-dose CRISPR–Cas9 therapy extends lifespan of mice with Hutchinson–Gilford progeria syndrome. Nature medicine. 25(3):419-422. doi: 10.1038/s41591-019-0343-4.

Carvalho GR, Creer S, Allen MJ, Costa FO, Tsigenopoulos CS, Goff-Vitry L, ... & Metfies K. 2009. Genomics in the Discovery and Monitoring of Marine Biodiversity. In Introduction to Marine Genomics. Series: Advances in Marine Genomics. Vol. 1. Cock JM, Tessmar-Raible K, Boyen C, Viard F. (Eds.). Springer. pp:1-32. doi: 10.1007/978-90-481-8639-6_1.

Chen JS, Ma E, Harrington LB, Da Costa M, Tian X, Palefsky JM, & Doudna JA. 2018. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science. 360(6387):436-439. doi: 10.1126/science.aar6245.

Chen M, Mao A, Xu M, Weng Q, Mao J, & Ji J. 2019. CRISPR-Cas9 for cancer therapy: Opportunities and challenges. Cancer letters. 447:48-55. doi: 10.1016/j.canlet.2019.01.017.

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, ... & Root DE. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology. 34(2):184-191. doi: 10.1038/nbt.3437.

Doudna JA, & Charpentier E. 2014. The new frontier of genome engineering with CRISPR-Cas9. Science. 346(6213). doi: 10.1126/science.1258096.

Ebina H, Misawa N, Kanemura Y, & Koyanagi Y. 2013. Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. Scientific reports. 3(1):1-7. doi: 10.1038/srep02510.

El-Mounadi K, Morales-Floriano ML, & Garcia-Ruiz H. 2020. Principles, applications, and biosafety of plant genome editing using CRISPR-Cas9. Frontiers in plant science. 11. doi: 10.3389/fpls.2020.00056.

Gomez-Chiarri M, Guo X, Tanguy A, He Y, & Proestou D. 2015. The use of-omic tools in the study of disease processes in marine bivalve mollusks. Journal of invertebrate pathology. 131: 137-154. doi: 10.1016/j.jip.2015.05.007.

Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, ... & Zhang F. 2017. Nucleic acid detection with CRISPR-Cas13a/C2c2. Science. 356(6336):438-442. doi: 10.1126/science. aam9321.

Hu T, Chitnis N, Monos D, & Dinh A. 2021. Next-generation sequencing technologies: An overview. Human Immunology. doi: 10.1016/j.humimm.2021.02.012.

Hua JT, Ahmed M, Guo H, Zhang Y, Chen S, Soares F, ... & He HH. 2018. Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19. Cell. 174(3): 564-575. doi: 10.1016/j.cell.2018.06.014.

Jiang W, Bikard D, Cox D, & Zhang F. 2013. Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 31(3): 233-239. doi: 10.1038/nbt.2508.

Kennedy EM, Kornepati AV, Goldstein M, Bogerd HP, Poling BC, Whisnant AW et al. 2014. Inactivation of the human papillomavirus E6 or E7 gene in cervical carcinoma cells by using a bacterial CRISPR/Cas RNA-guided endonuclease. Journal of virology. 88(20):11965–11972. doi: 10.1128/JVI.01879-14.

Kennedy EM, Bassit LC, Mueller H, Kornepati AVR, Bogerd HP, Nie T et al. 2015. Suppression of hepatitis B virus DNA accumulation in chronically infected cells using a bacterial CRISPR/Cas RNA-guided DNA endonuclease. Virology. 476:196–205. doi: 10.1016/j.virol.2014.12.001.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, & Cooper DN. 2007. Single base-pair substitutions in exon–intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Human mutation. 28(2):150-158. doi: 10.1002/humu.20400.

Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, ... & Wijmenga C. 2013. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 9(1):e1003201. doi: 10.1371/journal.pgen.1003201.

Li H, & Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25(14):1754-1760. doi: 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, ... & Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25(16):2078-2079. doi: 10.1093/bioinformatics/btp352.

Liu G & Luan Y. 2014. Identification of Protein Coding Regions in the Eukaryotic DNA Sequences Based on Marple Algorithm and Wavelet Packets Transform. Abstract and Applied Analysis. 2014:14. doi: 10.1155/2014/402567.

Luthra R, Chen H, Roy-Chowdhuri S, & Singh RR. 2015. Next-generation sequencing in clinical molecular diagnostics of cancer: advantages and challenges. Cancers. 7(4):2023-2036. doi: 10.3390/cancers7040874.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, ... & 1000 Genomes Project Consortium. 2013. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome research. 23(5):749-761. doi: 10.1101/gr.148718.112.

Mullaney JM, Mills RE, Pittard WS, & Devine SE. 2010. Small insertions and deletions (INDELs) in human genomes. Human molecular genetics. 19(R2):131-136. doi: 10.1093/hmg/ddq400.

Narzisi G, & Schatz MC. 2015. The challenge of small-scale repeats for indel discovery. Frontiers in bioengineering and biotechnology. 3(8). doi: 10.3389/fbioe.2015.00008.

Oye, KA, Esvelt K, Appleton E, Catteruccia F, Church G, Kuiken T, Lightfoot SB, McNamara J, Smidler A, Collins JP. 2014. Regulating gene drives. Science. 345(6197):626−628. doi: 10.1126/science.1254287.

Rath J. 2017. Safety and Security Risks of CRISPR/Cas9. Ethics Dumping: Case Studies from North-South Research Collaborations. 107. doi: 10.1007/978-3-319-64731-9_13.

RStudio Team. 2020. RStudio: Integrated Development for R. RStudio. PBC. Boston, MA. Retrieved from http://www.rstudio.com/.

Schierding W, Antony J, Cutfield WS, Horsfield JA, & O'Sullivan JM. 2016. Intergenic GWAS SNPs are key components of the spatial and regulatory network for human growth. Human molecular genetics. 25(15):3372-3382. doi: 10.1093/hmg/ddw165.

Sovová T, Kerins G, Demnerová K, Ovesná J. 2016. Genome editing with engineered nucleases in economically important animals and plants: state of the art in the research pipeline. Current Issues in Molecular Biology. 21:41–62. doi: 10.21775/cimb.021.041.

Vignal A, Milan D, SanCristobal M, & Eggen A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics selection evolution. 34(3):275-305. doi: 10.1051/gse:2002009.

Wang K, Li M, & Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 38(16):e164. doi: 10.1093/nar/gkq603.

Yang Y, Wang L, Bell P, McMenamin D, He Z, White J, ... & Wilson JM. 2016. A dual AAV system enables the Cas9-mediated correction of a metabolic liver disease in newborn mice. Nature biotechnology. 34(3):334-338. doi: 10.1038/nbt.3469.

Yang H, Jia M, Geornaras I, Woerner DR, Morley PS, Belk KE. 2017. Expand the capabilities of a CRISPR-Cas9 system for sequence-specific elimination of foodborne pathogens in beef production. Final report submitted to National Cattlemen's Beef Association by the Center for Meat Safety & Quality, Colorado State University, Fort Collins, CO. 24 p.

Yang H, Jia M, Geornaras I, Woerner DR, Morley PS, Belk KE. 2018. Construct a phage-mediated system to deliver CRISPR-Cas9 antimicrobials for sequence-specific elimination of foodborne pathogens in beef production. Final report submitted to National Cattlemen's Beef Association by the Center for Meat Safety & Quality, Colorado State University, Fort Collins, CO. 32p.

Yin H, Song CQ, Dorkin JR, Zhu LJ, Li Y, Wu Q, ... & Anderson DG. 2016. Therapeutic genome editing by combined viral and non-viral delivery of CRISPR system components in vivo. Nature biotechnology. 34(3):328-333. doi: 10.1038/nbt.3471.

Zhang XH, Tee LY, Wang XG, Huang QS, & Yang SH. 2015. Off-target effects in CRISPR/Cas9-mediated genome engineering. Molecular Therapy-Nucleic Acids. 4:e264. doi: 0.1038/mtna.2015.37.

Zhao RF. 2012. ENCODE: Deciphering Function in the Human Genome. National Human Genome Research Institute. Retrieved from: https://www.genome.gov/27551473/genome-advance-of-the-month-encode-deciphering-function-in-the-human-genome/.

Zwane AA, Schnabel RD, Hoff J, Choudhury A, Makgahlela ML, Maiwashe A, ... & Taylor JF. (2019). Genome-wide SNP discovery in indigenous cattle breeds of South Africa. Frontiers in genetics. 10:273. doi: 10.3389/fgene.2019.00273.