Spring 5-2020

Using Bioinformatics Tools to Evaluate Potential Risks of Food Allergy and to Predict Microbiome Functionality

Mohamed Abdelmoteleb
University of Nebraska - Lincoln, mohamed.attia@huskers.unl.edu

USING BIOINFORMATICS TOOLS TO EVALUATE POTENTIAL RISKS OF FOOD ALLERGY AND TO PREDICT MICROBIOME FUNCTIONALITY

by

Mohamed Abdelmoteleb

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Food Science and Technology

(Bioinformatics)

Under the Supervision of Professor Rick Goodman

Lincoln, Nebraska

May, 2020

USING BIOINFORMATICS TOOLS TO EVALUATE POTENTIAL RISKS OF FOOD ALLERGY AND TO PREDICT MICROBIOME FUNCTIONALITY

Mohamed Abdelmoteleb, Ph.D.

University of Nebraska, 2020

Advisor: Rick Goodman

Novel foods and Genetically Engineered (GE) organisms are being developed for nutritional, industrial, and environmental applications. Dietary interventions have been used recently to mitigate methane emissions in ruminants. In this project, bioinformatics tools have been used to answer two main questions. The first question is the potential allergy risks for consumption of novel foods and GE organisms. The second question is the effects of dietary interventions on microbiome functionality related to methane production in ruminants.

To answer the first question, regulatory authorities in the United States and Europe now expect an evaluation of new proteins in novel foods or genetically engineered organisms to be evaluated for possible allergy and Celiac disease (CeD) risk. Two microalgal species, a fungus, House Cricket, and GE Canola have been tested to evaluate potential IgE cross-reactivity. Whole genome sequencing, genomic, transcriptomic, proteomic, and immunoinformatics techniques have been used to predict potential cross-reactivity. Bioinformatics tools helped us to characterize their proteomes and critically evaluate matches to putative or proven allergens. The two microalgal species and *Fusarium* sp. had matches to putative allergens, which are extensively conserved in allergenic, and non-allergenic species, leading to the need for critical evaluation of the CODEX guidelines. Shrimp allergic patients may experience cross-

reactions if they consume crickets. There is no reason to suspect that the GE canola would elicit allergic reactions or would induce toxic responses. In addition, we developed a sequence searchable celiac database to identify peptides and proteins for risk assessment of novel food proteins.

Concerning the second question, we studied the effect of dietary nitrate and sulfate on finishing cattle performance and methane emissions. To address to question, 16S sequencing and metagenomics were used for better understanding of rumen microbiome composition and functionality. Sulfate and nitrate combination helped to reduce methane emissions, with a reduction in average daily gain, dry matter intake and gain:feed. Ruminal bacterial composition illustrated high abundance of phyla with less hydrogen production, and genera with high H2 utilization capability in fatty acids' formation, sulfate and nitrate reduction instead of methane production. Metagenomics demonstrated a significant decrease in enzymes linked to methanogenesis in COMBO diet.

ACKNOWLEDGMENTS

I sincerely appreciate the guidance and intellectual simulation from my advisor Dr. Rick Goodman. I really appreciate all your advice, continuous kind help, inputs, patience, trust, guidance during this project and critical revision of the manuscript.

I sincerely thank Dr. Samodha Fernando for his help, guidance and support, and revision of the microbiome section. Thank you, Dr Fernando for giving me the opportunity to work in your lab to get experience in many experimental and computational tools.

I would like to thank Dr. Joseph Baumert, Dr. Amanda Ramer-Tait and Dr. Chi Zhang for serving on my committee. Thank you, Dr. Baumert for your help whenever I needed and nice discussions in FARRP lab meeting. Thank you, Dr. Amanda Ramer-Tait for your amazing teaching in your class and the basic background I learned in Immunology. Thank you, Dr. Chi Zhang for helping me expand my knowledge of bioinformatics and answering my questions whenever I needed any help.

Special thanks and appreciation to the Government of Egypt for funding my PhD studies for 4 years. Thank you, my Professors, colleagues at Botany Department, Faculty of Science, Mansoura University on your help, simulation, and supporting my application for the External Mission. I'd like to thank also Mansoura University for supporting me as a candidate for this External Mission in the field of Bioinformatics.

I sincerely thank Dr. Josh Herr for helping me with troubleshooting computational questions I had in this project. Appreciation to Holland Computing Center for providing supercomputing resources and the patient technical assistance from Dr. Jean-Jack Reitoven, Dr. Natasha Pavlovikj and Carrie Brown.

I am grateful to have all my lab mates in FARRP lab, Dr. Samodha Fernando's lab, and Dr. Chi Zhang's Lab. Thank you for your support, kind help and encouragement during the work of this project.

Special appreciation to Lee Palmer for helping me in proteomic data analysis. Thank you, Car Reen Kok, Chris Anderson, Wesley Tom, Nirosh Aluthge and Waseem Abbas for your kind help in fixing bugs I had in some software tutorials.

Thank you, my parents, brothers, and my family for your continuous support and help during my life and always being there for me.

Lastly, thank you my lovely wife, and my lovely son, Omar, for supporting me, never giving up on me and being there for me at any time.

GRANT FUNDING

PhD financial support was funded by Egyptian government scholarship (GM:1031) for 4 years. Research funding was provided by the Food Allergy Research & Resource Program (FARRP) at the University of Nebraska. This research is also partially funded by Fermentalg and Sustainable Products. This work is supported also by Animal Nutrition, Growth and Lactation grant no. 2018-67015-27496, Effective Mitigation Strategies for Antimicrobial Resistance grant no. 2018-68003-27545, and Multi-state research project no. 1000597 from the USDA National Institute of Food and Agriculture awarded to SCF.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

LIST OF FIGURES

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 1

LITERATURE REVIEW

1.1. Introduction

Novel food ingredient sources are being developed to meet the growing demand for dietary interventions in industrialized countries due to the increasing human population, concerns for animal welfare, and environmental impacts of traditional sources of these foods. In addition, food production is challenged by changing weather patterns, loss of arable land, and evolving plant pests and disease. Considering the safe use of novel food sources, it is imperative to evaluate possible risks including their allergenic potential. Many of the novel sources have been consumed in some geographic regions and there may be a history of safe use, although use and safety or risk are rarely well documented e.g. insects (Palmer et al, 2020). Some new sources are truly novel, with no history of safe human consumption, for example various microbial sources such as microalgal, fungal or yeast sources have been introduced as foods or food ingredients (Schonknecht et al, 2013). In addition, genetically modified (GM) plants are increasingly used for industrial applications and food production.

Without a history of use, the potential risk for food allergy in novel foods cannot be based on population data and must therefore be assessed using other means. Regulatory authorities in industrialized countries have rules governing the use of novel foods and the potential risks of food allergy is an important health issue that requires consideration in the assessment of these novel foods (Goodman et al, 2016). The primary health concern is whether the new food represents a risk as an allergenic source for at least a proportion of the general population, primarily those allergic to similar proteins.

Dietary intervention strategies have been also used to mediate some environmental and climatic issues. Greenhouse Gases (GHGs) and global warming are one of the major concerns. The United States is second among the top 10 emitters of global greenhouse gases (GHGs). Only China and the US generate more than one third of the total emissions from GHGs. Methane production through enteric fermentation in ruminants accounts for 27% of the total global methane emission (EPA, 2020). Methane is a greenhouse gas with 28 times global warming potentiality than that of CO_2 (Myhre et al, 2013). Methane losses in enteric fermentation in cattle accounts for 2-12% of total gross energy intake, which otherwise could be used for improving cattle performance and milk production (Hristov et al. 2013). Methanogenesis is driven by ruminal microbial communities including methanogens (mainly archaea), bacteria, fungi and viruses. As diet can change the composition of microbial communities, dietary intervention can be used to reduce greenhouse gas emissions from cattle by controlling microbial populations (Russell 2002). Understanding the impact of dietary interventions on the microbiome composition and functionality may help to increase food resources and environmentally safe livestock production.

1.2. Evaluation of potential risks of food allergy for novel foods and genetically engineered (GE) organisms

The US recognizes eight major allergenic sources (peanut, tree nuts, milk, eggs, crustacean shellfish, finned fish, soybeans and wheat), the European Union recognizes 14 (adding barley, rye and oats to cereals, a reduced number of tree nuts, mustard, sesame seeds, lupin, molluscan shell fish and Sulphur dioxide) (Taylor and Hefle, 2006). Allergen management of conventional packaged foods is through appropriate labeling to

warn allergic consumers of the specific contents so they can voluntarily avoid the food. The United States, European Union and many other countries require labeling of all ingredients, not just the allergenic ingredients. Additional rules exist in many countries for managing and identifying potential cross-contact between allergens and foods which do not contain allergens in their ingredient lists. For foods which are improperly labelled, countries such as the U.S. can force a recall of the food (Allen et al, 2014).

Understanding food allergy risks requires knowledge of the proteins in various foods that commonly and rarely cause food allergy. While risks of food allergy normally only address the specific foodstuff or ingredient, only a few protein types from a source cause most reactions. Peanut (*Arachis hypogaea*) is a source of many severe allergic reactions, and the dominant allergens are thought to be the most abundant seed storage proteins, which have been designated Ara h 1, a vicilin; Ara h 2, a 2S albumin; Ara h 3, a legumin-like globulin and Ara h 6, another abundant 2S albumin. Ara h 2 and Ara h 6 are proteins stabilized by four intrachain disulfide-bonds and they are readily soluble in saliva, making them available for immediate reactivity (Bublin and Breiteneder, 2014). In addition, they are not rapidly digested at acidic pH by pepsin, suggesting stability in the stomach. Eleven other peanut proteins are recognized as allergens, though it is clear they are less potent than Ara h 1, 2, 3 and 6 and only a few individuals have IgE antibodies to them. However, peanuts also produce a few thousand other proteins that are not recognized as allergens. Protein homologues of the dominant peanut allergens are found in other legumes and tree nuts and are the major allergens for most people with clinical allergy. Other peanut allergens, such as peanut agglutinin, profilin (Ara h 5), PR-10 protein (Ara h 8), lipid transfer proteins (Ara h 9, 16 and 17) oleosins (Ara h 10, 11, 14

and 15), and defensins (Ara h 12 and 13) are minor allergens and are recognized as less abundant and less potent, posing less significant risks of allergy in the population (Bublin and Breiteneder, 2014).

1.2.1. Novel foods

Recently, some novel food sources have been developed for several nutritional, environmental and industrial purposes. What is or should be done to evaluate the safety of these products? Are there specific ways to predict allergenicity? Or is the risk primarily one of IgE cross-reactivity?

1.2.1.1. Microalgae, fungal or yeast sources

Microalgae is considered a potential food source due to its high protein content and other nutritional components e.g. amino acids, vitamins, dietary fiber, and a variety of antioxidants, bioactive materials, and chlorophylls. *Chlorella* is commonly consumed particularly in East Asian countries such as Japan, Taiwan, and Korea. Some algal species are not considered as novel in Europe and the US e.g. *Chlorella vulgaris* and *Chlorella pyrenoidosa*, since historically they have been consumed in foods in many countries (Wells et al. 2017). Other unicellular algae have been developed recently for use in food products but has not yet been consumed by humans (Schonknecht et al, 2013).

Fungal sources have been used in several food products. Quorn is one of the most common examples, which contains mycoprotein derived from *Fusarium venenatum* and produced via fermentation. Quorn products have been consumed in the United Kingdom for 30 years and since 2002 in the US (Finnigan et al, 2019). Other strains of *Fusarium* with different compositions are now under development.

1.2.1.2. Insects

Insects may be one of the novel food sources due to their high content in proteins, nutrients, and other environmental factors (Payne et al, 2016; Rumpold and Schluter, 2013; van Huis et al, 2016, Hall et al, 2017). The use of insects as food (entomophagy) dates to the early development of humans. Over 2000 species are reportedly consumed in 113 countries mainly in Africa, Asia and Latin America (van Huis et al, 2013; EFSA, 2015). In the US and Europe, entomophagy has been marginalized. However, this situation is changing currently. The European Union has identified some insects including cricket, mealworm, wax worm and locust as novel food sources for human consumption. Some studies reported that consumers in western countries may be classified within four groups ranging from strongly disgusted to strong acceptors toward insect consumption (Cunha et al, 2014; Cunha et al, 2015). Food developers are beginning to use mealworm and cricket as protein sources in processed foods (Broekman et al, 2017).

1.2.2. Genetically engineered (GM) organisms

The direct introduction of genes into plants raised questions regarding food safety and evaluation processes of potential allergenicity and toxicity of foods derived from the GM plants. The safety assessment of genetically engineered (GE) organisms has served as a model for assessing allergenicity risk of novel foods in the US. Risks and assessment steps for GE organisms were broadly discussed in the early 1990's (Federal Register Docket No. 92N-0139, Vol 57, No. 104, May 29, 1992) and (Metcalfe et al, 1996). A primary health related concern has been whether a new gene in a GE organism encodes an allergen or a potentially cross-reactive protein that would act as an allergen for those

who are already allergic. Advisory groups were convened by the Food an Agricultural Organization (FAO) and World Health Organization (WHO) panels in 1996 and 2000. Primary questions were around the source of the gene(s) transferred to a new GE crop and a sequence comparison to known allergenic proteins. If the source was known to cause allergies then the protein from the transferred source once expressed should be tested using sera from subjects allergic to the source to understand whether the protein is a cause of allergy (Goodman et al, 2008a). Additionally, if the protein has an eight amino acid identity match to an allergen in a BLASTP or FASTA alignment, sera from subjects allergic to the source should be tested for IgE binding (Hileman et al, 2002). It is also recommended to test the stability of the protein with a simple test-tube assay with pepsin at pH 1.2 (Astwood et al, 1996). The additional characteristic of relative abundance of potent food allergens compared to other proteins is often forgotten but seems right when viewing dominant allergens in peanuts and tree nuts (Astwood and Fuchs, 1996}. A panel of scientists gathered by FAO/WHO in 2001 suggested using unproven additional tests in bioinformatics, higher pH in pepsin digestion, targeted human IgE test using samples of 50 serum from subjects with unrelated allergies and unproven animal model tests. This was a more stringent demand for tests than had been required before as reviewed in 2005 (Goodman and Hefle, 2005). The CODEX Alimentarius Commission Guideline in 2003 (CODEX, CAC/GL 44-2003) corrected the overly ambitious FAO/WHO 2001 guideline. The CODEX 2003 guideline was reaffirmed in 2009 (CODEX 2009).

The CODEX Alimentarious Commission of the Food and Agricultural Organization and World Health Organization of the United Nations had recommended a weight-of-evidence (WoE) approach with a set of experimental tools for an overall

assessment of the allergenic potential cross-reactivity for genetically engineered organisms (CODEX, 2009). Bioinformatics, as one of these approaches, is a screening process that can describe the degree of similarity between a novel protein and known allergenic proteins and if the identity is above the threshold, serum testing would be performed using sera from subjects allergic to the matched allergen (Goodman, 2008b).

While dealing with novel food sources, it is appropriate to evaluate possible health risks to consumers, including the IgE mediated allergenic potential (Naigeli et al, 2017; Sicherer and Sampson, 2018) and celiac disease eliciting potential based on peptides of gluten from wheat, barley and rye grains that are recognized by T cells or induce intestinal toxicity. Celiac disease (CD) is an autoimmune disorder in the upper small intestine occurring in genetically susceptible individuals, triggered mainly by gluten and related prolamins. The disease is considered one of the most common genetic autoimmune diseases which affects 1.4% of the global population. Well characterized haplotypes in the human leukocyte antigen (HLA) class II region (either DQ2 or DQ8) confer a large part of the genetic susceptibility to celiac disease (Ruiz-Carnicer et al, 2019; Sollid et al. 2012). The AllergenOnline.org databases includes allergens and putative allergens as well as celiac eliciting peptides in searchable format of these updated and curated databases (www.allergenonline.org/celiachome.shtml). For allergy the criteria from CODEX is commonly used, >35% identity over 80 amino acids. For Celiac disease, exact peptide matches to published native or deamidated peptides that cause T cell proliferation in the context of MHC Class II, DQ2 or DQ8 is a primary criterion for concern (Ruiz-Carnicer et al, 2019; Sollid et al. 2012).

1.2.3. Bioinformatics approaches in evaluation of allergy risk assessment

The development of rapid genomic, transcriptomic and proteomic methods and auto-annotations of proteins as "allergens" in NCBI is leading to regulatory demands for whole genome analysis or transcriptome analysis of whole novel food organisms, not just GE organisms. Some regulators or scientific advisors are recommending using predicted proteins from the whole organism genome or transcriptome be compared to allergen databases using the CODEX guidelines to predict possible risks of food allergy. Importantly, the CODEX guideline was not intended to evaluate the full-proteome or predicted protein dataset of a whole organism and the criteria of >35% identity over 80 has not been validated for whole proteome comparisons and are likely over-predictive for cross-reactivity.

How do we judge potential risks of IgE cross-reactivity of all the proteins expressed by whole organisms? Few organisms used in foods have broad protein sequence records. It would be very complex and expensive to define the proteome of organisms or even the tissues that are consumed de novo. Thus, developers are left with the choice of using a full genome predicted proteome if the genetic sequence is known for the source species or developing specific cDNA libraries from specific tissues. In either case, predicting which genes or transcripts are actively translated into protein, and the doses of each protein can only be estimated. The predicted proteomes can then be compared to known allergens using diverse bioinformatics cutoffs to estimate protein identity matches across broad taxa to estimate how many proteins would likely require serum IgE tests to consider possible risks of cross-reactivity (Goodman, 2006; Goodman, 2008b).

Until June 2019, the amino acid sequence comparison could be performed at the NCBI website (https://www.ncbi.nlm.nih.gov/protein) using BLASTP, limited by keywords (allergen, allergy) limits. The general unlimited BLASTP now available online is much less useful when results show broad taxonomic matches to all classifications as the user must perform literature research to understand the relevance of any matches. The UniProt database (https://www.uniprot .org/) also has a BLASTP function, but without a keyword search. Therefore, searches with a special allergen database are far more useful. It is efficient and productive to use a peer reviewed allergen database such as AllergenOnline.org (Goodman et al, 2016). However, because all databases have a lag-time in updating, a recommended final step is to use BLAST comparison to a downloaded NCBI protein sequences with a restricted keyword delimiter of ‘‘allergen” using a tool which allows a coded keyword search. Therefore, AllergenOnline.org has developed an alternative approach to identify putative and proven allergens from the NCBI (Goodman and Hefle, 2005). Both the specialized allergen databases and the list of proteins identified as “allergen” in NCBI contain many sequences that are members of highly conserved protein groups and are associated with “allergen” simply based on modest sequence identities, without proof of allergy. Thus, it should be expected that many potentially false positive sequence alignments will be identified as the result of a full-genomics screen.

The AllergenOnline.org databases includes allergens and putative allergens as well as celiac eliciting peptides in searchable format of these updated and curated databases (http://www.allergenonline.org/celiachome.shtml). For allergy the criteria from CODEX is commonly used, >35% identity over 80 amino acids. If the protein causes

basophil activation or skin prick test positive reactions, it is considered as an allergen and added to AOL (Goodman et al, 2016). It is important to note that many of the sequences in the www.Allergenonline.org database have only been demonstrated to show IgE binding from people with allergies and have not clearly been proven to be the cause of clinically defined allergies of any kind (Scala et al, 2018; Faber et al, 2017; Ruethers et al, 2018). Furthermore, the database has not collected information on the dose of the protein in the source material, or stability in heating or to digestion by pepsin. One of the requirements by regulators in the US and the EU is to assure that the new food products with processed, cultured species including algae, molds, and insects are safe for allergic subjects (van Putten et al, 2006, van der Spiegel, 2013, EFSA, 2015). The AOL uses FASTA comparison with the criteria of matches being >35% identity over 80 amino acids as was set by the CODEX Allergenicity guideline in 2003. That guidance should, however, be viewed as highly conservative and precautionary based on historical experiences of cross-reactivity and clinical co-reactivity. As noted by various researchers, in vitro IgE cross reactivity is common for proteins sharing >70% amino acid identity over nearly their full-lengths, but cross-reactivity is extremely rare for proteins sharing less than 50% identity (Aalberse, 2000). It is also important to consider other aspects of protein structure and IgE binding to understand cross-reactivity (Aalberse et al, 2001). Allergens that are cross-reactive by shared IgE binding in a laboratory test may cause clinical reactivity, but that is not certain. Reactions can be very mild to severe depending on the sensitivity of the consumer, the number of IgE binding epitopes, the affinity of binding, the amount of protein, and the route of exposure. Proteins that cause cross-reactions can be grouped into protein families, although there are many non-allergic

proteins within any of the identified protein groups. For instance, Bet v 1 homologues are commonly expressed in many plant food sources, and a few cause allergies but most are heat inactivated by mild cooking (Radauer and Breiteneder, 2006; Bohle et al, 2006). Cupins include vicilins and legumins that are major seed storage proteins and important plant food allergens (Radauer and Breiteneder, 2007; Scala et al, 2018). Importantly, the Cupin superfamily includes many proteins from many non-allergenic sources including human proteins as seen in Figure 2 (Radauer and Breiteneder, 2007). As with the important muscle allergen tropomyosins from crustaceans, IgE cross-reactions are common among crustaceans but less so with mollusks and insects or house dust mites, all of which share >60% identity. However, homologues in birds and mammals show more than 52% identity to shrimp tropomyosin, but do not share clinical cross reactivity (Faber et al, 2016; Ruethers et al, 2018).

It has been suggested that *E*-scores (expectation scores) generated from the FASTA algorithm are useful evaluation criteria to be considered in order to make a more informed decision as to whether a protein has the potential to cause allergenic cross-reactivity along with the current criterion of >35% identity over 80 amino acid threshold (Thomas et al, 2005; Ladics et al, 2007; Silvanovich et al, 2009; Cressman et al, 2009). The E-score reflects the measure of relatedness among protein sequences and can help separate the potential random occurrence of aligned sequences from those alignments that may share structurally relevant similarities. A very small E-score (e.g., 10e-7) reflects a likely functional similarity and may suggest a biologically relevant similarity for allergy or potential cross-reactivity, while large E-scores (>1.0) are typically associated with

alignments that do not represent a biologically relevant similarity (Pearson 2000, 2014, 2016; Henikoff and Henikoff 1992, 1996).

The end-result of the bioinformatics comparison with allergens is a decision about the need for specific serum testing and if so, the specific allergic population that should be used to collect serum samples (Goodman et al, 2005). But, since appropriate serum testing is not trivial, correct interpretation of bioinformatics findings are important. It is often difficult to obtain relevant human samples and the experimental design required to conduct a test assessing possible cross-reactivity can be complex. Therefore, comparisons using FASTA searches with the CODEX guidelines to identify potentially risky proteins requiring further testing needs to be critically re-evaluated (Siruguri et al, 2015).

1.3. Methane mitigation in ruminants

Methane production through enteric fermentation in ruminants is an environmental as well as a nutritional concern. Enteric fermentation includes hydrolysis of plant organic matter into soluble organic molecules; followed by acidogenesis into alcohols and acetogenesis into volatile fatty acids. These steps are controlled by rumen microbiota including bacteria, fungi, and viruses (Russell 2002, Shah 2014). The final step in eneric fermentation is methanogenesis which includes two main routes. The first route is the conversion of H_2 and CO_2 into methane through the most abundant hydrogenotrophic archaea *Methanobrevibacter* and other hydrogenotrophic genera e.g. *Methanimicrococcus Methanosphaera*, and *Methanobacterium*. The other is utilization of methylamines and methanol in methane production by less abundant methylotrophs e.g. Methanosarcinales, Methanosphaera, Methanomassiliicoccaceae (Morgavi et al, 2012). Byproducts e.g. volatile fatty acids and methane are emitted at the end of fermentation

with a significant consequence of reduced cattle performance and efficiency. At the heart of methane production are microbes, and these microbes are known to change based on substrate availability in the diet. Therefore, understanding the relationship between diet, methane, and microbial community will help identify microbial species associated with methane to develop new intervention strategies.

1.3.1. Novel dietary interventions explored for methane mitigation in ruminants

Dietary interventions have been widely explored as methane mitigation strategies (Beauchemin et al, 2007a; Beauchemin et al, 2007b, Buddle et al, 2011; Johnson and Johnson, 1995; McAllister and Newbold, 2008). Several dietary intervention strategies have been studied to mitigate methane emissions, including the use of inhibitors, electron acceptors, ionophores, inclusion of grain over fibrous feed, and defaunation of protozoa (Hristov et al. 2013). Boadi et al, (2004) showed that lipids can reduce methane production in cattle. Unsaturated lipids have been reported as H2 sink alternative competing for H2 with methanogens (Poulsen et al, 2013). Ionophore supplementations have also been screened as a methane mitigation tool (Schelling, 1984). However, the impacts of monensin utilization to reduce methane were indicated as a short term (Johnson and Johnson, 1995). The use of 3-nitrooxypropanol has been reported to decrease methane formation (Duval and Kindermann 2012). In addition, tannin supplementation has shown a 20% reduction on methane production (Khiaosa-ard et al. 2015). Supplementation of steam flaked processed corn in beef cattle has been reported to reduce methane emissions (Hales et al, 2012). Nitrate and supplementations have been shown to dramatically decrease methane emissions in sheep and dairy cows (Van

Zijderfeld et al, 2010). However, Troy et al, (2015) have reported that dietary nitrates had no impact on methane emissions.

1.3.2. Effect of dietary interventions on ruminal microbiome composition

The methane mitigation strategies by dietary intervention influence the ruminal microbial community composition, which is the main driver of methanogenesis. Several studies have explored the effects of diet on microbial communities. However, these studies have not tightened the gap in our knowledge between measurements of methane emissions and dietary impacts on the microbial community structure (Johnson and Johnson, 1995; McAllister et al, 1996; Fernando et al, 2010; Hook et al, 2010). Pesta, 2015 explored the effect of fat supplementation on microbial community composition and methane production. Knoell, 2016 studied the effects of forage quality, and modified distiller's grains plus solubles (MDGS) supplementation on the microbial composition and methane production in growing and finishing cattle. Tapio et al, 2017 have reported a positive correlation between decrease in the abundance of three bacterial OTUs and reducing methane emissions. Two OTUs are dominated by less H2-producing bacteria.

The two main routes of methanogesis are controlled by methenaogens. Some studies found a correlation between *Methanobrevibacter* SGMT clade and methane emissions. Methanobrevibacter SGMT and SGMT clades can utilize high concentrations of H_2, as they have methyl coenzyme M reductase isozymes (McrI and McrII) (Zhou et al, 2011; Danielsson et al, 2012; Shi et al, 2014 and Danielsson, 2016). Other studies have shown weak correlations between relative abundance of methanogens and methane

emissions in dairy cows and sheep (Morgavi et al, 2012; Zhou et al, 2011; Danielsson et al, 2012; Danielsson 2016; Kittelmann et al, 2014 and Shi et al, 2014).

1.3.3. Effect of dietary interventions on ruminal microbiome function

Other studies focused on identifying the effects of dietary supplementations on the ruminal microbiome function. Shabat et al, 2016 analyzed the microbial composition, gene content, and metabolomic composition in 146 milking cows. They reported specific enrichment of metabolic pathways which are correlated with higher methane yield. She et al, (2014) used metagenomic and metatranscriptomic sequencing to identify differences in microbiome function in sheep with low and high methane yield. They demonstrated a similar abundance of methanogens and methanogenesis pathway genes in high and low methane emitters. However, they observed significant increase in the transcriptional profiles of methanogenesis related genes in sheep with high methane yield.

1.3.4. Bioinformatics approaches for identifying microbiome composition and functionality

The lack of current understanding of the rumen microbial ecosystem and the interactions between microbial species is due partially due to lack of sensitivity of the previous experimental tools e.g. classical culturing techniques. Recent genomic tools provide precise characterization and quantification of rumen microbes. Although polymerase chain reaction techniques helped in quantification than culture, they are limited to species with species-specific probes. Similarly, next-generation sequencing technology quantify each species, however only those species that are recorded in the database. Bioinformatics tools are used to determine the taxonomic and the functional

profile of rumen microbes (Franzosa et al., 2015). Metagenomic sequencing of the 16S rRNA subunit of the prokaryotic ribosome has been utilized to identify bacterial and archaeal operational taxonomic units (Kim et al, 2011; Li et al, 2012, Henderson et al, 2015). 16S technology has been widely used to fully characterize the rumen microbiome composition. In addition, shotgun metagenomic sequencing provides direct data for functional and metabolic attributes of the microbiome (Jovel et al, 2016). Functional profiling can be predicted through combining 16S and shotgun metagenomics approaches, 16S function analysis is considered as inferred data, while shotgun sequencing represents direct data for functional aspects of the microbes (Jovel et al, 2016). These tools generate valuable data, but they are restricted to species-level taxonomic identification. However, strain-level identification may provide deep insights into biological questions (Franzosa et al, 2015).

To accurately describe the functional activity, not only the potential, the use of multiomic tools is recommended e.g. transcriptomics, proteomics, and metabolomics (Franzosa et al., 2015). Transcriptomics represent a benefit of giving the option to carry out metagenomic and metatranscriptomic sequencing (Giannoukos et al., 2012). In addition, transcriptomics can help to characterize RNA viruses in the rumen (Culley et al., 2006; Zhang et al., 2006). Proteomic analysis can be used to determine the functional activity, in which the mass and abundance of peptides are calculated using methods based on mass spectrometry; post-translation modifications may also be detected (Altelaar et al, 2012). Metaproteomics provide the functional changes that may occur despite no observable differences in microbiome profile (Franzosa et al., 2015).

Metabolomics detect metabolites and small molecules within the microbial community, suggesting the importance of molecules in the mediation of microbial-host interactions (Franzosa et al., 2015). Ultimately, there is no one tool that can completely describe both the taxonomy and function of the microbial community. Therefore, integration of multiple bioinformatics tools provides the most reasonable description of the microbial community (Franzosa et al., 2015).

1.4. Conclusions and future perspectives

The genomes of many novel foods have not been fully characterized. Therefore, combination of genomic, transcriptomic, and proteomic techniques may help the regulatory agencies, food developers and allergic patients to evaluate potential allergy risks. Bioinformatics approach of comparing the genomes and proteomes of novel foods to proven and putative allergens may provide critical evaluations regarding false positive matches. Genomics and transcriptomics might help in building protein databases for novel foods, that can be used for validating proteomic data generated from mass spectroscopy.

Similarly, using multiple bioinformatics tools may provide better understanding of rumen microbiome structure, abundance, and function, and their impacts on cattle performance and methane emissions. Both 16S sequencing and metagenomics will be used to explore the effects of diet on the ruminal microbiome composition and functionality. Understanding the ecological and mechanistic insights of dietary interventions, microbes and methane production may help improving livestock industry.

Overall, bioinformatics tools will be addressed to answer two main questions. The first question is the potential allergy risks for consumption of novel foods including microalgae, fungi, insects, and GM organisms. The second question is the effects of dietary interventions on microbiome functionality related to methane production in ruminants.

1.5. References

Aalberse, R.C., Kleine Budde, I., Stapel, S.O., van Ree, R., 2001. Structural aspects of cross-reactivity and its relation to antibody affinity. Allergy 56 Suppl 67:27-29.

Aalberse. R.C., 2000. Structural biology of allergens. J. Allergy Clin. Immunol. 106:228-238.

Allen, K.J., Turner, P.J., Pawankar, R., et al, 2014. Precautionary labelling of foods for allergen content: are we ready for a global framework. World Allergy Organ J. 7(1):10.

Altelaar, A.F., Munoz, J., Heck, A.J., 2013. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet. 14(1):35–48.

Astwood, J.D., Fuchs, R.L., 1996. Allergenicity of foods derived from transgenic plants. Monogr Allergy. 32:105–120.

Astwood, J.D., Leach, J.N., Fuchs R.L., 1996. Stability of food allergens to digestion in vitro. Nat Biotechnol. 14(10):1269–1273.

Beauchemin, K.A., McGinn, S.M., Martinez, T.F., McAllister, T.A., 2007a. Use of condensed tannin extract from quebracho trees to reduce methane emissions from cattle. Journal of Animal Science 85:1900–1906.

Beauchemin, K.A., McGinn, S.M., Petit, H., 2007b. Methane abatement strategies for cattle: lipid supplementation of diets. Canadian Journal of Animal Science 87:431–440.

Boadi, D., Benchaar, C., Chiquette, J., Massé, D., 2004. Mitigation strategies to reduce enteric methane emissions from dairy cows: Update review. Canadian Journal of Animal Science 84(3): 319-335

Bohle, B., Zwolfer, B, Heratizadeh, A., Jahn-Schmid, B., Antonia, Y.D., et al, 2006. Cooking birch pollen-related food: divergent consequences for IgE- and T cell-mediated reactivity in vitro and in vivo. J. Allergy Clin. Immunol.118(1):242-249.

Broekman, H.C., Knulst, A.C., de Jong, G., Gaspari, M., et al, Is mealworm or shrimp allergy indicative for food allergy to insects? 2006, Mol Nutr Food Res. 61(9).

Bublin, M., Breiteneder, H., 2014. Cross-reactivity of peanut allergens. Curr Allergy Asthma Rep. 14(4):426.

Buddle, B.M., Denis, M., Attwood, G.T., et al, 2011. Strategies to reduce methane emissions from farmed ruminants grazing on pasture. Vet J. 188(1):11–17.

CODEX Alimentarius Commission. Foods derived from modern biotechnology 2009. 2nd Edition, World Health Organization, Food and Agricultural Organization of the United Nations. Rome, Italy.

Cressman, R.F., Ladics, G., 2009. Further evaluation of the utility of “sliding window” FASTA in predicting cross-reactivity with allergenic proteins. Regul. Toxicol. Pharmacol. 54(3 Suppl):S20-25.

Culley, F.J., Pennycook, A.M., Tregoning, J.S., Hussell, T., Openshaw, P.J., 2006. Differential chemokine expression following respiratory virus infection reflects Th1- or Th2-biased immunopathology. J Virol. 80:4521–4527.

Cunha, L.M., Gonc¸alves, A.T. S., Varela, P., Hersleth, M. et al, 2015. Adoption of insects as a source for food and feed production: a cross-cultural study on determinants of acceptance. "11th Pangborn Sensory Science Symposium", Gothenburg, Sweden, 2015.

Cunha, L.M., Moura, A.P., Costa-Lima, R., 2014. Consumers’ associations with insects in the context of food consumption: comparisons from acceptors to disgusted. “Insects to Feed the World”, Netherlands.

Danielsson, R., 2016. Methane production in dairy cows. P. 45.

Danielsson, R., A. Schnurer, V. Arthurson, Bertilsson J., 2012. Methanogenic population and CH4 production in Swedish dairy cows fed different levels of forage. Appl Environ Microbiol. 78:6172–6179.

Duval, S., Kindermann, M., 2012. Use of nitooxy organic molecules in feed for reducing enteric methane emissions in ruminants, and/or to improve ruminant performance. International Patent Application WO 2012/084629 A1. World Intellectual Property Organization, Geneva, Switzerland

EFSA, 2015. Risk profile related to production and consumption of insects as food and feed. EFSA J. 13: 4257–4317.

Faber, M.A., Pascal, M., El Karbouchi, O., Sabato, V., Hagendorens, M.M., Decuyper, I.I., Bridts, C.H., Ebo, D.G., 2017. Shellfish allergens: tropomyosin and beyond. Allergy 72:842-848.

Fernando, S.C., Purvis, H.T., Najar, F.Z., et al, 2010. Rumen microbial population dynamics during adaptation to a high-grain diet. Appl Environ Microbiol. 76(22):7482–7490.

Finnigan, T.J.A., Wall, B.T., Wilde, P.J., Stephens, F.B., Taylor, S.L., Freedman, M.R, 2019. Mycoprotein: The future of nutritious nonmeat protein, a symposium review., Curr. Dev. Nutr. 3(6), nzz021.

Franzosa, E.A., Huang, K., James F. Meadow, J.F., 2015. Dirk Gevers, Katherine P. Lemon, Brendan J. M. Bohannan, Curtis Huttenhower Identifying personal microbiomes using metagenomic codes. PNAS 112(22):E2930-E2938

Giannoukos, G., Ciulla, D.M., Huang, K. et al, 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. 13:r23.

Goodman, R.E., 2008b. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. Food Chem. Toxicol. 46:24-34.

Goodman, R.E., Ebisawa, M., Ferreira, F., Sampson, H.A., van Ree, R., Vieths, S., Baumert, J.L., Bohle, B., Lalithambika, S., Wise, J., Taylor, S.L., 2016. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. Mol. Nutr. Food Res., 60, 1183-1198.

Goodman, R.E., Hefle, S.L., 2005. Gaining perspective on the allergenicity assessment of genetically modified food crops. Expert Rev. Clin. Immunol. 1(4):561-578.

Goodman, R.E., Hefle, S.L., Taylor, S.L., van Ree, R., 2005. Assessing genetically modified crops to minimize the risk of increased food allergy: a review. Int. Arch. Allergy Immunol. 137(2):153-166.

Goodman, R.E., Vieths, S., Sampson, H.A., Hill, D., Ebisawa, M., Taylor, S.L. van Ree, R., 2008a. Allergenicity assessment of genetically modified crops—what makes sense? Nat Biotechnol. 26(1):73-81.

Hales, K.E., N.A. Cole, MacDonald, J.C., 2012. Effects of corn processing method and dietary inclusion of wet distillers grains with solubles on energy metabolism, carbon−nitrogen balance, and methane emissions of cattle. J. Anim. Sci. 90:3174–3185.

Hall, F.G., Jones, O.G., O'Haire, M.E., & Liceaga, A.M., 2017. Functional properties of tropical banded cricket (Gryllodes sigillatus) protein hydrolysates. Food Chem. 224:414–422.

Hall, F., Johnson, P.E., Liceaga, A., 2018. Effect of enzymatic hydrolysis on bioactive properties and allergenicity of cricket (Gryllodes sigillatus) protein. Food Chem. 262:39-47.

Henderson, G., Cox, F., Ganesh, S. et al, 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep 5:14567.

Henikoff, J.G., Henikoff S., 1996. Blocks database and its applications. 1996. Methods Enzymol. 266:88-105.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89 (22):10915 -10919.

Hileman, R.E., Silvanovich, A., Goodman, R.E., Rice, E.A., Holleschak, G., Astwood, J.D., Hefle, S.L, 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. Int. Arch. Allergy Immunol. 128(4):280-291.

Hristov, A.N., J.L. Firkins, J. Dijkstra, E. Kebreab, G. Waghorn, H. P. Makkar, A. T. Adesogan, W. Yang, C. Lee, P. J. Gerber, B. Henderson, and J. M. Tricarico. 2013. Special Topics—Mitigation of methane and nitrous oxide emissions from animal operations: I. A review of enteric methane mitigation options. J. Anim. Sci. 91:5045–5069.

Hristov, A.N., J.L. Firkins, J. Dijkstra, E. Kebreab, G. Waghorn, H. P. Makkar, A. T. Adesogan, W. Yang, C. Lee, P. J. Gerber, B. Henderson, and J. M. Tricarico. 2013. Special Topics—Mitigation of methane and nitrous oxide emissions from animal operations: I. A review of enteric methane mitigation options. J. Anim. Sci. 91:5045–5069.

Johnson, K.A., and D.E. Johnson. 1995. Methane emissions form cattle. J. Anim. Sci. 73:2483–2492.

Jovel, J., Patterson, J., Wang, W., et al, 2016. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Front Microbiol. 7:459.

Khiaosa-Ard, R., Metzler-Zebeli, B.U., Ahmed, S., et al, 2015. Fortification of dried distillers grains plus solubles with grape seed meal in the diet modulates methane mitigation and rumen microbiota in Rusitec. J Dairy Sci 98(4):2611–2626.

Kim, B., Jeon, Y., Chun, J., 2013. Current Status and Future Promise of the Human Microbiome. Pediatric Gastroenterology, Hepatology & Nutrition 16(2):71-79

Kittelmann, S., Pinares-Patino, C.S., Seedorf, H., et al, 2014. Two different bacterial community types are linked with the low-methane emission trait in sheep. Plos One 9(7):e103171.

Knoell, A. L., 2016. The effect of diet on the bovine rumen microbial community structure and composition and its effects on methane production in growing and finishing cattle [thesis). University of Nebraska-Lincoln.

Ladics, G.S., Bannon, G.A., Silvanovich, A., Cressman, R.F., 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA

search for the elucidation of potential identities to known allergens. Mol. Nutri. Food Res. 51(8):985-998.

McAllister, T.A., Newbold, C.J., 2008. Redirecting rumen fermentation to reduce methanogenesis. Australian Journal of Experimental Agriculture 48:7-13.

McAllister, T.A., Okine, E.K., Mathison G.W., Cheng, K.G. 1996. Dietary, environmental and microbiological aspects of methane production in ruminants. Can. J. Anim. Sci. 76:231- 243.

Metcalfe, D.D., Astwood, J.D., Townsend, R., Sampson, H.A., Taylor, S.L., Fuchs, R.L., 1996. Assessment of the allergenic potential of foods derived from genetically engineered crop plants. Crit. Rev. Food Sci. Nutr. 36, Supplement S165-186.

Morgavi, D. P., C. Martin, J.P. Jouany, and M. J. Ranilla, 2012. Rumen protozoa and methanogenesis: not a simple cause-effect relationship. British Journal of Nutrition 107:388-397.

Myhre, G., Shindell, D., Bréon, F.M., Collins, W., et al, 2013. Anthropogenic and Natural Radiative Forcing. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Palmer, L.K., Marsh, J.T., Lu, M., Goodman, R.E., Zeece, M.G., Johnson, P.E., 2020. Shellfish Tropomyosin IgE Cross-Reactivity Differs Among Edible Insect Species. Mol Nutr Food Res. e1900923.

Payne, C.L.R., Scarborough, P., Rayner, M., Nonaka, K., 2016. A systematic review of nutrient composition data available for twelve commercially available edible insects, and comparison with reference values. Trends. Food Sci. Technol. 47:69–77.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132: 185-219.

Pearson, W.R., 2014. BLAST and FASTA similarity searching for multiple sequence alignment. Methods Mol. Biol. 1079:75-101.

Pearson, W.R., 2016. Finding protein and nucleotide similarities with FASTA. Curr. Protoc. Bioinformatics. 53:3.9.1-25.

Pesta, A. C., 2015. Dietary strategies for the mitigation of methane production by growing and finishing cattle. Ph.D. Diss. Univ. of Nebraska-Lincoln.

Poulsen, M., Schwab, C., Jensen, B.B., et al. 2013. Methylotrophic methanogenic Thermoplasmata implicated in reduced methane emissions from bovine rumen [published correction appears in Nat Commun. 2013,4:1947). Nat Commun.,4:1428.

Radauer, C., Breiteneder, H., 2007. Evolutionary biology of plant food allergens. J. Allergy Clin. Immunol. 120(3):518-525.

Radauer, C., Breiteneder, H., 2006.
Pollen allergens are restricted to few protein families and
show distinct patterns of speciesdistribution. J. Allergy Clin. Immunol. 117(1):141-7.

Ruethers, T., Taki, A.C., Johnston, E.B., Nugraha, R., Le, T.T.K., Kalic, T., McLen, T.R., Kamath, S.D., Lopata, A.L., 2018. Seafood allergy: A comprehensive review of fish and shellfish allergens. Mol. Immunol. 100:28-57.

Ruethers, T., Taki, A.C., Johnston, E.B., Nugraha, R., Le, T.T.K., Kalic, T., McLen, T.R., Kamath, S.D., Lopata, A.L., 2018. Seafood allergy: A comprehensive review of fish and shellfish allergens. Mol. Immunol. 100:28-57.

Ruiz-Carnicer Á, Comino I., Segura V., et al, 2019. Celiac Immunogenic Potential of α-Gliadin Epitope Variants from Triticum and Aegilops Species. Nutrients.,11(2):220.

Rumpold, B.A., Schluter, O.K.,2013. Nutritional composition and ¨ safety aspects of edible insects. Mol. Nutr. Food Res. 57:802–823.

Russell, J. B., 2002. Rumen Microbiology and Its Role in Ruminant Nutrition. Published by James B. Russell, Ithaca, NY.

Scala, E., Villalta, D., Meneguzzi, M., Giani, M., Asero, R. An atlas of IgE sensitization patterns in different Italian areas., 2018. A multicentre cross-sectional study. Eur. Ann. Allergy Clin. Immunol. 50(4): 148-155.

Scala, E., Villalta, D., Meneguzzi, M., Giani, M., Asero, R., 2018. An atlas of IgE sensitization patterns in different Italian areas. A multicentre cross-sectional study. Eur. Ann. Allergy Clin. Immunol. 50(4): 148-155.

Schelling, G.T., 1984. Monensin mode of action in the rumen. J Anim Sci.58(6):1518–1527.

Schonknecht, G., Chen, W.H., Ternes, C.M., et al. 2013. Gene transfer from Bacteria and Archaea facilitated evolution of an extremophilic eukaryote, Science 339:1207-1210.

Shabat, S.K., Sasson, G., Doron-Faigenboim, A., et al, 2016. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. The ISME journal 10(12):2958–2972.

Shah, A., Mahmood, F., Maroof, Q., et al, 2014. Microbial ecology of anaerobic digesters: the key players of anaerobiosis. The Scientific World Journal P. 183752.

Shi, W.B., Moon, C.D., Leahy, S.C., et al, 2014. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res. 24:1517–25.

Silvanovich, A., Bannon, G., McClain, S., 2009 The use of E-scores to determine the quality of protein alignments. Regul. Toxicol. Pharmacol. 54(3 Suppl):S26-31.

Sollid, L.M., Qiao, S., Anderson, R.P., Gianfrani, C., Koning, F., 2012. F.Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. Immunogenetics 64:455–460

Tapio, I., T. J. Snelling, F. Strozzi, R. J. Wallace, R. J., 2017. The ruminal microbiome associated with methane emissions from ruminant livestock. Journal of animal science and biotechnology. 8:7.

Taylor, S.L., Hefle, S.L., 2006. Food allergen labeling in the USA and Europe. Curr Opin Allergy Clin Immunol. 6(3):186–190.

Thomas, K., Bannon, G., Hefle, S., Herouet, C., Holsapple, M., Ladics, G., Macintosh, S., Privalle, L., 2005. In silico methods for evaluating human allergenicity to novel proteins: International bioinformatics workshop meeting report, 23-24 February. Toxicol. Sci. 88(2):307-310.

Troy, S., Duthie, C.A., Hyslop, J., et al, 2015. Effectiveness of nitrate addition and increased oil content as methane mitigation strategies for beef cattle fed two contrasting basal diets. Journal of Animal Science 93(4):1815-23.

United States Environmental Protection Agency (EPA), 2020. Draft Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2018. Complete Report. Pages 719.

Siruguri, V., Bharatraj, D.K, Vankudavath, R.N., Mendu, V.V., Gupta, V., Goodman, R.E., 2015. Evaluation of Bar, Barnase, and Barstar recombinant proteins expressed in genetically engineered Brassica juncea (Indian mustard) for potential risks of food allergy using bioinformatics and literature searches. Food Chem. Toxicol. 83: 93-102.

van der Spiegel, M., Noordam, M.Y., van der Fels-Klerx, H.J., 2013. Safety of novel protein sources (insects, microalgae, seaweed, duckweed, and rapeseed) and legislative aspects for their application in food and feed production. Compr. Rev. Food. Sci. Food Saf., 12:662–678.

van Huis, A., 2016. Edible insects are the future? Proc. Nutr. Soc. 75:294–305.

van Huis, A., Itterbeeck, J.V., Klunder, H., Mertens, E. et al, 2013. Edible Insects: Future Prospects for Food and Feed Security, Food and Agriculture Organization of the United Nations (FAO), Rome.

van Putten M.C., Frewer, L.J., Gilissen, L.J.W.J., Gremmen, B, Peijnenburg, A.A.C.M., Wichers, H.J., 2006. Novel foods and food allergies: A review of the issues. Trends Food Sci. & Technol. 17:289-299.

Van Zijderveld, S.M., Gerrits, J.J., Apajalahti, J.A., et al, 2010. Nitrate and sulfate: Effective alternative hydrogen sinks for mitigation of ruminal methane production in sheep. J. Dairy Sci. 93:5856-5866.

Wells, M.L., Potin, P., Craigie, J.S., Raven, J.A. et al. 2017. Algae as nutritional and functional food sources: revisiting our understanding. J Appl Phycol. 29(2):949–982.

Zhang, T., Breitbart, M., Lee, W.H., Run, J.Q., Wei, C.L., et al, 2006. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 2006:4:e3.

Zhou, M., Chung, Y.H., Beauchemin, K.A., et al, 2011. Relationship between rumen methanogens and methane production in dairy cows fed diets supplemented with a feed enzyme additive. J Appl Microbiol. 111:1148–58.

CHAPTER 2

EVALUATING POTENTIAL RISKS OF FOOD ALLERGY OF NOVEL FOOD SOURCES BASED ON COMPARISON OF PROTEINS PREDICTED FROM GENOMES AND COMPARED TO WWW.ALLERGENONLINE.ORG

This chapter is in progress to be submitted for peer review: Mohamed Abdelmoteleb, Chi Zhang, Brian Furey, Mark Kozubal, Marion Champeaud, Richard E. Goodman

2.1. Abstract

Potential proteins from three new novel food sources (*Chlorella variabilis*, *Galdieria sulphuraria*, and a *Fusarium* sp.) were evaluated for potential allergic cross-reactivity by comparing the predicted amino acid sequences from their genomes against the allergens in the www.AllergenOnline.org (AOL) database using CODEX limits of >35% identity over 80 amino acids. The results contain matches to hundreds of highly conserved proteins that would trigger serum IgE testing if the proteins were in GE crops. To address the inequality of extensively conserved sequences, predicted proteins from curated genomes of 23 highly diverse species of animals including humans, plants and arthropods were compared to AOL sequences. The compiled identities of this extensive data collection were used to critically evaluate the CODEX identity limits and *E*-scores, with comparison to documented cases of cross-reactivity. Many allergens are defined by IgE binding alone without consideration regarding elicitation of allergic reactions or abundance in the sources. Proteins that are highly conserved across diverse taxa are unlikely to pose risks of clinical cross-reactivity. These results provide essential data for redefining allergens in AOL and for providing guidance on more flexible sequence identity matches for risk assessment.

2.2. Introduction

There is a future expectation toward growing demand for novel foods worldwide to feed increased human populations. While considering novel food sources, it is imperative to evaluate possible risks, including their allergenic potential (Goodman et al, 2016). Three organisms *Chlorella variabilis, Galdieria sulphuraria*, and a *Fusarium* sp. are being developed as single-cell novel protein sources. *Chlorella* is a genus of single-celled green algae which contains high concentrations of protein (51%–60% of dry matter), amino acids, vitamins, dietary fiber, and a variety of antioxidants, bioactive materials, and chlorophylls. Green algae have a history of sustainable production and consumption. (Klamcyzynska and Mooney, 2017). *Chlorella vulgaris* and *Chlorella pyrenoidosa* are not considered novel in the EU since they have been historically consumed by humans (Regulation EC No. 258/97). In the United States they are recognized as GRAS by the FDA as an algae commonly consumed in foods in many countries (Wells et al. 2017). Recently the genome of *Chlorella variabilis*, NC64A was completed and was used here as a model genome (Blanc et al, 2010). The unicellular red algae, *Galdieria sulphuraria*, isolated from extreme environments (from pH 0 to 4, and up to 56 °C) is being proposed as an edible alga with a high content of protein and other important dietary nutrients. It can be grown via fermentation and is being developed for use in food products (Schonknecht et al, 2013), but has not yet been commonly consumed by humans.

The *Fusarium* genus of fungus is already used in several food products with the brand name, Quorn. Quorn is produced and marketed as a human food by Marlow Foods, Ltd. Quorn contain mycoprotein which is derived from *Fusarium venenatum*, grown by

fermentation (Finnigan et al, 2019). Other strains of *Fusarium* with differing compositions are now under development. Products of Quorn have been consumed as a non-meat protein source in the United Kingdom for 30 years and since 2002 in the US. There are a few case reports of food allergy to Quorn (Katona and Kaminski, 2002; Hoff et al, 2002). Some IgE binding is likely due to allergy to inhalation allergens of *Fusarium sp*. (Weber and Levetin, 2014). Some consumers of Quorn have experienced transient GI symptoms without IgE antibody production and a very small number have experienced possible IgE mediated food allergic reactions including one reported fatal reaction (Tee et al, 1993; Hoff et al, 2003a, 2003b; Yeh et al, 2016; Jacobson and DePorter, 2018). However, many common food sources have caused at least one fatal food allergic reaction. As long as ingredients in foods are clearly labeled, consumers who are aware of their allergies can avoid consumption and reactions (Ramsey et al, 2019; Gowland and Walker, 2015). Recently international governmental regulators are considering updating and modernizing the safety evaluation processes to improve evaluations and improve the timeliness and accuracy of decisions (Slikker et al, 2018; Elles et al, 2019).

The primary health concern is whether the new food represents a risk as an allergenic source for at least a proportion of the general population, primarily those allergic to similar proteins. Based on our years of use and development of AllergenOnline.org, it appears that the CODEX guidelines are far too conservative to judge proteins that match evolutionarily conserved allergens, especially when applied to whole genomes. We have performed this study in part to understand the extent of over-predictions.

Our hypothesis is that bioinformatics approach of comparing the genomes of novel foods to AOL does identify matches to extensively conserved sequences in different allergenic and non-allergenic sources which must be critically evaluated before a conclusion of a risk of allergenicity can be drawn. There are three objectives in this study. First, to identify proteins from the genome of three species that might represent a risk of allergy based on comparison of the predicted proteins against allergens in the AllergenOnline.org database using the CODEX criteria of >35% identity over 80 amino acids. Second, to consider the inequality of extensively conserved sequences using predicted proteins from the genomes of 23 highly diverse allergenic and non-allergenic species in the same comparison recording the functionality and abundance where possible to consider possible risks. The third objective is to critically evaluate the limits of CODEX guidelines if used as a whole genome analysis. Can the CODEX criteria be modified to be more predictive of risk?

2.3. Materials and methods

2.3.1. Preparation of protein sequences of the three targeted genomes

The predicted proteins for the genomes of *Chlorella variabilis* NC64A and the genomes of 23 highly diverse species have been downloaded from different databases including the NCBI genome library (https://www.ncbi.nlm.nih.gov/genome), EnsemblPlants (http://plants.ensembl.org/index.html), and Phytozome V. 12, the Plant Genomics Resource (https://phytozome.jgi.doe.gov/pz/portal.html#) which summarized in Table 1. For species without published genomes as of October 2018, we downloaded all predicted protein sequences from the NCBI protein library. The bioinformatics

pipeline has been completed using a lab cluster on the Holland Computer Center server at the University of Nebraska.

2.3.2. Prediction of new *Galdieria* sp. and *Fusarium* sp. proteins based on genomic DNA sequences

For *Galdieria* sp, the company Fermentalg provided the DNA sequences which were identified using Illumina sequencing (2x150 bp reads). The sequencing quality was checked using FastQC (Andrews 2010) and cleaned using PRINSEQ (prinseq.sourceforge.net) by trimming of bases with low quality scores. Two assemblers were used; SPAdes with 21, 33, 55 and 77 k-mer values (Bankevich et al, 2012), and Trinity using 25 k-mer (https://github.com/trinityrnaseq/trinityrnaseq/wiki). Post assembly polishing was performed using Pilon (Walker et al, 2014). The quality of assembly was checked using Quast (Gurevich et al, 2013). The percent of mapping was evaluated using BWA mapper (Li and Durbin, 2009). Genes were predicted using the *Galdieria* model from AUGUSTUS (Stanke and Morgenstern, 2005). Potential tRNA were predicted using tRNAscan-SE (Lowe and Chan, 2016) and rRNA were predicted using barrnap (https://github.com/tseemann/barrnap). Functional annotation was conducted by a combination of AUGUSTUS software and BLASTP comparison for the predicted proteins against the published *Galdieria sulphuraria* genome (https://www.ncbi.nlm.nih.gov/genome/?term=Galdieria+sulphuraria) from the NCBI library. Sequences were compiled into FASTA files for comparison to the AllergenOnline.org database. The published *Galdieria sulphuraria* genomic sequences of Schonknecht *et al*, (2013) for genome ASM34128v1 were checked.

Sustainable Bioproducts, Inc. provided the genomic sequences for the *Fusarium sp.* they are proposing to use as a food product. They performed genomic sequencing using Pacbio (for long-reads) and Illumina (2x250 bp reads) for short, high quality reads of the cultured species. These sequences were compiled and evaluated for the highest accuracy and completeness. The reads were evaluated using FASTQC. Sequences were compiled using assemblers MaSuRCA with 22 k-mer value (Zimin et al, 2013) and SPAdes (Bankevich et al, 2012) used K-mers of 21, 33, 55, 77, 99 and 127. Post assembly polishing was performed using Pilon (Walker et al, 2014). Pacbio reads were mapped using Minimap2 (Li 2016), and illumina reads were mapped using Bowtie2 (Langmead and Salzberg, 2012). Genes were predicted using the *Fusarium* model (King et al, 2015) from AUGUSTUS (Stanke and Morgenstern, 2005), mitochondrial genes were predicted using Prodigal (Hyatt et al, 2010), tRNA were predicted using tRNA scan-SE (Lowe and Chan, 2016) and rRNA were predicted using Barrnap software (https://github.com/tseeman/barrnap/). Functional annotation was accomplished using the ERGO software package of IgenBio. The overall sequence completeness was further evaluated by comparison to the genomes of strains of *Fusarium* which had been previously characterized to provide a framework for understanding completeness (Niehaus et al, 2016).

2.3.3. FASTA comparison for the predicted protein sequences to Allergenonline.org version 16, 18B, and 19

The predicted protein sequences from the three novel foods and 23 different species were compared to allergens in version 16, 18B and 19 of www.AllergenOnline.org by overall FASTA 35. FASTA version 35 was installed on the Holland Computing

Center server to allow batch searches that mimic the individual protein searches available on the AllergenOnline website. Different *E*-score thresholds (10, 1, 0.001, 10e-7, 10e-30, 10e-50, 10e-75, 10e-100) were used to check the significance of matches. The same scoring matrix was used (BLOSUM 50) that is used on the public AllergenOnline.org website. Sequence matches to proteins in AllergenOnline.org were compiled in an Excel worksheet with a record of the highest match identity for each predicted protein. The matches were reviewed to identify those >35% identity over 80 or more amino acid segments.

2.3.4. BLASTP comparison of predicted protein sequences to the non-redundant NCBI Protein database that is a compilation of sequences from GenBank, RefSeq, TPA, SwissProt, PIR, PRF and PDB

Predicted protein sequences of *Galdieria sulphura*ria, *Fusarium* sp. and *Chlorella variabilis* as well as the 23 other species used in this study (Table 1) were used to search the general protein database using the current version of BLASTP in 2018 and early in 2019. The website is https://blast.ncbi.nlm.nih.gov/BLAST.cgi. The current version of BLASTP outputs have changed markedly (July 2019). Searches without keyword limits allows the highest identity matches to be viewed for evaluation of the common conservation of the protein sequences. The previous selection criteria using keyword limits such as “allergy” or “allergen” has been removed. That change speeds the search but eliminates interesting screening options. Our use of BLASTP of species targets from the 23 species and of the matched allergens from out AllergenOnline.org provides guidance on the relevance of low-identity matches including >35% identity over 80 amino acids.

2.4. Results

2.4.1. Prediction of *Galdieria* sp. and *Fusarium* sp. proteins based on genomic DNA sequences

For *Galdieria* sp., the number of reads after checking quality and trimming are 26.4M reads from Illumina. Assembly metrics are: 1998 contigs, largest contig 294001B, N50 54420B, N75 16958B, L50 66, L75 164, and GC% 40.27 for SPAdes; and for Trinity are 2890 contigs, largest contig 154130B, N50 24717B, N75 11797B, L50 292, L75 677, and GC% 40.30. The reads were mapped at 99.67% for SPAdes, and 99.59% for Trinity. The number of predicted proteins for *Galdieria sulphuraria* was 5701 from SPAdes and 11976 from Trinity.

For *Fusarium* sp., the quality of the sequences included 340k reads after trimming and correcting from Pacbio, and 56.5 M reads from Illumina. Assembled sequences included 89 contigs, with the largest contig being 4.9 MB, N50 for 3.2 MB, N75 for 2.3 MB and L50 6, L75 10 and 0 Ns with a GC content of 48.3%. Pacbio reads mapped at 99.95% using Minimap2 software. Illumina reads mapped at 99.81% using Bowtie2 software. The number of predicted proteins were 14239.

1.2.2. Comparison of all possible proteins from the genome of the three novel foods against AOL

Total number of matches and unique matches (beyond the limits of CODEX guidelines) resulting from FASTA comparison of all predicted proteins from the genomes of the three novel foods to AOL sequences have been summarized in Table 2. The resulting matches of the other 23 species are summarized in Table 3. A significant decrease in the number of matches that exceed the criteria of >35% identity over 80 AA

was recorded by reducing the size of the *E*-score threshold in different species. The normal default *E*-score for FASTA or for BLAST is 10, but 1 provides more strict alignments. However, the purpose of comparisons made for food safety is to identify proteins with high fidelity alignments that might indicate high evolutionary conservation to predict possible shared IgE antibody binding and clinical allergy based on shared IgE binding. Experiences in clinical allergy at many clinics demonstrate that matches of protein sequences using relatively large *E*-scores results in extreme over-prediction of possible cross-reactivity. As shown in Table 2, the three species of interest that have rarely been reported to cause allergies show high numbers of protein matches greater than 35% identity over 80 AA at 1.00e-07.

As a comparator, we tested all predicted proteins from the genomes of 23 species ranging from humans to fungi to fish and many species of plants. There are rare to common reports of allergy to some of these species (e.g. *Candida albicans and Arabidopsis thaliana*), whilst others are clearly sources of allergic reaction (e.g. peanut and soybean). Our intent was to identify an *E*-score limit that would likely represent proteins of likely risk for cross-reactivity using FASTA alignments and the CODEX limit of >35% identity over 80 AA.

2.4.3. Identities of all possible proteins from the genome of the three novel food sources and 23 common species matches to AOL

The results in Table 2 illustrate that the algae (*Chlorella variabilis* NC64A) has sequence matches to between 14 and 991 proteins in AOL, depending on which *E*-score limit was used. Even at the moderate *E*-score of 1e-07, there were 159 proteins that suggest potential cross-reactivity. A similar trend is seen when comparing all predicted

proteins from the 23 diverse species of organisms as shown in Table 3; there were high numbers of potential matches to allergens in most of the species. Pistachio had the lowest number, but few total proteins were predicted from nucleotide sequences for pistachio and for pecan (Table 3).

To examine these matches further, the highest scoring aligned proteins of *Chlorella variabilis* were compared to all proteins in AOL version 18B as shown in Table 4. The highest scoring allergen match was to cyclophilin of *Daucus carota*. However, cyclophilin is highly conserved and matches homologous proteins in 20 species out of the 23 studied species over CODEX limits. Likewise, heat shock protein 70 of the *Aedes aegypti* mosquito is highly conserved and showed sequence matches to proteins in 22 species. Most of the matched allergens are conserved across many species of the 23 chosen here. Many of the high scoring matches are to house-keeping proteins including cyclophilins, heat shock proteins, 60S ribosomal protein, triosephosphate isomerase, aldolase, gliadins. However, the percent identities are not high compared to BLASTP matches to homologues from a variety of protein sources and from species that are not likely to represent risks. There are a few examples in Table 4 where sequence matches have been found to bona fide allergens that are not matched widely amongst the 23 species. These are, however, generally amongst the lowest scoring matches in Table 4 with identity matches closer to the bottom range 35% identity, and with modest *E*-scores. Those include matches to thioredoxin of fungi at 39-40% identity and venom allergen 5 of a wasp at 35.8% identity.

Similarly, Table 5 illustrates that *Galdieria* sp. had matches to 59 weak or putative allergens and 6 very low scoring matches to food allergens (tropomyosin, vicilin,

and convicilin) with an *E*-score >0.02. Due to high sequence identity of evolutionary homologues, it was clearly overly predictive for possible risks of allergic cross-reactivity. The searches were rerun using an *E*-score of 1e-7 which resulted in the removal of proteins that are clearly unlikely to cause cross reactivity. The results are shown in Table 5. As with those from *Chlorella*, the identified allergens, which represent important protein classes of allergens, are either highly conserved, or have identity matches that show very low identities of proteins meaning they are unlikely to be a significant risk for cross-reactivity. This was demonstrated by further comparing the matched allergens to the NCBI Protein database using BLASTP.

The predicted proteins of the Quorn fungal genome-predicted proteome, another species of *Fusarium*, was tested as for background evaluation in a similar manner. The results are shown in the supplementary material (APPENDIX I). We found 181 matches to weak or putative allergens and 12 low scoring matches to food allergens (e.g. tropomyosin, glycinin, vicilin, and convicilin) with very low sequence identity over short AA segments. As with *Chlorella* and *Galdieria*, these could be classed as either being part of a highly conserved protein family, or having limited identity leading to the conclusion that they do not pose a significant risk of cross-reactivity. Products from Quorn have been safely consumed for over 30 years with very few clear cases of IgE mediated allergy.

2.4.4. Summary Examples of FASTA comparisons using all predicted proteins from the 23 studied species

Initially, the term of "major" allergen has been used to represent food allergens that are thought to cause severe clinically important reactions in the big 8 major allergens (e.g.

LTPs, vicilins, glycinins, tropomyosins, arginine kinases, and 2S albumins) (Jonhson et al, 2016). Other allergens of less commonly reported clinical reactions are denoted as a "minor" allergen. Predicted proteins from all 23 species were compared to AllergenOnline.org using an *E*-score cutoff of 10e-07. Wheat proteins matched 312 putative allergens, but only one major allergen in eight different sources. Soybean proteins had matches to 243 putative allergens and 32 matches to major allergens (vicilins and conglycinins of soybean, walnut, pecan and pistachio). Human proteins had 206 matches to weak or putative allergens and only one matched to a clinically relevant allergen, lipid transfer protein, though that was a modest identity match to LTP from pomegranate (42.3% identity with an *E*-score of 3.7e-19). Searching AllergenOnline.org with the pomegranate LTP shows many higher identity matches, often >55% ID with E-scores of smaller than 1e-20 to 1.1e-25. A number of the LTPs have reported evidence of cross-reactive laboratory IgE binding, but there are few cases of multiple allergic reactions to diverse sources of LTPs. The literature search identified many proteins that are unlikely to represent a risk of cross-reactivity as the protein sequences are conserved across broad taxonomic categories with no history of cross-reactivity.

2.4.5. Evaluation of the limits of CODEX guidelines looking for matches of >35% identity.

2.4.5.1. Identification of known allergens in AllergenOnline.org database at specific *E*-score limits for significance

The best *E*-score threshold for identification of known allergens in AllergenOnline.org database has been characterized. Table 6 illustrates the identified allergens in different allergenic species at representative *E*-scores of 10e-7, 10e-30, and

10e-100. All known major and minor allergens in AllergenOnline.org database were detected in all 23 species using *E*-scores of 10, 1, 0.001, and 10e-7. However, literature search results show many of those matched proteins that are highly unlikely to represent risks of cross-reactivity. A few important allergens were missed in FASTA searches when the *E*-score is reduced below 10e-7.

2.4.5.2. Major allergens of higher risk of cross-reactivity

The distribution of clinically important allergens (lipid transfer proteins, vicillins, glycinins, 2S albumins, tropomyosin, and arginine kinase) in the 23 species is shown in Table 7. The number of matches to clinically important allergens was related to taxonomic relationships, as these major allergens are not highly conserved in sequence and structure across extensive evolutionary distances for example above the level of taxonomic order and certainly not above the level of class. Lipid transfer proteins, vicillins and glycinins are highly conserved in beans, soybeans, apple, peach, and papaya. Major allergens in crustacean shellfish include tropomyosins and arginine kinases are highly conserved in human, drosophila, bovine, salmon, and cod.

2.4.5.3. Minor allergens and noise of CODEX limits

In this section, the focus was on the putative or minor allergens of lower risk of cross-reactivity. In this study of 23 species and 3 novel foods, most of the potential minor allergens were identified with sequence identities of less than 50%. Whilst the CODEX recommendation is to use a threshold of 35%, we wished to investigate the impact of using a higher threshold and its ability to eliminate noise from the search results on whole genomes. Significant matches were found to 170 minor different allergens which are highly conserved in at least 10 different studied allergenic and non-allergenic species out

of the 23 species. Table 8 illustrates the list of minor allergens which are highly conserved between different species in this study

2.5. Conclusion

It is becoming more common to use a whole genome/proteome bioinformatic approach to identify potential proteins in a wide variety of species. Some regulatory agencies or risk assessment scientists have suggested using such predicted proteins against allergen databases to identify possible risks of food allergy. The CODEX guideline (>35% identity over 80 amino acids to any known allergen) has been in place since before 2003. The comparison to www.AllergenOnline.org was made available in 2005 to assess individual proteins. The guideline suggests that a positive identity match would require serum IgE binding with samples from subjects allergic to the matched allergen. Many more proteins are identified as allergens in 2019 compared to 2005 (2129 proteins from 284 species compared to 1189 proteins from 208 species listed in the History section of AllergenOnline.org). With the expanded use of predicted protein sequences from genomes, transcriptomes or proteomes, for predicting possible risk we wanted to test the method broadly looking for false and true positive matches.

To this end, predicted proteins from the genomes of 23 diverse highly allergenic and low- or non-allergenic species including plant sources, fungi, fish, insect and other animal sources as well as human sequences against the www.AllergenOnline.org database using standard CODEX criteria as well as full-FASTA alignments to provide identity matches. A wide variety of *E*-score criteria was used to assess the impact of this parameter on the ability to reduce false positives whilst avoiding false negatives. Many housekeeping proteins across many species had moderate to high identities to minor

putative allergens in AOL. However, many of these proteins are highly conserved in all eukaryotes and as a consequence would be expected to be found in a search using standard CODEX criteria. In contrast, major allergens are not highly conserved in sequence and structure and were not identified using the search parameters except in closely related species.

For those highly conserved proteins identified across many species, there are nonetheless differences in the levels of AA sequence identity conservation that impact their potential for shared clinical cross-reactivity. Moreover, differences in protein abundance and potency are significantly different between species, affecting the allergenic potential of the biomass. In applying higher identity matches than CODEX criteria to the searches, we were still able to successfully identify all know allergens in the trial group. In particular, decreasing the *E*-score threshold significantly reduced the number of false-positive hits. We propose that an *E*-score threshold of 10e-7 is the optimum for identification of important allergens in this type of study.

Considering these results, three predicted proteomes from three novel foods were assessed against the AOL database. As for the 23 test species, a number of highly conserved minor allergens were identified. It was therefore concluded that *Chlorella variabilis, Galdieria sulphuraria* and *Fusarium* sp. do not represent a significant risk of allergenicity to the general population.

This study demonstrates that the current bioinformatics guideline for evaluating potential risks of food allergy for novel proteins protects allergic consumers, but also has the potential to produce many false-positive matches. The CODEX criteria work fairly well for isolated proteins in GE organisms, but in some cases the matches are overly

conservative. Importantly, some proteins shorter than 80 AA with higher identity matches, especially >50% identity, are likely predictive for some potent allergens such as Ara h 2. The results demonstrated a real need for critical evaluation of the limits and cut-offs chosen for this type of assessment. These must be set sufficiently low to capture all potential allergens so as not to put consumers at risk, but not so low as to make the reasoned assessment of allergenic potential impossible.

There may also be additional assessments that can be made on the databases of allergens, rather than simply classifying all potentially allergenic sequences together as one group; we might for example need to rank allergens into major allergens and minor allergens according to their risk based on clinical findings. The level of conservation across species also needs to be taken into account. These housekeeping genes are usually non-allergenic, but in some specific cases can be minor allergens, but this does not mean that all similar proteins pose a risk of allergenicity. Tighter criteria or addition of steps to consider abundance and end uses could improve the risk assessment.

Table 1. Sources for Predicted Protein Sequences from Genomes of Different Species

Species	Source
Chlorella variabilis NC64A	https://www.ncbi.nlm.nih.gov/genome/?term=Chlorella+variabilis+%5Borgn%5D
Human (*Homo sapiens*)	ftp://ftp.ncbi.nih.gov/genomes/H sapiens/protein/
Baker's yeast (*Saccharomyces cerevisiae*)	http://downloads.yeastgenome.org/sequence/S288C reference/orf protein/
Candida albicans SC5314	http://www.candidagenome.org/download/sequence/C albicans SC5314/Assembly22/current/
Cod (*Gadus morhua*)	ftp://ftp.ensembl.org/pub/release-86/fasta/gadus morhua/pep/
Chicken (*Gallus gallus*)	ftp://ftp.ensembl.org/pub/release-86/fasta/gallus gallus/pep/
Bovine (*Bos taurus*)	ftp://ftp.ensembl.org/pub/release-86/fasta/bos taurus/pep/
Drosophila melanogaster	ftp://ftp.flybase.net/genomes/Drosophila melanogaster/dmel r6.09 FB2016 01/fasta/
Salmon (*Salmo salar*)	ftp://ftp.ncbi.nih.gov/genomes/Salmo salar/protein/
Papaya (*Carica papaya*)	ftp://ftp.ncbi.nih.gov/genomes/Carica papaya/protein/
Soybeans (*Glycine max*)	ftp://ftp.ncbi.nih.gov/genomes/Glycine max/protein/
Apple (*Malus domestica*)	ftp://ftp.ncbi.nih.gov/genomes/Malus domestica/protein/
Rice (*Oryza sativa*)	ftp://ftp.ncbi.nih.gov/genomes/Oryza sativa Japonica Group/protein/
Peanut (*Arachis hypogaea*)	ftp://ftp.ncbi.nih.gov/genomes/Arachis hypogaea/protein/
Peach (*Prunus persica*)	ftp://ftp.ncbi.nih.gov/genomes/Prunus persica/protein/
Beans (*Phaseolus vulgaris*)	https://www.ncbi.nlm.nih.gov/genome/?term=Phaseolus+vulgaris+%5Borgn%5D
Potato (*Solanum tuberosum*)	ftp://ftp.ncbi.nih.gov/genomes/Solanum tuberosum/protein/
Wheat (*Triticum aestivum*)	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org Taestivum er
Maize (*Zea mays*)	ftp://ftp.ncbi.nih.gov/genomes/Zea mays/protein/
Arabidopsis thaliana	https://www.ncbi.nlm.nih.gov/genome/?term=arabidopsis++thaliana+%5Borgn%5D
Almond (*Prunus dulcis*)*	https://www.ncbi.nlm.nih.gov/protein/?term=prunus+dulcis
Pecan (*Carya illinoinensis*)*	https://www.ncbi.nlm.nih.gov/protein/?term=carya+illinoinensis
Pistachio (*Pistacia vera*)*	https://www.ncbi.nlm.nih.gov/protein/?term=pistacia+vera+%5Borgn%5D
English Walnut (*Juglans regia*)	ftp://ftp.ncbi.nih.gov/genomes/Juglans regia/protein/

* Species without complete published genomes till October 2018

Table 2. Total Number of Matches and Unique Matches (>35% Sequence Identity over 80 AA Alignment Length) at Different E-Scores in The Three Novel Foods

Species	Subject Hits	10	1	0.001	1.00e-07	1.00e-30	1.00e-50	1.00e-75	1.00e-100
***Chlorella variabilis* NC64A**	**Total**	277988	82613	9043	**3201**	413	119	57	**35**
	Unique	991	752	297	**159**	64	39	21	**14**
Galdieria sp.	**Total**	67989	17792	3202	**1222**	170	97	50	**32**
	Unique	101	96	85	**73**	39	32	12	**8**
Fusarium sp.	**Total**	192772	65321	13320	**5867**	646	317	135	**88**
	Unique	508	466	326	**232**	125	95	44	**30**

Table 3. Total and Unique Matches for Predicted Proteins from 23 Different Allergenic and Non-Allergenic Species.

Species	Subject Hits	10	1	0.001	1.00e-07	1.00e-30	1.00e-50	1.00e-75	1.00e-100
Homo sapiens	**Total**	6200050	2460980	510958	175239	19346	7860	2516	1817
(Human)	**Unique**	14997	13534	8546	5565	2556	1538	912	557
Saccharomyces cerevisiae	**Total**	71691	24440	5320	2043	384	243	200	158
(Baker's yeast)	**Unique**	225	214	164	132	68	52	40	32
***Candida albicans* SC5314**	**Total**	185065	73070	18846	7712	599	292	174	140
(Yeast)	**Unique**	648	621	482	327	113	75	45	39
***Gadus morhua* (Cod)**	**Total**	339873	118495	24766	10932	1910	991	354	248
	Unique	850	806	638	502	268	182	108	72
***Bos Taurus* (Bovine)**	**Total**	431730	162370	33305	13860	2131	760	350	245
	Unique	1280	1190	865	680	356	227	125	71
Gallus gallus	**Total**	1067198	463688	112614	41907	6624	3397	865	450
(Chicken)	**Unique**	2964	2731	1798	1261	636	423	269	153
Drosophila melanogaster	**Total**	735514	325747	85437	35174	3969	2413	1037	566
(Fruit fly)	**Unique**	3180	2959	2045	1306	503	286	168	117
Salmo salar	**Total**	2105661	931620	240600	93818	11910	5695	1318	973
(Salmon)	**Unique**	7039	6489	4416	2892	1217	720	487	320
Carica papaya	**Total**	330257	113307	30307	16765	5066	2665	621	149
(Papaya)	**Unique**	1140	1097	991	877	501	363	175	69
Glycine max	**Total**	916939	324720	85635	46849	12620	6459	1760	523
(Soybeans)	**Unique**	3055	2951	2612	2208	1250	881	407	179
Malus domestica	**Total**	745067	263553	74541	41863	13796	5996	1614	484
(Apple)	**Unique**	2867	2760	2432	2037	1039	720	320	146
***Oryza sativa* (Rice)**	**Total**	612090	174766	30203	17632	5038	2488	648	279
	Unique	1710	1578	1255	981	523	328	163	63
Arachis hypogaea	**Total**	1193850	414633	109245	59692	15021	8476	2356	739
(Peanut)	**Unique**	4175	4033	3529	2971	1506	1076	486	218
Prunus persica	**Total**	422277	157454	45557	26298	10252	5115	1433	346
(Peach)	**Unique**	1701	1637	1416	1201	713	517	264	111
Phaseolus vulgaris	**Total**	451134	149740	42236	25113	7048	3626	1005	265
(Beans)	**Unique**	1548	1485	1346	1181	701	488	220	89
Solanum tuberosum	**Total**	462504	171829	50277	27881	7858	4100	1000	294
(Potato)	**Unique**	1880	1822	1626	1374	723	512	242	86
Triticum aestivum	**Total**	5068723	1295317	213159	112949	25380	9739	2927	1436
(Wheat)	**Unique**	9064	8557	7331	6267	3290	1904	799	384
***Zea mays* (Maize)**	**Total**	1126007	346921	60378	30418	9059	4833	1156	528
	Unique	3094	2869	2208	1661	813	574	242	127
Arabidopsis thaliana	**Total**	692802	240908	61433	30158	8702	4575	1083	292
(Mustard)	**Unique**	2293	2205	1911	1618	834	613	283	112
Prunus dulcis	**Total**	13102	4540	2619	2323	699	392	26	4
(Almond)	**Unique**	54	54	52	50	45	25	15	5
Carya illinoinensis	**Total**	5086	2303	1273	796	440	301	74	52
(Pecan)	**Unique**	32	32	32	20	17	15	13	13
Pistacia vera	**Total**	3755	729	285	245	126	42	21	11
(Pistachio)	**Unique**	8	8	8	8	7	7	7	6
Juglans regia	**Total**	666338	235167	66984	36964	11699	6744	1573	386
(English Walnut)	**Unique**	2592	2505	2291	1933	1006	723	343	138

Table 4. FASTA Comparison of Predicted Proteins of *Chlorella variabilis NC64A* to AOL V18B (*E*-Score: 10e-07). The amino acid sequences of all proteins predicted from the genome of the species were used to search this version of the AllergenOnline database to find identity matches with proteins listed as allergens or putative allergens in the database using full-length FASTA searches with different *E*-scores, those from matches at 1e-7 are shown here.

AllergenOnline Version 18B	Highest %Seq_id	Align length	*E*-score	Conservation # of species of 23 maximum
gid|1941|cyclophilin [Daucus carota]	78.8	170	9.00E-75	**20**
gid|1926|cyclophilin [Catharanthus roseus]	76.8	168	2.70E-54	**18**
gid|2708|heat shock cognate 70 [Aedes aegypti]	73.4	305	4.70E-103	**22**
gid|2591|heat shock-like protein [Tyrophagus putrescentiae]	73.3	659	6.10E-168	**22**
gid|2291|Der f 33 allergen [Dermatophagoides farinae]	73.2	455	4.20E-155	**23**
gid|166|triosephosphat-isomerase [Triticum aestivum]	72.6	248	2.10E-105	**14**
gid|2301|glyceraldehyde-3-phosphate dehydrogenase [Triticum aestivum]	70.7	334	1.40E-100	**21**
gid|338|60S ribosomal protein L3 (Allergen Asp f 23] [Aspergillus fumigatus]	67.1	386	2.00E-118	**22**
gid|1033|cytochrome_c_[Curvularia_lunata]	66	103	1.5E-30	
gid|863|cyclophilin [Aspergillus fumigatus]	64.6	161	4.10E-47	**18**
gid|706|Lactoylglutathione lyase (Methylglyoxalase] (Aldoketomutase] (Glyoxalase I] (Glx I] (Ketone-aldehyde mutase] (S-D-lactoylglutathione methylglyoxal lyase] (Allergen Ory s ?] (Allergen Glb33] (PP33] [Oryza sativa]	62.9	283	1.00E-40	**13**
gid|543|60S acidic ribosomal protein P2 [Fusarium culmorum]	62.4	109	2.50E-23	**14**
gid|2076|heat shock protein 70 [Dermatophagoides farinae]	59.6	401	1.80E-71	**10**
gid|1092|manganese superoxide dismutase-like protein [Pistacia vera]	58.4	202	4.60E-54	**17**
gid|848|60S acidic ribosomal P1 phosphoprotein Pen b 26 [Penicillium brevicompactum]	57.6	85	1.20E-11	**6**
gid|648|major allergenic protein Mal f4 [Malassezia furfur]	57.5	320	2.60E-89	**20**
gid|2255|putative chitinase [Musa acuminata]	56.7	261	1.50E-65	**13**
gid|1707|aldolase A [Thunnus albacares]	56.4	353	7.70E-77	**19**
gid|587|Chain A, Latex Profilin Hevb8 [Hevea brasiliensis]	56.1	132	2.80E-35	**1**
gid|489|putative nuclear transport factor 2 [Davidiella tassiana]	55.4	112	5.90E-25	**14**
gid|2592|aldehyde dehydrogenase-like protein [Tyrophagus putrescentiae]	54.8	489	1.50E-89	**20**
gid|1248|eukaryotic translation initiation factor [Forcipomyia taiwana]	54.3	129	1.30E-43	**21**
gid|2463|ElF1 superfamily transcriptions factor [Triticum aestivum]	54.3	81	1.90E-22	**19**
gid|2262|transaldolase [Penicillium chrysogenum]	51.4	313	2.70E-74	**3**
gid|1960|aldolase a, fructose-bisphosphate 1 [Salmo salar]	50.6	350	9.50E-68	**18**
gid|509|group 15 allergen protein [Dermatophagoides farinae]	50	120	9.30E-12	**21**
gid|651|allergen [Malassezia sympodialis]	50	140	2.60E-27	**20**
gid|64|Minor allergen Alt a 7 (Alt a VII] [Alternaria alternata]	50	200	3.20E-42	**15**
gid|126|minor allergen beta-fructofuranosidase precursor [Lycopersicon esculentum] [Solanum lycopersicum (Lycopersicon esculentum]]	49.3	140	1.30E-41	**13**
gid|775|RecName: Full=Serine carboxypeptidase 2; AltName: Full=Serine carboxypeptidase II; AltName: Full=Carboxypeptidase D; AltName: Full=CPDW-II; Short=CP-WII; Contains: RecName: Full=Serine carboxypeptidase 2 chain A; AltName: Full=Serine carboxypeptidase II c [Triticum aestivum]	49.2	195	2.20E-55	**15**
gid|1542|peroxiredoxin [Triticum aestivum]	49.1	216	1.90E-60	**4**
gid|1544|troponin C [Tyrophagus putrescentiae]	49	147	8.60E-24	**22**
gid|650|allergen [Malassezia sympodialis]	48.9	131	4.60E-33	**16**
gid|1338|ragweed homologue of Art v 1 precursor [Ambrosia artemisiifolia]	48.8	84	2.10E-09	**21**
gid|951|Der f Mal f 6 allergen [Dermatophagoides farinae]	48.7	160	1.60E-27	**20**
gid|65|aldehyde dehydrogenase (NAD+] [Alternaria alternata]	46.3	480	1.10E-107	**19**
gid|64|Allergen Alt a 7 [Alternaria alternata]	45.7	138	1.00E-27	**9**
gid|2371|seed maturation-like protein precursor [Sesamum indicum]	44.5	330	3.10E-50	**15**
gid|2551|Par h I precursor [Parthenium hysterophorus]	44.4	81	2.00E-07	**18**

gid|18|Actinidain protease-like [Actinidia deliciosa]	43.8	356	9.40E-59	**19**
gid|775|serine carboxypeptidase II [Triticum aesivum]	43.8	153	2.70E-33	**10**
gid|647|allergen [Malassezia sympodialis ATCC 42132]	42.7	82	3.30E-14	**3**
gid|154|LMM glutenin 3 [Triticum aestivum]	42.5	167	6.40E-09	**17**
gid|1206|Sal k 3 pollen allergen [Salsola kali]	42.3	769	6.00E-94	**15**
gid|496|ferritin heavy chain-like protein [Dermatophagoides farinae]	42.1	183	3.90E-22	**19**
gid|496|ferritin [Dermatophagoides farinae]	42.1	164	4.40E-15	**8**
gid|151|Alpha/beta gliadin-like protein product [Triticum aestivum]	41.8	134	1.10E-07	**20**
gid|150|omega-5 gliadin [Triticum aestivum]	41.7	396	3.00E-21	**21**
gid|322|beta-xylosidase [Aspergillus niger]	41.7	132	2.70E-22	**11**
gid|333|Taka-amylase A (Taa-G1] precursor [Aspergillus oryzae]	41.7	103	2.00E-12	**1**
gid|588|prohevein [Hevea brasiliensis]	41.3	121	3.00E-18	**11**
gid|1565|collagen alpha-2(I] chain precursor [Bos taurus]	41.2	131	1.70E-07	**19**
gid|244|Pen c 1; alkaline serine protease [Penicillium citrinum]	41.2	250	2.90E-39	**1**
gid|154|LMW glutenin-like protein product [Triticum aestivum]	40.9	235	7.70E-07	**19**
gid|325|PPIase [Aspergillus fumigatus]	40.8	130	5.30E-17	**19**
gid|588|hevein [Hevea brasiliensis]	40.8	98	3.40E-19	**11**
gid|63|Protein disulfide-isomerase (PDI] (Allergen Alt a 4] [Alternaria alternata]	40.7	81	8.70E-10	**16**
gid|322|xylosidase [Aspergillus niger]	40.6	256	1.20E-36	**12**
gid|357|trypsin [Blomia tropicalis]	40.5	237	1.00E-25	**7**
gid|850|catalase [Penicillium citrinum]	40.4	483	2.60E-42	**21**
gid|2027|allergen [Malassezia sympodialis ATCC 42132]	39.7	816	1.10E-68	**21**
gid|876|thioredoxin [Aspergillus fumigatus]	39.6	91	2.40E-14	**16**
gid|243|allergen Pen n 18 [Penicillium chrysogenum]	39.1	266	6.10E-36	**2**
gid|2709|lysosomal aspartic protease [Aedes aegypti]	38.9	522	7.30E-71	**19**
gid|330|manganese superoxide dismutase [Aspergillus fumigatus]	38.9	208	3.00E-29	**4**
gid|150|D-type LMW glutenin subunit [Triticum aestivum]	38.6	176	8.00E-07	**21**
gid|2278|thioredoxin h [Triticum aestivum]	38.4	86	1.20E-11	**19**
gid|2080|glutathione transferase [Triticum aestivum]	38.4	159	3.00E-15	**15**
gid|162|27K protein [Triticum aestivum]	38.2	186	4.00E-30	**17**
gid|785|Bromelain precursor (Allergen Ana c 2] [Ananas comosus]	38.2	152	2.70E-23	**17**
gid|1776|thioredoxin [Plodia interpunctella]	38.1	97	3.00E-12	**19**
gid|150|omega-gliadin, partial [Triticum aestivum]	37.7	408	9.90E-11	**18**
gid|833|vacuolar serine protease [Rhodotorula mucilaginosa]	37.7	297	2.80E-33	**3**
gid|1171|subtilisin precursor [Bacillus licheniformis]	37.6	282	2.50E-14	**2**
gid|18|actinidin_[Actinidia_deliciosa]	37.5	307	2.1E-28	
gid|160|glutenin [Triticum aestivum]	37.4	123	5.10E-07	**10**
gid|151|Gliadin-like protein product [Triticum aestivum]	37.1	170	8.10E-07	**21**
gid|789|art v 2 allergen [Artemisia vulgaris]	37.1	140	5.10E-09	**7**
gid|1175|prepro AprM [Bacillus sp.]	37	146	7.20E-12	**1**
gid|875|calcium-binding protein [Ambrosia artemisiifolia]	36.7	139	4.00E-12	**16**
gid|987|allergen Bla g 6.0301 [Blattella germanica]	36.6	101	2.20E-08	**12**
gid|853|MPA3 allergen [Periplaneta americana]	36.6	243	6.60E-09	**7**
gid|355|cysteine protease precursor [Blomia tropicalis]	36.4	129	4.00E-12	**3**
gid|152|gamma-gliadin [Triticum aestivum]	36.3	204	1.50E-07	**19**
gid|793|thioredoxin [Aspergillus fumigatus]	36	86	1.70E-12	**14**
gid|962|putative Cup a 4 allergen [Hesperocyparis arizonica]	36	139	1.90E-09	**14**
gid|276|Venom allergen 5 (Antigen 5] (Ag5] (Allergen Pol f 5] (Pol f V] [Polistes fuscatus]	35.8	123	3.50E-11	**1**
gid|1171|RecName: Full=Subtilisin Carlsberg; Flags: Precursor [Bacillus licheniformis]	35.6	264	9.40E-09	**2**
gid|2576|enamine/imine deaminase [Dermatophagoides farinae]	35.5	124	4.80E-23	**21**
gid|151|alpha-type gliadin precursor protein [Triticum aestivum]	35.5	290	1.80E-07	**14**
gid|1174|RecName: Full=Subtilisin Savinase; AltName: Full=Alkaline protease [Bacillus lentus]	35.4	164	6.70E-20	**3**
gid|1959|enolase [Salmo salar]	35.3	428	7.60E-20	**15**
gid|1743|troponin C [Crangon crangon]	35.2	145	3.00E-10	**16**
gid|2335|chymotrypsin-like protein [Blattella germanica]	35.1	265	1.30E-17	**7**

Table 5. FASTA Comparison of All Proteins Representing *Galdieria* sp. Genome to AllergenOnline (*E*-score=10 or smaller). Amino acid sequences of predicted proteins from this red alga to allergens and putative allergens in AllergenOnline.org. The highest percent identities are shown with alignment lengths and smallest *E*-scores. The right-hand column shows number of the 23 common species that also have an identity score over 35% identity to the allergens in the left column.

AllergenOnline Version 18B	Highest % Seq_id	Align length	*E*-score	Conservation in # 23 species maximum
gid|2591|Putative heat shock-like protein [Tyrophagus putrescentiae]	72.3	653	1.00E-207	**22**
gid|2291|Putative Der f 33-like protein [Dermatophagoides pteronyssinus]	70.8	452	8.20E-150	**23**
gid|338|Putative 60S ribosomal protein L3 (Allergen Asp f 23]	69.4	385	5.80E-124	**22**
gid|2301|Putative glyceraldehyde-3-phosphate dehydrogenase [Triticum aestivum]	68.9	315	4.00E-92	**21**
gid|863|Putative cyclophilin [Aspergillus fumigatus]	65.5	145	5.70E-40	**18**
gid|1033|Allergen cytochrome c [Curvularia lunata]	63.1	103	1.70E-26	**0**
gid|2708|Putative heat shock cognate 70 [Aedes aegypti]	62.3	657	2.00E-172	**22**
gid|1959|Allergen enolase [Salmo salar]	62	437	2.10E-112	**15**
gid|1941|Putative cyclophilin [Daucus carota]	59.2	169	8.00E-41	**20**
gid|166|Putative triosephosphat-isomerase [Triticum aestivum]	59	249	1.30E-65	**14**
gid|2236|Putative transaldolase [Cladosporium cladosporioides]	59	317	2.90E-73	**6**
gid|1707|Allergen aldolase A [Thunnus albacares]	58.9	358	3.00E-84	**19**
gid|651|Putative allergen [Malassezia sympodialis]	58.8	102	1.90E-24	**20**
gid|509|Putative 98kDa HDM allergen [Dermatophagoides farinae]	56.3	87	6.40E-12	**20**
gid|1092|Putative manganese superoxide dismutase-like protein [Pistacia vera]	55.1	207	8.30E-52	**17**
gid|62|Putative RecName: Full=60S acidic ribosomal protein P2; AltName: Full=Minor allergen Alt a 5; AltName: Full=Allergen Alt a 6; AltName: Full=Allergen Alt a VI; AltName: Allergen=Alt a 5	54.8	115	1.30E-19	**10**
gid|1983|Putative 60S acidic ribosomal phosphoprotein P1 [Penicillium crustosum]	52.7	112	2.20E-22	**16**
gid|64|Putative Minor allergen Alt a 7 (Alt a VII]	52	202	9.40E-39	**15**
gid|1026|Allergen allergen [Malassezia sympodialis ATCC 42132]	50	106	2.00E-18	**0**
gid|325|Allergen PPIase [Aspergillus fumigatus]	48.9	135	2.80E-19	**19**
gid|1926|Allergen cyclophilin [Catharanthus roseus]	47.6	170	5.80E-32	**18**
gid|1544|Putative troponin C [Tyrophagus putrescentiae]	45.9	146	3.80E-27	**22**
gid|1248|Putative eukaryotic translation initiation factor [Forcipomyia taiwana]	45.2	325	4.10E-64	**21**
gid|1206|Allergen Sal k 3 pollen allergen [Salsola kali]	45.1	765	8.60E-77	**15**
gid|2849|Allergen Chain A, Beta-amylase	44	470	1.80E-59	**0**
gid|2582|Putative alcohol dehydrogenase [Curvularia lunata]	43.4	339	1.00E-61	**8**
gid|951|Allergen Der f Mal f 6 allergen [Dermatophagoides farinae]	43.4	143	2.30E-19	**20**
gid|496|Allergen ferritin heavy chain-like protein [Dermatophagoides pteronyssinus]	42.5	179	2.60E-25	**19**
gid|63|Putative Protein disulfide-isomerase (PDI] (Allergen Alt a 4]	42.4	92	7.50E-10	**16**
gid|65|Putative aldehyde dehydrogenase (NAD+] [Alternaria alternata]	42.3	506	5.60E-75	**19**
gid|246|Putative elongation factor 1 beta-like [Penicillium citrinum]	42.1	235	3.30E-36	**20**
gid|2076|Putative heat shock protein 70 [Dermatophagoides farinae]	40.4	560	2.60E-61	**10**
gid|850|Putative catalase [Penicillium citrinum]	39.9	489	4.90E-71	**21**
gid|2293|Allergen Der f 31 allergen [Dermatophagoides farinae]	39.6	144	1.90E-12	**11**
gid|1617|Putative alpha/beta gliadin precursor [Triticum aestivum]	39.1	161	1.10E-12	**13**
gid|2592|Putative aldehyde dehydrogenase-like protein [Tyrophagus putrescentiae]	38.5	405	3.90E-59	**20**
gid|251|Putative peroxisomal membrane protein [Penicillium citrinum]	37.8	172	2.50E-16	**2**
gid|650|Putative allergen [Malassezia sympodialis]	37.5	144	7.50E-22	**16**
gid|160|Allergen high molecular weight glutenin subunit 1Ax1 [Triticum aestivum]	36.4	110	3.30E-07	**4**
gid|2215|Allergen RecName: Full=Glutathione S-transferase 1; AltName: Full=GST class-sigma	36.3	204	1.60E-19	**4**
gid|799|Allergen NADP-dependent mannitol dehydrogenase [Davidiella tassiana]	36.2	246	4.20E-24	**5**
gid|1577|Allergen Sal k 4.03 allergen [Salsola kali]	35.8	148	6.10E-12	**0**
gid|2576|Putative enamine/imine deaminase [Dermatophagoides farinae]	35.7	126	2.40E-12	**21**
gid|1171|Allergen subtilisin precursor [Bacillus licheniformis]	35.4	178	4.20E-14	**2**
gid|2551|Putative Par h I precursor [Parthenium hysterophorus]	35.2	145	9.60E-12	**18**

Table 6. Identification of Known Allergens in Allergenonline Database with Different E-Score Threshold

Species	10E-07	10E-30	10E-100
Peanut (*Arachis hypogaea*)	Ara h1, Ara h2, Ara h3, Ara h4, Ara h6, Ara h7, Ara h8, profilin, lipid transfer proteins, oleosin, conarachin, glycinin	Ara h1, Ara h2, Ara h3, Ara h4, Ara h6, Ara h7, Ara h8, profilin, lipid transfer proteins, oleosin, conarachin, glycinin	Ara h1, Ara h3, Arah4, conarachin, glycinin
Apple *Malus domestica*	Mal d3 (LTP, non-specific lipid tranfer protein), Mal d1, profilin 1, allergen AP15, allergen ribonuclease like PR, Mal d2 (thaumatin like protein)	Mal d3 (LTP, non-specific lipid tranfer protein), Mal d1, profilin 1, allergen AP15, allergen ribonuclease like PR, Mal d2 (thaumatin like protein)	Mal d2 (thaumatin like protein)
Chicken (*Gallus gallus*)	Gal d2 (Ovalbumin), serum albumin, Gal d3 (ovotransferrin), Gal d1 (ovomucoid), myosin light chain, parvalbumin, Gal d4 (lysozyme C)	Gal d2 (Ovalbumin), serum albumin, Gal d3 (ovotransferrin), Gal d1 (ovomucoid), myosin light chain, parvalbumin, Gal d4 (lysozyme C)	Gal d2 (Ovalbumin), serum albumin, Gal d3 (ovotransferrin), Gal d1 (ovomucoid)
Soybeans (*Glycine max*)	Glycinin (A3B4, A-1a-B-X, G3 A2B1-a) subunits, beta-conglycinin (alpha, beta) subunits, 2S albumin, Gly m1 (trypsin inhibitor, putative kunitz trypsin inhibitor), profilin, Gly m Bd28k, Gly m Bd 30K)	Glycinin (A3B4, A-1a-B-X, G3 A2B1-a) subunits, beta-conglycinin (alpha, beta) subunits, 2S albumin, Gly m1 (trypsin inhibitor, putative kunitz trypsin inhibitor), profilin, Gly m Bd28k, Gly m Bd 30K)	Glycinin (A3B4, A-1a-B-X, G3 A2B1-a) subunits, beta-conglycinin (alpha, beta) subunits
Bovine *(Bos taurus)*	Lactotransferrin, collagen alpha-2, bovine serum albumin, alpha lactoglobulin, beta casein isoform, Bos d3 (calcium binding protein), kappa casein, beta lactoglobulin, Bos d2, alpha S casein	Lactotransferrin, collagen alpha-2, bovine serum albumin, alpha lactoglobulin, beta casein isoform, Bos d3 (calcium binding protein), kappa casein, beta lactoglobulin, Bos d2	Lactotransferrin, collagen alpha-2, bovine serum albumin
Candida albicans	Cand a1 (alcohol dehydrogenase), Cand a3 (enolase1), IgE -binding protein	Cand a1 (alcohol dehydrogenase), Cand a3 (enolase1), IgE -binding protein	Cand a1 (alcohol dehydrogenase), Cand a3 (enolase1), IgE -binding protein
Cod (*Gadus morhua*)	Gad m1 (parvalbumin) * No detection of Gad m2, Gad m3	Gad m1 (parvalbumin)	-
Carica papaya	Car p1 (allergen papain precusror)	Car p1 (allergen papain precusror)	Car p1 (allergen papain precusror)
Almond (*Prunus dulcis*)	Prunin 2, prunin 1, prunin du amandin, pru du (2.01A, 2.01B, 2.02, 2.02B), pru du 6, pru du 1.01, profilin, pronin 2, pru du 4.02	Prunin 2, prunin 1, prunin du amandin, pru du (2.01A, 2.01B, 2.02, 2.02B), pru du 6, pru du 1.01, profilin, pronin 2	Prunin 2, prunin 1, prunin du amandin, pru du (2.01A, 2.01B, 2.02, 2.02B)
Rice *(Oryza sativa)*	Lactoylglutathione lyase, RA 16, putative allergenic protein, RA5B, seed allergenic protein (RAG2, RAG1), expansin-B, polcalcin (Ph1p7)	Lactoylglutathione lyase, RA 16, putative allergenic protein, RA5B	Lactoylglutathione lyase
Pecan (*Carya illinoinensis*)	11S legumin, putative allergen	11S legumin	11S legumin
Phasoleus vulgaris	Non-specific lipid transfer protein (1b, 1a) precursors	Non-specific lipid transfer protein (1b, 1a) precursors	-
Pistacio *(Pistacia vera)*	2S albumin, 11S globulin, manganese superoxide dismutase-like protein, vicilin, Pis v 2.0201 allergen 11S globulin precusor, Pis v 2.0101 allergen 11S globulin precursor	11S globulin, manganese superoxide dismutase-like protein, vicilin, Pis v 2.0201 allergen 11S globulin precusor, Pis v 2.0101 allergen 11S globulin precusor	11S globulin, manganese superoxide dismutase-like protein, vicilin, Pis v 2.0201 allergen 11S globulin precusor, Pis v 2.0101 allergen 11S globulin precusor
Peach (*Prunus persica*)	Pru P 1.0301, thaumatin like protein, non-specific LTP, pru du 4.02, pru p1, pru p 2.01B, pru p 2.02, pru p 1.0201, pru du 2.01A	Pru P 1.0301, thaumatin like protein, non-specific LTP, pru du 4.02, pru p1, pru p 2.01B, pru p 2.02, pru p 1.0201, pru du 2.01A	
Salmon *(Salmo salar)*	Fructose biphosphate adolase A, enolase 3-2, aldolase A, enolase, parvalbumin (beta 1, beta 2, beta)	Fructose biphosphate adolase A, enolase 3-2, aldolase A, enolase, parvalbumin (beta 1, beta 2, beta)	Fructose biphosphate adolase A, enolase 3-2, aldolase A, enolase
Potato *(Solanum tuberosum)*	Patatin, aspartic protease inhibitor II, profilin, cysteine protease inhibitor, proteinase inhibitor	Patatin, aspartic protease inhibitor II, profilin, cysteine protease inhibitor	Patatin
Walnut *(Juglan regia)*	2S albumin, seed storage protein, non-specific LTP	Seed storage protein, non-specific LTP	-
Wheat *(Triticum aestivum)*	Putative hypothetical protein, thioredoxin peroxidase, non-specific LTP, HMW glutenin-like protein, alpha amylase inhibitor like protein, thaumatin like protein, alpha amylase inhibitor (0.28, 0.19), E1F1 superfamily transcription factor, serine proteinase inhibitor like allergen, profilin, endosperm transfer cell specific PR60 precurosr, serpin, serine carboxypeptidase, glyceraldehyde-3 phosphatedehydrogenase, triosephosphate isomerase, putative 27K protein, serine carboxypeptidase2, alpha/beta gliadin, alpha purothionin subtilisim, chymotrypsin inhibitor WSCI, pre-alpha gliadin, beta gliadin	Putative hypothetical protein, thioredoxin peroxidase, non-specific LTP, HMW glutenin-like protein, alpha amylase inhibitor like protein, thaumatin like protein, alpha amylase inhibitor (0.28, 0.19), E1F1 superfamily transcription factor, serine proteinase inhibitor like allergen, profilin, endosperm transfer cell specific PR60 precurosr, serpin, serine carboxypeptidase, glyceraldehyde-3 phosphatedehydrogenase, triosephosphate isomerase, putative 27K protein, serine carboxypeptidase2, alpha/beta gliadin, alpha purothionin subtilisim, chymotrypsin inhibitor WSCI, pre-alpha gliadin, beta gliadin	Serpin, serine carboxypeptidase, glyceraldehyde-3 phosphatedehydrogenase, triosephosphate isomerase

Table 7. Distribution of Matches to Clinically Important Allergens in the 23 Species

LTPs	Vicilins	Glycinins	Tropomyosins	Arginine kinase	2S albumins
Peanut	Papaya	Soybeans	Drosophila	Human	Pistachio
Kidney beans	Corn	Kidney beans	Salmon	Chicken	Potato
Walnut	Drosophila	Peanut	Atlantic cod	Bovine	Soybeans
Soybeans	Pistachio	Salmon	Chicken	Salmon	Walnut
Apple	Soybeans	Walnut	Human	Atlantic cod	Peanut
Papaya	Peanut	Chicken	Bovine	Drosophila	
Rice	Almond	Human			
Wheat	Pecan	Potato			
Peach	Walnut				
Potato	Potato				
Bovine	Apple				
Human	Peach				
Corn	Human				
Arabidopsis	Salmon				
Almond					

Table 8. List of Minor Allergens Which Are Highly Conserved Between Species Under Study and Beyond CODEX Guidelines.

Minor allergen	Conservation	Minor allergen	Conservation
gid|2291|Der_f_33_allergen_[Dermatophagoides_farinae]	23	gid|2076|Der_f_28_allergen_[Dermatophagoides_farinae]	13
gid|1544|troponin_C_[Tyrophagus_putrescentiae]	22	gid|2134|Chain_B,_2.70_A_Crystal_Structure_Of_The_Amb_A_11_Cysteine_Protease|_A_Major_Ragweed_Pollen_Allergen|_In_Its_Proform_[Ambrosia_artemisiifolia]	13
gid|2591|heat_shock-like_protein_[Tyrophagus_putrescentiae]	22	gid|2255|putative_chitinase_[Musa_acuminata]	13
gid|2708|heat_shock_cognate_70_[Aedes_aegypti]	22	gid|2439|glutathione_S-transferase_[Betula_pendula]	13
gid|338|60S_ribosomal_protein_L3_(Allergen_Asp_f_23)_[Aspergillus_fumigatus]	22	gid|2594|chitinase_[Zea_mays]	13
gid|1248|eukaryotic_translation_initiation_factor_[Forcipomyia_taiwana]	21	gid|409|papain_precursor_[Carica_papaya]	13
gid|1338|ragweed_homologue_of_Art_v_1_precursor_[Ambrosia_artemisiifolia]	21	gid|49|phytocystatin_[Actinidia_deliciosa]	13
gid|150|D-type_LMW_glutenin_subunit_[Triticum_aestivum]	21	gid|592|beta-1,3-glucanase_[Hevea_brasiliensis]	13
gid|150|omega-5_gliadin_[Triticum_aestivum]	21	gid|619|pollen_allergen_Jun_o_4_[Juniperus_oxycedrus]	13
gid|151|Gliadin-like_protein_product_[Triticum_aestivum]	21	gid|698|calcium-binding_protein_[Olea_europaea]	13
gid|2027|allergen_[Malassezia_sympodialis_ATCC_42132]	21	gid|706|Lactoylglutathione_lyase_(Methylglyoxalase)_(Aldoketo mutase)_(Glyoxalase_I)_(Glx_I)_(Ketone-aldehyde_mutase)_(S-D-lactoylglutathione_methylglyoxal_lyase)_(Allergen_Ory_s_?)_(Allergen_Glb33)_(PP33)_[Oryza_sativa]	13
gid|2301|glyceraldehyde-3-phosphate_dehydrogenase_[Triticum_aestivum]	21	gid|844|thioredoxin_h1_protein_[Zea_mays]	13
gid|2576|enamine/imine_deaminase_[Dermatophagoides_farinae]	21	gid|1096|pollen_allergen_Pla_o_2_[Platanus_orientalis]	12
gid|509|group_15_allergen_protein_[Dermatophagoides_pteronyssinus]	21	gid|152|gamma-gliadin_B_precursor_[Triticum_aestivum]	12
gid|850|catalase_[Penicillium_citrinum]	21	gid|152|putative_gamma-gliadin_[Triticum_aestivum]	12
gid|1337|TCTP_[Alternaria_alternata]	20	gid|1747|pollen_allergen_CPA63_[Cryptomeria_japonica]	12
gid|151|Alpha/beta_gliadin-like_protein_product_[Triticum_aestivum]	20	gid|1884|putative_allergen_Pru_du_2.01A_[Prunus_dulcis_x_Prunus_persica]	12
gid|1941|cyclophilin_[Daucus_carota]	20	gid|2134|cysteine_protease_[Ambrosia_artemisiifolia]	12
gid|246|elongation_factor_1_beta-like_[Penicillium_citrinum]	20	gid|234|isoflavone_reductase_related_protein_[Pyrus_communis]	12
gid|2592|aldehyde_dehydrogenase-like_protein_[Tyrophagus_putrescentiae]	20	gid|2461|hypothetical_protein_[Triticum_aestivum]	12
gid|509|98kDa_HDM_allergen_[Dermatophagoides_farinae]	20	gid|2479|lipid-transfer_protein_7k-LTP_precursor_[Solanum_lycopersicum]_[Solanum_lycopersicum_(Lycopersicon_esculentum)]	12
gid|648|major_allergenic_protein_Mal_f4_[Malassezia_furfur]	20	gid|2579|Manual_Entry_Cha_o_3_[Chamaecyparis_obtusa]	12
gid|651|allergen_[Malassezia_sympodialis]	20	gid|262|polygalacturonase_[Platanus_x_acerifolia]	12
gid|951|Der_f_Mal_f_6_allergen_[Dermatophagoides_farinae]	20	gid|285|peanut_agglutinin_precursor;_prePNA_[Arachis hypogaea]	12
gid|152|gamma-gliadin_[Triticum_aestivum]	19	gid|322|xylosidase_[Aspergillus_niger]	12
gid|154|LMW_glutenin-like_protein_product_[Triticum_aestivum]	19	gid|345|allergenic_isoflavone_reductase-like_protein_Bet_v_6.0102_[Betula_pendula]	12
gid|1542|RecName:_Full=1-Cys_peroxiredoxin_PER1;_AltName:_Full=Rehydrin_homolog;_AltName:_Full=Thioredoxin_peroxidase_[Triticum_aestivum]	19	gid|36|putative_pectate_lyase_precursor_[Ambrosia_artemisiifolia]	12
gid|1565|collagen_alpha-2(I)_chain_precursor_[Bos_taurus]	19	gid|38|Pollen_allergen_Amb_a_3_(Amb_a_III)_(Allergen_Ra3)_[Ambrosia_artemisiifolia_(elatior)]	12
gid|1707|aldolase_A_[Thunnus_albacares]	19	gid|424|pollen_allergen_[Chamaecyparis_obtusa]	12
gid|1776|thioredoxin_[Plodia_interpunctella]	19	gid|448|isoflavone_reductase-like_protein_CJP-6_[Cryptomeria_japonica]	12
gid|18|Actinidain_protease-like_[Actinidia_deliciosa]	19	gid|449|allergen_Cry_j_2_[Cryptomeria_japonica]	12
gid|2278|thioredoxin_h_[Triticum_aestivum]	19	gid|449|pollen_allergen_[Cryptomeria_japonica]	12
gid|2463|EIF1_superfamily_transcriptions_factor_[Triticum_aestivum]	19	gid|466|pre-pro-cucumisin_[Cucumis_melo]	12
gid|2709|lysosomal_aspartic_protease_[Aedes_aegypti]	19	gid|477|FAD-linked_oxidoreductase_BG60_[Cynodon_dactylon]	12
gid|325|PPIase_[Aspergillus_fumigatus]	19	gid|563|RecName:_Full=Hydrophobic_seed_protein;_Short=HPS;_AltName:_Allergen=Gly_m_1_[Glycine_max]	12
gid|496|ferritin_heavy_chain-like_protein_[Dermatophagoides_pteronyssinus]	19	gid|582|latex_protein_allergen_Hev_b_7_[Hevea_brasiliensis]	12
gid|65|aldehyde_dehydrogenase_(NAD+)_[Alternaria_alternata]	19	gid|582|putative_latex_allergen_hev_b_7.02_[Hevea_brasiliensis]	12
gid|694|Ole_e_5_olive_pollen_allergen_[Olea_europaea]	19	gid|585|ENSP-like_protein_[Hevea_brasiliensis]	12
gid|150|omega-gliadin,_partial_[Triticum_aestivum]	18	gid|593|small_rubber_particle_protein_[Hevea_brasiliensis]	12
gid|151|gliadin_[Triticum_urartu]	18	gid|644|MF1_[Malassezia_furfur]	12
gid|18|RecName:_Full=Actinidain;_Short=Actinidin;_AltName:_Allergen=Act_c_1;_Flags:_Precursor_[Actinidia_chinensis]	18	gid|660|Manioc_Glu_[Manihot_esculenta]	12
gid|1926|cyclophilin_[Catharanthus_roseus]	18	gid|695|allergen_Ole_e_10_[Olea_europaea]	12
gid|1960|aldolase_a,_fructose-bisphosphate_1_[Salmo_salar]	18	gid|699|beta-1,3-glucanase-like_protein_[Olea_europaea]	12
gid|2551|Par_h_I_precursor_[Parthenium_hysterophorus]	18	gid|699|Chain_A,_Solution_Structure_Of_The_C-Terminal_Domain_Ole_E_9_[Olea_europaea]	12
gid|467|pathogen-related_protein_1_[Cucumis_melo_var._inodorus]	18	gid|749|beta-1,_3-glucananse_[Musa_acuminata_AAA_Group]_[Musa_acuminata_AAA_Group]	12

gid|518|aldehyde_dehydrogenase_(NAD+)_[Davidiella_tassiana]	**18**	gid|773|putative_leucine-rich_repeat_protein_[Triticum_aestivum]	**12**
gid|863|cyclophilin_[Aspergillus_fumigatus]	**18**	gid|84|Zm13_[Zea_mays]	**12**
gid|1092|manganese_superoxide_dismutase-like_protein_[Pistacia_vera]	**17**	gid|891|Sal_k_1_pollen_allergen_[Salsola_kali]	**12**
gid|154|LMM_glutenin_3_[Triticum_aestivum]	**17**	gid|897|polygalacturonase_[Lilium_longiflorum]	**12**
gid|162|27K_protein_[Triticum_aestivum]	**17**	gid|979|Amb_a_1-like_protein_[Artemisia_vulgaris]	**12**
gid|688|villin_1_[Nicotiana_tabacum]	**17**	gid|987|allergen_Bla_g_6.0301_[Blattella_germanica]	**12**
gid|72|putative_nuclear_transport_factor_2_[Alternaria_alternata]	**17**	gid|152|gamma-gliadin_precursor_[Triticum_aestivum]	**11**
gid|785|Bromelain_precursor_(Allergen_Ana_c_2)_[Ananas_comosus]	**17**	gid|1605|Ole_e_11.01_allergen_precursor_[Olea_europaea]	**11**
gid|152|gamma_gliadin_precursor_[Triticum_aestivum	**16**	gid|1884|putative_allergen_Pru_p_2.01B_[Prunus_dulcis_x_Prunus_persica]	**11**
gid|154|low_molecular_weight_glutenin_[Triticum_aestivum]	**16**	gid|2003|kiwellin_[Actinidia_arguta]	**11**
gid|1743|troponin_C_[Crangon_crangon]	**16**	gid|2238|metallothionein_type_2_[Coffea_arabica]	**11**
gid|1983|60S_acidic_ribosomal_phosphoprotein_P1_[Penicillium_crustosum]	**16**	gid|2293|Der_f_31_allergen_[Dermatophagoides_farinae]	**11**
gid|594|latex_allergen_[Hevea_brasiliensis]	**16**	gid|322|beta-xylosidase_[Aspergillus_niger]	**11**
gid|63|Protein_disulfide-isomerase_(PDI]_(Allergen_Alt_a_4]_[Alternaria_alternata]	**16**	gid|331|60S_acidic_ribosomal_protein_P2_(Allergen_Asp_f_8]_(AfP2]_[Aspergillus_fumigatus]	**11**
gid|650|allergen_[Malassezia_sympodialis]	**16**	gid|389|oleosin_[Corylus_avellana]	**11**
gid|875|calcium-binding_protein_[Ambrosia_artemisiifolia]	**16**	gid|586|enolase,_isoform_1_[Hevea_brasiliensis]	**11**
gid|876|thioredoxin_[Aspergillus_fumigatus]	**16**	gid|588|hevein_[Hevea_brasiliensis]	**11**
gid|1206|Sal_k_3_pollen_allergen_[Salsola_kali]	**15**	gid|588|prohevein_[Hevea_brasiliensis]	**11**
gid|165|serpin_[Triticum_aestivum]	**15**	gid|688|villin_2_[Nicotiana_tabacum]	**11**
gid|1959|enolase_[Salmo_salar]	**15**	gid|694|allergen_Ole_e_5_[Olea_europaea]	**11**
gid|2080|glutathione_transferase_[Triticum_aestivum]	**15**	gid|73|60S_acidic_ribosomal_protein_P1_(Allergen_Alt_a_12]_(Alt_a_XII]_[Alternaria_alternata]	**11**
gid|2371|seed_maturation-like_protein_precursor[Sesamum_indicum]	**15**	gid|749|Chain_A,_Crystal_Structure_At_1.45-_Resolution_Of_The_Major_Allergen_Endo-Beta-1|3-Glucanase_Of_Banana_As_A_Molecular_Basis_For_The_Latex-Fruit_Syndrome_[Musa_acuminata]"	**11**
gid|64|Minor_allergen_Alt_a_7_(Alt_a_VII]_[Alternaria_alternata]	**15**	gid|1268|Pas_n_1_allergen_precursor_[Paspalum_notatum]	**10**
gid|775|RecName:_Full=Serine_carboxypeptidase_2;_AltName:_Full=Serine_carboxypeptidase_II;_AltName:_Full=Carboxypeptidase_D;_AltName:_Full=CPDW-II;_Short=CP-WII;_Contains:_RecName:_Full=Serine_carboxypeptidase_2_chain_A;_AltName:_Full=Serine_carboxypeptid	**15**	gid|160|glutenin_[Triticum_aestivum]	**10**
gid|151|alpha-gliadin,_partial_[Triticum_monococcum_subsp._aegilopoides]	**14**	gid|160|high_molecular_weight_glutenin_subunit__10_[Triticum_aestivum]	**10**
gid|151|alpha-type_gliadin_precursor_protein_[Triticum_aestivum]	**14**	gid|1748|pollen_allergen_CJP-8_[Cryptomeria_japonica]	**10**
gid|166|triosephosphat-isomerase_[Triticum_aestivum]	**14**	gid|1884|thaumatin-like_protein_2_[Prunus_persica]	**10**
gid|1697|hevamine_[Hevea_brasiliensis]	**14**	gid|2076|heat_shock_protein_70_[Dermatophagoides_farinae]	**10**
gid|247|68_kDa_allergen_[Penicillium_chrysogenum]	**14**	gid|2594|RecName:_Full=Endochitinase_A;_AltName:_Full=Seed_chitinase_A;_Flags:_Precursor_[Zea_mays]	**10**
gid|343|allergen_[Betula_pendula]	**14**	gid|344|peptidylprolyl_isomerase_(cyclophilin]_[Betula_pendula]	**10**
gid|489|putative_nuclear_transport_factor_2_[Davidiella_tassiana]	**14**	gid|520|minor_allergen,_ribosomal_protein_P2_Davidiella_tassiana]	**10**
gid|543|60S_acidic_ribosomal_protein_P2_[Fusarium_culmorum]	**14**	gid|567|allergen_Gly_m_Bd_28K_[Glycine_max]	**10**
gid|566|Bd_30K_[Glycine_max]	**14**	gid|584|major_latex_allergen_Hev_b_4_[Hevea_brasiliensis]	**10**
gid|793|thioredoxin_[Aspergillus_fumigatus]	**14**	gid|586|Enolase_2_(2-phosphoglycerate_dehydratase_2]_(2-phospho-D-glycerate_hydro-lyase_2]_(Allergen_Hev_b_9]_[Hevea_brasiliensis]	**10**
gid|927|Per_a_6_allergen_[Periplaneta_americana]	**14**	gid|601|Humj1_[Humulus_japonicus]	**10**
gid|962|putative_Cup_a_4_allergen_[Hesperocyparis_arizonica]	**14**	gid|62|RecName:_Full=60S_acidic_ribosomal_protein_P2;_AltName:_Full=Minor_allergen_Alt_a_5;_AltName:_Full=Allergen_Alt_a_6;_AltName:_Full=Allergen_Alt_a_VI;_AltName:_Allergen=Alt_a_5_[Alternaria_alternata]	**10**
gid|126|minor_allergen_beta-fructofuranosidase_precursor_[Lycopersicon_esculentum]_[Solanum_lycopersicum_(Lycopersicon_esculentum]]	**13**	gid|658|thaumatin-like_protein_precursor_Mdtl1_[Malus_domestica]	**10**
gid|151|pre-alpha-/beta-gliadin_A-III_[Triticum_aestivum]	**13**	gid|660|allergenic-related_protein_Pt2L4_[Manihot_esculenta]	**10**
gid|1617|alpha/beta_gliadin_precursor_[Triticum_aestivum]	**13**	gid|775|serine_carboxypeptidase_II_[Triticum_aestivum]	**10**
gid|18|actinidin_[Actinidia_deliciosa]	**13**	gid|891|pectin_methylesterase_allergenic_protein_[Salsola_kali]	**10**

2.6. References

Andrews S., 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bankevich, A., Nurk, S., Antipov, D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Blanc, G., Duncan, G., Agarkova, I., Borodovsky, M., Gurnon, J., Kuo, A., Lindquist, E., Lucas, S., Pangilianan, J, Polle, J., et al., 2010. The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. The Plant Cell., 22(9):2943–2955.

Ellens, K.W., Levac, D., Pearson, C., Savoie, A., Strand, N., Louter, J.,Tubekuys C., 2019. Canadian regulatory aspects of gene editing technologies. Transgenic Res., 28(Suppl 2):165-168.

Finnigan, T.J.A., Wall, B.T., Wilde, P.J., Stephens, F.B., Taylor, S.L., Freedman, M.R., 2019. Mycoprotein: The future of nutritious nonmeat protein, a symposium review. Curr. Dev. Nutr. 3(6), nzz021.

Goodman, R.E., Ebisawa, M., Ferreira, F., Sampson, H.A., van Ree, R., Vieths, S., Baumert, J.L., Bohle, B., Lalithambika, S., Wise, J., Taylor, S.L., 2016. AllergenOnline: a peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reaqctivity. Mol. Nutr. Food Res. 60, 1183-1198.

Gowland, M.H., Walker, M.J., 2015. Food allergy, a summary of eight cases in the UK criminal and civil courts: effective last resort for vulnerable consumers? J. Sci. Food Agric. 95(10):1979-1990.

Gurevich, A., Saveliev V., Vyahhi N., Tesler G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072-1075.

Hoff, M., Ballmer-Weber, B.K., Niggemann, B., Cistero-Bahima, A., San Miguel-Moncin, M., Conti, A., Haustein, D., Vieths, S., 2003b. Molecular cloning and immunological characterization of potential allergens from the mould Fusarium culmorum. Mol. Immunol. 39(15):965-975.

Hoff, M., Trueb, R.M., Ballmer-Weber, B.K., Vieths, S., Wuethrich, B., 2003a Immediate-type hypersensitivity reaction to ingestion of mycoprotein (Quorn) in a patient allergic to molds caused by acidic ribosomal protein P2. J. Allergy Clin. Immunol. 111(5):1106-1110.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(1):119.

Jacobson, M.F., DePorter, J., 2018. Self-reported adverse reactions associated with mycoprotein (Quorn-brand) containing foods. Ann. Allergy Asthma Immunol., 120(6):626-630.

Johnson, P.E., Sayers, R.L., Gethings, L.A., et al., 2016. Quantitative Proteomic Profiling of Peanut Allergens in Food Ingredients Used for Oral Food Challenges. Anal Chem., 88(11):5689–5695.

Katona, S.J., Kaminski, E.R., 2002. Sensitivity to Quorn mycoprotein (Fusarium venenatum) in a mould allergic patient. J. Clin. Pathology 55(11):87-88.

King, R., Urban, M., Hammond-Kosack, M.C.U., Hassani-Pak, K., Hammond-Kosak, K.E.H., 2015. The completed genome sequence of the pathogenic ascomycete fungus Fusarium graminearum. BMC Genomics, 16:544

Klamczynska, B., Mooney, W.D., 2017. Heterotrophic microalgae: a scalable and sustainable protein source. Sustainable Protein Sources. [doi: http://dx.doi.org/10.1016/B978-0-12-802778-3.00020-2)

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4):357–359.

Li, H., 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32:(14): 2103–2110.

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

Lowe, T.M., Chan, P.P., 2016. tRNAscan-SE On-Line: Search and contextural analysis of Transfer RNA genes. Nucl Acids Res, 44:W54-57.

Niehaus, M., Munsterkotter, M., Proctor, R.H., et al., 2016. Comparative “Omics” of the Fusarium fujikuroi species complex highlights differences in genetic potential and metabolite synthesis. Genome Biol Evol, 8(11):3574-3599.

Ramsey, N.B., Duffey, D., Anagnostou, K., Coleman, N.E., Davis, C.M., 2019. Epidemiology of anaphylaxis in critically ill children in the United States and Canada. J. Allergy Clin. Immunol. Pract., 7:2241-2249.

Schonknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Brautigam, A. Baker, B.J., Banfield, J.F., Garavito, R.M., et al., 2013. Gene transfer from Bacteria and Archaea facilitated evolution of an extremophilic eukaryote. Science 339:1207-1210.

Slikker, W. Jr., de Souza Lim, T.A., Archella, D, de Silva, J.B. Jr., Barton-Maclaren, T., Bo, L. Buvinich, D., Chaudhry, Q., Chuan, P., Deluyker, H., et al., 2018. Emerging technologies for food and drug safety. Regul. Toxicol. Pharmacol., 98:115-128.

Stanke, M., Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic acids research. 33(Web Server issue), W465–W467.

Tee, R.D., Gordon, D.J., Welch, J.A., Newman Taylor, A.J., 1993. Investigation of possible adverse allergic reactions to mycoprotein (‘Quorn’). Clin. Exp. Allergy 23(4):257-260.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al., 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLOS One. 9(11): e112963.

Weber, R.W., Levetin, E., 2014. Allergen of the month-Fusarium. Annal. Allergy Asthma Immunol. 112(5):A11.

Wells, M.L., Potin, P., Craigie, J.S., Raven, J.A. et al., 2017. Algae as Nutritional and Functional Food Sources: Revisiting Our Understanding. J. Appl. Phycol., 29(2):949–982.

Yeh, C.C., Tai, H.Y., Chou, H., Wu, K.G., Shen, H.D., 2016. acuolar serine protease is a major allergen of Fusarium proliferatum and an IgE-cross reactive pan-fungal allergen. Allergy Asthma Immunol. Res. 8(5):438-444

Zimin, A.V., Marcais, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. The MaSuRCA genome assembler. Bioinformatics, 29(21):2669-2677.

2.7. Acknowledgement and Financial Support

The authors thank Fermentalg and Sustainable Products, Inc for funding this project. The authors thank Holland Computing Center, University of Nebraska-Lincoln for the computational resources.

CHAPTER 3

TRANSCRIPTOMICS AND PROTEOMICS EVALUATION OF POTENTIAL IgE CROSS-REACTIVITY FOR CONSUMPTION OF HOUSE CRICKET (*Acheta domesticus*)

This chapter is in progress to be submitted for peer review: Mohamed Abdelmoteleb, Lee K. Palmer, Justin T. Marsh, Philip E. Johnson and Richard E. Goodman

3.1. Abstract

Insects have been consumed for millenia in many countries, although rarely in Europe and U.S. New foods are being developed now from crickets and mealworms for markets in Europe and North America. Recently regulators in the United States began asking developers to assure that new food products containing cultured, processed insects are safe for crustacean allergic subjects, based on comparisons of genomic, transcriptomic or proteomic data. The potential cross-reactivity for House Cricket (*Acheta domesticus*) is the focus of this study. The transcriptome of cricket was compiled using multiple *de novo* assemblers. Predicted transcripts were compared to AllergenOnline.org V18B using BLASTX to find potentially significant alignments with known and putative allergens including tropomyosin (TM) and arginine kinase (AK). Abundance of mRNA of these proteins in cricket were estimated using RNA-seq quantification with RSEM software. Different isoforms of TM and AK were predicted. Predicted protein sequences were used to evaluate proteomic data of *Aceta domesticus* obtained by LC-MSMS to confirm the presence from a likely food preparation. Probable IgE epitopes were predicted using five immunoinformatics programs and compared to published epitopes from shrimp. Very high sequence identity, high abundance of

transcripts, and common IgE epitopes of tropomyosin, arginine kinase was reported between insects and cockroach, HDM, and other crustaceans. Based on recent research in the Netherlands, crustacean-allergic consumers are likely to experience cross-reactions if they consume foods containing proteins from meal worms. We sought to understand possible risks of cricket for crustacean allergic consumers.

3.2. Introduction

There is a future expectation toward growing demand for food and animal derived foods worldwide primarily based on protein and micronutrient availability. Yet animal derived protein has a high environmental cost. People are considering insects as a possible efficient protein source due to their high content in proteins, nutrients (iron, zinc, riboflavin, pantothenic acid, biotin, folic acid, and polyunsaturated fatty acids), and other environmental concerns (Payne et al, 2016; Rumpold and Schluter, 2013; van Huis et al, 2016, Hall et al, 2017).

The European Union has identified some insects as legal novel food sources, nominating several species of insects as potential human food sources, including house cricket (*Acheta domesticus*), banded cricket (*Gryllodes sigillatus*), field cricket (*Gryllus assimilis*), yellow mealworm (*Tenebrio molitor*), lesser mealworm (*Alphitobius diaperinus*), wax moth (*Galleria mellonella*), silkworm moth (*Bombyx mori*), and migratory locust (*Locusta migratoria*) (Ribeiro et al, 2018). Recent studies of consumer preferences have shown that within Western cultures consumers may be classified within four groups ranging from strong acceptors to strongly disgusted regarding the concept of

using insects for food or feed, presenting diverse perspectives on the future acceptance of dietary in Western diets (Cunha et al, 2014; Cunha et al, 2014).

When considering novel food sources, possible risks must be evaluated including their allergenic potential. One of the requirements by regulators in the United States is to assure that the new food products with cultured and processed insects are safe for crustacean allergic subjects and those allergic to house dust mites (van der Spiegel, 2013, EFSA, 2015). The allergenic risk in regard to the novel food insects might arise due to potential cross-reactivity with other arthropods, especially crustaceans as one of the most common triggers of food allergy in the western countries (Stanhope et al, 2015; Schluter et al, 2017) and house dust mite (HDM) as one of the most frequent indoor allergens which develop allergic respiratory reactions (Sheehan et al, 2015).

Tropomyosin (TM) and arginine kinase (AK) are the major cross-reacting allergens across all invertebrates including among crustacean, insects and mollusks. Both proteins are also conserved in mammalian and avian species, but with divergent amino acid sequences. Tropomyosin is a myofibrillar protein which consists of a coiled-coil dimer with 33–38 kDa monomeric molecular masses. TM is present in the muscle and involved in movement and posture. In various species there may be 2 or more isoforms of TMwith slightly different function, sequences, and expression (Ayuso et al, 2002; Pedrosa et al, 2015). AK is an important metabolic enzyme (356 AA, 40 kDa) for energy metabolism of shellfish. The water soluble, heat-labile arginine kinase was characterized as a novel allergen in shrimp by 2D immunoblotting and mass spectrometry (Yu et al, 2003). In addition, arginine kinase was identified as an allergen in octopus and crab (Shen et al, 2012; Shen et al, 2011).

One of the perplexing questions in allergy research is to understand the characteristics that differentiate allergens from nonallergens (Vogel and Marcotte, 2012). It has been hypothesized that allergens are the abundant and/or stable proteins in allergenic food sources. However, the experimental evidence to accept or refute this hypothesis is limited (Chan et al, 2015). Statistical comparisons for the abundance of allergens versus nonallergens using genomic and proteomic scales are still lacking. Quantification of RNA-seq to infer the protein level is not rigorously accurate. However, the general conclusion for the levels of non-allergens and allergens are likely valid. In a study, the HDM, *Dermatophagoides pteronyssinus* (DP) proteome was evaluated using RNA-seq methods, thermophilic stability, in addition to a combined chemical denaturation and mass spectrometry approach to assess the abundance of all proteins. Non-allergens had a wide range of expression levels, and allergens trend toward the highly expressed proteins in this range (Ogburn et al, 2017).

The identification of epitopes is an important tool for the characterization of cross-reactivity, allergenicity, and the possible inhibitory potential of allergens and subsequently understanding the interaction mechanisms and recognition in the allergic reaction. Identification of the allergenic epitopes of proteins among different species should provide a clear evidence for study of the relationship between allergenicity and protein structure (Motoyama et al 2007; Yu et al, 2003; Fu et al, 2018). The protein sequences of TM in 14 shrimp species and AK in 12 species can be downloaded in the NCBI library. Tropomyosin epitopes have been identified in 5 different shrimp species, including *Litopenaeus vannamei* (Ayuso et al, 2010), *Penaeus monodon* (Zheng et al, 2011), *Penaeus chinensis* (Fu et al, 2018), *Penaeus indicus* (Shanti et al, 1993), and,

Farfantepenaeus aztecus (Ayuso et al, 2002), but AK epitopes only in *Penaeus chinensis* (Fu et al, 2018), and *Litopenaeus vannamei* (Matsuo et al, 2015). Nevertheless, the epitopes of allergenic TM and AK in these novel insects have not been identified. Identifying of the allergenic epitopes of shrimp proteins among different species provides more evidence for study of the relationship between protein structure and allergenicity (Fu et al, 2018).

Classic methods used to identify the allergenic epitopes are costly, time-consuming, and require experienced practitioners (Fu et al, 2018). Recently, immunoinformatics has become a useful tool for predicting epitopes from immunological proteins (Bian, 2003; Li et al, 2005). The critical amino acids in epitopes in previous studies mostly appeared as aromatic or charged amino acids in junction with nearby amino acids to influence the folding, hydrophilic properties of proteins or ability of direct binding to IgE (Wangorsch et al, 2007; Scealy et al, 2006). While introducing a mutated critical amino acid, the epitope stability and binding ability may change. Prediction of epitopes will depend upon detection of physicochemical properties, conservation and relative frequency of different amino acids in the target epitopes (Hopp et al, 1981; Kyte et al, 1982).

Food developers are beginning to use specific cultured insects (mealworm and cricket) as sources of protein in processed foods. Recent studies in Europe demonstrated both IgE and clinical allergic cross-reactivity for some shrimp allergic subjects and for those cultivating mealworm when exposed to proteins of mealworm or shrimp (Broekman et al, 2017). Two food safety authorities, the EFSA for the European Union and ANSES for Argentina, are advising those with shrimp allergy to avoid consuming

mealworm (EFSA, 2015; ANSES, 2015). However, regulators in the United States are asking for assurance of safe use of processed, cultured insects, based on comparisons of genomic, transcriptomic or proteomic data and the Food and Drug Authority (FDA) has not specifically approved of any precautionary labeling for this potential hazard.

Therefore, the objectives of this study are: 1) to evaluate the potential cross-reactivity of tropomyosin and arginine kinase for house cricket proteins and 2) to assess the abundance of tropomyosin and arginine kinase in house cricket using RNA-seq analysis and non-targeted proteomics; comprehensive characterization of the highly abundant tropomyosin and arginine kinase isoforms using transcriptomic and proteomic resources; and 3) to characterize the probable IgE binding epitopes for the two cricket proteins compared to the known allergens, TM and AK in crustacean shellfish, cockroaches, and HDM.

3.3. Methodology

3.3.1. Literature search and systematic review for studies of IgE binding and allergy

The literature about insects and food allergy is scarce. Therefore, in patients with food allergy to crustaceans or with allergy to HDM, the allergic risk after consuming insects needs to be reviewed, and systematically assessed. A systematic search of four databases was performed to understand the potential safe use of *Acheta domesticus:* PubMed database (http://www.ncbi.nlm.nih.gov/sites/entrez), Scopus (https://www-scopus-com.libproxy.unl.edu/), Google Scholar (https://scholar-google-com.libproxy.unl.edu/) and Web of Science (https://apps-webofknowledge-com.libproxy.unl.edu/), conducted on March 6, 2019. Inclusion criteria were to find studies with serum IgE binding assays and positive clinical cases of reported food allergy to crickets. Queries

included “cricket AND (allerg* OR hypersensitiv* OR anaphyla* OR cross-reactiv*)”. A complete review was performed for all publications to identify, characterize allergens and identify positive IgE reactions.

3.3.2. Preparation of SRA reads, transcriptome assembly and alignment of the predicted transcripts against AllergenOnline.org V18B database

Allergens from *Acheta domesticus* have not been completely characterized. This study was intended to investigate the allergenic potential of cricket proteins that would be consumed in food, focusing on the two major allergens: tropomyosin and arginine kinase. The genome and proteome of cricket have not been reported though initial transcriptome work was published (Drinnenberg et al, 2014). The transcriptome of the cricket was reported by the Malik lab (SRR1552491; https://www.ncbi.nlm.nih.gov/sra/?-term=SRR1552491). The published sequence read archive (SRA) was downloaded from the NCBI library, checked for quality, assembled, aligned against AllergenOnline.org V18B database (Goodman et al, 2016) focusing on TM and AK protein sequences. The quality of the reads was checked using Fastqc (Andrews, 2010), and trimming of low-quality bases was performed using Prinseq (Schmieder and Edwards, 2011). De novo assembly using rnaSPAdes (Bakevich et al, 2012), Trinity (Grabherr et al, 2011), Velvet and Oases (Zerbino et al, 2008) were used to increase the confidence in the transcriptomic predictions. The quality of assembly was assessed using Quast (Gurevich et al, 2013). The predicted contiguous transcripts were compared using BLASTX to the AllergenOnline.org V18B database, focusing on tropomyosin, and arginine kinase protein sequences.

3.3.3. Examining the transcriptional profile of TM and AK allergens

The predicted transcripts from assembly were processed using RSEM (Li and Dewey, 2011) to quantitate the expression levels of these transcripts in fragments per kilobase of transcript per million mapped reads (FPKM). The (Transcripts Per Kilobase Million) TPM and (Fragments Per Kilobase Million) FPKM values were recorded.

3.3.4. Characterization of TM and AK isoforms

Multiple sequence alignments (MSA) of the predicted transcripts for TM and AK from the three assemblers (rnaSPAdes, Velvet, and Trinity) were conducted using NCBI Multiple Sequence Alignment Viewer (https://www.ncbi.nlm.nih.gov/projects/msaviewer/) and MUSCLE software (Edgar 2004). The amino acid sequences for TM and AK were predicted using multiple programs e.g. BLASTX, Prodigal (Hyatt et al, 2010), and Transeq (https://www.ebi.ac.uk/Tools/st/emboss_transeq/). Then, MSA of the predicted proteins was conducted using NCBI MSA Viewer and MUSCLE to predict potential TM and AK isoforms. There are two newly published partial sequences for TM on the NCBI (Accession numbers: QCI56568.1 and QCI56569.1). These sequences were used to validate the prediction of TM isoforms.

3.3.5. Proteomic analysis for the TM and AK allergens

3.3.5.1. Sample preparation and protein content determination

Adult house cricket (*Acheta domesticus*) was provided by Lee Palmer, University of Nebraska-Lincoln, Dept. of Food Science & Technology and stored at -20 °C. Samples were ground to a powder with a mortar and pestle at room temperature. The powder was

suspended in 1X phosphate buffered saline (PBS) at pH 7.4, 0.01M (diluted from 10X PBS stock solution, Fisher, product # BP3994, lot # 167923) at the ratio of 1:10 w/v and shaken for 2 hr to extract proteins. The suspensions were centrifuged at 12,000 rpm, 22°C for 30 min to remove particulates. Supernatants were collected, aliquoted and stored at -20°C. All extractions were aliquoted and stored at -20°C until further use.

3.3.5.2. Mass spectrometry

Sample extracts of the same type in PBS buffer were pooled, total protein concentrations were quantified using a 2D Quant-KitTM (GE Healthcare), and the samples were prepared for LC-MS/MS. Samples of three µg of protein were diluted to 10.5 µl in OptimaTM LC-MS grade water (Fisher), added 15 µl of 50 mM ammonium bicarbonate, and reduced with 1.5 µl of 100 mM dithiothreitol (DTT; ACROS, Fair Lawn, NJ). The mixtures were centrifuged (16 k x g for 5 minutes), heated (95 °C for 5 minutes), and put on ice (30 seconds). Samples were alkylated with 3 µl of 100 mM iodoacetamide (Sigma, St. Louis. MO) for 20 minutes in the dark at room temperature. One µl of 100 ng/µl trypsin was added to each sample and the mixtures were kept at 37°C for 3 hours, then 1 µl of trypsin was added and mixed and solutions were maintained at 30°C overnight. Supernatants were frozen at -20°C before analysis. SDS-PAGE was used to verify digestion of samples using an estimated 0.75 µg of each sample before and after digestion.

The protein digests were diluted with OptimaTM water with acidified glycogen phosphorylase to produce 100 fmol rabbit glycogen phosphorylase B per 200 ng tryptic peptides. Separation was accomplished by 1D liquid chromatography using a 5 µl injection of tryptic peptides in the Ultimate 3000RSL® liquid chromatography (UPLC)

system, equipped with a Hypersil Gold C18 1.9 µm, 100 x 1 mm analytical reversed phase column. Mass spectrometry was performed using a Thermo Fisher Q-exactive plusTM with the following MS settings: scan range resolution 70,000, 200-2000 m/z, min AGC target 1.5x103, intensity threshold 2.5x104 with MS2 acquisition of the 10 most abundant targets of each MS1 scan and a 3s dynamic exclusion window. The MS2 spectra were acquired using a resolution of 70,000 with an AGC target of 1x106, maximum fill time of 60 ms and a normalized collision energy of 27 mV. Data analysis was performed using PEAKS 8.5 (Bioinformatics Solutions Inc., Waterloo, ON, Canada) using the predicted protein sequences database for TM and AK.

3.3.6. Prediction of IgE epitopes for TM and AK

The amino acid sequences of TM and AK obtained from published transcriptomic studies were used to predict allergenic IgE epitopes of the cricket proteins. The complete sequences of TM and AK were analyzed using five immunoinformatics based computational approaches including SVMTrip (http://sysbio.unl.edu/SVMTriP/), BCPred (Chen et al, 2007), ABCpred (http://www.imtech.res.in/raghava/abcpred/dataset.html), BepiPred 1.0 (http://www.cbs.dtu.dk/services/BepiPred/), and Immunomedicine Group (http://imed.med.ucm.es/Tools/antigenic.pl). The predictions were performed using default parameters of each program. Results of the five immunoinformatics tools were compared, and allergenic epitopes predicted by no less than two tools are considered to be candidates. Additionally, published gE epitopes determined by others for TM and AK of other species were compared to predictions obtained here.

3.4. Results and Discussion

3.4.1. Literature Search and systematic review

Literature searches were conducted using four different search engines with the pre-determined queries “cricket AND (allerg* OR hypersensitiv* OR anaphyla* OR crossreactiv*) on March 2019. Pubmed, scopus, google scholar and web of science identified 32, 41, 61 and 38 articles respectively. Abstracts of all articles were reviewed, and duplicate entries removed, identifying 10 relevant articles. References, method of detection and experimental conclusion are summarized in Table 1. These studies suggested a positive proportion of reactions between IgE sera from patients with food allergy to crustaceans, cockroaches, HDM and insects, and protein extracts from crickets. Therefore, literature review showed a possibility of IgE cross-reactivity between cricket and other crustaceans.

3.4.2. Prediction of potential cross-reactivity for *Acheta domesticus*

The *Acheta domesticus* transcriptome was assembled using de novo assemblers. The assemblers (rnaSPAdes, Velvet and Trinity) have been used for transcriptome assembly as there is no published reference genome for *Acheta domesticus*. The assembly metrics are: 17496 contigs, largest contig 24020B, N50 2561B, N75 1442B, L50 3779, L75 7937, and GC% 39.71 for rnaSPAdes; for Velvet are 43409 contigs, largest contig 23557B, N50 2120B, N75 1293B, L50 10922, L75 21851, and GC% 40.42 and for Trinity are 26952 contigs, largest contig 15808B, N50 1667B, N75 960B, L50 6630, L75 13776, and GC% 40.73. The predicted transcripts were compared to tropomyosins and arginine kinases in AllergenOnline V19 database for prediction of potential cross-

reactivity. Matches of these proteins to allergens over the CODEX guidelines of :>35% identity over 80 AA were identified. Potential cross-reactivity between TM and AK in house cricket and AllergenOnline V19 was reported in Tables 2 and 3, respectively. Very high sequence identity matches (>80%) of cricket TM and AK were found to TM and AK in cockroaches, crustacean shellfish (e.g. Shrimp, Crab, and Lobster) and HDM. The predicted TM and AK transcripts from the three assemblers with their expression levels as TPM and FPKM values are shown in Tables 4, and 5, respectively. Several transcripts which had significant matches to TM and AK in different crustaceans were highly abundant. This suggests a high possibility of cross-reactivity between cricket, seafoods, cockroaches and HDM. Tropomyosin and arginine kinase have been described as highly cross-reactive allergens among crustacean, insects and mollusks (Binder et al, 2001; Yu et al, 2003). Since house dust mites and other arthropods (e.g. crickets), have a 75–85% TM sequence identities, IgE cross-reactive binding is expected for those with allergy to edible insect species and possibly clinical reactivity (Hall et al, 2018). In addition, two recent studies reported that cricket contains tropomyosin that may cross-react with those in shrimp and other crustaceans (Abdelmoteleb et al, 2018, Palmer et al, 2020).

3.4.3. Characterization of *Acheta domesticus* TM and AK isoforms

Multiple sequence alignments of the predicted transcripts of *Acheta domesticus* TM and AK from different assemblers were shown in Figure 1. Predicted transcripts from the three different assemblers clustered together into branched nodes with high sequence identity suggesting high confidence in our transcriptomic predictions for TM and AK in this cricket. Multiple sequence alignment of the translated proteins of TM and AK were conducted to predict the presence of isoforms. Sequence alignments of the translated

proteins of *A. domesticus* TM suggested the presence of multiple isoforms as shown in Figures 2. However, MSA for AK translated proteins suggested the presence of 1 or 2 isoforms as illustrated in Figures 3. The predicted isoforms for TM were compared to two recently published partial TM sequences, (Accession numbers: QCI56568.1 and QCI56569.1). Predictions for two TM isoforms were confirmed through pairwise alignments with the published TM sequences as shown in Figure 4. Predicted isoforms will be validated through proteomic analysis.

3.4.4. Proteomic evaluation of the predicted *A. domesticus* TM and AK sequences

The predicted protein sequences from Velvet and rnaSPAdes assemblers for TM and AK were used as a database to validate the transcriptomic predictions using proteomic data. The false discovery rate was statistically significant (0.1%). Both Velvet and rnaSPAdes identified 9 shared peptides in TM protein sequences (Table 6). For arginine kinases, Velvet identified 7 shared peptides in agreement with rnaSPAdes in addition to 2 unique peptides (Table 6). The predicted TM and AK protein sequences generated by Trinity have not been validated in proteomic analysis. Figures 5 and 6 illustrate the high-quality mass spectrum for the predicted TM and AK sequences.

3.4.5. Immunoinformatics predictions of possible epitopes of *A. domesticus* TM and AK and comparison to known IgE binding epitopes of shrimp.

Prediction of TM and AK epitopes were conducted using 5 different immunoinformatics software. Eight different epitopes have been identified in TM of shrimp (*Penaeus aztecus*). Epitopes 1, 7 and 8 are common in crustaceans only; epitopes 2, 3, 4 and 5 are common epitopes among crustaceans, insects, and mites, but not

mollusks; epitope 6 is highly conserved among crustaceans, mollusks, insects and mites (Ayusa et al, 2002). Figure 7 shows that epitopes 2, 3, 4, 5 and 6 have been characterized in the TM predicted sequence using at least 2 different immunoinformatics tools. Therefore, common IgE epitopes validate also the potential cross-reactivity. Immunoinformatics analysis didn't validate common IgE epitopes for AK using the five tools.

3.5. Conclusion

This study was primarily designed to examine the potential IgE cross-reactivity between TM and AK in crickets and those in crustaceans, cockroaches and HDM using transcriptomic and proteomic approaches. Basically, the literature review has reported a fair proportion of articles which showed some positive IgE immunoblotting assays. In addition, transcriptomic approaches have illustrated high sequence identities, high abundance of transcripts of TM and AK between cricket and cockroach, HDM, and other crustaceans. Transcriptomic and proteomic data suggested the presence of several TM isoforms. This is compatible with Palmer et al (2020) as they identified multiple TM isoforms using LC-MSMS and PCR mapping. The LC-MSMS confirmed the predicted amino acid sequences of a TM and AK through high quality mass spectroscopy data. Therefore, shrimp allergic patients may experience cross-reactions if they consume cricket or other insects. However, it is not yet possible to clearly determine risks for crustacean allergic subjects based only on the sequence information we generated in this study. There are still remaining questions behind this study: how can we protect people who have allergy to crustaceans; are there risks to those with airway allergy to cockroach

or house dust mites if they consume crickets, meal worms or other insects; and how should we educate and notify allergic consumers about those potential risks?

Table 1. Literature Search and Systematic Review

Study	Methodology	Results
Hall et al, 2018	IgE immunoblotting using shrimp allergic sera	Positive IgE serum results were observed between sera and tropomyosin in the unhydrolyzed cricket and crickets with 15-50% degree of hydrolysis. Negative results were observed in crickets with 60-85% degree of hydrolysis.
Kamemura et al, 2019	Enzyme-linked immunosorbent assay (ELISA) and IgE crosslinking-induced luciferase expression assay (EXiLE).	Potentiality of cricket allergens to induce allergic reactions in crustacean allergic patients
Francis et al, 2019	IgE serum testing	IgE reactivity against the cricket protein extracts showed two bands (40 and 14 kDa).
Pali-Schöll et al, 2019	IgE immunoblotting from patients allergic to crustaceans, house dust mite or flies	Positive IgE reactions to house cricket and desert locust proteins
Srinroch et al, 2015	IgE immunoblotting using sera from prawn-allergic patients and LC-MS/MS	Hexamerin1B (HEX1B) was identified as a novel allergen in field cricket (*Gryllus bimaculatus*). Cross-reactions was reported between arginine kinase in *G. bimaculatus* and *Macrobrachium* spp.
Prasad et al, 2009 R	2880 skin prick tests with 60 allergens were performed in 48 patients of nasobronchial allergy	Crickets represented 16.7% of the most common allergens.
Bagenstose et al, 1980	Skin tests, radioallergosorbent test (RAST), bronchial challenge, and in vitro histamine release	Skin tests suggested that cricket are potent allergens.
Lierl et al, 1994	Allergic asthmatic children Serum IgE testing from allergic asthamatic children	A significant proportion of allergic asthmatic children have positive IgE binding to protein extracts of cricket, moth, cricket, housefly, and grasshopper.
Berzhets et al, 2006	IgE testing using sera from 20 patients with severe and intermediate atopic asthma.	Allergens' extracts of cricket have specific binding activity.
Palmer et al, 2020	IgE Immunoblotting using sera/plasma from patients sensitized to shellfish or insects	Distinct patterns of cross-reactivity are reported with three insect species including cricket showing possible reactivity.

Table 2. Prediction of TM Sequence Identity Matches and Possible Cross-Reactivity to *Acheta domesticus* for those allergic to known allergens. The identity matches to cockroaches, crustacean shellfish (Shrimp, Crab, and Lobster), and HDM are shown.

Assembler	Species	% Sequence Identity	Alignment Length	E-Score
rnaSPAdes	gi|8101069|gid|353|tropomyosin [Blattella germanica]	98.095	105	1.64E-43
	gi|4378573|gid|211|tropomyosin [Periplaneta americana]	97.143	105	1.04E-42
	gi|19310971|gid|211|tropomyosin [Periplaneta fuliginosa]	92.857	84	4.63E-37
	gi|238477263|gid|1738|tropomyosin [Crangon crangon]	92.632	95	7.93E-34
	gi|7024506|gid|425|heat stable allergen tropomyosin [Charybdis feriatus]	92.632	95	8.35E-34
	gi|448278534|gid|2032|tropomyosin [Portunus pelagicus]	92.632	95	8.54E-34
	gi|151505279|gid|1111|tropomyosin [Scylla serrata]	92.632	95	1.12E-33
	gi|119674937|gid|1097|allergen tropomyosin [Portunus sanguinolentus]	92.632	95	1.12E-33
	gi|170791252|gid|1191|Lit v 1 tropomyosin [Litopenaeus vannamei]	92.632	95	1.23E-33
	gi|60892782|gid|911|tropomyosin [Penaeus monodon]	92.632	95	1.23E-33
	gi|2660866|gid|598|slow tropomyosin isoform [Homarus americanus]	92.632	95	1.54E-33
	gi|7024506|gid|425|heat stable allergen tropomyosin [Charybdis feriatus]	92.381	105	1.28E-40
	gi|448278534|gid|2032|tropomyosin [Portunus pelagicus]	92.381	105	1.43E-40
	gi|2353266|gid|493|tropomyosin [Dermatophagoides pteronyssinus]	90.385	104	3.44E-38
Velvet	gi|20387029|gid|628|troposmyosin [Lepisma saccharina]	95.652	92	4.43E-43
	gi|8101069|gid|353|tropomyosin [Blattella germanica]	94.737	133	5.71E-57
	gi|4378573|gid|211|tropomyosin [Periplaneta americana]	94.737	133	5.71E-57
	gi|19310971|gid|211|tropomyosin [Periplaneta fuliginosa]	94.737	133	6.71E-57
	gi|151505279|gid|1111|tropomyosin [Scylla serrata]	93.985	133	1.51E-57
	gi|119674937|gid|1097|allergen tropomyosin [Portunus sanguinolentus]	93.985	133	1.51E-57
	gi|170791252|gid|1191|Lit v 1 tropomyosin [Litopenaeus vannamei]	93.985	133	3.62E-57
	gi|238477263|gid|1738|tropomyosin [Crangon crangon]	93.985	133	3.43E-57
	gi|60892782|gid|911|tropomyosin [Penaeus monodon]	93.985	133	3.62E-57
	gi|448278534|gid|2032|tropomyosin [Portunus pelagicus]	93.985	133	2.25E-57
	gi|134305330|gid|941|tropomyosin [Eriocheir sinensis]	93.985	133	1.54E-57
	gi|151505281|gid|1097|tropomyosin [Portunus trituberculatus]	93.985	133	1.56E-57
	gi|2353266|gid|493|tropomyosin [Dermatophagoides pteronyssinus]	85.714	133	1.76E-50
	gi|2660866|gid|598|slow tropomyosin isoform [Homarus americanus]	85.052	194	1.31E-89
Trinity	gi|20387029|gid|628|troposmyosin [Lepisma saccharina]	97.203	143	2.06E-68
	gi|4378573|gid|211|tropomyosin [Periplaneta americana]	96.135	207	6.48E-105
	gi|19310971|gid|211|tropomyosin [Periplaneta fuliginosa]	95.804	143	2.86E-66
	gi|8101069|gid|353|tropomyosin [Blattella germanica]	94.686	207	3.06E-103
	gi|151505281|gid|1097|tropomyosin [Portunus trituberculatus]	93.706	143	3.61E-65
	gi|151505279|gid|1111|tropomyosin [Scylla serrata]	93.706	143	3.65E-65
	gi|119674937|gid|1097|allergen tropomyosin [Portunus sanguinolentus]	93.706	143	3.65E-65
	gi|7024506|gid|425|heat stable allergen tropomyosin [Charybdis feriatus]	93.007	143	1.24E-65
	gi|2660866|gid|598|slow tropomyosin isoform [Homarus americanus]	93.007	143	4.49E-65
	gi|170791252|gid|1191|Lit v 1 tropomyosin [Litopenaeus vannamei]	93.007	143	5.83E-65
	gi|60892782|gid|911|tropomyosin [Penaeus monodon]	93.007	143	5.83E-65
	gi|448278534|gid|2032|tropomyosin [Portunus pelagicus]	93.007	143	5.96E-65
	gi|238477263|gid|1738|tropomyosin [Crangon crangon] [Crangon crangon]	93.007	143	6.5E-65
	gi|2440053|gid|493|tropomyosin [Dermatophagoides pteronyssinus]	85.211	142	4.46E-58

Table 3. Prediction of AK Sequence Identity Matches and Possible Cross-Reactivity to *Acheta domesticus* for those allergic to known allergens. The identity matches to cockroaches, crustacean shellfish (Shrimp, Crab, and Lobster), insects and HDM are shown.

Assembler	Species	% Sequence Identity	Alignment Length	E-Score
rnaSPAdes	gi|86160922|gid|1303|arginine kinase [Blattella germanica]	98.913	92	7.44E-63
	gi|167782135|gid|926|arginine kinase [Periplaneta americana]	97.826	92	6.61E-62
	gi|15886861|gid|264|arginine kinase [Plodia interpunctella]	90	150	6.05E-90
	gi|375298903|gid|1958|arginine kinase [Scylla paramamosain]	89.103	156	1.30E-101
	gi|82658675|gid|1284|arginine kinase [Bombyx mori]	88	150	8.29E-90
	gi|115492980|gid|896|arginine kinase [Litopenaeus vannamei]	87.179	156	4.98E-99
	gi|308154236|gid|209|RecName: Full=Arginine kinase; Short=AK; AltName: Allergen=Pen m 2 [Penaeus monodon]	86.538	156	7.55E-99
	gi|37785884|gid|949|arginine kinase [Dermatophagoides farinae]	81.41	156	2.18E-90
	gi|238477265|gid|1739|arginine kinase [Crangon crangon]	85.256	156	7.36E-97
Velvet	gi|221602737|gid|1303|arginine kinase [Blattella germanica]	93.675	332	0
	gi|50428904|gid|926|arginine kinase [Periplaneta americana]	93.675	332	0
	gi|82658675|gid|1284|arginine kinase [Bombyx mori]	86.667	90	5.75E-55
	gi|15886861|gid|264|arginine kinase [Plodia interpunctella]	86.667	90	1.58E-53
	gi|375298903|gid|1958|arginine kinase [Scylla paramamosain]	86.232	138	2.24E-73
	gi|238477265|gid|1739|arginine kinase [Crangon crangon]	82.609	138	2.16E-70
	gi|308154236|gid|209|RecName: Full=Arginine kinase; Short=AK; AltName: Allergen=Pen m 2 [Penaeus monodon]	82.583	333	0
	gi|115492980|gid|896|arginine kinase [Litopenaeus vannamei]	82.583	333	0
	gi|37785884|gid|949|arginine kinase [Dermatophagoides farinae]	77.273	88	9.76E-52
Trinity	gi|15886861|gid|264|arginine kinase [Plodia interpunctella]	95.062	81	2.51E-42
	gi|82658675|gid|1284|arginine kinase [Bombyx mori]	95.062	81	2.16E-40
	gi|221602737|gid|1303|arginine kinase [Blattella germanica]	93.539	356	0
	gi|50428904|gid|926|arginine kinase [Periplaneta americana]	93.539	356	0
	gi|115492980|gid|896|arginine kinase [Litopenaeus vannamei]	87.654	81	1.28E-40
	gi|375298903|gid|1958|arginine kinase [Scylla paramamosain]	87.654	81	5.28E-40
	gi|308154236|gid|209|RecName: Full=Arginine kinase; Short=AK; AltName: Allergen=Pen m 2 [Penaeus monodon]	86.42	81	3.89E-40
	gi|37785884|gid|949|arginine kinase [Dermatophagoides farinae] [Dermatophagoides farinae]	86.42	81	1.01E-37
	gi|238477265|gid|1739|arginine kinase [Crangon crangon]	85.185	81	8.06E-39

Table 4. Expression Levels of TM Transcripts in *Acheta domesticus.* The TM expression levels were quantified as fragments per kilobase million (TPM) and fragments per kilobase million mapped reads (FPKM). Also, the sequence identity, alignment length and E-score for the best match for each transcript are illustrated.

Assembler	**Predicted Transcripts**	Best TM % Sequence Identity	Alignment Length	E-Score	**TPM**	**FKPM**
rnaSPAdes	NODE_23703_length_316_cov_2940.57_ID_66963	98.095	105	1.64E-43	**1042.97**	**1832.18**
	NODE_23791_length_314_cov_2902.65_ID_67023	98.077	104	2.18E-43	**1151.78**	**2023.32**
	NODE_25225_length_287_cov_3007.68_ID_66797	97.895	95	1.57E-36	**0**	**0**
	NODE_10996_length_969_cov_7.4156_ID_21523	95.423	284	2.06E-156	**4.95**	**8.7**
	NODE_27198_length_254_cov_2271.01_ID_68121	94.048	84	2.10E-38	**330.46**	**580.51**
	NODE_13715_length_714_cov_858.031_ID_34593	93.204	103	1.98E-39	**47.64**	**83.69**
	NODE_9246_length_1216_cov_2130.2_ID_65961	92.683	164	2.61E-86	**891.47**	**1566.04**
	NODE_25475_length_283_cov_2627.72_ID_55179	91.209	91	7.60E-53	**711.53**	**1249.94**
	NODE_18252_length_469_cov_1508.86_ID_48849	82.222	90	5.39E-38	**792.85**	**1392.8**
	NODE_15339_length_606_cov_5.8918_ID_30411	81.095	201	2.20E-109	**3.05**	**5.35**
	NODE_26599_length_263_cov_3.75652_ID_54239	67.816	87	1.23E-33	**6.2**	**10.9**
Velvet	Locus_147_Transcript_1/28_Confidence_0.013_Length_279	95.652	92	4.43E-43	**0**	**0**
	Locus_147_Transcript_20/28_Confidence_0.187_Length_664	95.489	133	3.29E-58	**121.73**	**121.71**
	Locus_147_Transcript_16/28_Confidence_0.213_Length_1513	93.985	133	3.36E-54	**60.5**	**60.49**
	Locus_147_Transcript_2/28_Confidence_0.213_Length_837	93.814	194	3.01E-102	**1.45**	**1.45**
	Locus_147_Transcript_23/28_Confidence_0.067_Length_913	93.133	233	4.56E-122	**7.01**	**7.01**
	Locus_147_Transcript_6/28_Confidence_0.387_Length_649	92.893	197	1.26E-97	**2303.14**	**2302.8**
	Locus_147_Transcript_24/28_Confidence_0.000_Length_334	90.909	110	3.63E-45	**0**	**0**
	Locus_16_Transcript_4/15_Confidence_0.125_Length_266	90.909	88	3.76E-51	**0**	**0**
	Locus_16_Transcript_3/15_Confidence_0.125_Length_247	90.244	82	2.17E-47	**4.75**	**4.75**
	Locus_147_Transcript_9/28_Confidence_0.360_Length_868	89.894	188	4.10E-85	**4.98**	**4.97**
	Locus_147_Transcript_8/28_Confidence_0.307_Length_1134	86.577	149	4.84E-68	**20.19**	**20.19**
	Locus_16_Transcript_7/15_Confidence_0.500_Length_1261	83.2	250	2.73E-129	**757.25**	**757.14**
	Locus_16_Transcript_5/15_Confidence_0.625_Length_1003	79.2	250	1.19E-122	**514.54**	**514.47**
	Locus_16_Transcript_6/15_Confidence_0.625_Length_1534	79.2	250	5.18E-120	**838.38**	**838.26**
	Locus_147_Transcript_27/28_Confidence_0.133_Length_689	78.571	182	8.95E-94	**7.59**	**7.59**
	Locus_147_Transcript_28/28_Confidence_0.027_Length_505	75.817	153	9.11E-75	**0**	**0**
	Locus_7010_Transcript_2/2_Confidence_0.538_Length_368	68.367	98	4.10E-39	**6.91**	**6.91**
	Locus_147_Transcript_12/28_Confidence_0.053_Length_1627	65.06	83	7.31E-25	**0**	**0**
	Locus_16_Transcript_2/15_Confidence_0.100_Length_316	63.855	83	9.91E-28	**418.6**	**418.53**
	Locus_16_Transcript_1/15_Confidence_0.125_Length_310	63.415	82	6.80E-30	**1.37**	**1.37**
Trinity	c68359_g1_i1	76.433	157	3.45E-78	**2.7**	**2.05**
	c66363_g1_i1	72.093	86	2.59E-28	**25.62**	**19.53**
	c54565_g1_i1	71.084	83	1.06E-24	**34.83**	**26.54**
	c38530_g1_i1	65.854	82	5.16E-29	**1.01**	**0.77**

Table 5. Expression Levels of AK Transcripts in *Acheta domesticus.* The TM expression levels were quantified as fragments per kilobase million (TPM) and fragments per kilobase million mapped reads (FPKM). Also, the sequence identity, alignment length and E-score for the best match for each transcript are illustrated.

Assembler	Predicted Transcripts	Best AK % Sequence Identity	Alignment Length	E-Score	**TPM**	**FPKM**
rnaSPAdes	NODE 25380 length 284 cov 3.95618 ID 51755	98.913	92	7.44E-63	**5.09**	**8.94**
	NODE 6033 length 1878 cov 724.79 ID 15823	96.154	156	5.69E-109	**340.18**	**597.59**
	NODE 7681 length 1492 cov 862.831 ID 21967	90.659	182	8.85E-113	**435.61**	**765.24**
	NODE 16361 length 551 cov 2.60811 ID 32593	90	150	6.05E-90	**1.5**	**2.63**
	NODE 14285 length 672 cov 8.73709 ID 28223	86.905	168	8.52E-100	**4.15**	**7.29**
Velvet	Locus 2 Transcript 1586/2132 Confidence 0.003 Length 1210	93.399	303	0	**7.76**	**7.75**
	Locus 2 Transcript 1587/2132 Confidence 0.003 Length 1291	92.079	303	0	**1.65**	**1.65**
	Locus 2 Transcript 1588/2132 Confidence 0.003 Length 1255	91.935	310	0	**1.43**	**1.43**
	Locus 2 Transcript 1589/2132 Confidence 0.003 Length 2053	93.675	332	0	**2110.81**	**2110.5**
	Locus 2 Transcript 1590/2132 Confidence 0.004 Length 3241	93.675	332	0	**14.97**	**14.97**
	Locus 2 Transcript 1591/2132 Confidence 0.003 Length 2090	93.675	332	0	**3.87**	**3.87**
	Locus 2 Transcript 1592/2132 Confidence 0.002 Length 1482	93.399	303	0	**4.38**	**4.38**
	Locus 2 Transcript 1593/2132 Confidence 0.000 Length 386	90.361	83	2.21E-56	**2.38**	**2.38**
	Locus 2 Transcript 1594/2132 Confidence 0.003 Length 2523	93.675	332	0	**11.21**	**11.21**
	Locus 2 Transcript 1595/2132 Confidence 0.000 Length 680	87.778	180	8.53E-111	**0**	**0**
	Locus 2 Transcript 1596/2132 Confidence 0.002 Length 2137	93.675	332	0	**274.1**	**274.06**
	Locus 2 Transcript 1597/2132 Confidence 0.000 Length 293	91.209	91	8.58E-55	**0**	**0**
	Locus 2 Transcript 17/2132 Confidence 0.001 Length 266	93.182	88	4.23E-62	**0**	**0**
Trinity	c36629 g1 i1	95.062	81	2.51E-42	**1372.36**	**1045.97**
	c1045 g1 i1	93.539	356	0	**5**	**3.81**
	c23233 g1 i1	88.235	187	1.42E-116	**2.28**	**1.74**
	c43648 g1 i1	86.905	168	4.44E-100	**4.38**	**3.34**
	c45134 g1 i1	93.396	106	5.41E-75	**1.18**	**0.9**
	c7742 g1 i1	93.539	356	0	**1382.79**	**1053.92**

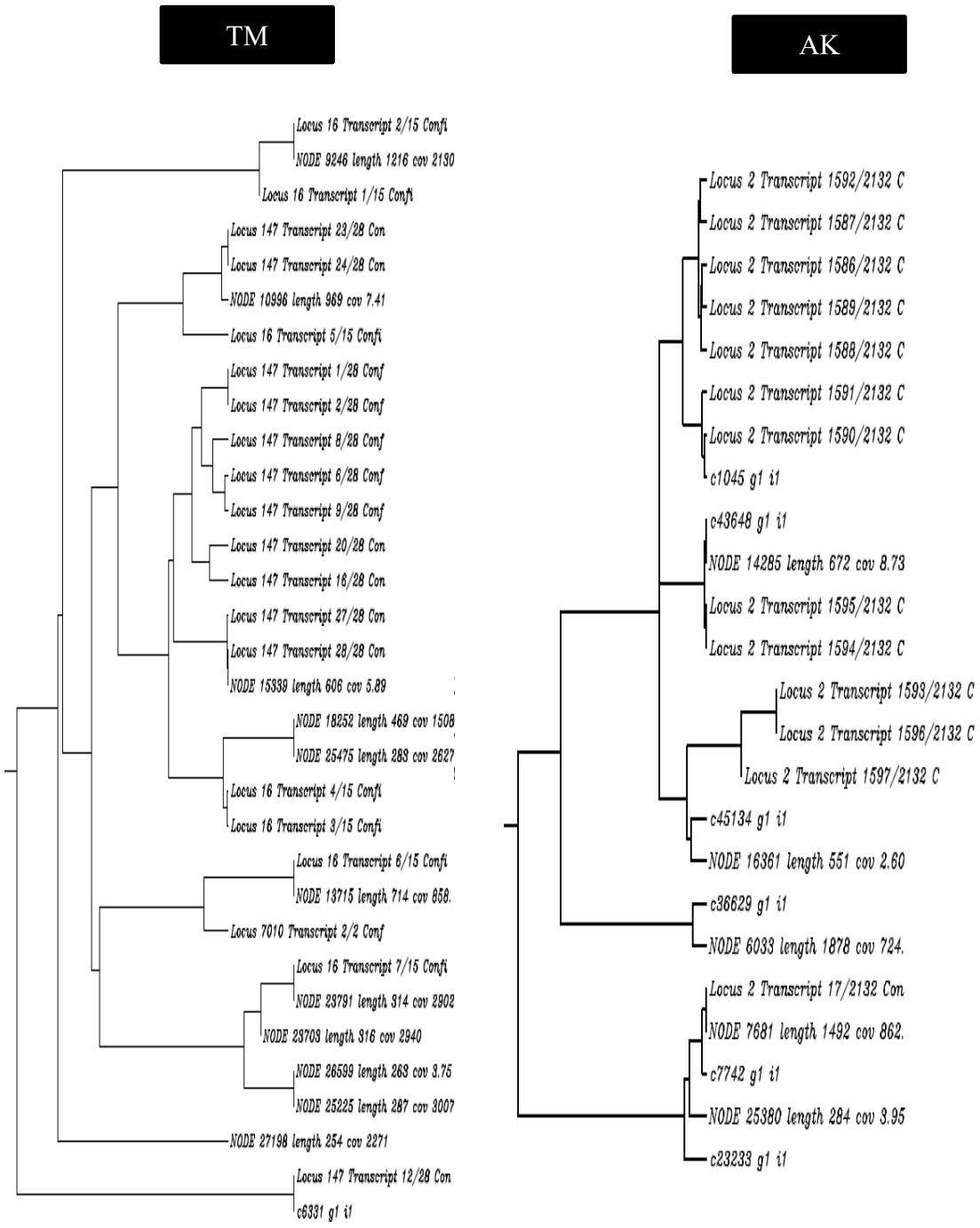


Figure 1. Multiple Sequence Alignments of The Predicted Transcripts for TM And AK. Transcripts from different assemblers are clustered in closely related branches suggesting high quality transcriptomic predictions.

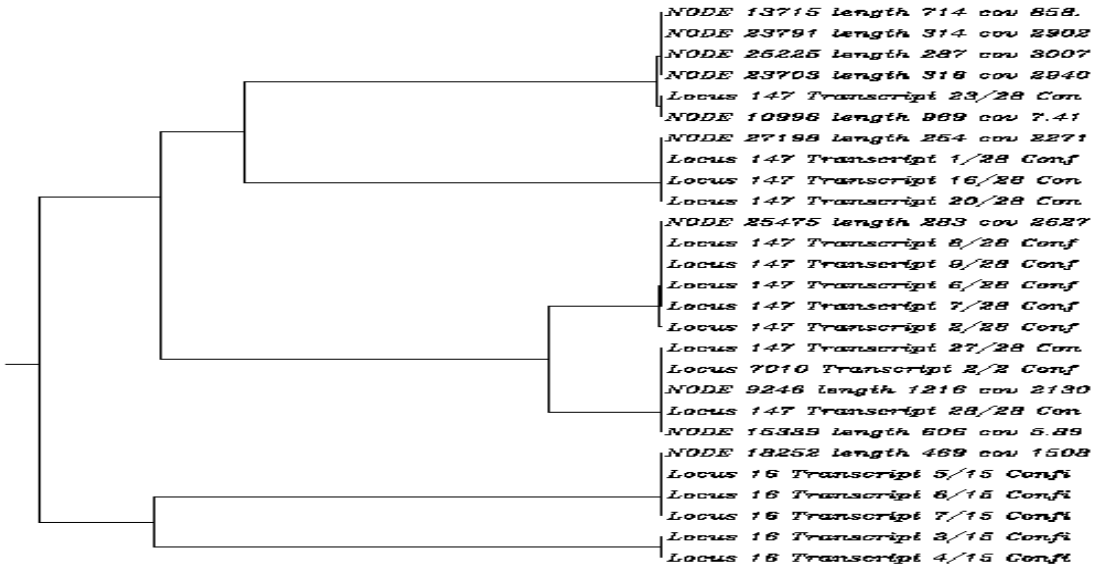


Figure 2. Multiple Sequence Alignment of The Predicted Proteins for TM. Five to six clustered branches with high scoring identity matches validate the presence of multiple TM isoforms.

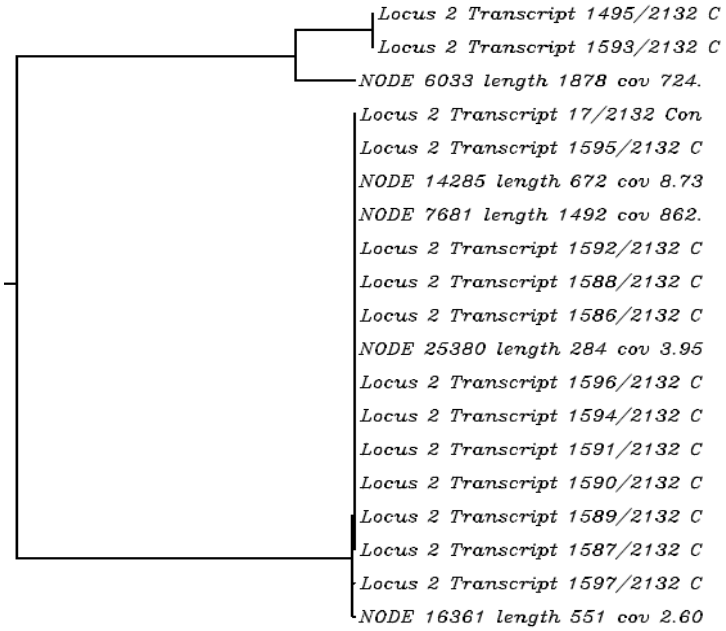


Figure 3. Multiple Sequence Alignment of The Predicted Proteins for AK. One or two clustered branches with high scoring identity matches validate the presence of 1-2 AK isoforms.

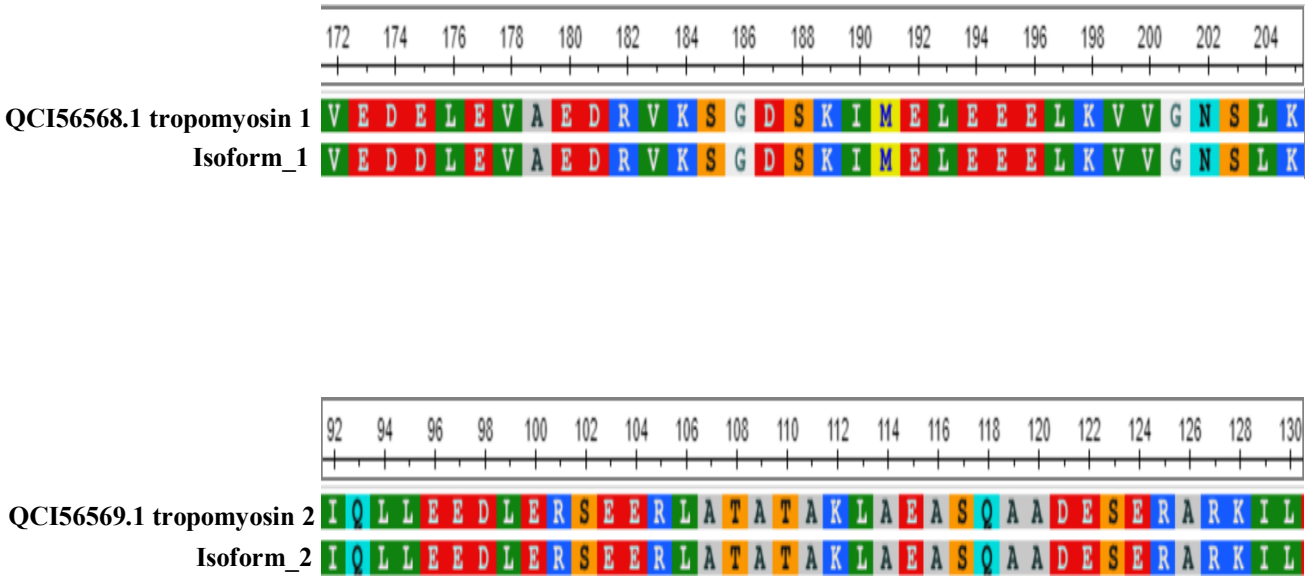


Figure 4. Validation of Two TM Isoforms Using Pairwise Alignment to Published TM Partial Sequences.

Table 6. Validation of The Predicted TM and AK Sequences in *Acheta domesticus* Using LC-MSMS. The translated protein sequences from Velvet and rnaSPAdes assemblers were validated by identifying peptides generated by mass spectroscopy (using PEAKS software).

Tropomyosins	Velvet	rnaSPAdes
#Peptides ID	9	9
#Peptides shared	9	9
#Peptides unique	0	0

Arginine kinases	Velvet	rnaSPAdes
#Peptides ID	9	7
#Peptides shared	7	7
#Peptides unique	2	0

False Discovery Rate: 0.1%

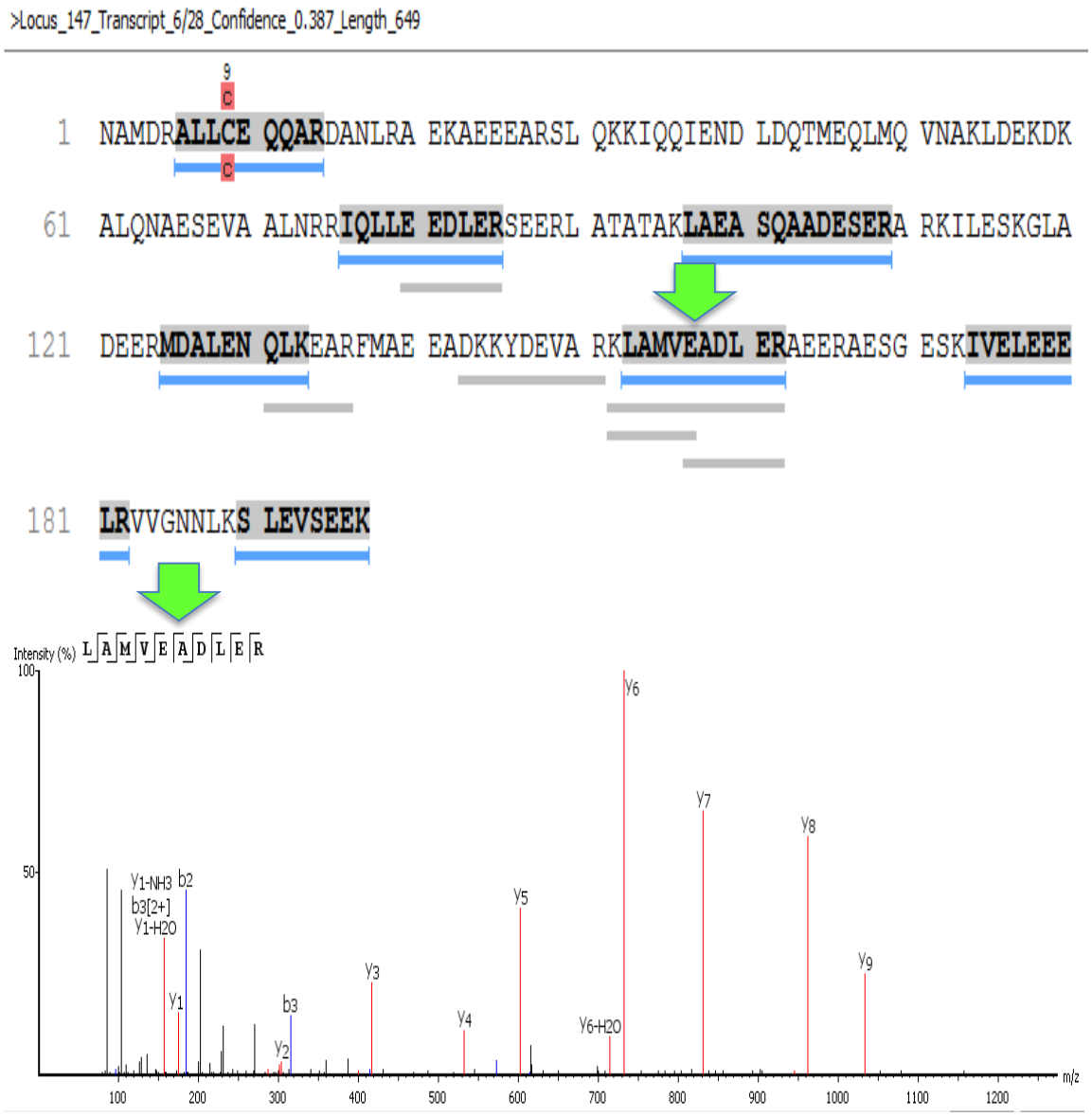


Figure 5. Validation of Predicted *A. domesticus* TM Sequence Using LC-MSMS. The identified peptides were detected in the upper TM sequence (FDR = 0.1). The figure shows the electrospray mass spectrum of the highlighted peptides

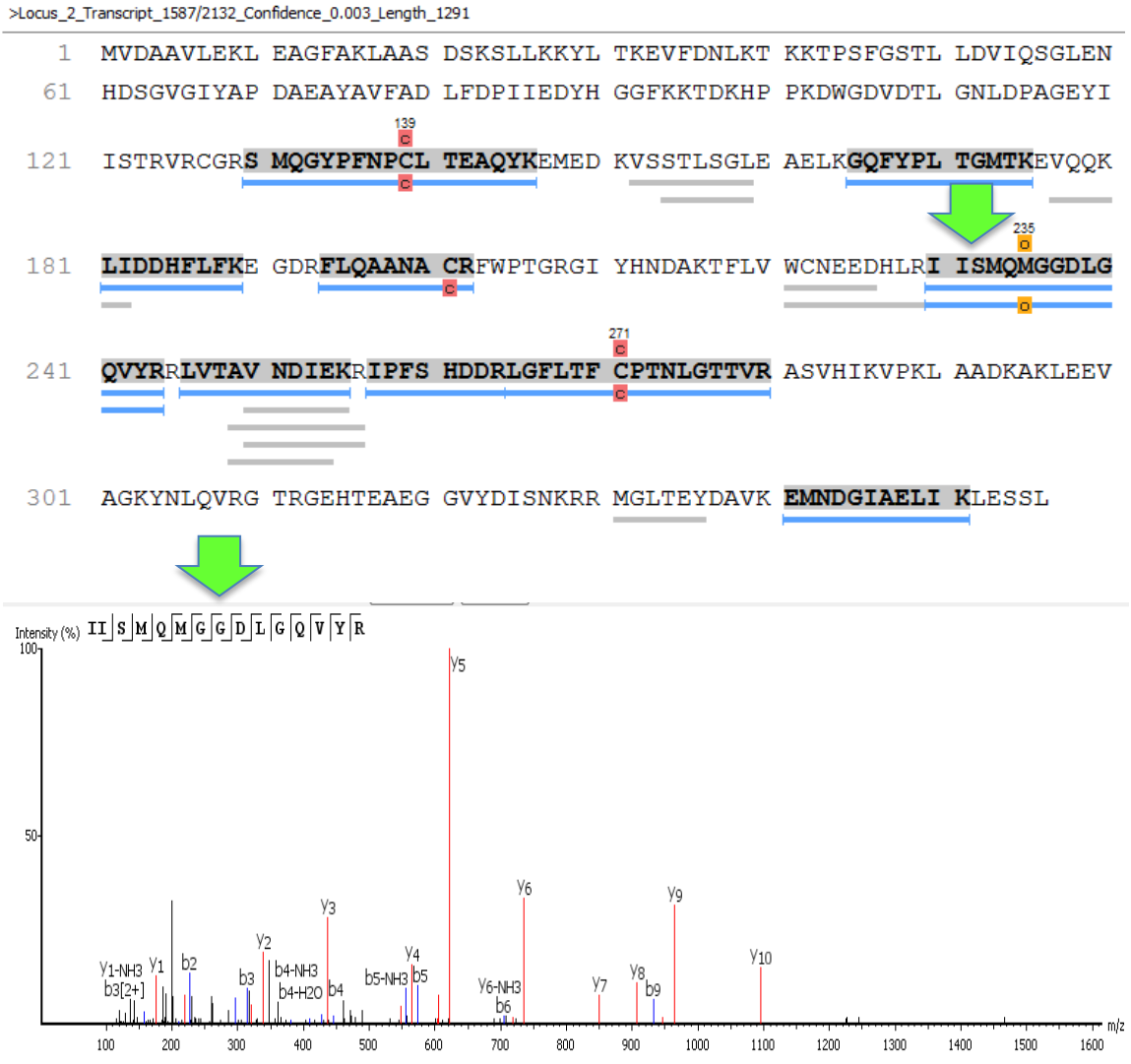


Figure 6. Validation of Predicted *A domesticus* AK Sequence using LC-MSMS. The identified peptides were detected in the upper AK sequence (FDR = 0.1). The figure shows the electrospray mass spectrum of the highlighted peptide.

Shrimp (*Penaeus aztecus*) TM Pen a 1 Epitopes

Epitope 1	Epitope 2	Epitope 3	Epitope 4	Epitope 5	Epitope 6	Epitope 7	Epitope 8
43-55 (VHNLQKRMQQLEN)	87-101 (ALNRRIQLLEEDLER)	137-141 (DEERM)	144-151 (LENQLKEA)	187-197 (ESKIVELEEEL)	249-259 (LQKEVDRLEDEL)	266-273 (KYKSITDE)	273-281 (ELDQTFSEL)

> House Cricket (*Acheta domesticus)* Predicted TM

MDAIKKKMQAMKLEKDNAMDRALLCEQQARDANLRAEKAEEEARSLQKKIQTIENELDQT

QEQLGQVNAKLEEKDKALQLAESEVA**ALNRRIQlleedler**seerlATATAKLAEASQAADESERQR

KILENRSLA**DEERM**DA**LENQLKEA**RFLAEEADKKYDEVARKLAMVEADLeraeeraesg**esKIVE**

LEEELRVVGNNLKSLEVSEEKANQREEEYKQQIKNLTTRLKeaearaefaerSVQK**LQKEVDRLED**

ELVHEKEKYKFICDDLDMTFTELIGN

Epitopes	SVMTriP	BCPREDS	ABCpred	BepiPred-2.0	Immunomedicine group
ALNRRIQlleedler	☑	☑		☑	☑
DEERM	☑		☑	☑	
LENQLKEA	☑		☑		☑
ESKIVELEEEL		☑		☑	☑
LQKEVDRLEDEL	☑			☑	☑

Figure 7. Prediction of Common IgE Epitopes Between Shrimp and House Cricket (*Acheta domesticus*). Epitopes 2, 3, 4, 5 and 6 were common between shrimp and cricket**.**

3.6. References

Abdelmoteleb, M., Palmer, L., Pavlovikj, N., Marsh, J., Johnson, P., Goodman, R. E., 2018. Bioinformatics and Proteomics Evaluations to Consider IgE Binding Assays for Potential Cross-Reactivity between House-Cricket (Acheta domesticus) Used in Food, Crustaceans and Cockroaches. J. Allergy Clin. Immunol. Suppl. 141(2):AB263.

Andrews, S., 2010. FastQC A Quality Control tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

ANSES 2015. Opinion of the French Agency for Food, Environmental and Occupational Health and Safety (ANSES) on the use of insects as food and feed and the review of scientific knowledge on the health risks related to the consumption of insects. ANSES opinion request No. 2014-SA-0153. ANSES: Maisons-Alfort Cedex. ANSES/PR1/9/01-06, 38 pp.

Ayuso, R., Lehrer, S. B., Reese, G., 2002. Identification of continuous, allergenic regions of the major shrimp allergen Pen a 1 (tropomyosin). Int. Arch. Allergy Immunol. 127(1):27−37.

Ayuso, R., Reese, G., Leong-Kee, S., Plante, M., & Lehrer, S. B., 2002. Molecular basis of arthropod cross-reactivity: IgE-binding cross-reactive epitopes of shrimp, house dust mite and cockroach tropomyosins. Int. Arch. Allergy Immunol. 129(1):38–48.

Ayuso, R., Sanchez-Garcia, S., Lin, J., Fu, Z., Ibanez, M. D., Carrillo, T., Blanco, C., Goldis, M., Bardina, L., Sastre, J., Sampson, H. A., 2010. Greater epitope recognition of shrimp allergens by children than by adults suggests that shrimp sensitization decreases with age. J. Allergy Clin. Immunol. 125(6): 1286−1293.e3.

Bagenstose, A.H., Mathews, K.P., Homburger, H.A., Saaveard-Delgado, A.P., 1980 Inhalant allergy due to crickets. J Allergy Clin Immunol., 65(1):71–74.

Bakevich, A., Nurk. S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al, 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455-477.

Berzhets, V.M., Radikova, O.V., Khlgatian, S.V., Berzhets, A.I., et al., 2006. Evaluation of Specific Activity of Preparations of Allergens from Synanthropic Insects. Epidemiol Immunobiol, (7):74–78

Bian, H., Reidhaar-Olson JF, Hammer J., 2003. The use of bioinformatics for identifying class II restricted T-cell epitopes. Methods 29(3):299−309.

Broekman, H.C., Knulst, A.C., de Jong, G., Gaspari, M., et al., 2017. Is mealworm or shrimp allergy indicative for food allergy to insects? Mol Nutr Food Res. 61(9)

Chan, T.F., Ji, K.M., Yim, A.K., Liu, X.Y., Zhou, J.W., Li, R.Q., et al., 2015. The draft genome, transcriptome, and microbiome of Dermatophagoides farinae reveal a broad spectrum of dust mite allergens. J Allergy Clin Immunol. 135(2):539–48.

Chen, J., Liu, H., Yang, J., Chou, K., 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 33(3):423–428.

Cunha, L.M., Gonc¸alves, A.T. S., Varela, P., Hersleth, M. et al., 2015. Adoption of insects as a source for food and feed production: a cross-cultural study on determinants of acceptance. "11th Symposium Pangborn Sensory Science ", Gothenburg, Sweden.

Cunha, L.M., Moura, A.P., Costa-Lima, R., 2014. Consumers' associations with insects in the context of food consumption: comparisons from acceptors to disgusted. "Insects to Feed the World", Netherlands.

Drinnenberg, I.A., deYoung, D., Henikoff, S., & Malik, H.S., 2014.Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. eLife 3(e03676).

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

EFSA, 2015. Risk profile related to production and consumption of insects as food and feed. EFSA J. 13: 4257–4317.

Francis, F., Doyen, V., Debaugnies, F., et al., 2019. Limited cross reactivity among arginine kinase allergens from mealworm and cricket edible insects. Food Chem., 276:714–718.

Fu, L., Wang, J., Ni, S., Wang, C., Wang Y., 2018. Identification of Allergenic Epitopes and Critical Amino Acids of Major Allergens in Chinese Shrimp (Penaeus chinensis) by Immunoinformatics Coupled with Competitive-Binding Strategy. J. Agric. Food Chem. 66:2944−2953.

Goodman, R.E., Ebisawa, M., Ferreira, F., Sampson, H.A., van Ree, R., Vieths, S., Baumert, J.L., Bohle, B., Lalithambika, S., Wise, J., Taylor, S.L., 2016. AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross-reactivity. Mol Nutr Food Res. 60(5):1183-1198.

Grabherr, M. G., Haas, B. J., Yassour, M., et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. biotech. 29:7.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8):1072–1075.

Hall, F.G., Jones, O.G., O'Haire, M.E., Liceaga, A.M., 2017. Functional properties of tropical banded cricket (Gryllodes sigillatus) protein hydrolysates. Food Chem. 224:414–422.

Hall, F., Johnson, P.E., Liceaga, A., 2018. Effect of enzymatic hydrolysis on bioactive properties and allergenicity of cricket (Gryllodes sigillatus) protein. Food Chem. 262:39-47.

Hopp, T.P., Woods, K.R., 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U. S. A. 78 (6):3824−3828.

Hyatt, D., Chen G.L., Locascio, P.F, et al., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(1):119.

Kamemura, N., Sugimoto, M., Tamehiro, N., et al., 2019. Cross-allergenicity of crustacean and the edible insect Gryllus bimaculatus in patients with shrimp allergy. Mol Immunol., 106:127–134.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157(1):105−32.

Li, B., Dewey, C., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.

Li, G. F., Wang, Y., Zhang, Z.-S., Wang, X.-J., Ji, M.-J., Zhu, X., Liu, F., Cai, X.-P., Wu, H.-W., Wu, G.-L., 2005. Identification of Immunodominant Th1-type T cell Epitopes from Schistosoma japonicum 28 kDa Glutathione-S-transferase, a Vaccine Candidate. Acta Biochim. Biophys. Sin. 37(11):751−758.

Lierl, M.B., Riordan, M.M., Fischer, T.J., 1994. Prevalence of insect allergen-specific IgE in allergic asthmatic children in Cincinnati, Ohio. Ann Allergy, 72(1):45–50.

Matsuo, H., Yokooji, T., Taogoshi, T., 2015. Common food allergens and their IgE-binding epitopes. Allergol. Int. 64(4):332−343.

Motoyama, K., Suma, Y., Ishizaki, S., Nagashima, Y., Shiomi, K., 2007. Molecular cloning of tropomyosins identified as allergens in six species of crustaceans. J. Agric. Food Chem. 55(3):985−91.

Ogburn, R.N., Randall, T.A., Xu. Y., Roberts, J.H., Mebrahtu, B., Karnuta, J.M., Rider, S.D., Kissling, G.E., Mueller G.A., 2017. Are dust mite allergens more abundant and/or more stable than other Dermatophagoides pteronyssinus proteins? J Allergy Clin Immunol. 139(3):1030-1032.

Pali-Schöll, I., Meinlschmidt, P., Larenas-Linnemann, D., et al., 2019. Edible insects: Cross-recognition of IgE from crustacean- and house dust mite allergic patients, and reduction of allergenicity by food processing. World Allergy Organ J., 12(1):100006.

Palmer, L.K., Marsh, J.T., Lu, M., Goodman, R.E., Zeece, M.G., Johnson, P.E., 2020. Shellfish Tropomyosin IgE Cross-Reactivity Differs Among Edible Insect Species. Mol Nutr Food Res., e1900923.

Payne, C.L.R., Scarborough, P., Rayner, M., Nonaka, K., 2016. A systematic review of nutrient composition data available for twelve commercially available edible insects, and comparison with reference values. Trends. Food Sci. Technol. 47:69–77.

Pedrosa, M., Boyano-Martínez, T., García-Ara, C., & Quirce, S., 2015. Shellfish allergy: A comprehensive review. Clin. Rev. Allergy Immunol. 49(2):203–216.

Prasad, R., Verma, S.K., Dua, R., Kant, S., Kushwaha, R.A., Agarwal, S.P., 2009. A study of skin sensitivity to various allergens by skin prick test in patients of nasobronchial allergy. Lung India, 26(3):70–73.

Rumpold, B.A., Schluter, O.K., 2013. Nutritional composition and ¨ safety aspects of edible insects. Mol. Nutr. Food Res. 57:802–823.

Scealy, M., Mackay, I.R., Rowley, M.J. , 2006. Amino acids critical for binding of autoantibody to an immunodominant conformational epitope of the pyruvate dehydrogenase complex subunit E2: identification by phage display and site-directed mutagenesis. Mol. Immunol. 43(6):745−53.

Schluter, O., Rumpold, B., Holzhauser, T., Roth, A. et al, 2017. Safety aspects of the production of foods and food ingredients from insects. Mol. Nutr. Food Res. 61:1–14.

Schmieder R, Edwards R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27(6):863-864.

Shanti, K.N., Martin, B.M., Nagpal, S., Metcalfe, D.D., Rao, P.V., 1993. Identification of tropomyosin as the major shrimp allergen and characterization of its IgE-binding epitopes. J. Immunol. 151(10):5354−5363.

Sheehan, W.J., Phipatanakul, W., 2015. Difficult to control asthma: epidemiology and its link with environmental factors. Curr. Opin. Allergy Clin. Immunol. 15:397–401.

Shen, H.W., Cao, M.J., Cai, Q.F., Ruan, M.M., Mao, H.Y., Su, W. J., and Liu, G.M., 2012. Purification, cloning, and immunological characterization of arginine kinase, a novel allergen of Octopus fangsiao. J. Agric. Food Chem. 60:2190-2199.

Shen, Y., Cao, M.J., Cai, Q.F., Su, W.J., Yu, H.L., Ruan, W.W., and Liu, G.M., 2011. Purification, cloning, expression and immunological analysis of Scylla serrata arginine kinase, the crab allergen. J. Sci. Food Agri. 91:1326-1335.

Srinroch, C., Srisomsap, C., Chokchaichamnankit, D., Punyarit, P., Phiriyangkul, P., 2015. Identification of novel allergen in edible insect, Gryllus bimaculatus and its cross-reactivity with Macrobrachium spp. Allergens. Food Chem., 184:160–166.

Stanhope, J., Carver, S., Weinstein, P., 2015. The risky business of being an entomologist: a systematic review. Environ. Res. 140:619–633.

van der Spiegel, M., Noordam, M.Y., van der Fels-Klerx, H. J., 2013. Safety of novel protein sources (insects, microalgae, seaweed, duckweed, and rapeseed) and legislative aspects for their application in food and feed production. Compr. Rev. Food. Sci. Food Saf. 12:662–678.

van Huis, A., 2016. Edible insects are the future? Proc. Nutr. Soc. 75:294–305.

Vogel, C., Marcotte, E.M., 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 13(4):227–32.

Wangorsch, A., Ballmer-Weber, B.K., Rö sch, P., Holzhauser, T., Vieths, S., 2007. Mutational epitope analysis and cross-reactivity of two isoforms of Api g 1, the major celery allergen. Mol. Immunol. 44(10):2518−27.

Yao, B., Zhang, L., Liang, S., Zhang, C., 2012. SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity. PLoS ONE 7(9): e45152.

Yu, C. J., Lin, Y.F., Chiang, B.L., Chow, L.P., 2003. Proteomics and immunological analysis of a novel shrimp allergen, Pen m 2. J. Immunol. 170(1):445−53.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Zheng, L. N., Lin, H., Pawar, R., Li, Z. X., Li, M. H., 2011. Mapping IgE binding epitopes of major shrimp (Penaeus monodon) allergen with immunoinformatics tools. Food Chem. Toxicol. 49(11):2954−60.

CHAPTER 4

BIOINFORMATICS ANALYSIS OF ALLERGENICITY, TOXICITY AND POTENTIAL HORIZONTAL GENE TRANSFER (HGT) TO MICROBES, OF A NUTRITIONALLY ENHANCED GENETICALLY ENGINEERED CANOLA

4.1.Abstract

A genetically engineered canola was produced by a biotechnology seed company as a nutritionally enhanced food product. The potential allergenicity, toxicity of proteins expressed by genes transferred into canola by genetic engineering and the potential horizontal gene transfer (HGT) of the transferred DNA in the canola to microbes of a nutritionally enhanced genetically engineered canola has been evaluated in this study. Potential open reading frames at the entire DNA insert in chromosome A02 and A05 of canola were predicted by predicting potential peptides in all six reading frames using the ORFfinder tool and a Perl program written by our lab. Bioinformatics evaluations of potential allergenicity were performed using www.AllergenOnline.org, version 18B using full-length FASTA and sliding 80mer FASTA searches; and the NCBI Protein database using BLASTP with a keyword limit (allergen). Evaluation of identity matches to toxins was accomplished using BLASTP with keyword search limits (toxic and toxin). The potential HGT from canola plant to microbes was analyzed by literature search and BLASTN search of T-DNA inserts in AO2 and AO5 against all published bacterial and archaeal genomes (including incomplete genomes) above EFSA guidelines. The lack of significant identity matches to allergens and toxins illustrated that if any of the ORFs were expressed as proteins, there is no reason to suspect that would elicit allergic reactions or would induce toxic responses. Literature searches did not show evidence of relevant cases for HGT from plants to microbes. Bioinformatics analysis raises no

concerns that the T-DNA inserts in transgenic canola would be transferrable to bacteria or archaea through HGT.

4.2. Introduction

An international agricultural biotechnology company, NuSeed collaborated with the Commonwealth Scientific and Industrial Research Organization (CSIRO) of Australia to develop an Omega-3 producing canola variety that has been genetically engineered (GE) to produce high levels of docosahexaenoic acid in seed. They asked for our help in performing specific bioinformatics evaluation for food safety. The transgenic canola event produced docosahexaenoic acid (DHA) representing up to 10% by weight in *Brassica napus* L. seeds. The transgenic canola was obtained through gene transformation of seven microalgae and yeast genes: Micromonas pusilla delta-6 desaturase, Pyramimonas cordata delta-5 elongase, Pavlova salina delta-5 desaturase, Pichia pastoris omega-3 desaturase, Pavlova salina delta-4 desaturase, Lachancea kluyveri delta-12 desaturase, Pyramimonas cordata delta-6 elongase (Micpu-Δ6D, Pyrco-Δ5E, Pavsa-Δ5D, Picpaω3D, Pavsa-Δ4D, Lackl-Δ12D and Pyrco-Δ6E, respectively) in the DHA biosynthetic pathway. The herbicide resistant gene Phosphinothricin N-Acetyltransferase (PAT), originally from *Streptomyces viridochromogenes* was used as a selectable marker (Colgrave et al, 2019). The developer had a full genome sequence determination of the transgenic plant using whole-genome and PCR-amplicon sequencing. The, DHA canola was characterized to have one insert on chromosome A02 and another insert on chromosome A05. Both inserts were required to achieve high DHA production in canola seeds. The insert on A02 had genes Micpu-Δ6D, Pyrco-Δ5E, Pavsa-Δ5D and Picpa-ω3D, and didn't have genes PavsaΔ4D, Lackl-Δ12D and Pyrco-Δ6E and

PAT; the insert replaced a 15-bp sequence (GTAGCACGACAAGTT) in the 3' UTR of a gene (HPP) located on chrUn_random of B. napus (Darmor) reference genome at position 118589903-118591677 and on chromosome A02 of B. rapa (Chiifu) reference genome at position 18569298-18571066. The insert on chromosome A05 had two complete eight-gene sets which formed a palindromic structure with RB-LB:LB-RB orientation; the insert replaced a 20-bp sequence (CACGGTGGAGGTCACCATGT) in the 2nd exon of the PTI (Pto-Interacting Protein) gene located on chromosome A05 of B. napus (Darmor) reference genome at position 17267746-17270700 (Colgrave et al, 2019).

The methods used for the safety assessment are consistent with the process outlined by the CODEX Alimentarious Commission (2009) for evaluation of the potential safety of crops developed through genetic engineering. However, certain regulatory authorities typically request an evaluation of potential proteins (open reading frames) in all six potential coding frames throughout the inserted DNA segments. Recently, EFSA Panel 2017 published a new report describing new guidelines for the risk assessment and monitoring of genetically engineered (GE) plants that is incompletely described. The sequence identity between DNA inserts in the GE plant and the DNA present in microbial genomes, is required to define the probability for horizontal gene transfer from plants to microbes. Sequence similarity searches should be performed using BLAST or FASTA with listing all default parameters (*E*-value, word size, match/mismatch scores and gap costs). In addition, assessing HGT to bacteria or archaea can be conducted using complete bacterial and archaeal sub-division of databases e.g.

National Center for Biotechnology Information (NCBI) or the European Nucleotide Archive (ENA).

Homologous recombination becomes increasingly inefficient with decreasing length of sequences with high identity (de Vries and Wackernagel, 2002; Monier et al, 2007; EFSA, 2009; Overballe-Petersen et al, 2013). All matches with a threshold of 95% identity in alignments of at least 200 bp in length should be reported and considered further for the potential of HGT based on homologous recombination. The analysis should be presented in a graphic summary that depicts the results against the insert and flanking region, if relevant, with the information of its genetic elements. A summary table indicating all microbial target organisms for pair or higher order sequence stretches with the potential for double homologous recombination should be provided. The table should report the position of the alignment in the microbial target sequence, the length and percentage of identity, the annotation of the hit and the orientation of the alignment against the microbial target sequence.

Although there is no specific published evidence demonstrating horizontal DNA transfer from an eukaryotic plant chromosome to a microbe as noted in the 2017 EFSA recommendation, the EFSA is still asking developers to evaluate the inserted DNA sequence against the genomes of bacteria and archaea. The EFSA recognized that illegitimate recombination with transfer of recombinant DNA from the GE plant to microbes is extremely unlikely. Therefore, the focus is on homologous recombination (HR) where the insert DNA has high identity (>95% for 200 nucleotides) with microbial DNA, and similar match on the opposite side of recombinant DNA in the GE plant to consider HR as a possibility.

The objective of this study is to perform an evaluation of the potential allergenicity and toxicity of potential proteins that might be expressed from unexpected transcription and translation of DNA throughout the DNA inserts in canola line. The second objective is to perform an evaluation of the potential for horizontal gene transfer (HGT) from the DHA expressing canola line to living microbial organisms based on knowledge of DNA transfer mechanisms and identical DNA sequence segments identified between the insert DNA and that of known microbial genomic sequences.

4.3.Materials and methods

4.3.1.Prediction of hypothetical Open Reading Frames (ORFs)

An over-prediction method has been used assuming that the longest Start-to-Stop segments or Stop-to-Stop segments might be transcribed and translated in arriving at a prediction of all possible ORFs of 30 or more amino acids as potential proteins. The choice of Start-to-Stop (ATG to any of the three TGA, TAG or TGA) and Stop-to-Stop where the Stop (TGA, TAG or TGA) are converted to ATG as a pseudo-start, and the three Stop codons end the segment. The two methods are expected to identify different numbers of potential peptides. Potential ORFs were identified using the ORFfinder tool on the website of the NCBI (https://www.ncbi.nlm.nih.gov /orffinder/) with the DNA sequences identified by the sponsor as the inserted DNA in chromosomes AO2 (12,110 bp) and AO5 (46,614 bp). Alternatively, a Perl program has been used as a different ORF finding program to predict consecutive segments of 30 or more amino acid or longer translation products in a batch-wise fashion.

4.3.2. FASTA3 overall search of AllergenOnline.

The sequences of hypothetical peptides for A02 and A05 were searched for identity matches against AllergenOnline.org version 18B with *E*-scores of 10 and of 1 using FASTA, version 35.04 (15 January 2009). This pipeline was conducted using batch-mode at Holland Computing Center, UNL. The sequence comparisons of hypothetical ORFs to potential and proven allergens was conducted in September 2018.

4.3.3.FASTA3 of AllergenOnline.org by 80 AA segments.

This search was used only for predicted potential ORFs that had identity matches >35% over segments of 80 amino acids as identified during FAST3 searches. The rationale for the short-window is that this might help in identifying structural motifs, much shorter than the intact protein, which might contain a conformational IgE binding epitope. The AllergenOnline.org search compensates for sequences shorter than 80 AA that might have very high identities over shorter segments by allowing FASTA alignments with 29 aa matches to be identified as well as full-80 AA alignments. The algorithm is explained on the website (www.allergenonline.org). The 80mer window search was conducted in September 2018.

4.3.4. Comparisons of ORFs with the NCBI Protein database by BLASTP

4.3.4.1. BLASTP of NCBI Entrez without a keyword limit.

The BLASTP program is available on the NCBI Entrez website (http://www.ncbi.nlm.nih.gov/BLAST/). The purpose of this BLASTP search was to compare the putative peptide sequences were all evaluated against all protein sequences to determine the prevalence of common homologues. The *E*-score is influenced by the length of the BLASTP alignment as well as identities of the AA sequences and the

scoring matrix (BLOSUM 62). Smaller *E*-scores represent more significant alignments. However, the length of alignment and percent identity are the most important estimators of significant matches. All BLASTP searches using different keywords were conducted in September 2018.

4.3.4.2. BLASTP of NCBI Entrez with "allergen" as keyword limit.

BLASTP search was used comparing the putative peptide sequences against the entire Entrez Protein database, with a limit option selected to query entries for "allergen", to align only with proteins identified as allergens or associated with allergy. The purpose of this BLASTP search is to ensure that a significant match with a newly discovered allergenic sequence that has not been entered into the current version of AllergenOnline.org is not overlooked. Evaluation of the *E*-score, the length of the alignment and the percent identity of any identified match is necessary to judge the significance of any alignment using BLASTP.

4.3.4.3. BLASTP of NCBI Entrez with "toxin" and "toxic" as keywords limit.

The purpose of this BLASTP search was to identify matches to known toxic proteins (toxins) and if alignments share significant identities, to determine potential risks that would require further testing for all putative peptides. There are no fully inclusive databases of toxins. Due to the widely diverse actions of toxins, there are no uniform databases of toxins. Using a keyword limit of "toxin" or "toxic" minimizes but does not eliminate false positive identities. Thus, matched sequences must be further evaluated by searching without keyword limits and sometimes searching with the matched "toxin" to consider exposure and evidence of toxicity.

4.3.4.4. Judging significance of bioinformatics results and performing secondary check for validity.

The very conservative bioinformatics estimate of potential allergenic cross-reactivity was defined by CODEX (2003/2009). That is based on an assumption that shared IgE binding and triggering basophil or mast cell triggering might be identified as a sequence that shares >35% identity over 80 amino acids with any known allergen. The overall FASTA alignment that was performed for each hypothetical ORF provides an overall identity match, length of alignment and *E*-score value. The 80mer window search on the public website www.allergenonline.org requires individual sequence input. It cannot be done efficiently in batch mode. The output tables of data from the batch-mode FASTA were inspected. Sequences that were longer than 80 AA, with >35% identity were taken as positive findings. Sequences that showed an apparent match of >35% identity over 80 or more amino acids were manually entered in the public online version of AllergenOnline.org to test whether the highest scoring 80mer had an identity >35%. For toxins the criteria are not as well defined. Toxic proteins have different modes of action, different AA sequence lengths and the ability for sequence similar (homologous proteins) to share toxicity can vary but is usually restricted to proteins having >50% identity. The findings must be considered relative to matches with other common proteins. Thus, the searches are done using the NCBI Protein database with keyword limits. For sequences that are longer than 30 amino acids and having >30% identity to a protein with a keyword association of toxin or toxic, the significance of the match can be judged by comparing the searched sequence (ORF) vs. the NCBI Protein database without any keyword. If there are a number of alignments with higher sequence

identities for the ORF with common proteins, the ORF is unlikely to represent a toxin. In addition, the sequence of the toxin/toxic associated protein can be compared with the NCBI database without keyword limit to judge whether the “toxin/toxic” protein has many high scoring matches to common proteins. On rare occasions, publications would have to be reviewed to evaluate potential toxicity of the keyword selected protein, toxicity of the source and reactivity of sequence similar proteins.

4.3.5. Horizontal gene transfer from plants to microbes

4.3.5.1. Scientific literature review on horizontal gene transfer from plants to microbes.

The PubMed database (http://www.ncbi.nlm.nih.gov/pubmed) maintained by the U.S. National Library of Medicine was used as the primary data source for scientific literature on HGT. The primary question is whether there is evidence of HGT from eukaryotes including plants to bacteria or archaea.

4.3.5.2. Sequence comparison to microbial genomic sequences

The complete and incomplete sequences of bacteria and archaea from the NCBI Prokaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/) from GenBank was loaded onto Holland Computing Center’s server at the University of Nebraska-Lincoln in October 2018. The complete genomic sequences are from annotated, reference genomes. The incomplete genomes are un-annotated, draft genomes of assembled contig datasets of chromosomes, plasmids and organelles. The incomplete dataset includes less-certain genomes and they have not been annotated for likely functionality of genes and proteins. Matches to the complete genomes are verifiable using

nucleotide BLASTN on the NCBI website and more likely to show possible HGT targets. The incomplete genomes have greater uncertainty and are not verifiable by normal BLASTN in the nucleotide database of NCBI.

In this study the entire DNA sequence of each of the two inserts in DHA canola were compared to bacterial complete and incomplete genome sequences and to complete and incomplete archaeal genome sequences available from NCBI on 19 October 2018. Default parameters of BLASTN were used for the alignments (*E*-score limit 10, Word size 10, Matrix match +1, mismatch -2, gap penalties existence = 0, extension = 2.5). BLASTN matched insert of at least 200 nucleotides were scored as positive, if their identities were 95% identical or more to the microbial DNA.

4.4. Results and discussion

4.4.1. Prediction of ORFs

Hypothetical ORF AA sequences were predicted from the full-DNA sequences in both the A02 and the A05 insertion sites. The number of Start-Stop and Stop-Stop ORFs were 47 and 90 in A02; while 90 and 368 in A05 respectively. The number of ORFs identified using Start-to-Stop predictions is lower than with Stop-to-Stop as expected. Clearly there are many potential ORFs that would be expected to be found in such large segments of DNA. One consideration is how many ORFs might be found, how many might be translated into peptides or proteins and how can we evaluate these for food safety? Most eukaryotic organisms have few overlapping expressed genes or produced proteins from the same linear segment of DNA. The inserted DNA in A02 and A05 are packed with 4 genes (A02) or 16 genes (8 genes in an end-to-end duplex) of A05. There is little chance that most potential ORFs in the inserted DNA to be expressed. Transcripts

(mRNA) are determined in cells based on transcription start and stop sites as well as other regulatory sequences. In addition, translation products (proteins) occur for reading frames that have appropriate ribosomal binding sites and other factors, severely limiting the number of proteins that occur from segments of linear DNA sequences.

4.4.2. Sequence comparison of the putative ORFs from DHA canola to allergens and toxins.

All putative peptide sequences (ORFs) were compared to known allergens using both a full-length FASTA alignment search for all sequences of 30 AA or longer. Those with significant identity scores (>35% identity over 80 AA) were individually tested using the sliding window of 80 AA comparison against AllergenOnline.org, version 18B. Additionally, a BLASTP search was performed against the NCBI database using keyword search limits of "allergen", "toxin" and "toxic". Significant results for all comparisons for each putative ORF are shown in Tables (1, 2, 3, 4, 5 and 6) with separate Tables for each search.

4.4.2.1. Full length FASTA3 vs. AllergenOnline.org with putative peptides.

Results of the full length FASTA3 searches of putative peptides against AllergenOnline.org, version 18B is the most important step for uncovering potential risks of allergy. Significant matches of predicted start-stop ORFs in AO2 and AO5 to allergens are reported in Tables 1 and 2. Start-Stop ORFs in A02 and A05 had one and four matches respectively to 2S albumins in walnut with low sequence identities. Stop-stop ORFs didn't show any significant matches to known allergens. None of the full-length FASTA alignments were significant in terms of uncovering any risk of potential allergenicity or cross-reactivity based on matches to allergens (Tables 1 and 2). The length of putative

peptides of less than 30 AA are unlikely to elicit a reaction even if bound by two IgE antibodies. Scoring results for the putative peptides showing alignments with *E*-scores less than 10 are shown and demonstrate no significant matches with any allergen. Their percent identities are markedly below the level that is likely to indicate cross-reactivity (< 50% identity, Aalberse, 2000) and it is also below the 35% identity level over 80 or more aa that was suggested by Codex (2003) as a match that may possibly be cross-reactive. Thus, there is only a small likelihood that any of the eight proteins are sufficiently similar to an allergen to suspect they might trigger allergic responses in allergic subjects due to cross-reactivity. There is no reason to suggest serum IgE tests would be useful to evaluation safety of this product further.

4.4.2.2. Sliding 80-amino acid window FASTA3 vs. AllergenOnline.org version 18B.

Results of the comparisons of the amino acid sequences of the putative peptides against all the sequences in AllergenOnline.org version 18B database were negative. This is a very stringent bioinformatics evaluation for potential risks of allergy and cross-reactivity based on the CODEX Alimentarius guidelines (2009). The lack of any match for each protein indicates low risk for allergy from these proteins.

4.4.2.3. BLASTP of NCBI Protein Database with and without keyword limits for each putative ORF in each insert.

The full-length sequences of the putative peptides were compared to all sequences in NCBI-Entrez database to find the most evolutionarily conserved proteins with results presented (Tables 3, 4 for AO2 and 5 and 6 for AO5). The scoring alignments with *E*-scores of the top one to three protein alignments identified by BLASTP were considered in some detail to determine if there is significant homology to proteins of sources with

likely safe human exposure or unsafe exposure, and when compared to results from searches with keywords (allergen, allergy, toxin, toxic), provides a relative evaluation of potential risks. The results from BLASTP comparison to all proteins were neutral, but the ubiquitous nature of the proteins without obvious indications of harm suggesting they are generally safe, abundant enzymes.

4.4.2.3.1. BLASTP of NCBI Entrez using keywords "allergen".

The full-length amino acid sequences of the putative ORF peptides were compared to sequences in NCBI Entrez, which were designated as "allergen" in the NCBI database in late September 2018. The alignment results with keyword "allergen" returned only one possible ORF of greater than 35% identity over 80 AA (ORF34 as a single copy in AO2; also present as four copies in AO5 since AO5 has two complete and reversed insertion DNA copies). The others were all negative The ORF34 peptide was compared to AllergenOnline.org version 18B and showed a slightly higher identity match (41% over 80 AA). The single copy in AO2 and four copies in AO5 are highly unlikely to be transcribed and translated as they are between two inserted genes. Thus, the probability of allergy or allergic cross-reactivity to hypothetical proteins identified by ORF analysis is extremely small based on observations of Aalberse (2000) and Goodman et al. (2008).

4.4.2.3.2. BLASTP of NCBI Entrez with "toxin", "toxic" and no keyword.

The putative peptide sequences from the junctions of DNA in canola were compared to sequences in NCBI-Entrez, which were designated by keywords for toxin or toxic and then without a keyword. The matches identified the closest overall matches from the NCBI Protein Database in early September 2018 from all three categories. The

alignment results with these keywords did not return any significant alignments that suggest possible harm to consumers. Taken together with the previously conducted bioinformatics searches of the eight, intended, expressed proteins, there does not appear to be a basis to suspect that the transgenic canola represents any risk of harm for consumers.

4.4.2.3.3. Bioinformatics summary for the hypothetical peptides (ORFs) throughout the two DNA inserts.

None of the results from the bioinformatics searches of the amino acid sequences from the putative peptides at the junctions of inserted DNA in canola carry significant risks of allergy or toxicity compared to commonly consumed proteins from a diverse variety of food sources.

4.4.3. Potential horizontal gene transfer from plants to microbes

4.4.3.1. PubMed Searches

The scientific literature database PubMed, was searched for information about possible horizontal gene transfer from plants to microbes including bacteria and archaea. Sixty articles were identified that suggest that evolution occurs by direct DNA transfer, including the uptake of DNA in microbes. Nielsen was the lead investigator on two studies testing potential gene transfer of the neomycin (Nielsen et al, 1998 and Nielsen et al, 2000). In those cases, the question was whether antibiotic resistance afforded by NPTII could be transferred to a bacteria, using the highly transformable *Actinetobacter* sp. The bacteria used had already been transfected with a plasmid containing the NPTII with either a 10 bp or a 200 bp deletion. The DNA used in the experiment was from herbicide tolerant transgenic sugar beet of Monsanto that contains an intact, plant DNA

encoded NPTII. The *Acinetobacter* sp. with a 10 bp deletion in the encoded NTPII did have a very low rate of recovery of complemented mutation at (Nielsen et al, 1998), the 200 bp deleted form of bacteria did not recover NTPII resistance in much larger scale exposure. The conclusion was that HGT from plant DNA is very unlikely even when an advantage like NTPII under exposure to the antibiotic, would offer an advantage. In many cases the potential risks have focused on the potential transfer of antibiotic resistance genes from a transgenic crop to soil or gut microbes (Nielsen et al, 1998). In most cases DNA sequence similarities were identified that authors suggested evidence for possible HGT. However, in most cases the authors concluded that the identities were not perfect and could represent very old transfers. In a few cases there were specific functional advantages that were identified, such as transfer of antibiotic resistance or transfer of adherence proteins that would allow an advantage to the putative gene recipient. For instance, the ability of four varieties of *Xanthomonas* sp. to infect common beans was associated with HGT of TAL genes between bacterial species, but not from plant to bacteria (Ruh et al, 2017). In most cases where HGT seemed plausible the DNA was most likely transferred between bacteria by plasmids through conjugation, or by bacteriophages (Hasegawa et al, 2018; McCullor et al, 2018). Although a number of mechanisms have been proposed that would allow HGT to occur, there is little direct evidence of direct transfer of DNA from a eukaryote to a microbial recipient. In plausible cases, the DNA was likely transferred to the microbe through replication systems including plasmids or bacteriophages. Additionally, transfer of naked DNA would require sequence matches of the donor and recipient DNA that allow recombination of double stranded DNA.

4.4.3.2. Sequence comparison of canola DNA to microbial genomic sequences.

The insert DNA sequences in chromosome A02 and in chromosome A05 were compared to complete and incomplete genomic sequences by BLASTN 2.7.1+. Alignments were considered positive if at least 200 nucleotides long and with 95% identity or greater. Ninety-five alignments of at least 200 nucleotides and 95% identity were found for single position matches to microbial DNA from both the A02 and the A05 inserts, although almost all of those were to incomplete genomic DNA contiguous sequences (Contigs). Since the matches were mostly to Contigs, we could not verify the full genome match through the NCBI Protein database. No qualified identified matches meeting the 95% identity criteria for at least 200 nucleotides were identified with archaea genomic DNA. The taxonomic identity of the sequence matching to insert DNA was recorded along with the percent identity, the length of the alignment, start and stop positions of the insert DNA and the *E*-scores were recorded. EFSA panel 2017 recommended that the results should be presented in a graphic summary that depicts the matches against the insert and flanking region, if relevant, with the information of its genetic elements. Therefore, genetic elements, location of matches for potential HGT targets against the A02 T-DNA insert (12,110 bp) and DHA canola and its flanking canola sequences were illustrated in Figures 1A and 1B. In addition, Figures 2A, 2B and 2C showed the genomic structure of first half of insert AO5 (1-25000), with genetic elements and possible HGT targets based on DNA sequence identity marked nucleotide 1-25000. The results for the right side of A05 T-DNA insert (25,000-52,000) are shown in Figures 3A, 3B and 3C.

4.4.3.3. Evaluation of potential horizontal gene transfer sequences.

Bioinformatics analyses of the nucleotide sequences in the two inserts of DHA canola were compared to DNA sequences in complete and incomplete genomes of bacteria and archaea. A number of matches of 95% or more were found for 200 bp or longer. Only one pair of sequence matches was found from the A02 insert that aligned with a two segments of DNA in *Pseudomonas putida*, with a two bp gap. That would result in possible HGT of 1,166 bp if HGT occurred in that species. The other identity matches in A02 were not paired with another identity match and thus would not result in a legitimate recombinational event. The DNA insert of A05 continuous sequences contain multiple segments of sequence in different microbes that had high identity matches with at least two pairs of segments to 20 complete or incomplete genomes. The other pairs were approximately two to four hundred bp apart. None of the sequences that might form an HGT unit appear to encode a gene that would logically provide a benefit to a microbe such as antibiotic resistance or an adhesion molecule. Some of the species that were identified, including *Xanthomonas* sp, *Agrobacterium* sp, *Pseudomonas* sp. and *Streptomyces* sp. have been identified as being able to take in DNA either through conjugation and plasmid transfer, or in some cases, intake of naked double-stranded DNA. Importantly, it appears that most cases of potential HGT would not result in a gain in fitness for the bacterial species, and most are likely to interrupt potentially functional gene sequences. Those are most plausibly disadvantageous for the bacteria.

4.5. Conclusions.

Bioinformatics analyses were performed previously by the Goodman laboratory and submitted to regulatory agencies by NuSeed on the eight proteins intentionally added to allow production of DHA. In addition, putative peptides (ORFs) located at the five DNA junctions present due to insertion of the DNA in two chromosomes in GE canola line to produce DHA were evaluated and submitted to regulators in Australia and the United States. The current evaluation was to consider potential identity matches of hypothetical ORFs throughout the inserted DNA in the two insertion sites, with those of known allergens and toxins. All new potential (hypothetical) ORFs with codons for 30 or more amino acids were analyzed. No significant homologies were identified at the junctions of the introduced DNA and the endogenous canola DNA. Based on the evidence, cross-reactive IgE binding and food toxicology tests are not scientifically justified to further evaluate safety of this canola line as there is no evidence that new proteins that represent possible allergens or toxins have been introduced (Goodman et al, 2008). The current bioinformatics analyses demonstrated that the development of the genetically modified DHA canola has not produced any new open reading frames that are expected to result in the expression of new proteins beyond those encoded by the transgenes. Searches of all potential ORFs did not uncover possible alignments that suggest possible risks of allergy or toxicity.

PubMed searches did identify a few publications from scientists who have previously suggested possible hypothetical risks of HGT to transfer antibiotic resistance genes into environmental microbes (Droge et al, 1998). Yet the evidence for natural transformation between a eukaryote such as a plant and a microbe is quite rare and tests

showing transfers have only been successful at low rates, when very strict conditions of high concentrations of DNA and microbes are present and in the absence of competing microbes or natural environmental matrices (Nielsen et al, 2000). Taken together, this bioinformatics analysis raises no concerns that the DNA from the two inserts in this transgenic canola would be transferrable to bacteria or archaea through horizontal gene transfer in a way that would adversely impact the environment.

Table 1. A02 Start-to-Stop ORF with a CODEX Significant Alignment to An Allergen in AOL V18B. The amino acid sequences of A02 Start-to-Stop ORFs were compared to AOL to find significant matches (>35% sequence identity over 80 AA alignment length) to allergens using full-length FASTA and 80mer AA searches.

ORF#	Strand DNA	Frame (1-6)	Start First:Last nucleotide	GI # Matched Sequence	Length-ORF AA of ORF | AA aligned	E-Score	Percent Identity	Best 80mer Alignment Percent ID
ORF34	-	2	9976 | 9632	31321942_2S albumin seed storage protein, partial [Juglans nigra]	114 | 96	7.8e-5	38.5	38.5%

Table 2. A05 Start-to-Stop ORF with a CODEX Significant Alignment to An Allergen in AOL V18B. A05 Start-to-Stop ORFs were compared to AOL to find significant matches (>35% sequence identity over 80 AA alignment length) to allergens using full-length FASTA and 80mer AA searches.

ORF#	Strand DNA	Frame (1-6)	Start First:Last nucleotide	GI # Matched Sequence	Length-ORF AA of ORF | AA aligned	E-Score	Percent Identity	Best 80mer Alignment Percent ID
ORF28	+	1	36640:36984	31321942_2S albumin seed storage protein, partial [Juglans nigra]	114 | 96	7.8e-5	38.5	41.4%
ORF48	+	2	17927:18271	31321942_2S albumin seed storage protein, partial [Juglans nigra]	114 | 96	6.2e-5	38.5	41.4%
ORF121	-	1	28686:28342	31321942_2S albumin seed storage protein, partial [Juglans nigra]	114 | 96	6.2e-5	38.5	41.4%
ORF208	-	3	9973:9629	31321942_2S albumin seed storage protein, partial [Juglans nigra]	114 | 96	7.8e-5	38.5	41.4%

Table 3. A02 Start-to-Stop ORFs Comparisons with the NCBI Protein database by BLASTP using Keyword Limits: Toxin, Toxic, Allergen and no keyword. A02 Start-to-Stop ORFs were compared to NCBI Protein database using different keywords to identify relevant allergens and toxins.

Query_id	Keyword	Subject_id	Pct_identity	Align length	E-value
lcl|ORF6:5170:5973	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	33.054	239	9.44E-25
	NO Keyword	ACR53360.1_delta-5 elongase [Pyramimonas cordata]	100	267	0
lcl|ORF15:6581:8029	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	41.463	41	0.002
	Toxic	KNG44469.1_hypothetical protein TW65_08823 [Stemphylium lycopersici]	50	26	9.8
	NO Keyword	A4KDP0.1_RecName: Full=Acyl-lipid (8-3]-desaturase; AltName: Full=AN Delta(5]-fatty-acid desaturase; AltName: Full=Acyl-lipid 5-desaturase; AltName: Full=Delta-5 desaturase	100	425	0
lcl|ORF33:11440:10193	Allergen	ACU89247.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Desulfomicrobium baculatum DSM 4028]	22.581	93	2.3
	Toxin	EDX65010.1_fatty acid desaturase [Bacillus cereus 03BB108]	32.353	68	0.071
	Toxic	KHQ50621.1_putative hydrocarbon oxygenase protein [Mameliella alba]	40	50	0.028
	NO Keyword	XP_002494184.1_Hypothetical protein PAS_chr4_0743 [Komagataella phaffii GS115]	100	415	0
lcl|ORF52:2052:661	Allergen	CCX31489.1_Similar to Uncharacterized membrane protein C1322.03; acc. no. O94543 [Pyronema omphalodes CBS 100304]	47.368	38	4.2
	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	40	45	0.01
	Toxic	ODM29929.1_fatty acid desaturase [Marinobacter adhaerens]	43.902	41	0.000717
	NO Keyword	XP_003056992.1_predicted protein [Micromonas pusilla CCMP1545]	100	463	0

Table 4. A02 Stop-to-Stop ORFs Comparisons with the NCBI Protein database by BLASTP using Keyword Limits: Toxin, Toxic, Allergen and no keyword. A02 Stop-to-Stop ORFs were compared to NCBI Protein database using different keywords to identify relevant allergens and toxins.

Query_id	Keyword	Subject_id	Pct_identity	Align length	E-value
lcl|ORF10:5023:5973	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	32.636	239	2.2E-24
	Toxic	KKB31490.1_Spermidine N(1]-acetyltransferase [Bacillus thuringiensis serovar mexicanensis]	29.851	67	6.5
	No Keyword	ACR53360.1_delta-5 elongase [Pyramimonas cordata]	100	267	0
lcl|ORF27:6458:8029	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	41.463	41	0.002
	Toxic	KLK99258.1_uroporphyrinogen-III synthase [Bacillus pumilus]	35.593	59	0.67
	No Keyword	A4KDP0.1_RecName: Full=Acyl-lipid (8-3]-desaturase; AltName: Full=AN Delta(5]-fatty-acid desaturase; AltName: Full=Acyl-lipid 5-desaturase; AltName: Full=Delta-5 desaturase	100	425	0
lcl|ORF68:11680:10193	Allergen	ACU89247.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Desulfomicrobium baculatum DSM 4028]	22.581	93	3
	Toxin	EDX65010.1_fatty acid desaturase [Bacillus cereus 03BB108]	32.353	68	0.11
	Toxic	KHQ50621.1_putative hydrocarbon oxygenase protein [Mameliella alba]	40	50	0.032
	No Keyword	XP_002494184.1_Hypothetical protein PAS_chr4_0743 [Komagataella phaffii GS115]	100	415	0
lcl|ORF102:2124:661	Allergen	CCX31489.1_Similar to Uncharacterized membrane protein C1322.03; acc. no. O94543 [Pyronema omphalodes CBS 100304]	47.368	38	4.2
	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	40	45	0.012
	Toxic	ODM29929.1_fatty acid desaturase [Marinobacter adhaerens]	45	40	0.000874
	No Keyword	XP_003056992.1_predicted protein [Micromonas pusilla CCMP1545]	100	463	0

Table 5. A05 Start-to-Stop ORFs Comparisons with the NCBI Protein database by BLASTP using Keyword Limits: Toxin, Toxic, Allergen and no keyword. A05 Start-to-Stop ORFs were compared to NCBI Protein database using different keywords to identify relevant allergens and toxins.

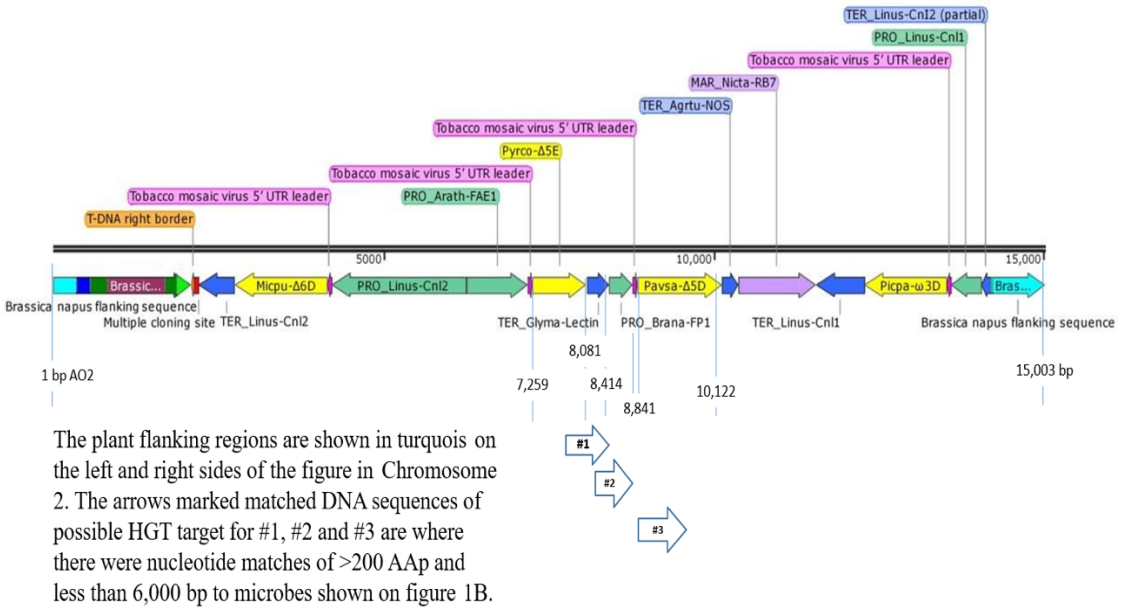
Query_id	Keyword	Subject_id	Pct_identity	Align Length	E-value
lcl|ORF6:5167:5970	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	33.054	239	9.44E-25
	No keyword	ACR53360.1_delta-5 elongase [Pyramimonas cordata]	100	267	0
lcl|ORF15:20626:21492	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	39.111	225	1.61E-37
	Toxin	4RGN_B_Chain B, Structure Of Staphylococcal Enterotoxin B Bound To Two Neutralizing Antibodies, 14g8 And 6d3	31.884	69	3.7
	No keyword	ACR53359.1_delta-6 elongase [Pyramimonas cordata]	100	288	0
lcl|ORF17:22399:22950	Allergen	AEV97129.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Niastella koreensis GR20-10]	64	25	9.3
	Toxin	CCK25597.1_GNAT family toxin-antitoxin system, toxin component [Streptomyces davaonensis JCM 4913]	42.857	161	2.41E-29
	Toxic	AXS75741.1_phosphinothricin acetyltransferase [Expression vector p390-blpR-cmcas9-gfp]	85.714	168	8.40E-105
	No keyword	WP_003988626.1_N-acetyltransferase [Streptomyces viridochromogenes]	100	183	7.86E-133
lcl|ORF27:35176:36423	Allergen	ACU89247.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Desulfomicrobium baculatum DSM 4028]	22.581	93	2.3
	Toxin	EDX65010.1_fatty acid desaturase [Bacillus cereus 03BB108]	32.353	68	0.071
	Toxic	KHQ50621.1_putative hydrocarbon oxygenase protein [Mameliella alba]	40	50	0.028
	No keyword	XP_002494184.1_Hypothetical protein PAS_chr4_0743 [Komagataella phaffii GS115]	100	415	0
lcl|ORF40:6578:8026	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	41.463	41	0.002
	Toxic	KNG44469.1_hypothetical protein TW65_08823 [Stemphylium lycopersici]	50	26	9.8
	No keyword	A4KDP0.1_RecName: Full=Acyl-lipid (8-3]-desaturase; AltName: Full=AN Delta(5]-fatty-acid desaturase; AltName: Full=Acyl-lipid 5-desaturase; AltName: Full=Delta-5 desaturase	100	425	0
lcl|ORF45:14057:15400	Toxin	ANA38154.1_acyl-CoA desaturase [Acinetobacter baumannii]	28.276	290	2.51E-12
	Toxic	KEI69057.1_CrtR [Planktothrix agardhii NIVA-CYA 126/8]	32.143	56	1.3
	No keyword	A0PJ29.1_RecName: Full=Acyl-lipid (7-3]-desaturase; AltName: Full=Acyl-lipid 4-desaturase; AltName: Full=Delta-4 desaturase; Short=PsD4Des	100	447	0
lcl|ORF47:16460:17710	Toxin	KKC53285.1_fatty acid desaturase [Bacillus sp. UMTAT18]	27.586	58	0.1
	Toxic	KEJ96575.1_fatty acid desaturase [Sulfitobacter pseudonitzschiae]	33.333	54	0.000482
	No keyword	BAD08375.1_delta 12-fatty acid desaturase [Lachancea kluyveri]	100	416	0
lcl|ORF77:44564:45955	Allergen	CCX31489.1_Similar to Uncharacterized membrane protein C1322.03; acc. no. O94543 [Pyronema omphalodes CBS 100304]	47.368	38	4.2
	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	40	45	0.01
	Toxic	ODM29929.1_fatty acid desaturase [Marinobacter adhaerens]	43.902	41	0.000717
	No keyword	XP_003056992.1_predicted protein [Micromonas pusilla CCMP1545]	100	463	0
lcl|ORF113:40035:38587	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	41.463	41	0.002
	Toxic	KNG44469.1_hypothetical protein TW65_08823 [Stemphylium lycopersici]	50	26	9.8
	No keyword	A4KDP0.1_RecName: Full=Acyl-lipid (8-3]-desaturase; AltName: Full=AN Delta(5]-fatty-acid desaturase; AltName: Full=Acyl-lipid 5-desaturase; AltName: Full=Delta-5 desaturase	100	425	0
lcl|ORF118:32556:31213	Toxin	ANA38154.1_acyl-CoA desaturase [Acinetobacter baumannii]	28.276	290	2.51E-12
	Toxic	KEI69057.1_CrtR [Planktothrix agardhii NIVA-CYA 126/8]	32.143	56	1.3
	No keyword	A0PJ29.1_RecName: Full=Acyl-lipid (7-3]-desaturase; AltName: Full=Acyl-lipid 4-desaturase; AltName: Full=Delta-4 desaturase; Short=PsD4Des	100	447	0
lcl|ORF120:30153:28903	Toxin	KKC53285.1_fatty acid desaturase [Bacillus sp. UMTAT18]	27.586	58	0.1
	Toxic	KEJ96575.1_fatty acid desaturase [Sulfitobacter pseudonitzschiae]	33.333	54	0.000482
	No keyword	BAD08375.1_delta 12-fatty acid desaturase [Lachancea kluyveri]	100	416	0
lcl|ORF150:2049:658	Allergen	CCX31489.1_Similar to Uncharacterized membrane protein C1322.03; acc. no. O94543 [Pyronema omphalodes CBS 100304]	47.368	38	4.2
	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	40	45	0.01
	Toxic	ODM29929.1_fatty acid desaturase [Marinobacter adhaerens]	43.902	41	0.000717
	No keyword	XP_003056992.1_predicted protein [Micromonas pusilla CCMP1545]	100	463	0
lcl|ORF186:41446:40643	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	33.054	239	9.44E-25
	No keyword	ACR53360.1_delta-5 elongase [Pyramimonas cordata]	100	267	0
lcl|ORF195:25987:25121	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	39.111	225	1.61E-37
	Toxin	4RGN_B_Chain B, Structure Of Staphylococcal Enterotoxin B Bound To Two Neutralizing Antibodies, 14g8 And 6d3	31.884	69	3.7
	No keyword	ACR53359.1_delta-6 elongase [Pyramimonas cordata]	100	288	0

lcl|ORF197:24214:23663	Allergen	AEV97129.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Niastella koreensis GR20-10]	64	25	9.3
	Toxin	CCK25597.1_GNAT family toxin-antitoxin system, toxin component [Streptomyces davaonensis JCM 4913]	42.857	161	2.41E-29
	Toxic	AXS75741.1_phosphinothricin acetyltransferase [Expression vector p390-blpR-cmcas9-gfp]	85.714	168	8.40E-105
	No keyword	WP_003988626.1_N-acetyltransferase [Streptomyces viridochromogenes]	100	183	7.86E-133
lcl|ORF207:11437:10190	Allergen	ACU89247.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Desulfomicrobium baculatum DSM 4028]	22.581	93	2.3
	Toxin	EDX65010.1_fatty acid desaturase [Bacillus cereus 03BB108]	32.353	68	0.071
	Toxic	KHQ50621.1_putative hydrocarbon oxygenase protein [Mameliella alba]	40	50	0.028
	No keyword	XP_002494184.1_Hypothetical protein PAS_chr4_0743 [Komagataella phaffii GS115]	100	415	0

Table 6. A05 Stop-to-Stop ORFs Comparisons with the NCBI Protein database by BLASTP using Keyword Limits: Toxin, Toxic, Allergen and no keyword. A05 Stop-to-Stop ORFs were compared to NCBI Protein database using different keywords to identify relevant allergens and toxins.

Query_id	Keyword	Subject_id	Pct_identity	Align length	E-value
lcl|ORF10:5020:5970	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	32.6	239	2.20E-24
	Toxic	KKB31490.1_Spermidine N(1)-acetyltransferase [Bacillus thuringiensis serovar mexicanensis]	29.9	67	6.5
	No Keyword	ACR53360.1_delta-5 elongase [Pyramimonas cordata]	100.0	267	0
lcl|ORF27:20479:21492	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	39.1	225	8.07E-37
	Toxin	4RGN_B_Chain B, Structure of Staphylococcal Enterotoxin B Bound To Two Neutralizing Antibodies, 14g8 And 6d3	31.9	69	4.7
	No Keyword	ACR53359.1_delta-6 elongase [Pyramimonas cordata]	100.0	288	0
lcl|ORF29:22354:22950	Allergen	AEV97129.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Niastella koreensis GR20-10]	64.0	25	9.6
	Toxin	CCK25597.1_GNAT family toxin-antitoxin system, toxin component [Streptomyces davaonensis JCM 4913]	42.9	161	3.77E-29
	Toxic	AXS75741.1_phosphinothricin acetyltransferase [Expression vector p390-blpR-cmcas9-gfp]	85.7	168	9.12E-105
lcl|ORF50:34936:36423	No Keyword	WP_003988626.1_N-acetyltransferase [Streptomyces viridochromogenes]	100.0	183	5.06E-133
	Allergen	ACU89247.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Desulfomicrobium baculatum DSM 4028]	22.6	93	3
	Toxin	EDX65010.1_fatty acid desaturase [Bacillus cereus 03BB108]	32.4	68	0.11
	Toxic	KHQ50621.1_putative hydrocarbon oxygenase protein [Mameliella alba]	40.0	50	0.032
	No Keyword	XP_002494184.1_Hypothetical protein PAS_chr4_0743 [Komagataella phaffii GS115]	100.0	415	0
lcl|ORF222:40158:38587	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	41.5	41	0.002
	Toxic	KLK99258.1_uroporphyrinogen-III synthase [Bacillus pumilus]	35.6	59	0.67
	No Keyword	A4KDP0.1_RecName: Full=Acyl-lipid (8-3]-desaturase; AltName: Full=AN Delta(5]-fatty-acid desaturase; AltName: Full=Acyl-lipid 5-desaturase; AltName: Full=Delta-5 desaturase	100.0	425	0
lcl|ORF91:13925:15400	Toxin	ANA38154.1_acyl-CoA desaturase [Acinetobacter baumannii]	27.5	284	3.51E-12
	Toxic	KEI69057.1_CrtR [Planktothrix agardhii NIVA-CYA 126/8]	32.1	56	1.9
	No Keyword	A0PJ29.1_RecName: Full=Acyl-lipid (7-3]-desaturase; AltName: Full=Acyl-lipid 4-desaturase; AltName: Full=Delta-4 desaturase; Short=PsD4Des	100.0	447	0
lcl|ORF94:16220:17710	Toxin	EKD45279.1_hypothetical protein ACD_69C00356G0002 [uncultured bacterium]	28.3	60	9.1
	Toxic	KEJ96575.1_fatty acid desaturase [Sulfitobacter pseudonitzschiae]	33.3	54	0.000639
	No Keyword	BAD08375.1_delta 12-fatty acid desaturase [Lachancea kluyveri]	100.0	416	0
lcl|ORF140:44492:45955	Allergen	CCX31489.1_Similar to Uncharacterized membrane protein C1322.03; acc. no. O94543 [Pyronema omphalodes CBS 100304]	47.4	38	4.2
	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	40.0	45	0.012
	Toxic	ODM29929.1_fatty acid desaturase [Marinobacter adhaerens]	45.0	40	0.000874
	No Keyword	XP_003056992.1_predicted protein [Micromonas pusilla CCMP1545]	100.0	463	0
lcl|ORF222:40158:38587	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	41.5	41	0.002
	Toxic	KLK99258.1_uroporphyrinogen-III synthase [Bacillus pumilus]	35.6	59	0.67
	No Keyword	A4KDP0.1_RecName: Full=Acyl-lipid (8-3]-desaturase; AltName: Full=AN Delta(5]-fatty-acid desaturase; AltName: Full=Acyl-lipid 5-desaturase; AltName: Full=Delta-5 desaturase	100.0	425	0
lcl|ORF235:32688:31213	Toxin	ANA38154.1_acyl-CoA desaturase [Acinetobacter baumannii]	27.5	284	3.51E-12
	Toxic	KEI69057.1_CrtR [Planktothrix agardhii NIVA-CYA 126/8]	32.1	56	1.9

	No Keyword	A0PJ29.1_RecName: Full=Acyl-lipid (7-3]-desaturase; AltName: Full=Acyl-lipid 4-desaturase; AltName: Full=Delta-4 desaturase; Short=PsD4Des	100.0	447	0
lcl|ORF238:30393:28903	Toxin	EKD45279.1_hypothetical protein ACD_69C00356G0002 [uncultured bacterium]	28.3	60	9.1
	Toxic	KEJ96575.1_fatty acid desaturase [Sulfitobacter pseudonitzschiae]	33.3	54	0.000639
	No Keyword	BAD08375.1_delta 12-fatty acid desaturase [Lachancea kluyveri]	100.0	416	0
lcl|ORF284:2121:658	Allergen	CCX31489.1_Similar to Uncharacterized membrane protein C1322.03; acc. no. O94543 [Pyronema omphalodes CBS 100304]	47.4	38	4.2
	Toxin	KKO86649.1_fatty acid desaturase [Corynebacterium ulcerans]	40.0	45	0.012
	Toxic	ODM29929.1_fatty acid desaturase [Marinobacter adhaerens]	45.0	40	0.000874
	No Keyword	XP_003056992.1_predicted protein [Micromonas pusilla CCMP1545]	100.0	463	0
lcl|ORF365:41593:40643	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	32.6	239	2.20E-24
	Toxic	KKB31490.1_Spermidine N(1]-acetyltransferase [Bacillus thuringiensis serovar mexicanensis]	29.9	67	6.5
	No Keyword	ACR53360.1_delta-5 elongase [Pyramimonas cordata]	100.0	267	0
lcl|ORF382:26134:25121	Allergen	XP_005650877.1_hypothetical protein COCSUDRAFT_58870 [Coccomyxa subellipsoidea C-169]	39.1	225	8.07E-37
	Toxin	4RGN_B_Chain B, Structure of Staphylococcal Enterotoxin B Bound To Two Neutralizing Antibodies, 14g8 And 6d3	31.9	69	4.7
	No Keyword	ACR53359.1_delta-6 elongase [Pyramimonas cordata]	100.0	288	0
lcl|ORF384:24259:23663	Allergen	AEV97129.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Niastella koreensis GR20-10]	64.0	25	9.6
	Toxin	CCK25597.1_GNAT family toxin-antitoxin system, toxin component [Streptomyces davaonensis JCM 4913]	42.9	161	3.77E-29
	Toxic	AXS75741.1_phosphinothricin acetyltransferase [Expression vector p390-blpR-cmcas9-gfp]	85.7	168	9.12E-105
	No Keyword	WP_003988626.1_N-acetyltransferase [Streptomyces viridochromogenes]	100.0	183	5.06E-133
lcl|ORF405:11677:10190	Allergen	ACU89247.1_alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen [Desulfomicrobium baculatum DSM 4028]	22.6	93	3
	Toxin	EDX65010.1_fatty acid desaturase [Bacillus cereus 03BB108]	32.4	68	0.11
	Toxic	KHQ50621.1_putative hydrocarbon oxygenase protein [Mameliella alba]	40.0	50	0.032
	No Keyword	XP_002494184.1_Hypothetical protein PAS_chr4_0743 [Komagataella phaffii GS115]	100.0	415	0



The plant flanking regions are shown in turquois on the left and right sides of the figure in Chromosome 2. The arrows marked matched DNA sequences of possible HGT target for #1, #2 and #3 are where there were nucleotide matches of >200 AAp and less than 6,000 bp to microbes shown on figure 1B.

Figure 1A. Genomic Structure of A02 Insert, with Genetic Elements Marked Nucleotide 1-15003. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

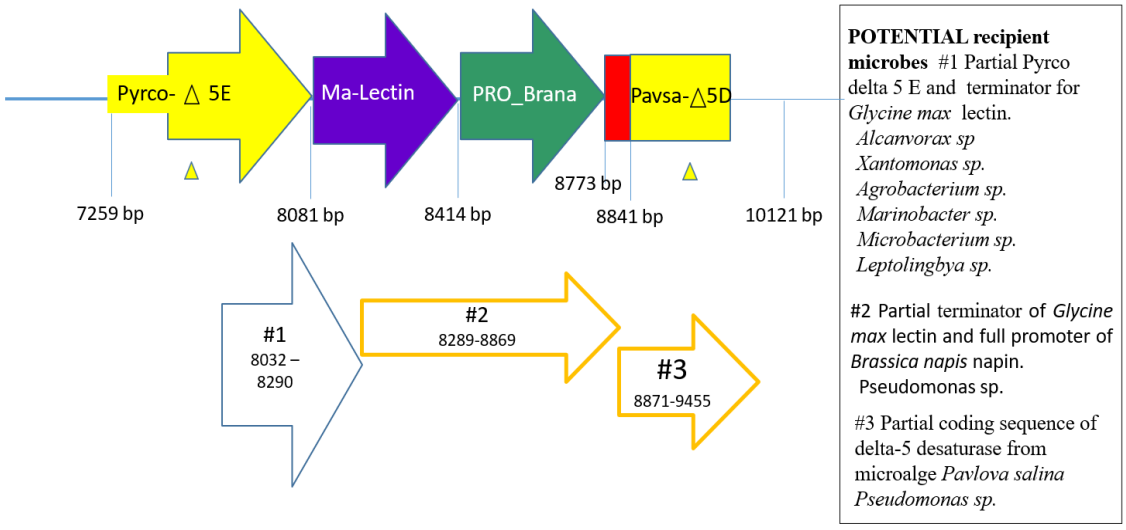


Figure 1B. Expanded Graphic Image for A02 with Matched DNA Segments of Hypothetical HGT Targets. Primary colored arrows represent coding genes (4) and the red indicates a TMV 5'-UTR leader. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s). Note #2 and #3 white boxes may indicate possible discontinuous transfer and possible change in microbe genome. Others (#1) are unlikely to change the bacterial genome.

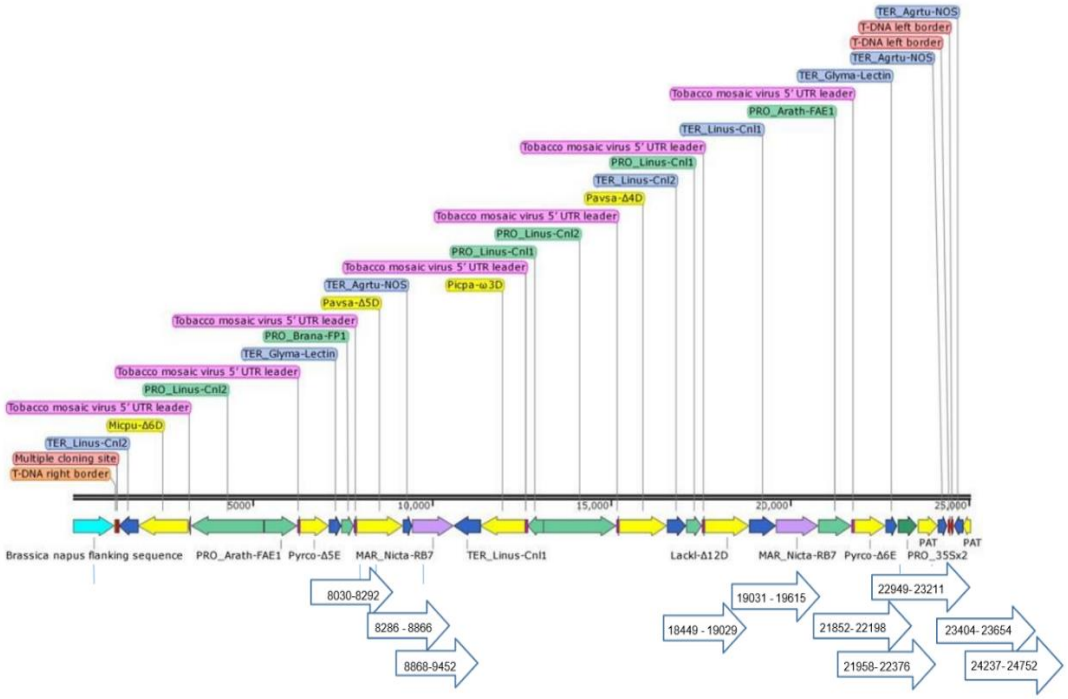


Figure 2A. Genomic Structure of First Half of Insert A05, with Genetic Elements Marked Nucleotide 1-25000. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

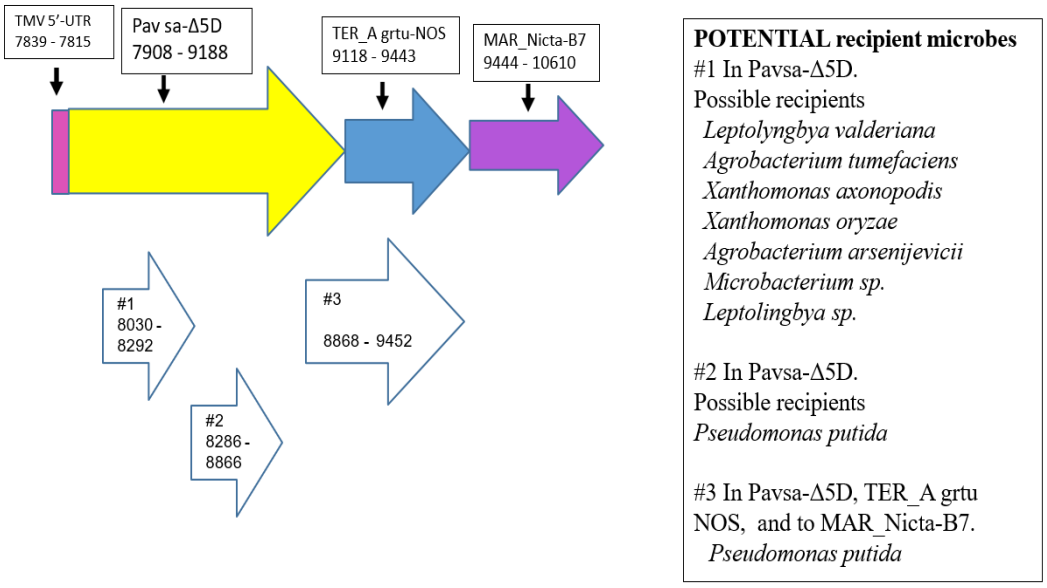


Figure 2B. Possible HGT Targets Left Side of Chromosome AO5 Based on DNA Sequence Identity. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

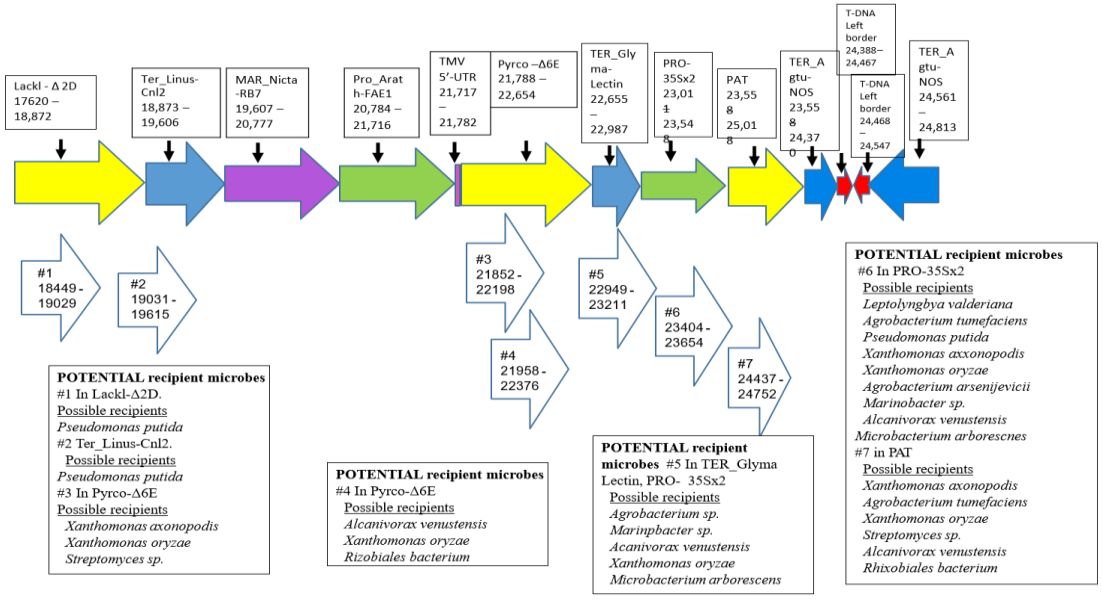


Figure 2C. Possible HGT Targets Right Side of Chromosome AO5 Based on DNA Sequence Identity. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

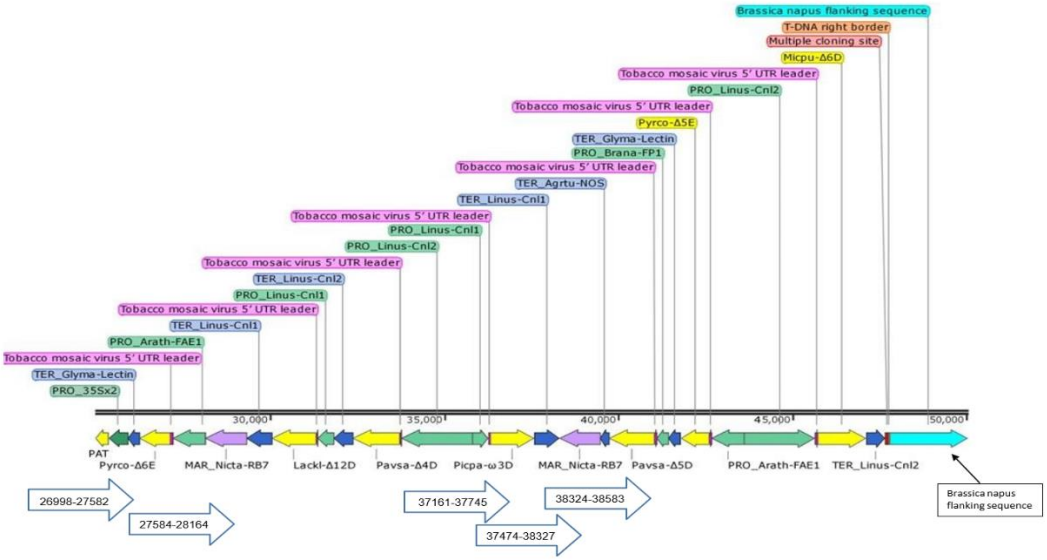


Figure 3A. Genomic Structure of Second Half of Insert AO5, with Genetic Elements Marked Nucleotide 25,000-52,000. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

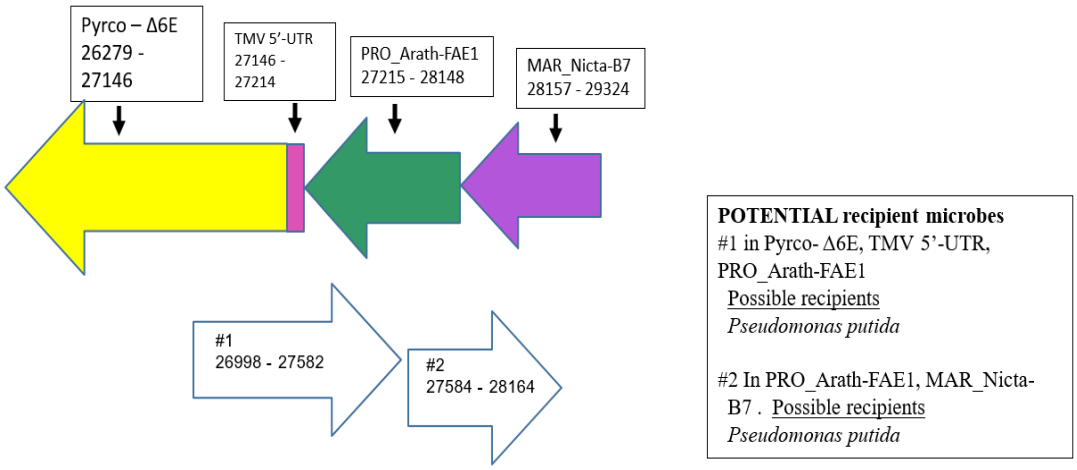


Figure 3B. Possible HGT Targets Right Side of Chromosome AO5 Based on DNA Sequence Identity. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

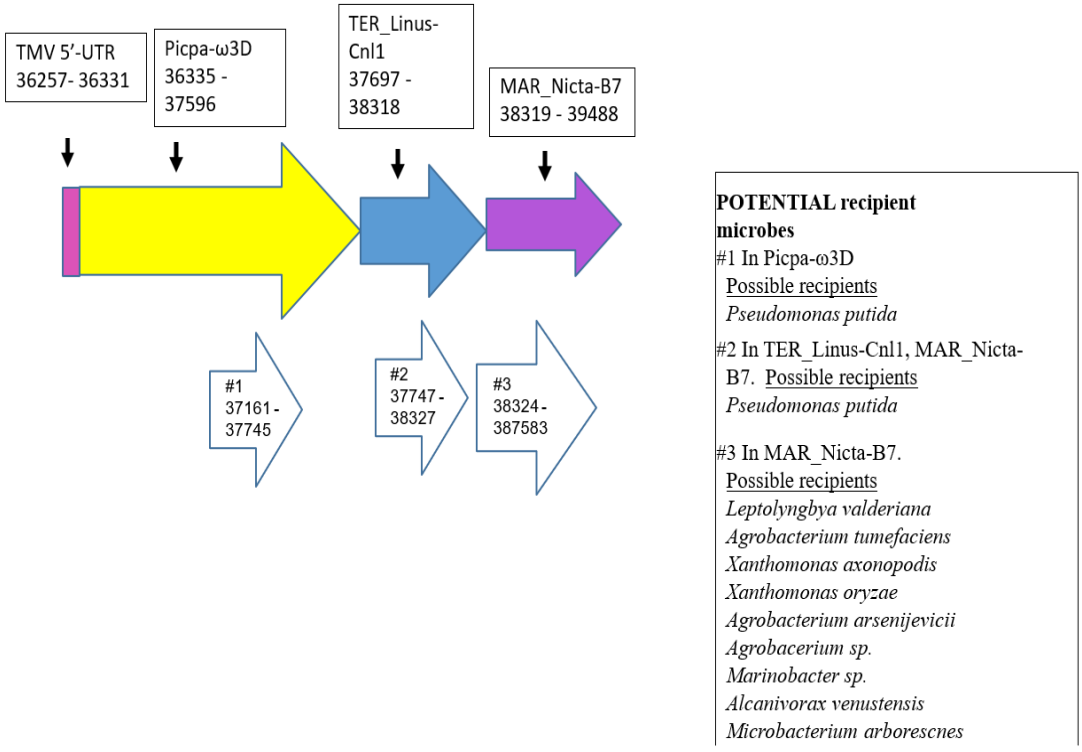


Figure 3C. Possible HGT Targets Right Side of Chromosome AO5 Based on DNA Sequence Identity. White arrows indicated possible HGT target sequences with >95% identity match to microbe(s).

4.6. References

Aalberse, R.C., 2000. Structural biology of allergens. J Allergy Clin Immunol 106:228-238.

Codex Alimentarius Commission., 2003. Alinorm 03/34: Joint FAO/WHO Food Standard Programme, Codex Alimentarius Commission, Twenty-Fifth Session, Rome, Italy 30 June-5 July, 2003. Appendix III, Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants and Appendix IV, Annex on the assessment of possible allergenicity, pp. 47-60.

Colgrave, M.L., Byrne K., Pillai S.V., et al, 2019. Quantitation of seven transmembrane proteins from the DHA biosynthesis pathway in genetically engineered canola by targeted mass spectrometry. Food Chem Toxicol., 126:313–321.

De Vries, J. and Wackernagel, W., 2002. Integration of foreign DNA during natural transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. Proceedings of the National Academy of Sciences of the USA, 99, 2094-2099.

Droge, M., Puhler A., Selbitschka, W., 1998. Horizontal gene transfer as a biosafety issue: a natural phenomenon of public concern. J Biotechnol 64(1):75-90.

EFSA (European Food Safety Authority), 2009. Consolidated presentation of the joint Scientific Opinion of the GMO and BIOHAZ Panels on the "Use of Antibiotic Resistance Genes as Marker Genes in Genetically Modified Plants" and the Scientific Opinion of the GMO Panel on "Consequences of the Opinion on the Use of Antibiotic Resistance Genes as Marker Genes in Genetically Modified Plants on Previous EFSA Assessments of Individual GM Plants. The EFSA Journal 2009, 1108, 1-8.

EFSA (European Food Safety Authority), Gennaro A,Gomes A, Herman L, Nogue F, Papadopoulou N, Tebbe C, 2017. Technical report on the explanatory note on DNA sequence similarity searches in the context of the assessment of horizontal gene transfer from plants to microorganisms. EFSA supporting publication 2017:EN-1273. 11 pp.

Goodman, R.E., 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. Food Chem Toxicol 46:S24-S34

Goodman, R.E., Ebisawa M, Ferreira F, Sampson HA, van Ree R, Vieths S, Baumert JL, Bohle B, Lalithambika S, Wise J, Taylor SL., 2016. AllergenOnline: A peer-reviewed, curated allergen database to assess novel food proteins for potential cross reactivity. Mol Nutr Food Res. (60:1183-1198.

Goodman, R.E., Vieths S, Sampson HA, Hill D, Ebisawa M, Taylor SL, van Ree R., 2008. Allergenicity assessment of genetically modified crops—what makes sense? Nat Biotechnol 26(1):73-81.

Hasegawa, H., Suzuki, E., Maeda, S., 2018. Horizontal plasmid transfer by transformation in Escherichia coli: Environmental factors and possible mechanisms. Front Microbiol 9:2365.

McCullor, K., Postoak, B., Rahman, M., King, C., McShan, W.M., 2018. Genomic sequencing of high-efficiency transducing Streptococcal bacteriophage A25: consequences of escape from lysogeny. J Bacteriol 200(23): pii:e00358-18.

Monier, J.-M, Bernillon, D., Kay, E., Faugier, A., Rybalka, O., Dessaux, Y., Simonet, P., and Vogel, T.M., 2007. Detection of potential transgenic plant DNA recipients among soil bacteria. Environmental Biosafety Research, 6, 71-83.

Nielsen, K.M., Bones, A.M., Smalla, K., van Elsas, J.D., 1998. Horizontal gene transfer from transgenic plants to terrestrial bacteria – a rare event? FEMS Microbiol Rev 22:79-103.

Nielsen, K.M., van Elsas, J.D., Smalla, K., 2000. Transformation of Acinetobacter sp. Strain BD413(pFG4ΛnptII) with Transgenic Plant DNA in Soil Microcosms and Effects of Kanamycin on Selection of Transformants. Appl Envir Microbiol. 66(3):1237-1242.

Overballe-Petersen, S., Harms, K., Orlando, L.A., Mayar, J.V., et al., 2013. Bacterial natural transformation by highly fragmented and damaged DNA. Proceedings of the National Academy of Sciences of the USA, 110, 19860-19865.

Ruh, M., Briand, M., Bonneau, S., Jacques, M.A., Chen, N.W.G, 2017. Xanthomonas adaptation to common bean is associated with horizontal transfers of genes encoding TAL effectors. BMC Genomics 18(1):670.

CHAPTER 5

DEVELOPMENT OF A SEQUENCE SEARCHABLE CELIAC DATABASE OF PEPTIDES AND PROTEINS FOR RISK ASSESSMENT OF NOVEL FOOD PROTEINS

This chapter is in progress to be submitted for peer review: Plaimein Amnuaycheewa, Mohamed Abdelmoteleb, John Wise, Barbara Bohle, Fatima Ferreira, Afua O. Tetteh, Steve L. Taylor, and Richard E. Goodman

5.1. Abstract

Celiac disease (CeD) is a genetically-restricted autoimmune enteropathy induced by prolamins (glutens) in grain of wheat, barley, rye, and oats. Consumers with MHCII DQ2 or DQ8 are at risk, though 1.4% of the global population has clinically proven CeD while 40% of DQ2+ or DQ8+ subjects do not. CeD subjects must avoid gluten to remain disease-free and regulatory authorities in Europe and the United States now expect an evaluation of new proteins in genetically modified crops or in novel foods to be evaluated for possible CeD risk. A database of 1,016 gluten peptides was developed in 2012 from published evidence of stimulating CD4+ T cells from CeD subjects or causing intestinal toxicity. A peptide sequence amino acid (AA) matching program was developed and a FASTA3 algorithm search added to show overall comparison to 68 representative gluten proteins that would require further testing if novel proteins match CeD peptides or proteins above identified criteria. The database was updated in 2018, removing peptides shorter than 9 AA and adding newly identified CeD peptides and proteins. Bioinformatics comparisons were performed with homologous proteins from Pooideae and from non-

Pooideae monocots, dicots and animal proteins to determine predictive matches for risk assessment

5.2. Introduction

Novel proteins and novel complex foods are being introduced into the human diet through creation of genetically engineered organisms, by addition of isolated proteins or by the introduction of new foods from novel organisms without previous documented history of safe human consumption (van Putten et al, 2006). Prior to marketing, novel proteins and novel foods should undergo a safety evaluation to ensure safe consumption by those with specific food allergies and for those with celiac disease (CeD). The 2003 Codex Alimentarius Commission guideline calls for evaluating genes (proteins) transferred from wheat and its relatives into a different species to be evaluated for potential risks of eliciting CeD as part of the overall food safety evaluation (CODEX 2003). The Food Allergy Research and Resource Program at the University of Nebraska developed a database of specific CeD peptides and proteins and provide bioinformatics tools to identify proteins that would possess probable risks of eliciting CeD. We used the exact peptide match and FASTA comparisons to evaluate sequences of proteins with known risks of CeD and homologous proteins from non-CeD eliciting sources to evaluate their use as a screening tool with low rates of false positive and false negative results.

Celiac disease is a T cell-mediated adverse reaction to ingested glutens, which are prolamins in wheat (including kamut and spelt), barley, rye, oat, and hybrids such as Triticale. The disease manifests primarily as an autoimmune disease in the upper small intestine, but it has significant extra-intestinal and overall health consequences including

malnutrition. The disease affects approximately 1.4% of the global population and is considered one of the most common genetically restricted autoimmune diseases (Rubio-Tapia et al, 2012; Singh et al, 2018; Cukrowska et al, 2017). In Europe and the UK, more than 90% of the patients express major histocompatibility complex (MHC) receptors HLA-DQ2.5 (DRB1*301-DQA1*0501-DQB1*0201) and between 5-10% of the patients express HLA-DQ8 (DRB1*04-DQA1*0301-DQB1*0302) (Polvi et al, 1998; Romanos et al, 2009; Sollid 2017). The percentage of CeD patients in the US carrying HLA-DQ2.5 or HLA-DQ8 has been estimated to be 82% or 16%, respectively (Fasano et al, 2003). A multicenter European study reported some variation in HLA genes with nearly 0.4% of CeD patients carrying DR5-DQ7 (DRB1*11/12-DQA1*0505-DQB1*0301) or DR7-DQ2 (DRB1*07-DQA1*0201-DQB1*0202) which can form heterozygous DQ2.5 (DQA1*0505-DQB1*0202) (Karell et al, 2003). These MHC receptors bind peptides with specific amino acid sequences and present them to CD4+ T cells that are effective elicitors of CeD. The MHC restriction is predictive but is not the definitive determinant since nearly 40% of the general population carry HLA-DQ2 or HLA-DQ8 genes, but only 1.4% of the population exhibit CeD (Jabri and Sollid, 2017). Meta-analyses of genome-wide association studies have revealed that CeD patients also commonly express variants of 39 non-HLA, immune-related genes that contribute to pathology including CTLA4, CD80, CD28, IL2, IL21, CCR4 and TLR7 (Hunt et al, 2008; Dubois et al, 2010; Trynka et al, 2012). While the MHC restriction limits the peptides that can be presented, it is essential to consider the impact of the endogenous human intestinal tissue transglutaminase (TG2) enzyme when screening food proteins as both native sequences

and those that are deamidated by TG2. The TG2 itself becomes a target of the activated T cells.

Gluten is defined as a macropolymer of prolamins that are rich in proline and glutamine amino acids. The unique proline-glutamine composition contributes to the visco-elastic properties of grain flour important for bread making, but the sequences also confers resistance to proteolytic digestion in the gastrointestinal tract (Di Sabatino and Corazza, 2009). Importantly, many digestion resistant gluten peptides are reported to translocate across intestinal epithelium either via modulation of epithelial permeability by stimulating CXCR3 receptors or via transcellular absorption. The peptides bind with genetically restricted major histocompatibility complex receptors HLA-DQ2.5 or DQ8 on antigen presenting cells (APCs) in the lamina propria. The MHCII bound peptides are then presented and activate pro-inflammatory CD4+ T cells (Lammers et al, 2008; Tripathi et al, 2009; Groschwitz and Hogan, 2009; Fasano 2011; Perez-Gregorio et al, 2005). Other gluten peptides can mediate intestinal inflammation through innate immune activation. A 13-amino acid gliadin peptide (LGQQQPFPPQQPY) was found to induce secretion of IFN-γ, TNF-α and IL-15 from intestinal epithelial cells, macrophages, and dendritic cell (DCs) (Londei et al, 2005; Jabri and Sollid, 2009). The IL-15 cytokine promotes proliferation and survival of NK cells and CD8+ T cells, thus promoting intraepithelial lymphocytosis and inflammation (Londei et al, 2005). IL-15 induces the expression of *MHC* class I related chain (MIC) on enterocytes and the counter-ligand natural killer group 2D (NKG2D) on the intraepithelial lymphocytes. The T cell receptor-independent interaction between *MIC and* NKG2D leads to apoptosis of the enterocyte *resulting in* destruction of the epithelial layer and villous atrophy (Roberts et al, 2001;

Meresse et al, 2004; Tang et al, 2009). *IL-15 together with retinoic acid was found to induce the expression of IL-23, which mediates the* differentiation of proinflammatory Th17 cells (DePaolo et al, 2011). IL-15 also impairs the suppressor activity of Treg cells by activating the phosphatidylinositol 3-kinase pathway (Ben Ahmed et al, 2009; Zanzi et al, 2011). The role of IL-15 in mediating CeD pathogenesis is well documented in refractory CeD patients who exhibit villous atrophy without recent ingestion of gluten. In such cases, IL-15 plays a central role in sustaining the destructive intraepithelial lymphocytes (IELs) and suppression of IL-15 effectively mitigates severe inflammation (Mention et al, 2003; Malamut et al, 2010). Building this celiac database included focusing on induced cytokine expression. Our search for peptides to include in this CeD database included focusing on induction of specific cytokines when stimulated with these peptides.

An exact peptide sequence matching algorithm was developed which searches to identity 100% identity matches with included CeD peptides. A full FASTA sequence alignment program was also developed with a database of representative gluten proteins to provides comparative sequence alignments with the parental proteins (68 in the 2012, 72 in 2017) for predicting potential risks of CeD in cases where some active peptides may have been missed. The database was tested both in 2012 and in 2017 following the update, with representative proteins from Pooideae and from non-Pooideae plants as well as proteins from fungi, bacteria and animal sources. Tests were performed by comparing the amino acid sequences of proteins with exact peptide matches and with FASTA alignments between each of the sequences of the Pooideae prolamins to evolutionary homologues from outside of Pooideae with no history of causing CeD.

The celiac database, bioinformatics tools, and established criteria are available for public use at http://www.allergenonline.org/celiachome.shtml for evaluation of any protein for potential risks of the proteins for risks to CeD consumers. The European Food Safety Authority (EFSA) recently developed a guideline stating the any new protein expressed in a GMO must be evaluated for safety to CeD consumers (Hanspeter et al, 2017). Initially they included reference to the AllergenOnline.org Celiac database, but now are recommending testing for exact identity matches to four amino acid peptides with specific allowed variation. Tests of the new proposal by us and by Ping Song et al, (2018) have demonstrated that that method has poor selectivity and a high false positive rate. Our tests with the current databased, as presented here, show a high predictive rate with 12% to 26% of proteins from banana to swine having at least one match per protein. Searches with the CeD database in Allergenonline.org have much higher true positive and lower false positive matches as test results report here.

5.3. Methods

5.3.1. Literature review and collection of CeD reactive peptides

The first version of the database was released in 2012 following searches and review of the PubMed literature database of the National Library of Medicine (http:// www.ncbi.nlm.nih.gov/pubmed) using keywords “celiac” and “coeliac” to identify studies investigating proteins and peptides capable of eliciting CeD pathogenesis. Overall, 68 relevant publications between November 1984 and October 2012 were used to select 1,016 gluten peptides of 8 to 55 AA long that stimulated CD4+ T cells of the restricted to MHC class II molecule DQ2.5, DQ2.2, DQ8 or DQ9 or were shown to elicit

toxic reactions in intestines of CeD subjects (Table 1). A positive reaction of CD4+ CeD T cells was proliferation that showed greater than a 2-fold stimulatory index upon presentation of peptide in the context of an appropriate MHCII or release of IFN-γ. Of the 997 peptides, 445 were in native form and 552 were in predicted deamidated peptide sequences. Many studies demonstrated that DQ2.5 preferentially binds to peptides having a 9-mer binding core with negatively charged anchors at positions P4, P6 or P7 whereas the DQ8 allele preferentially binds peptides with negatively charged anchors at positions at P1 and P9. In addition, but to a lesser extent, DQ2.5 and DQ8 alleles preferentially bind to peptides with proline (P) at positions P1 and P6, respectively (Sollid 2017; Vartdal et al, 1996; van de Wal et al, 1996; Kim et al, 2004; Kwok et al, 1996; Henderson et al, 2007). Digestion resistant gluten peptides lack polar acidic amino acids and are rich in proline and glutamine (Q). The position of specific amino acids in these peptides allows or inhibits deamidation by human TG2 in appropriately spaced Q residues, changing them to glutamic acid (E). These optimum sequences increase the binding avidity for HLA molecules allowing stimulation of gluten-specific T cells (Sollid 2017; Kim et al, 2004; van de Wal et al, 1998; Arentz-Hansen et al, 2000; Vader et al, 2002; Stepniak et al, 2010). The TG2 deamination is important for the selection of T cell epitopes, since most of the DQ2.5 recognized epitopes are in the deamidated form (Dorum et al, 2010). Interestingly, the DQ8 molecule recognizes gluten epitopes in both native and deamidated forms. The DQ8 receptor binding was shown to be to a native gluten epitope that is presented to the T cell receptor with a negative charge on β57 of the CDR3β loop, while the DQ8 molecule binding a deamidated prolamin epitope is present in the neutral CDR3β loop (Hovhannisyan et al, 2008). The specificity of TG2

deamidation of gluten peptides has not been conclusively demonstrated, but some residues are more effectively modified in peptides with the configuration of QXP where X represents most amino acids other than P (Vader et al, 2002; Dorum et al, 2010; Fleckenstein et al, 2002). An exceptionally immunogenic peptide is a decamer α-gliadin (p123-132: QLIPCMDVVL), which was found to possess a unique ability to induce HLA-A2 specific CD8+ T cells isolated from biopsies of CeD patients carrying either DQ2 or DQ8, causing the T cells to undergo maturation to express Fas ligand and to secrete IFN-γ and granzyme B (Gianfrani et al, 2003).

Of the 1,016 originally identified peptides, 18 elicited pathological effects to the intestine without evidence of specific T cell activation. These are categorized as toxic peptides. Some of the toxic peptides overlap immunogenic peptides. These peptides appeared to trigger innate immune responses. The toxic properties reported in publications included one or more of the following: reduction in epithelial brush border alkaline phosphatase activity; increased intestinal permeability; reduction in enterocyte surface cell height (ECH) or reduction in villus height to crypt depth ratio (VH:CD); expression of epithelial apoptotic mediator ligand HLA-E molecule; maturation and migration of macrophage, DC, and CD4+ T cells to the lamina propria; or expression of inflammatory cytokines IFN-γ, TNF-α and IL-15 (Auricchio et al, 1982; Barone et al, 2011; Caputo et al, 2010; de Ritis et al, 1994; Sturgess et al, 1994; Mantzaris and Jewell, 1991; Wieser et al, 1986; Biagi et al, 1999; Maiuri et al, 1996; Londei et al, 2005; Jabri and Sollid, 2009).

5.3.2. Construction of the database

In 2012, the 1,016 identified CeD were searched against the non-redundant NCBI Protein database by BLASTP to identify the source or homologous proteins. The BLASTP default search algorithm parameters were used with an Expect threshold (*E*-score) of 10, matrix selection of BLOSUM62, gap costs of 11 for existence and 1 for extension. The conditional compositional score matrix adjustment was used with no filtering or masking selection. The BLAST results showed 425 native peptides from the 1,016 identified peptides had identity matches with 147 prolamins of the Pooideae grass subfamily. The 147 proteins were then aligned using the EMBL-EBI multiple sequence alignment program ClustalW2. Identical proteins were removed, and 68 non-redundant proteins were collected as representative for CeD proteins. Bread wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*) are the major sources of the CeD active prolamins which account for about 63% (43 out of 68 proteins) and 16% (11 out of 68), respectively (Table 1).

In 2012, the 1,016 identified CeD peptides ranged from 8-55 AA. These peptides and the 68 representative CeD proteins linked with the NCBI Protein accession numbers, were loaded in a MySQL relational database management system. The 1,016 peptides linked to publications available in the browse function of the database. Query proteins from database users could be entered in the search window and compared to the database to see if they contain exact identity matches to any of the 1,016 peptides by an exact sequence match program. The 68 representative source proteins viewed in the browse function and sequences of query proteins from users could be compared for identity

scores to each of these CeD proteins by full-length FASTA3 sequence alignment, version 35.04 (Pearson 2000). The peptide and protein database sections and complete references of the 68 publications were available at http://www.allergenonline.org/celiachome.shtml from January 2012 until November 2017 when the database was updated.

5.3.3. Update of the database

In 2017, an additional literature and database review was conducted by a panel of six scientists. As a result, 34 peptides were identified as being too short (<9 amino acids) to be presented to T cells or without having clear published evidence of reactivity were removed. The core nine-amino acid peptides listed in the 2017 EFSA guidance on allergenicity assessment of genetically modified plants were added along with their deamidated forms (Hanspeter et al, 2007; Sollid et al, 2012). Four additional publications were added, bringing the total to 72 references. Another barley prolamin and three oat prolamins were identified and the final number of the representative CeD proteins in the database increased to 72 (Table 1). Database version 2 was posted online in October, 2017 and the text was revised to the current form in January 2018.

5.3.4. Testing the database to define criteria for potential risks for eliciting CeD

Tests were conducted in 2012 and 2017 using both the exact word match and FASTA35 that are available for public use to test a variety of protein sequences. Tests were performed with glutens from known CeD causing species (wheat, barley, rye and oats) and with homologous proteins from grain sources outside of Pooideae that have a history of safe use without causing CeD (maize, millet, rice, sorghum and others). The analyses were conducted using query sequences to identify proteins in the NCBI protein

database from CeD sources and non-CeD sources using keywords: gluten, glutelin, glutenin, prolamin, prolamine, gliadin, hordein, secalin, avenin, zein, kafirin, coixin, canein and pennisetin. Each of the known CeD protein sequences were searched against the non-redundant NCBI protein database by BLASTP using the Expect threshold of 10 and with the exclusion of the Pooideae proteins (NCBI taxonomic identifier: 147368) and with exclusion of patented proteins. The resulting sequences were compiled and sorted into four groups as follows: 1) 2,666 prolamins from the Pooideae subfamily that may be considered possibly unsafe for CeD patients; 2) 1,059 prolamins and prolamin related proteins from the grass subfamilies of Chloridoideae, Ehrhartoideae, and Panicoideae, sources that are known to be safe for CeD individuals; 3) 1,050 prolamin-like proteins from the Dicotyledon class that are known to be safe for CeD patients; and 4) 48 unrelated proteins, obtained solely from the BLAST search; and considered safe for CeD patients (Table 2). Results of each of the query sequences from manual searches against the CeD database using both the exact peptide match and FASTA3 search were recorded with exact match hits and FASTA sequence homology scores (percent identity score, alignment overlap length, and *E*-score) derived from all the searches. Evaluation of the FASTA3 alignment scores were used to set minimum percent identity and *E*-scores that suggest risks of CeD for version 1. Similar searches were used with version 2 to validate the criteria focusing on 1) 5,786 prolamins from the Pooideae subfamily; 2) 1,755 prolamins and prolamin related proteins from the grass subfamilies of Chloridoideae, Ehrhartoideae, and Panicoideae; and 3) 4,724 prolamin-like proteins from the Dicotyledon class. A summary of the results was used to set final criteria.

5.3.5. Tests using hypothetical alanine-substituted alpha-gliadin

To further evaluate the utility of using a FASTA3 alignment to the 72 representative proteins, the sequence of the α-gliadin of *Triticum aestivum* (NCBI GI number: 7209265) which contains 53 overlapping CeD active peptides identified with the exact sequence matching program (Figure 2A). The sequence was altered by substitutions in amino acid sequence to eliminate all exact peptide matches. Two *in silico* modification trials are presented as representatives that do not have peptide identity matches to the CeD. In Figure 2B), 13 theoretical substitutions were made with addition of alanine (A) in place of 12 glutamine (Q) and one tyrosine (Y) residues. In Figure 2C), 11 substitutions were made with addition of alanine (A) in place of three serine (S), two glycines (G), four lysine (L), one proline (P) and one glutamine (Q) amino acid residues. The modified alpha-gliadin sequences were evaluated using both exact peptide match to verify loss of identities and with FASTA3 to test the utility and verification limits for FASTA3 sequence alignment comparisons.

5.4. Results and Discussion

In the review publications of CeD reactive peptides, wide differences were noted in specificity, sensitivity and severity of described reactions (Stepniak et al, 2005). For example, pure oat products that are not contaminated by wheat, barley or rye, were reported to be well-tolerated by the majority of CeD consumers (Picarelli et al, 2001; Rashid et al, 2007). However, avenin-reactive T cells that mediate the intestinal inflammation typical of CeD were identified from a number of CeD patients (Vader et al, 2003; Arentz-Hansen et al, 2004; Real et al, 2012; Hardy et al, 2015). Since our aim is to

include all known prolamin peptides with scientific evidence of CeD induction to ensure that all CeD individuals are protected by our bioinformatics search tools, the reported T-cell reactive avenin peptides are included in our database.

Version 1 (2012) and version 2 of the celiac database (2018) are summarized in Table 1. Both versions included peptides that were published as stimulating CD4+ T cell proliferation from CeD subjects, in the context of MHC DQ 2 or DQ8, or as peptides that cause toxic responses to intestinal villi from biopsies of CeD subjects. Version 2 with 1,013 peptides is slightly smaller than version 1 (1,016) even though some new peptides were added as peptides of less than 9 amino acids were eliminated as being too small to efficiently bind MHC and activate T cells. All peptides are found solely in the prolamin storage proteins of the Pooideae subfamily of grasses, not in other cereals known to be safe for CeD patients such as corn, rice, sorghum, and millets (Figure 1). Our recommendation for users of this database is that any query protein found to contain even a single match to one of the known 1,013 peptides could represent a risk of eliciting CeD in susceptible individuals. These proteins should be tested further before being introduced into a "gluten-free" food. Our tests demonstrated exact matches to the 1,013 CeD active peptides are found only in proteins from Pooideae sequences or in predicted deamidation products of those sequences. We also recognize that nearly 21% (562 of 2,666) of the gluten-like proteins evaluated from Pooideae do not contain any of the known CeD reactive peptides (Table 2). Those proteins might or might not be safe for CeD consumers as some T-cell reactive, or toxic peptides may remain undiscovered (Koning et al, 2005) We therefore proposed using the full-length FASTA sequence

alignment tool to identify query proteins that may lack an exact peptide match to our peptide dataset, but may include previously undefined CeD reactive peptides.

In order to demonstrate the utility of using a FASTA alignment, we substituted alanine residues for amino acids in positions of exact CeD peptides of a clearly reactive α-gliadin (NCBI GI number: 7209265). The substitutions were made so that each of the known 53 overlapping CeD active peptides were no longer native (Figure 2A, B and C). The resulting protein sequences (Figures 2B and 2C) were searched for exact matches to verify that all exact peptide matches are not identified. When these substituted sequences were searched with the full FASTA3 sequence alignment tool, the two modified sequences showed >95.5% identity to α-gliadin with *E*-scores smaller than 1.1e-78 and we suggest that these conservative substitutions might be recognized by the MHC DQ 2 or 8 and by T-cells of CeD patients. Without laboratory or clinical evidence of safety, it is prudent to flag these two sequences that are highly homologous to the representative CeD protein as needing further testing before including them in food not labeled as containing gluten. It is clear when using the full FASTA3 sequence alignment comparison tool that careful evaluation of matching data is required since the query sequence can align with any of the representative CeD protein sequences in regions harboring the antigenic determinants or in regions (AA 101 to 200) without the antigenic peptide determinants (Figure 2A). Only a high percent identity score obtained from alignment with the regions harboring the antigenic determinants is relevant to CeD.

We recognize that there are many glutin-like homologous proteins in other grass subfamilies outside of Pooideae and even in dicotyledonous plants that are known to have

a clear history of safe consumption for those with CeD. The results of our FASTA comparisons with a large number of these homologues were collected to provide identity scores, alignment overlap lengths, and *E*-scores that were used to set limits to differentiate conservative safety guidelines that are useful to identify possibly risky sequences. The results from these FASTA analyses were performed using our first and now the second version of the database as summarized in Table 2. Full FASTA alignments indicated that the 562 Pooideae prolamins lacking any exact match to the known CeD reactive peptides, but with high identity FASTA alignments up to 98.4% over at least a half-protein length (187/288) and an *E*-score of 2.7e-45, but also up to 79.3% identical for a full-length (290/288) alignment with *E*-score of 3.5e-63 to representative CeD proteins. In contrast, although a number of query sequences in non-Pooidaea grass subfamilies (group II) were found to align with full-length FASTA alignments to representative CeD proteins, none were more than 43% identical to the representative CeD proteins. Many of the query sequences in group II represent very short alignments with the representative CeD proteins and with the minimum *E*-score of 3.5e-17. In addition, full-length alignment comparison analyses of the prolamin-like sequences from Dicotyledons class (group III) resulted in even lower identity scores and larger *E*-score values while short overlaps (10/20) had up to 60% identities with *E*-scores as large as 8. Last, FASTA identity scores of the protein sequences from animals, fungi or bacteria (group IV) show that most of the 48 proteins from group IV are hypothetical proteins based on genomic data, none of the sources are related to cereals and no evidence exists that these proteins can trigger the adverse immune responses relevant to CeD. The results indicated that these 48 proteins could produce full-length (437/439)

alignments with up to 41.2% identity and with a smallest *E*-score of 8.7e-25. These were mostly very short alignments with half-protein lengths (11/20) aligned with a maximum of 72.7% identity over the short length and having a minimum *E*-score of 5.8e-03. We observed that the sequences from the groups II, III, and IV did not align without gaps in alignment to the representative CeD proteins.

The results obtained from the second analysis, using version 2 of the database in 2018 tested a total of 12,265 sequences. The results (Table 3) were consistent with those obtained in the 2012 analysis. Taken together, the full FASTA sequence alignment appears to be useful to identify proteins with possible CeD risks. This provides a safety assurance that even if all CeD active peptides are not known, a FASTA alignment to this celiac database that identify an alignment of 45% or higher identity over a 100 amino acid overlap to the representative CeD proteins, and also having an *E*-score of smaller than 1e-14 should be taken as a potential risk to those with CeD. A protein meeting the criteria that suggests risk could be evaluated further by T-cell activation tests using CeD reactive T cell clones and antigen presenting cells or tetramers of MHC DQ 2 and DQ8. A positive result in such tests would more fully demonstrate a risk of CeD from that protein. The proposed evaluation scheme and criteria that we have chosen to assess novel food proteins of potential risk for eliciting CeD is depicted in Figure 3. Final criteria for CeD risky proteins were identified as those proteins with FASTA3 identity matches >45% over 100 amino acid alignments and with *E*-scores smaller than 1e-14 as potentially risky proteins for those with CeD.

In conclusion, cereal grains from other non-Pooideae grass subfamilies have not exhibited a history of eliciting CeD. Those grains can be used as alternative nutrient sources for those with CeD. The exact peptide sequence matching tool is the most definitive tool for risk assessment of any novel or GMO proteins identified to contain any of the known CeD active peptides as they likely pose a high risk to induce CeD. Due to incomplete knowledge on the CeD antigenic peptides, and the chance for mutations that might remove exact matching sequences, but possibly not diminish CeD antigenicity, we recommend the use of a full FASTA3 sequence alignment tool as an important back-up comparison for risk assessment. Any proposed new food protein with a FASTA3 scores of > 45% identity over more than 100 amino acid overlap and with an *E*-score < 1e-14 appears to be of potential risk for eliciting CeD and should be critically evaluated further for the safe use for CeD individuals. Among the existing gluten databases, the AllergenOnline.org celiac database contains the largest number of identified CeD reactive sequences (Juhasz et al, 2015; Bromilow et al, 2017). Our celiac peptide and protein database provides an effective screening system for identification and analysis of CeD reactive peptides and proteins for a thorough food safety evaluation, while also avoiding the high rate of false positive findings that occur if a four amino acid segment search recommended by the European Food Safety Authority in 2017 (Naegeli et al, 2017) is used for evaluation (Song et al, 2018). We anticipate maintaining this curated database in the future and will be verifying the accuracy of predictions for future updates using a similar evaluation protocol.

Table 1. Statistics of the AllergenOnline.org celiac peptide and protein database construction and inclusion characteristics.

		Version 1 (Released in 2012)	Version 2 (Released in 2018)
References	Number of publication references	68	**72**
	Publication year of references	1984 to 2012	1984 to **2017**
Peptides	Number of peptides	1,016	**1,013**
	Number of native peptides	464	**465**
	Number of deamidated peptides	552	**548**
	Number of immunogenic peptides	998	**1,004**
	Number of CD4+ T cell reactive peptides	997	**1,003**
	Number of CD8+ T cell reactive peptides	1	1
	Number of toxic peptides (without T cell reactivity)	18	**9**
	Length of peptides (amino acid)	8 - 55	**9** – 55
	Averaged length of peptides (amino acid)	16 ± 4	16 ± 4
Proteins	Number of proteins	68	**72**
	Number of proteins in *Triticum aestivum*	43	43
	Number of synthetic constructs in *Triticum aestivum*	1	1
	Number of proteins in *Triticum monococcum*	2	2
	Number of proteins in *Hordeum vulgare*	11	**12**
	Number of proteins in *Secale cereale*	6	6
	Number of proteins in *Avena sativa*	3	**6**
	Number of proteins in *Avena nuda*	2	2
	Length of proteins (amino acid)	20 - 800	20 – 800

Version 1 was released in 2012, version 2 in 2018. Both were based on data from publications testing proteins and peptides for responses in humans or in cultures of human samples, for T cell activation or toxic responses from biopsies. Changes between versions are in bold font.

Table 2. FASTA Sequence Identity Scores and Alignments of The Representative Prolamin-Like Protein Groups Clustered by Source Organism Types That Were Tested with The Allergenonline.Org Ced Protein Database Version 1.

	Group	Number of proteins searched from NCBI	Contain exact CeD active peptides	Best FASTA identity score results		
				Alignment overlap length (CeD protein length)	% Identity to the CeD protein	*E*-score
I	Prolamins in Pooideae	2,104 *	Yes	827 (827)	100	2.8e-179
				287 (290)	100	7.8e-81
				842 (838)	98.1	1.4e-195
	Prolamins in Pooideae	562 *	No	20 (20)	95	2.9e-05
				187 (288)	98.4	2.7e-45
				290 (288)	79.3	3.5e-63
II	Prolamins and prolamin-like proteins in Chloridoideae, Ehrhartoideae, and Panicoideae	1,059 * ‡	No	54 (52)	40.7	6.7
				12 (20)	66.7	1.9
				268 (360)	41	3.5e-17
III	Prolamin-like proteins in Dicotyledons	1,050 *	No	68 (68)	33.8	2.3
				10 (20)	60	8.8
				121 (648)	30.6	1.8e-06
IV	Unrelated proteins, (animals, fungi and microbes)	48 Δ	No	29 (29)	58.6	3.8
				11 (20)	72.7	5.8e-03
				437 (439)	41.2	8.7e-25

* Proteins were identified from the NCBI protein database using keywords: gluten, glutelin, glutenin, prolamin, prolamine, gliadin, hordein, secalin, avenin, zein, kafirin, coixin, canein and pennisetin

‡ 35 proteins were obtained by BLAST searched the 68 representative celiac proteins against the NCBI Protein-Protein *(non-redundant sequences)* database with the exclusion of Pooideae (taxid: 147368)

Δ proteins were obtained by BLAST searches with the 68 representative celiac proteins against the NCBI Protein-Protein *(non-redundant sequences)* database with the exclusion of Pooideae (taxid: 147368).

Table 3. Repeat of The FASTA Sequence Identity Scores and Alignments of The Larger Representative Prolamin-Like Protein Groups Clustered by Source Organism Types That Were Tested with The Allergenonline Ced Protein Database Version 2.

Group		Number of proteins searched from NCBI	Contain exact CeD active peptides	Best FASTA identity score results		
				Alignment overlap length (CeD protein length)	% Identity to the CeD protein	*E*-score
I	Prolamins in Pooideae	4623 *	Yes	828 (828)	100	1.0e-177
				439 (290)	100	1.6e-165
				455 (455)	100	8.4e-153
	Prolamins in Pooideae	1163 *	No	291 (288)	98.6	3.7e-09
				264 (279)	98.9	1.1e-73
				266 (269)	98.5	3.6e-68
II	Prolamins and prolamin-like proteins in other monocots	1755 *‡	No	292 (250)	37.3	3.6e-09
				168 (181)	40.5	9.1e-09
				222 (222)	37.4	2.4e-08
III	Prolamin-like proteins in Dicotyledons	4724 *	No	305 (838)	32.1	1.6e-04
				372 (439)	28.8	9.5e-04
				253 (290)	29.2	9.3e-03

* Proteins were identified from the NCBI protein database using keywords: gluten, glutelin, glutenin, prolamin, prolamine, gliadin, hordein, secalin, avenin, zein, kafirin, coixin, canein and pennisetin

‡ 35 proteins were obtained by BLAST searched the 68 representative celiac proteins against the NCBI Protein-Protein *(non-redundant sequences)* database with the exclusion of Pooideae (taxid: 147368)

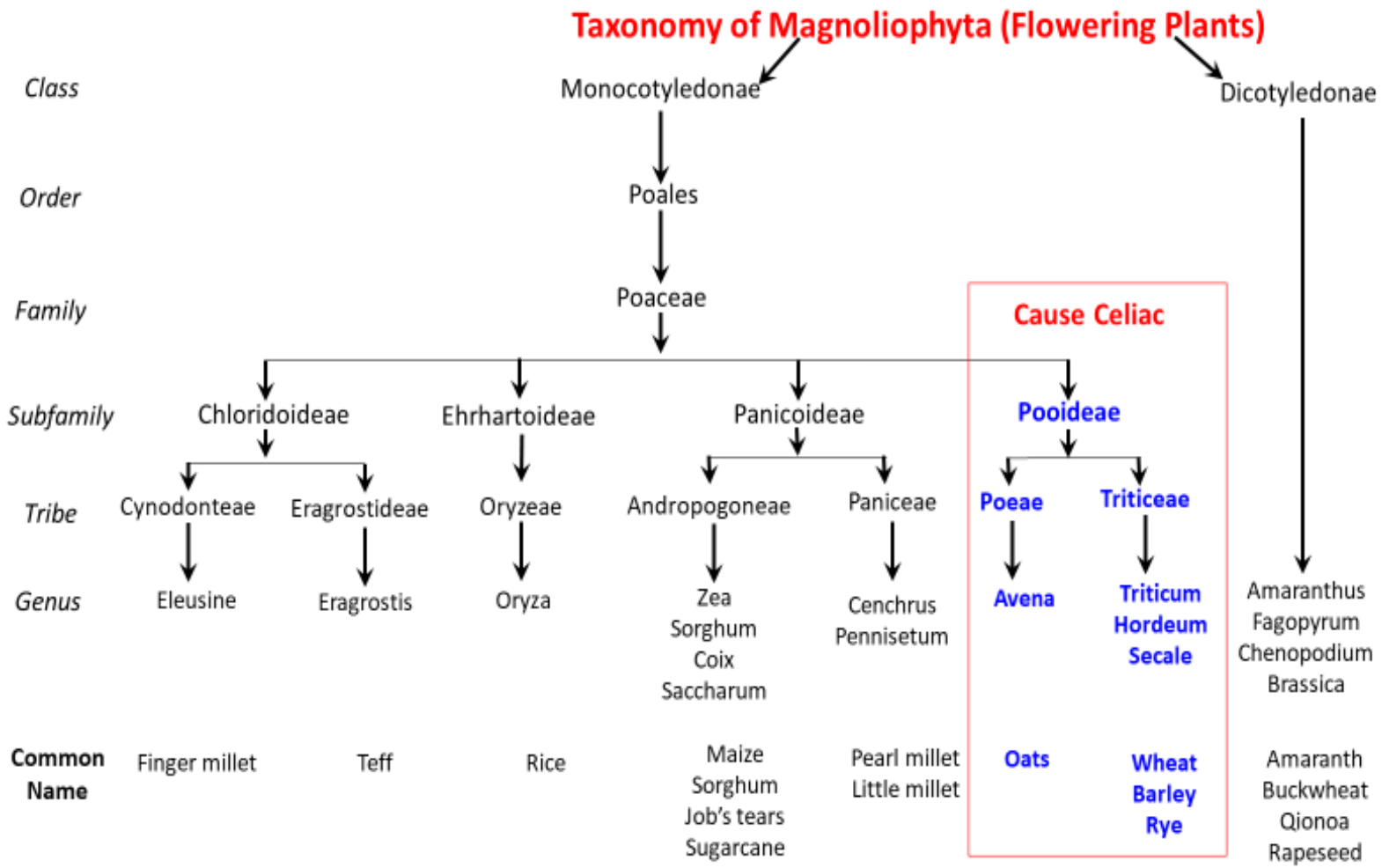


Figure 1. Taxonomic Tree of Cereals and Dicotyledonous Plants Based on NCBI Taxonomy. Published evidence of CD safe foods show reactions only to grains from members of the Pooideae sub-family of grasses.

A) Exact peptide match with first 100 AA of accession # CAB76964.1 alpha-gliadin [Triticum aestivum]

Exact peptide matches with additional sequences and alignment with CAB76964.1

B) FASTA alignment of 13 AA substituted with CAB76964.1 AllergenOnline.org/celiac Proteins

```
>>gi|7209265|emb|CAB76964.1| alpha-gliadin [Triticum aestivum]                    (290 aa)
 initn: 1996 init1: 1996 opt: 1996  Z-score: 1365.3  bits: 260.6 E(): 2.2e-072
Smith-Waterman score: 1996; 95.5% identity (95.5% similar) in 290 aa overlap (1-290:1-290)

               10        20        30        40        50        60        70        80        90       100
Sub    MVRVPVPQLQPQNPSQAQPQEQVPLVQQQQFPGQQAPFPPQQPYPQPQPFPSAQPYLQLQPFPQPQLPYPAPALPYPQPALPYPQPQPFRPAQPYPQSQP
       :::::::::::::::: :::::::::::::::::: ::::::::::::::::: :::::::::::::::::: : :::::: :::::::::::: ::::::::
gi|720 MVRVPVPQLQPQNPSQQQPQEQVPLVQQQQFPGQQQPFPPQQPYPQPQPFPSQQPYLQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPFRPQQPYPQSQP
               10        20        30        40        50        60        70        80        90       100

              110       120       130       140       150       160       170       180       190       200
Sub    QYSQPQQPISQQQQQQQQQQQQKQQQQQQQQQILQQILQQALIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILH
       :::::::::::::::::::::::::::::::::::::::: :::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|720 QYSQPQQPISQQQQQQQQQQQQKQQQQQQQQQILQQILQQQLIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILH
              110       120       130       140       150       160       170       180       190       200

              210       220       230       240       250       260       270       280       290
Sub    QQQQQQQAQQQQPLSQVSFQQPQQQYPSGAGSFQPSQANPQAQGSVAPQQLPQFEEIRNLALETLPAMCNVYIPPACTIAPVGIFGTNYR
       ::::::: :::::::::::::::::::: ::::::: :::::::: :::::::::::::::::::::::::::::: ::::::::::::::
gi|720 QQQQQQQQQQQQQPLSQVSFQQPQQQYPSGQGSFQPSQQNPQAQGSVQPQQLPQFEEIRNLALETLPAMCNVYIPPYCTIAPVGIFGTNYR
              210       220       230       240       250       260       270       280       290
```

C) FASTA alignment of 11 AA substituted with CAB76964.1 AllergenOnline.org/celiac Proteins

```
>>gi|7209265|emb|CAB76964.1| alpha-gliadin [Triticum aestivum]                    (290 aa)
 initn: 2027 init1: 2027 opt: 2027  Z-score: 1640.0  bits: 311.4 E(): 1.1e-087
Smith-Waterman score: 2027; 96.2% identity (97.9% similar) in 290 aa overlap (1-290:1-290)

               10        20        30        40        50        60        70        80        90       100
Sub    MVRVPVPQLQPQNPAQQQPQEQVPLVQQQQFPAQQQPFPPQQPYPQPQPFPAQQPYLQAQPFPQAQLPYPQPQAPYPQPQAPYPQPQPFRPQAPYPQSQP
       ::::::::::::::.::::::::::::::::::::.::::::: ::::: :::::::: :::::: :::::::::: :::::::
gi|720 MVRVPVPQLQPQNPSQQQPQEQVPLVQQQQFPGQQQPFPPQQPYPQPQPFPSQQPYLQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPFRPQQPYPQSQP
               10        20        30        40        50        60        70        80        90       100

              110       120       130       140       150       160       170       180       190       200
Sub    QYSQPQQPISQQQQQQQQQQQQKQQQQQQQQILQQIAQQQLIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILH
       :::::::::::::::::::::::::::::::::: ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
gi|720 QYSQPQQPISQQQQQQQQQQQQKQQQQQQQQILQQILQQQLIPCRDVVLQQHSIAYGSSQVLQQSTYQLVQQLCCQQLWQIPEQSRCQAIHNVVHAIILH
              110       120       130       140       150       160       170       180       190       200

              210       220       230       240       250       260       270       280       290
Sub    QQQQQQQQQQQQQPLSQVSFQQPQQQYPSGQASFQPSQQNPQAQGAVQPQQLPQFEEIRNLALETLPAMCNVYIPPYCTIAPVGIFGTNYR
       ::::::::::::::::::::::::::::::.:::::::::::::.::::::::::::::::::::::::::::::::::::::::::::::
gi|720 QQQQQQQQQQQQQPLSQVSFQQPQQQYPSGQGSFQPSQQNPQAQGSVQPQQLPQFEEIRNLALETLPAMCNVYIPPYCTIAPVGIFGTNYR
              210       220       230       240       250       260       270       280       290
```

Figure 2A, B, and C. Amino Acid Sequence Alignments of An A-Gliadin (NCBI GI number: 7209265) with 53 overlapping CeD reactive peptides identified with the exact sequence match tool (A), full FASTA sequence alignment results with homology scores of the α-gliadin theoretically substituted with 13 alanine residues (B), and with 11 alanine residues (C).

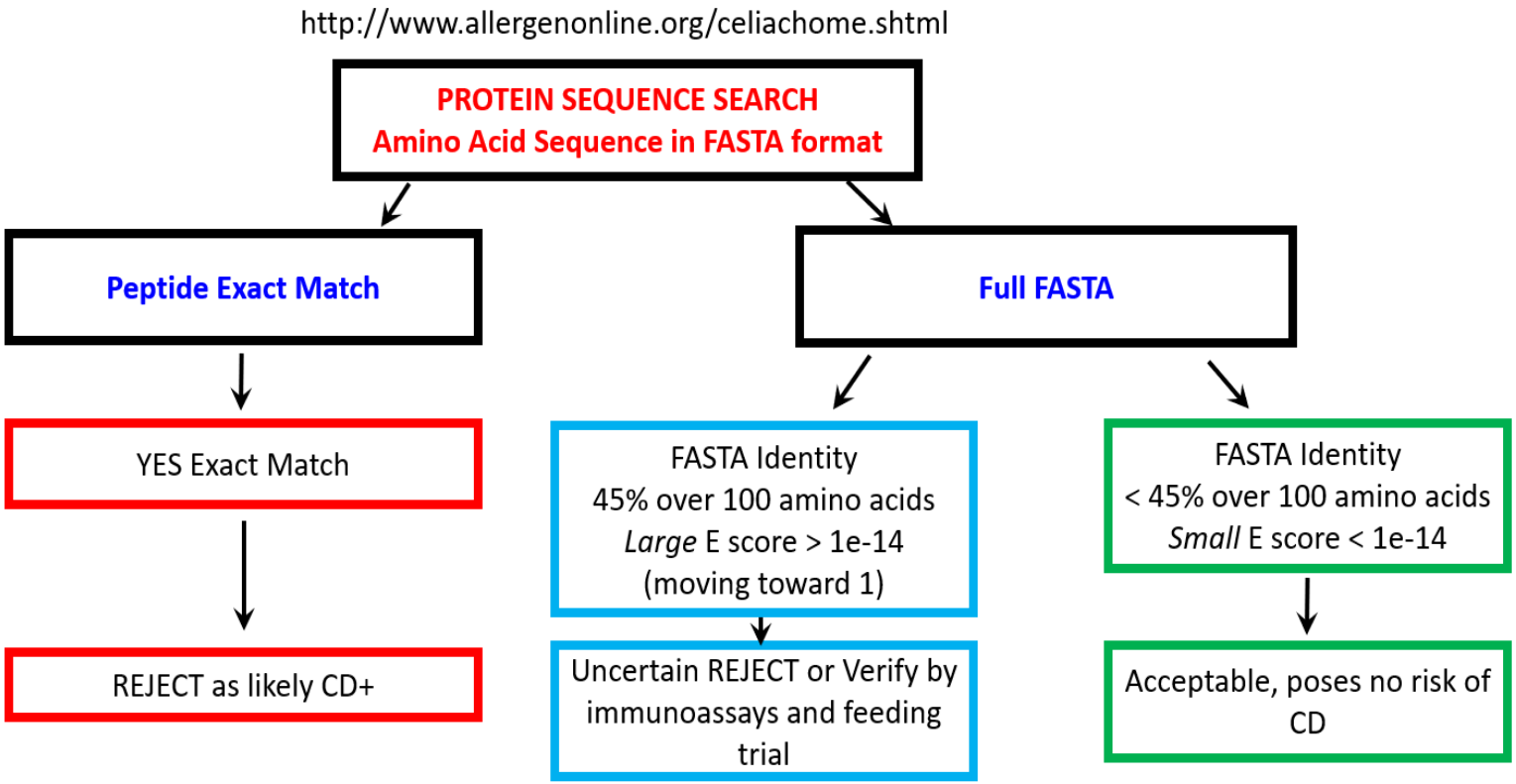


Figure 3. Evaluation Criteria to Predict the Likelihood of A Query Protein to Cause Elicitation of Ced. An exact match to any of the 1,013 peptides indicates likely rejection. Alternatively, a FASTA3 alignment with an *E*-score limit of 1e -14 and minimum alignment length >100 AA with an identity percent of the protein at 50% should trigger testing or rejection.

5.5. References

Arentz-Hansen, H., Fleckenstein, B., Molberg, O., et al, 2004. The molecular basis for oat intolerance in patients with celiac disease. PLoS Med 1:e1.

Arentz-Hansen, H., Korner, R., Molberg, O., et al, 2000. The intestinal T cell response to alpha-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase. J Exp Med 191:603-612.

Auricchio, S., De Ritis, G., De Vincenzi, M., et al, 1982. Effects of gliadin-derived peptides from bread and durum wheats on small intestine cultures from rat fetus and coeliac children. Pediatr Res 16:1004-1010.

Barone, M.V., Zanzi, D., Maglio, M., et al., 2011. Gliadin-mediated proliferation and innate immune activation in celiac disease are due to alterations in vesicular trafficking. PLoS One 6:e17039.

Ben Ahmed, M., Belhadj Hmida, N., Moes, N., et al, 2009. IL-15 renders conventional lymphocytes resistant to suppressive functions of regulatory T cells through activation of the phosphatidylinositol 3-kinase pathway. J Immunol 182:6763-6770.

Biagi, F., Ellis, H.J., Parnell, N.D., et al, 1999. A non-toxic analogue of a coeliac-activating gliadin peptide: A basis for immunomodulation? Aliment Pharmacol Ther 13:945-950.

Bromilow, S., Gethings, L.A., Buckley, M., et al., 2017. A curated gluten protein sequence database to support development of proteomics methods for determination of gluten in gluten-free foods. J Proteomics 163:67-75.

Caputo, I., Barone, M.V., Lepretti, M., et al, 2010. Celiac anti-tissue transglutaminase antibodies interfere with the uptake of alpha gliadin peptide 31-43 but not of peptide 57-68 by epithelial cells. Biochim Biophys Acta 1802:717-727.

Codex Alimentarius Commission, 2003. Alinorm 03/34: Joint FAO/WHO Food Standard Programme, Codex Alimentarious Commission, Twenty-Fifth Session, Rome, Italy, 30 June–5 July, 2003. Appendix III, Guideline for the conduct of food safety assessment of foods derived from recombinant-DNA plants, and Appendix IV, Annex on the assessment of possible allergenicity. In: Anonymous Joint FAO/WHO Food Standards Programme. Twenty-Fifth Session (FA). Rome, Italy. ed. 2003:47-60.

Cukrowska, B., Sowinska, A., Bierla, J.B., et al, 2017. Intestinal epithelium, intraepithelial lymphocytes and the gut microbiota - Key players in the pathogenesis of celiac disease. World J Gastroenterol 23:7505-7518.

de Ritis, G., Auricchio, S., Jones, H.W., et al, 1988. In vitro (organ culture) studies of the toxicity of specific A-gliadin peptides in celiac disease. Gastroenterology 94:41-49.

DePaolo, R.W., Abadie, V., Tang, F., et al, 2011. Co-adjuvant effects of retinoic acid and IL-15 induce inflammatory immunity to dietary antigens. Nature 471:220-224.

Di Sabatino, A., Corazza, G.R, 2009. Coeliac disease. The Lancet 373:1480-1493.

Dorum, S., Arntzen, M.O., Qiao, S.W., et al, 2010. The preferred substrates for transglutaminase 2 in a complex wheat gluten digest are peptide fragments harboring celiac disease T-cell epitopes. PLoS One 5:e14056.

Dubois, P.C.A., Trynka, G., Franke, L., et al, 2010. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet 42:295-302.

Fasano, A., Berti, I., Gerarduzzi, T., et al, 2003. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: A large multicenter study. Archives of Internal Medicine 163:286-292.

Fasano, A., 2011. Zonulin and its regulation of intestinal barrier function: The biological door to inflammation, autoimmunity, and cancer. Physiol Rev 91:151-175.

Fleckenstein, B., Molberg, O., Qiao, S-W, et al, 2002. Gliadin T cell epitope selection by tissue transglutaminase in celiac disease. J Biol Chem 277(37):34109-34116.

Gianfrani, C., Troncone, R., Mugione, P., et al, 2003. Celiac disease association with CD8+ T cell responses: identification of a novel gliadin-derived HLA-A2-restricted epitope. J Immunol 170:2719-2726.

Groschwitz, K.R., Hogan, S.P, 2009. Intestinal barrier function: Molecular regulation and disease pathogenesis. J Allergy Clin Immunol 124:3-20.

Hanspeter, N., Birch, A.N., Josep, C., et al, 2017. Guidance on allergenicity assessment of genetically modified plants. EFSA Journal 15:e04862.

Hardy, M.Y., Tye-Din, J.A., Stewart, J.A., et al, 2015. Ingestion of oats and barley in patients with celiac disease mobilizes cross-reactive T cells activated by avenin peptides and immuno-dominant hordein peptides. J Autoimmun 56:56-65.

Henderson, K.N., Tye-Din, J.A., Reid, H.H., et al, 2007. A structural and immunological basis for the role of human leukocyte antigen DQ8 in celiac disease. Immunity 27:23-34.

Hovhannisyan, Z., Weiss, A., Martin, A., et al, 2008. The role of HLA-DQ8 beta57 polymorphism in the anti-gluten T-cell response in coeliac disease. Nature 456:534-538.

Hunt, K.A., Zhernakova, A., Turner, G., et al, 2008. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 40:395-402.

Jabri, B., Sollid, L.M., 2017. T cells in celiac disease. J Immunol 198:3005-3014.

Jabri, B., Sollid, L.M., 2009. Tissue-mediated control of immunopathology in coeliac disease. Nat Rev Immunol 9:858-870.

Juhasz, A., Haraszi, R., Maulis, C., 2015. ProPepper: a curated database for identification and analysis of peptide and immune-responsive epitope composition of cereal grain protein families. Database (Oxford) 2015:10.1093/database/bav100. Print 2015.

Karell, K., Louka, A.S., Moodie, S.J., et al, 2003. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: Results from the European genetics cluster on celiac disease. Hum Immunol 64:469-477.

Kim, C.Y., Quarsten, H., Bergseng, E., et al, 2004. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc Natl Acad Sci USA 101:4175-4179.

Koning, F., Gilissen, L., Wijmenga, C., 2005. Gluten: A two-edged sword. Immunopathogenesis of celiac disease. Springer Semin Immunopathol 27:217-232.

Kwok, W.W., Domeier, M.L., Raymond, F.C., et al, 1996. Allele-specific motifs characterize HLA-DQ interactions with a diabetes-associated peptide derived from glutamic acid decarboxylase. J Immunol 156:2171-2177.

Lammers, K.M., Lu, R., Brownley, J., et al, 2008. Gliadin induces an increase in intestinal permeability and zonulin release by binding to the chemokine receptor CXCR3. Gastroenterology 135:194-204.

Londei, M., Ciacci, C., Ricciardelli, I., et al, 2005. Gliadin as a stimulator of innate responses in celiac disease. Mol Immunol 42:913-918.

Maiuri, L., Troncone, R., Mayer, M., et al, 1996. In vitro activities of A-gliadin-related synthetic peptides: Damaging effect on the atrophic coeliac mucosa and activation of mucosal immune response in the treated coeliac mucosa. Scand J Gastroenterol 31:247-253.

Malamut, G., El Machhour, R., Montcuquet, N., et al, 2010. IL-15 triggers an antiapoptotic pathway in human intraepithelial lymphocytes that is a potential new target in celiac disease-associated inflammation and lymphomagenesis. J Clin Invest 120:2131-2143.

Mantzaris, G., Jewell, D.P., 1991. In vivo toxicity of a synthetic dodecapeptide from A gliadin in patients with coeliac disease. Scand J Gastroenterol 26:392-398.

Mention, J., Ben Ahmed, M., Bègue, B., et al, 2003. Interleukin 15: A key to disrupted intraepithelial lymphocyte homeostasis and lymphomagenesis in celiac disease. Gastroenterology 125:730-745.

Meresse, B., Chen, Z., Ciszewski, C., et al, 2004. Coordinated induction by IL15 of a TCR-independent NKG2D signaling pathway converts CTL into lymphokine-activated killer cells in celiac disease. Immunity 21:357-366.

Naegeli, H., Birch, A.N., Casacuberta, J. et al, 2017. Guidance on allergenicity assessment of genetically modified plants. EFSA Journal 15(6):4862.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185-219.

Perez-Gregorio, M.R., Dias, R., Mateus, N., et al, 2018. Identification and characterization of proteolytically resistant gluten-derived peptides. Food Funct 9:1726-1735.

Picarelli, A., Di Tola, M., Sabbatella, L., et al, 2001. Immunologic evidence of no harmful effect of oats in celiac disease. Am J Clin Nutr 74:137-140.

Polvi, A., Arranz, E., Fernandez-Arquero, M., et al, 1998. HLA-DQ2-negative celiac disease in Finland and Spain. Hum Immunol 59:169-175.

Rashid, M., Butzner, D., Burrows, V., et al, 2007. Consumption of pure oats by individuals with celiac disease: A position statement by the Canadian Celiac Association. Can J Gastroenterol 2007;21:649-651.

Real, A., Comino, I., de Lorenzo, L., et al, 2012. Molecular and immunological characterization of gluten proteins isolated from oat cultivars that differ in toxicity for celiac disease. PLoS One 7:e48365.

Roberts, A.I., Lee, L.F., Schwarz, E.F., et al, 2001. NKG2D receptors induced by IL-15 costimulate CD28-negative effector CTL in the tissue microenvironment. J Immunol 167:5527-5530.

Romanos, J., van Diemen, C.C., Nolte, I.M., et al, 2009. Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology 137:834-840.

Rubio-Tapia, A., Ludvigsson, J.F., Brantner, T.L., et al, 2012. The prevalence of celiac disease in the United States. Am J Gastroenterol 107:1538-1544.

Singh, P., Arora, A., Strand, T.A., et al, 2018. Global prevalence of celiac disease: Systematic review and meta-analysis. Clin Gastroenterol Hepatol 16:823-836.

Sollid, L.M., Qiao, S.W., Anderson, R.P., et al, 2012. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. Immunogenetics 64:455-460.

Sollid, L.M., 2017. The roles of MHC class II genes and post-translational modification in celiac disease. Immunogenetics 69:605-616.

Song, P., Podevin, N., Mirsky, H., et al, 2018. Q-X1-P-X2 motif search for potential celiac disease risk has poor selectivity. Regulatory Toxicology and Pharmacology 99:233-237.

Stepniak, D., Vader, L.W., Kooy, Y., et al, 2005. T-cell recognition of HLA-DQ2-bound gluten peptides can be influenced by an N-terminal proline at p-1. Immunogenetics 57:8-15.

Sturgess, R., Day, P., Ellis, H.J., et al, 1994. Wheat peptide challenge in coeliac disease. Lancet 343:758-761.

Tang, F., Chen, Z., Ciszewski, C., et al, 2009. Cytosolic PLA2 is required for CTL-mediated immunopathology of celiac disease via NKG2D and IL-15. J Exp Med 206:707-719.

Tripathi, A., Lammers, K.M., Goldblum, S., et al, 2009. Identification of human zonulin, a physiological modulator of tight junctions, as prehaptoglobin-2. Proc Natl Acad Sci U S A 106:16799-16804.

Trynka, G., Hunt, K.A., Bockett, N.A., et al, 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43:1193-1201.

Vader, L.W., de Ru, A., van der Wal, Y., et al, 2002. Specificity of tissue transglutaminase explains cereal toxicity in celiac disease. J Exp Med 195:643-649.

Vader, L.W., Stepniak, D.T., Bunnik, E.M., et al, 2003. Characterization of cereal toxicity for celiac disease patients based on protein homology in grains. Gastroenterology 125:1105-1113.

van de Wal, Y., Kooy, Y., van Veelen, P., et al, 1998. Selective deamidation by tissue transglutaminase strongly enhances gliadin-specific T cell reactivity. J Immunol 161:1585-1588.

van de Wal, Y., Kooy, Y.M., Drijfhout, J.W., et al, 1996. Peptide binding characteristics of the coeliac disease-associated DQ(alpha1*0501, beta1*0201) molecule. Immunogenetics 44:246-253.

van Putten, M.C., Frewer, L.J., Gilissen, L.J.W.J., et al, 2006. Novel foods and food allergies: A review of the issues. Trends Food Sci Technol 17:289-299.

Vartdal, F., Johansen, B.H., Friede, T., et al, 1996. The peptide binding motif of the disease associated HLA-DQ (alpha 1* 0501, beta 1* 0201) molecule. Eur J Immunol 26:2764-2772.

Wieser, H., Belitz, H.D., Idar, D., et al, 1986. Coeliac activity of the gliadin peptides CT-1 and CT-2. Z Lebensm Unters Forsch 182:115-117.

Zanzi, D., Stefanile, R., Santagata, S., et al, 2011. IL-15 interferes with suppressive activity of intestinal regulatory T cells expanded in celiac disease. Am J Gastroenterol 106:1308-1317.

5.6. Acknowledgment and Financial Support

The authors thank the Royal Thai Government Scholarship for its support to P. Amnuaycheewa during his PhD studies, and the Government of Egypt for funding Mohamed Abdelmoteleb during his PhD studies. The authors thank Professor Frits Koning of Leiden University Medical Center for stimulating our research to develop the celiac database in 2009. We thank Dr. Joe Murray of Mayo Clinic for accepting PA into his laboratory to work on a mouse model of Celiac Disease in 2013.

CHAPTER 6

EFFECT OF DIETARY NITRATES AND SULFATES ON ENTERIC METHANE MITIGATION IN FINISHING CATTLE

This chapter is in progress to be submitted for peer review: Abdelmoteleb M., Allie Knoell, Samodha C. Fernando

6.1. Abstract

The objectives of this study were to investigate the effect of nitrate and sulfate supplementations on cattle performance and methane emissions in finishing diets and to identify the effect of nitrate and sulfate addition on rumen microbiota composition and function**.** One hundred and thirty one day feeding trial was conducted using 24 head of cattle (initial BW = 918 lb; SD = 79 lb) where the cattle received one of four treatments no supplementation (CT), 2.0% dietary nitrate (NT), 0.54% dietary sulfate (SF) or COMBO (SF+NT), with 6 steers per treatment. Performance, and CH_4:CO_2 emissions data were analyzed using MIXED procedure of SAS. Rumen samples were collected and analyzed through amplicon sequencing targeting the 16S rDNA gene V4 bacterial and V6 archaeal regions and through shotgun metagenomics. Microbiome richness and composition were analyzed using DADA2 and Phyloseq. Microbial genes involved in pathways linked to methanogenesis, nitrate, and sulfate metabolism were identified using metagenomic sequencing information. Gene prediction, functional profile and pathway mapping were conducted using the KEGG database. Diets with only sulfate or nitrate, diet had no impact on CH_4:CO_2 emission ratio, but nitrate and sulfate in combination decreased CH_4:CO_2 ratio significantly. A reduction in dry matter intake (DMI) ($P < 0.01$),

average daily gain (ADG) (P = 0.07) and gain:feed (G:F) (P = 0.09) was also recorded. Significant increase in bacterial phyla with less H_2 production e.g. Proteobacteria; and genera with H2 utilization capability e.g. propionate, lactate forming bacteria e.g. *Prevotella, Bifidobacterium, Megasphaera, Selenomonas, Lactobacillus*; nitrate and sulfate reducing bacteria e.g. *Selenomonas, Desulfovibrio* was observed in COMBO diet. Differential gene abundance in metabolic pathways demonstrated decrease of enzymes linked to methanogenesis in COMBO diet. This study provides evidence that methane emission is linked to diet type and differential gene abundance in the cattle rumen microbiome.

6.2. Introduction

Agriculture represents 9% of the total greenhouse gases (GHGs) emissions in the US (EPA, 2020). Methane production through enteric fermentation in ruminants accounts for 27% of the total global methane emissions. Methane is a greenhouse gas, with a global warming potential 28 times that of CO_2 (Myhre et al, 2013). Methane production through enteric microbial fermentation in ruminants is an environmental as well as a nutritional concern (Moss et al. 2000). As an environmental concern, ruminants account for 97% of the total methane produced by domesticated animals and 75% of the methane produced by ruminants is produced by cattle (Crutzen et al, 1986; Mangino et al, 2007). As a nutritional concern, methane losses can vary from 2 to 12% of total gross energy intake cattle should otherwise use for performance and milk production (Johnson and Ward, 1996; Johnson and Johnson, 1995; Hristov et al. 2013).

Methane production through enteric fermentation can be summarized in four steps. The first step is breakdown of complex organic matter (carbohydrates, proteins, lipids) into soluble organic molecules (sugars, amino acids, fatty acids) followed by acidogenesis into alcohols, and acetogenesis into fatty acids (Russell 2002). The first three steps are controlled by rumen microbiota including bacteria, protozoa, fungi, and viruses. The final step includes H_2 utilization, which produced in the first three steps, in conversion of fatty acids, ammonia, and CO_2 into methane by methanogens (mainly archaea) in a process called methanogenesis (Russell 2002, Shah 2014). At the heart of methane production are microbes, and these microbes are known to change based on substrate availability in the diet (Danielsson et al. 2017). As diet can change microbial communities, dietary intervention can be used to reduce greenhouse gas emissions from cattle by controlling microbial populations (Van Zijderveld et al, 2010). Therefore, understanding the relationship between diet, methane, and microbial community will help identify microbial species associated with methane to develop new intervention strategies.

Dietary intervention strategies for mitigation of methane have been explored. Many studies have been conducted to identify strategies to minimize methane production. In a review, Hristov et al, 2013 stated that feeding tannins has often shown up to a 20% decrease in methane emissions. Other strategies, such as processing corn as steam flaked rather than dry rolled has been shown to decrease methane emissions in beef cattle (Hales et al, 2012). However, although these strategies exist, they have not been widely implemented by producers. Other approaches e.g. hydrogen utilization through

microbiota involved in digestion, alternative hydrogen sink and anti-methanogens have been used to mitigate methane emissions (Van Zijderveld et al, 2010).

In this study, we focus on alternative H+ sink approach for methane mitigation. One of the most popular alternative H+ sink in literature the last few years are nitrate and sulfate. Nitrates may serve as a terminal electron acceptor and therefore may behave as alternate hydrogen sink and can be converted to ammonia (Ungerfeld and Kohn, 2006). Sulphates can also act as potent methane inhibitor in many anaerobic systems including rumen. Reduction of sulphate leads to production of hydrogen sulfide (H_2S) which appears to play a role of electron donor in the reduction of nitrite to ammonia by nitrate-reducing, sulfide-oxidizing bacteria (Ungerfeld and Kohn, 2006). There is a debate between different studies about their impact in methane mitigation. However, the major concern is that both are toxic. Excess nitrate will lead to nitrogen dioxide production, which is also a greenhouse gas, and causes cattle toxicity by conversion of hemoglobin into methemoglobin (Van Zijderveld et al, 2010). Similarly, excessive sulfate will increase hydrogen sulfide production, which is also toxic, and odorous (Sarturi et al, 2013).

The rumen microbial community composition is poorly characterized when identifying methane mitigation strategies. The ability to identify microbial community structure while simultaneously measuring methane will provide a better understanding of the microbial composition on various commonly fed finishing diets and provide a better understanding of potential dietary intervention strategies in finishing feedlot cattle. Mitigation of ruminal methanogenesis can be evaluated through inhibiting archaeal methanogens and their effects on bacterial communities. As some bacterial taxa are

known for their capability in hydrogen utilization, will help understanding reasons behind reduction of methane emissions (Russell, 2002). Therefore, improving our understanding not only on the efficacy of methods to decrease methane emissions, but also on potential detrimental effects on nutrient digestion and animal production performance, where bacteria play a crucial role. However, further research is required to evaluate effects on the ruminal archaeal bacterial community structure using high-throughput DNA sequencing (Danielsson et al. 2017), and to evaluate how shifts in the community composition may potentially be associated with methane emissions.

Combining metagenomics to explore the effects of diets on enzymes and microorganisms involved in methane metabolism could further reveal integrative information of rumen function. Shabat et al, 2016 measured feed efficiency in 146 milking cows and performed analysis of microbiome and metabolome composition. They observed specific enrichment of microbes and metabolic pathways in each of these microbiome groups resulted in better energy and carbon channeling to the animal with reducing methane emissions to the atmosphere. In a similar study, She et al, (2014) explored the mechanistic basis of methane production in 22 sheeps with high and low methane yield through deep metagenomic and meta-transcriptomic sequencing. They demonstrated that transcription of methanogenesis pathway genes was substantially increased in sheep with high methane yields with significant increase in rumen methanogens.

Therefore, the objective of this study was to determine whether nitrate and/or sulfate may be effective as a methane mitigation strategy in finishing diets; understand

the impact of diet modification on the microbiome richness and composition; and finally, how microbiota will affect metabolic methane emissions.

6.3. Methods

6.3.1. Animals and experimental design

All animal care and management practices were approved by the University of Nebraska-Lincoln Institutional Animal Care and Use Committee. Initially, 24 head of cattle were limit fed a growing diet (50% alfalfa hay and 50% Sweet Bran® at 2% of BW) to reduce variation in gut fill. Cattle (initial BW = 918 lb; SD = 79 lb) were assigned for 131-day randomly to one of four treatments of finishing diet no supplementation (CT), 2.0% dietary nitrate (NT), 0.54% dietary sulfate (SF) or COMBO (SF+NT), with 6 steers per treatment (Table 1). On d 131, cattle were transported to a commercial abattoir (Greater Omaha Packing, Omaha, NE) to be harvested. All carcass data were collected. Methane and CO_2 were collected and analyzed, and emissions values were calculated as described previously (Pesta, 2015). Briefly, gas samples were collected from each steer 9 times, every 14 d throughout the feeding period. Prior to feeding on d 60, cattle were esophageally tubed to obtain 45 mL of rumen contents for microbial community and VFA profile analysis (Paz et al, 2016). This experiment was structured as a randomized block design with 2 blocks (by location of Calan bunks). Performance, and emissions data were analyzed with the PROC MIXED procedure of SAS (SAS Institute Inc., Cary, N.C.) with cattle as the experimental unit. Treatments were analyzed as a 2 × 2 factorial with the model including the main effects of nitrate and sulfate as well as the nitrate × sulfate interaction. Change in CH_4:CO_2 throughout the finishing period was analyzed as a repeated measure with the repeated variable being

sampling time point and steer being the subject. Variability in the data was expressed as the standard error of means (SEM), $P < 0.05$ was considered to be statistically significant and $P \leq 0.10$ was considered a statistical trend.

6.3.2. 16S rRNA library preparation, sequencing, and bioinformatics analysis of the V4 Bacteria and V6 Archaea Regions

6.3.2.1. Rumen sampling and DNA Isolation

A representative sample of rumen contents (solid particles and rumen fluid) of 40 mL was collected by esophageal tubing. The samples collected were snap frozen in liquid nitrogen and placed in a -80℃ until used for DNA extraction. DNA was extracted from 1 - 2 g of rumen contents using the MoBio PowerMag™ Soil DNA Isolation Kit (Optimized for KingFisher® Flex protocol) (MoBio Laboratories, Carlsbad, CA) according to the manufacture's protocol. Quality of the DNA was evaluated using gel electrophoresis and was stored at -20℃ until used for community analysis.

6.3.2.2. Bacterial and Archaeal 16S rRNA library preparation

The V4 region of the 16S rDNA gene specific to bacterial communities was amplified using the Terra PCR Direct Polymerase Mix Kit (Takara Bio USA) and 515F and 806R primers (Kozich et al, 2013). The V6 region of the 16S rRNA gene was amplified using extracted total rumen DNA using universal archaeal specific primers 751F and 934R (Whiteley et al, 2012). The V4 and V6 region of the 16S rRNA gene was amplified in a 15 µL and 20 µL reaction volume respectively. A PCR reaction consisted of 1X of Power SYBR® Green PCR Master Mix (Applied Biosystems by Life TechnologiesTM, Massachusetts, USA), 1.7 µM of 341F and 0.2 µM of 518R primer,

approx. 50 ng of extracted total DNA. PCR conditions were as follows: 3 min at 98°C for initial denaturation, 25 cycles of 30 s at 98°C, 30 s at 55°C for bacteria and 30s for 50°C for archaea, and 45 s at 68°C; the profile was terminated after a final 4-min hold at 68°C.

Following amplification, the product was run on a 1.8 % agarose gel using gel electrophoresis (QD LE Agarose, Green Bio Research, Baton Rouge, LA) at 120 V for 55 minutes for initial size verification and to ensure amplification. Following amplification, a 0.6X SPRI was conducted according to manufactures protocol (Agencourt® AMPure®) to remove primer dimers. SPRI products were normalized using Invitrogen Sequal Prep™ Normalization Plate kit (Frederick, Maryland) to 1 – 2 ng according to the manufacturer's protocol and pooled. Library qPCR preparation, normalization, and pooling was conducted using the Eppendorf epMotion (M5073, Germany).

6.3.2.3. Sequencing and bioinformatics analysis

Resulting amplicons were sequenced using the Illumina MiSeq platform (paired-end 2x250) using a V2 500 cycle kit with the dual-index sequencing strategy according to Kozich et al. (2013). The Illumina adapters were already removed. Subfolders separated by barcode numbers were created; barcode sequences were removed, and the sequences were demultiplexed. The fastq sequence files were processed using amplicon sequence variant error correction with DADA2 (ASVs) (Caporaso et al, 2010). Primers and low-quality regions of sequences were trimmed off (denoised), and reads were merged with chimera removal using DADA2 (Caporaso et al, 2010). Taxonomic classification was performed via GreenGenes database (ver.13_8) and SILVA database (Silva 132 99% nb

classifier). The output files (count table, tree file, taxonomy file, ASVs sequences) to R program (v. 3.6.2).

6.3.2.4. Statistical analysis

The sequences were rarefied (bacteria, 9101 and archaea, 1004) to achieve an equal sampling depth rarefaction. Microbiome richness, and taxonomic analysis were analyzed using DADA2 (Callahan et al, 2016), Phyloseq (McMurdie and Holmes 2011). The rarefied sequences were used for calculation of alpha diversity using the Observed, Chao1, Shannon, and Simpson indices (Kuczynski et al, 2011). Alpha diversity indices were statistically analyzed using Shapiro-Wilk's test. To visually observe shifts in global bacterial and archaeal community structure and its influence by diet, principle coordinate analyses was performed to estimate the distance between samples utilizing the Bray-Curtis, weighted, and unweighted UniFrac distances (Lozupone et al, 2011). Permutational multivariate analysis of variance (PERMANOVA) was performed to analyze the effect of diet on the bacterial taxonomic ASVs via adonis2 function in vegan package. Each dot within the plots represents a community from an animal. Relative abundance of phyla, classes, order, families, and genera were visualized using QIIME2 (Caporaso et al, 2010) and Phyloseq. It is generated based on the factors of phylogenetic relationships and abundance. Heatmaps were created to visualize significantly differential ASVs using R heatmap.2 function (Ploner 2014) with the ASV relative abundance as input.

6.3.3. Metagenome sequencing, gene prediction, functional profile and metabolic pathway mapping

To investigate predictive functional attributes of microbial communities, microbial genes involved in pathways linked to methanogenesis, nitrate, and sulfate metabolism were identified with metagenomic sequencing. Gene prediction, functional profile and pathway mapping were conducted using KEGG database (Kanehisa et al, 2014).

6.3.3.1. Metagenome library preparation and sequencing

The extracted DNA from rumen samples (PowerMax Soil DNA Isolation Kit (MO BIO Laboratories, Inc.) was purified using the MinElute PCR Purification Kit (Qiagen). Metagenome libraries were constructed using the Nextera XT DNA Library Prep Kit (New England Biolabs) according to the manufacturer's protocols, and sequenced using Illumina HiSeq instrument.

6.3.3.2. Data collection and pre-processing

For metagenomic sequencing, the DNA reads were sequenced using Illumina HiSeq (2x150 bp reads). The raw data were downloaded and processed as follows: (1) FASTQC (Andrews 2010) was used to check the data quality; (2) the forward and reverse reads were merged into a single file to make the pre-processing easier; (3) bbmap was used for removal of Illumina adaptors; (4) Vsearch (Rognes 2016) was used to trim reads with an estimated error rate greater than 0.02 = 2%; (5) after removing and trimming, some reads would have been lost. The reads that are still paired were merged into a single file and single reads which lost their pair into another file. The included sequences were trimmed to 100 bp to eliminate inconsistencies in sequences and reduce

the bias caused by sequencing. The number of sequences across all samples ranged from 2,024,960 to 8,189,985 sequences with an average value of 5,257,308 sequences.

6.3.3.3. Metagenome assembly Analysis and taxonomic profile

Four assemblers were used, metaSPAdes (kmer = 21, 33 and 55) (Bankevich et al, 2012), Megahit (k-mer = 25) (Li et al, 2015), Soapdenovo (k-mer = 31) (Luo et al, 2012), and Ray Meta (k-mer = 31) (Boisvert et al, 2012). The quality of assembly was checked using MetaQuast (Gurevich et al, 2013). The percentage of mapping was evaluated using BWA mapper (Li and Durbin, 2009). The composition of microbial communities from metagenomic shotgun sequencing data was analyzed using Metaphlan 1.7 (Nicola et al, 2012). The script metaphlan_hclust_heatmap.py was used to generate hierarchical clustering and heatmap visualization of multiple MetaPhlAn profiles for different diets.

6.3.3.4. Gene prediction, functional profile and metabolic pathway mapping

Open Reading Frames (ORFs) were predicted from the predicted contigs using Prodigal. Functional annotation for the predicted ORFs was conducted through Diamond BLASTP comparison for the predicted proteins against the Gene/protein (KEGG GENES) database; and identification of the KEGG orthology for the predicted genes using Ortholog (KEGG ORTHOLOGY, (KO)) database. The KOs enzymes involved in methane, sulfate and nitrate metabolism were checked using KEGG MAPPER.

6.3.3.5. Statistical analysis

To visually observe shifts in metabolic functions between different diets, Non-metric multidimensional scaling (NMDS) was performed to estimate the distance between samples utilizing the Bray-Curtis distances from KEGG orthologs using R

(3.6.2) vegan package (Lozupone et al, 2011). PERMANOVA was performed to analyze the effect of diet on the metabolic functional pathways via adonis2 function in vegan package. Pathways were plotted into a heatmap using the microbiome R package (version 1.9.19) (Lahti et al, 2017).

Differential abundance for the predicted KEGG ortholog groups (KOs) associated with enzymatic functions in different diet treatments was compared across the four diets. Differential abundance of KOs enzymes was determined independently using the EdgeR R package (Robinson et al, 2010). Differential abundance between environments was considered significant if the difference was greater than two-fold and the FDR-adjusted p-value was < 0.01. R script has been written to check the relative abundance of KOs enzymes and methane yield for each diet.

6.4. Results

6.4.1. Performance and CH4:CO2 emissions

Initially, 24 cattle were fed treatments in a 2 × 2 factorial with factors being the inclusion of 0 or 2.0% dietary nitrate (NT) and 0 or 0.54% dietary sulfate (SF). This study was a part of a larger study to explore the effects of nitrate and sulfate on cattle performance and methane emissions (Pesta 2015). Cattle performance and methane production data for the 27 samples used in this study are summarized in Table 2. Inclusion of nitrate and/or sulfate increased DMI ($P < 0.01$). Significant main effects of nitrate and sulfate tended to increase ADG ($P = 0.05$), but interaction effect was a statistical trend ($P = 0.1363$). Additionally, no significant main effects were observed due to nitrate ($P > 0.8$) or sulfate ($P > 0.3$) on G:F, but G:F improved ($P = 0.07$) in diets containing both sulfate and nitrate.

As emissions, a nitrate × sulfate interaction was observed for CH_4:CO_2 ($P = 0.01$). In diets with only sulfate or nitrate, diet decreased CH_4:CO_2 emissions ($P < 0.01$), but nitrate and sulfate in combination significantly decreased CH_4:CO_2 ($P = 0.0921$). These observations were slightly different from the whole study. Pesta 2015 found that diet had no impact on CH_4:CO_2 emissions with only sulfate or nitrate, but nitrate and sulfate in combination decreased CH_4:CO_2. However, a reduction in dry matter intake (DMI) ($P < 0.01$), average daily gain (ADG) ($P = 0.07$) and gain:feed (G:F) ($P = 0.09$) was also reported.

6.4.2. Microbiome richness and composition.

6.4.2.1. Bacteria

Results demonstrated that nitrate and sulfate supplementations did alter the rumen global bacterial community. The bacterial community was significantly affected by diet between the common basal diet and nitrate/sulfate treatment diets ($P <0.01$). All indices are illustrated in Figure 1. Nitrate and sulfate decreased the diversity of the rumen microbiota: the Observed's, Chao's, Shannon's and Simpson's alpha indices were clearly statistically significant across different diets (p-values <0.01, <0.01, = 0.08 and <0.01 respectively). The PCOA plots were generated by utilizing Bray-Curtis unifrac as a measure of β-diversity. Two distinct clusters ($P < 0.01$) were observed with significant correlation to methane yield. One cluster of common basal diet was associated with high methane emissions whereas the other cluster of nitrate and/or sulfate treatments was associated with low methane production. Two nitrate samples with high methane production clustered away from the other nitrates. PERMANOVA results showed statistically significant distances between common, nitrate, sulfate and COMBO diet

(P=0.001). Distances between nitrate and sulfate diets (P=0.01); nitrate and COMBO (P=0.002) were also statistically significant. PERMANOVA did not show significant differences between sulfate and COMBO diets (P=0.0.18). Figure 2 shows a clear clustering of the bacterial community based on diet type, suggesting that the treatment diets did change the ruminal bacterial community from the basal common diet.

The abundance profile for taxonomic OTUs were significantly different between common diet and nitrates/sulfate supplementations (Figure 3). Highly abundant taxonomic ASVs in the common basal diet were associated with low abundant ASVs in case of other diet treatments and vice versa. In addition, a set of ASVs were less abundant in case of COMBO diet. On the phylum level, Bacteroidetes and Firmicutes were the highly abundant phyla in common diet, while Proteobacteria were highly abundant in COMBO diet. On the class level, Bacteroidia and Clostridia classes were highly common in the basal common diet, and Negativicutes were significantly abundant in COMBO. Bacteroidales and Clostridiales were highly abundant in common diet, but Selenomonodales was highly abundant in COMBO. Provetellaceae, Lachnospiraceae, and Ruminococaceae were abundant in common diet, and Veillonellaceae was highly abundant in COMBO. On the generic level, significant increase in bacterial genera with H2 utilization capability e.g. propionate, lactate forming bacteria e.g. *Prevotella*, *Megasphaera*, *Selenomonas*, *Lactobacillus* and *Bifidobacterium*; nitrate and sulfate reducing bacteria e.g. *Selenomonas*, *Desulfovibrio* was observed in COMBO diet (Figure 5).

6.4.2.2. Archaea

The Observed, Chao, Simpson indices of alpha diversity of ruminal archaeal communities were not statistically significant between different diets (Figure 6), however, Shannon index tend to be statistically significant (P = 0.02). In addition, PCOA analysis using weighted unifrac as a measure of β-diversity has been shown in Figure 7. Permanova pairwise results illustrated statistically significant distances between common, nitrate, sulfate and COMBO (P=0.001). Distances between nitrate and sulfate diets (P=0.02); nitrate and COMBO (P=0.01) tend to be statically significant. PERMANOVA shows that distances between sulfate and COMBO diets were a statistical trend (P=0.07). Weighted unifrac distances as a measure of beta diversity is illustrated in Figure 7. Methanobacteria (class), Methanobacteriales (order), Methanobacteriaceae (family), *Methanobrevibacter* (genus) were highly abundant in nitrate and COMBO diet, followed by sulfate diet and finally common diet as shown in Figure 8. Other less abundant genera e.g. *Methanimicrococcus, Methanosarcina* and *Methanosphaera* were also recorded.

6.4.3. Taxonomic profile, gene prediction, functional profile and metabolic pathway mapping

Four different assemblers (metaSPAdes, Magahit, Soapdenovo and Ray Meta) was used for short metagenome reads assembly. The assembly metrics for each assembler are shown in Table 3. The number of predicted KOs enzymes from metaSPAdes, Megahit, Soapdenovo and Meta Ray are 4443, 3394, 2425 and 1828, respectively. In

general, metaSPAdes and megahit were the best assemblers for prediction of functional profile in different diets.

Taxonomic profile, KEGG orthology and metabolic pathways have been compared between different diet treatments using the four assemblers. Soapdenovo was the best in taxonomic identification of bacterial communities from short metagenome reads. Figure 9 shows the taxonomic abundance profile of the significant bacterial communities between different diets. Bacterial genera with H+ utilization capability was also significantly abundant in COMBO diet e.g. *Butyrivibrio, Prevotella, Bifidobacterium, and Propionibacterium*.

Pathway Mapping and metabolic Enzymes were cataloged and mapped to pathways according to the KEGG database. Beta diversity was evaluated using Non-metric multidimensional scaling (NMDS) to observe shifts in metabolic functions across different diets. PERMANOVA results were statistically significant between all diets (P=0.006). Distinct clusters were found between the common basal diet and nitrate/sulfate treatments. However, samples in nitrate, sulfate, COMBO diets were scattered into two separate clusters (Figure 10). The KEGG orthologs groups involved in methane, nitrate and sulfate metabolism have been shown in Figures 11a, 11b and 11c. Total abundance of KOs enzymes involved in methane, nitrate and metabolism were compared between different diets (Figure 12). KOs enzymes were highly abundant in nitrate diet, followed by common basal diet, and sulfate diet. It is highly significant that total abundance of KOs enzymes decreased in COMBO diet. In methane metabolism, KOs enzymes are involved in the following steps: ribulose-P, xylulose-P, serine-P, serine biosynthesis, F420 biosynthesis, CO2 => acetyl-CoA, coenzyme M biosynthesis and

finally methanogenesis. Methanogenesis step is controlled by enzymes which are involved in conversion of CO_2, methanol, acetate, methylamines into methane.

The KEGG ortholog groups related to enzymes involved in methanogenesis and their abundance profile involved in methanogenesis step have been illustrated in Table 4 and Figure 13. Differential gene abundance in metabolic pathways has shown decrease of enzymes linked to methanogenesis in COMBO diet. All enzymes which are involved in conversion of methanol, acetate, CO_2, methylamines into methane were significantly decreased in COMBO diet. However, significant increase in acetate kinase enzyme [EC 2.7.2.1] has been observed in COMBO diet, followed by sulfate diet as they play a major role in the propanoate production.

The relative abundance of KOs enzymes and methane yield for different diets was shown in Figure 14. The methane yield was higher in common diet, then nitrate, sulfate and COMBO diet respectively. KOs enzymes involved in methane metabolism showed different patterns between different diets. Increase of methane KOs was observed in common diet with high methane yield. Significant reductions in the relative abundance of methane KOs enzymes and methane yield were also recorded in nitrate, sulfate and COMBO diets. However, significant increase in some methane KOs enzymes were observed in some samples of sulfate and COMBO diet. This is because of high abundance of acetate kinase enzymes involved in conversion of acetate to methane. This was interpreted in metaSPAdes and megahit assemblers. KOs enzymes involved in nitrate and sulfate metabolism were consistent among different diets, suggesting that sulphate plays a role of electron donor in the reduction of nitrite to ammonia and nitrate plays a role of electron donor in the reduction of sulfate.

6.5. Discussion

This study suggested that methane output and cattle performance is affected by diet type. Only combination between nitrate and sulfate helped to reduce methane emissions. However, a reduction in DMI, ADG, and G:F was also reported. In addition, VFAs tend to increase in case of COMBO diet as reported in the main study of this project (Pesta, 2015). Cattle performance observed in this experiment is similar to the results observed by Newbold et al, 2014, as they found that increasing nitrate decreased DMI without any impact on ADG. In addition, a reduction in DMI and sulfur toxicity was reported with increasing sulfate in diets (Sarturi et al, 2013). However, other studies reported no changes in DMI and ADG with nitrate and sulfate supplementations in sheep (Van Zijderfeld et al, 2010). Methane emissions recorded in our study were different from other studies. Some studies reported a dramatic decrease in methane levels with nitrate and sulfate supplementations in sheep and dairy cows (Van Zijderfeld et al, 2010). Other studies reported no impact of dietary nitrates on methane production (Troy et al, 2015). The VFAs production is also a point of date in literature. Some studies reported no change in acetate and propionate concentrations with sulfate and nitrate supplementations (Van Zijderfeld et al, 2010), while others reported increase in acetate:propionate ratio with dietary nitrate (Troy et al, 2015).

Methanogenesis includes two main pathways which are controlled by archaea: the hydrogenotrophic pathway in which archaea converts H_2 and CO_2 produced by the bacteria, protozoa, and fungi to methane; and conversion of methyl groups (which are derived from methylamines and methanol) into methane. The hydrogenotrophic pathway is controlled by the most abundant hydrogenotrophic archaea in rumen

Methanobrevibacter and other significant hydrogenotrophic genera e.g. *Methanimicrococcus Methanosphaera*, and *Methanobacterium*. Less abundant methylotrophs e.g. Methanosarcinales, Methanosphaera, Methanomassiliicoccaceae can utilize methylamines and methanol, and produce methane (Morgavi et al, 2012).

High-throughput sequencing of the 16S rRNA sequence to monitor microbial composition showed that sulfate and nitrate in combination significantly increase bacterial genera with H_2 utilization capability in fatty acids formation e.g. propionate, lactate forming bacteria e.g. *Prevotella, Bacteroides, Megasphaera, Selenomonas, Bifidobacterium, Lactobacillus*; nitrate and sulfate reducing bacteria e.g. *Selenomonas, Desulfovibrio*. In addition, some bacterial phyla with less H_2 production capability were increased in COMBO diet e.g. Proteobacteria. These events are correlated with decreasing methane emissions in case of nitrate and sulfate combination. In conclusion, COMBO diet reduced the production of methane by activating VFAs producing bacteria, nitrate and sulfate reducing bacteria. These results agree with a study which showed a correlation between reduction in the relative abundance of three ASVs and lower methane emissions. Two ASVs were characterized by less common H_2-producing bacteria. Lower abundance of Proteobacteria and some Bacteroidetes were associated with high methane emissions (Tapio et al, 2017).

Bacteria Members of the rumen microbiome consists of cellulolytic, amylolytic, and proteolytic organisms in the feed particles, rumen fluid, and the rumen epithelium. Bacteria are responsible for fermenting the feed ending up with volatile fatty acids (VFAs). Some bacterial organisms e,g, *Ruminococcus flavefaciens, Ruminococcus albus, and Fibrobacter succinogenes* (Flint et al, 2008) secrete enzymes (endoglucanases,

exoglucanases, and β-glucosidases, and hemicellulases) to digest cellulose (Cai et al, 2010). Other bacteria e.g. *Butyrivibrio fibrosolvens* and *Prevotella ruminocola* digest hemicellulose, xylan and pectin and utilize the byproducts as a source for energy (Cai et al, 2010).

This study reported that bacterial populations are highly correlated with methane emissioms more than archaeal communities. There is a debate in literature about the correlation between archaea and methane emissions. Some studies found no or weak correspondence between methanogens and methane emissions in dairy cows and sheep using metagenomics and qPCR techniques. (Morgavi et al, 2012; Zhou et al, 2011; Danielsson et al, 2012; Danielsson 2016; Kittelmann et al, 2014 and Shi et al, 2014). Other studies found positive correlations between methane emissions and Methanobrevibacter SGMT clade (Zhou et al, 2011; Danielsson et al, 2012; Shi et al, 2014 and Danielsson 2016). Methanobrevibacter SGMT and SGMT clade have methyl coenzyme M reductase isozymes (McrI and McrII), which enables the archaea to utilize H_2 at higher concentrations, against the RO clade that has only McrI. Another study found that animals dominated the *Methanobrevibacter gottschalkii* clade tending to have higher methane emissions (Tapio et al, 2017).

Evaluation of KEGG Ortholog groups of enzymes involved in methanogenesis step has shown a reduction in gene abundance of those assigned to conversion of methanol, acetate, CO_2, and methylamines to methane in COMBO diet. This was clearly correlated with lower methane emissions in COMBO diet. The results agreed with other studies which reported specific enrichment of metabolic pathways which are correlated with higher methane yield in milking cows (Shabat et al, 2016); and increase in

methanogenesis pathway genes was substantially increased with high methane yields in sheep (She et al, 2014).

Therefore, the dynamics between the archaeal and bacterial community composition are correlated with H_2 utilization and H_2 production by bacteria. This mechanistic and ecological understanding of the rumen microbiome might help to increase in food resources and environmentally friendly livestock agriculture.

Table 1. Composition of Finishing Diets 0 Or 2.0% Nitrate and 0 or 0.54% Sulfate

Ingredient	CT	NT	SF	SF+NT
Dry-rolled corn	35.75	35.75	35.75	35.75
High-moisture corn	35.75	35.75	35.75	35.75
MDGS	10	10	10	10
Alfalfa hay	7.5	7.5	7.5	7.5
Molasses	5	5	5	5
Ca(NO3)2	_	2.65	_	2.65
CaSO4	_	_	0.77	0.77
Urea	0.75	0.75	__	_

MDGS = modified distillers grains plus solubles.

Table 2. Effect of Dietary Nitrates and Sulfates on Methane Production and Cattle Performance

							Main effects		Interaction
	CT	NT	SF	SF+NT	SEM	F-test	Sulfate	Nitrate	Sulfate*Nitrate
CH4:CO2 ratio	0.6424	0.05269	0.04933	0.03985	0.007569	0.0141	0.0126	0.0516	0.8417
CO2 level	1809.6	1212.93	1389.35	1072.9	79.904	< 0.01	< 0.01	< 0.01	0.0164
CH4 level	116.26	63.8	67.7667	42.7667	11.5358	< 0.01	< 0.01	< 0.01	0.0921
DMI	16.2589	22.4283	23.9967	21.55	1.1507	< 0.01	< 0.01	0.0255	< 0.01
ADG	2.3622	3.11	3.1083	3.21	0.3091	0.0151	0.055	0.0541	0.1363
G:F	0.1439	0.1378	0.1288	0.1493	0.01051	0.2575	0.805	0.3208	0.0749

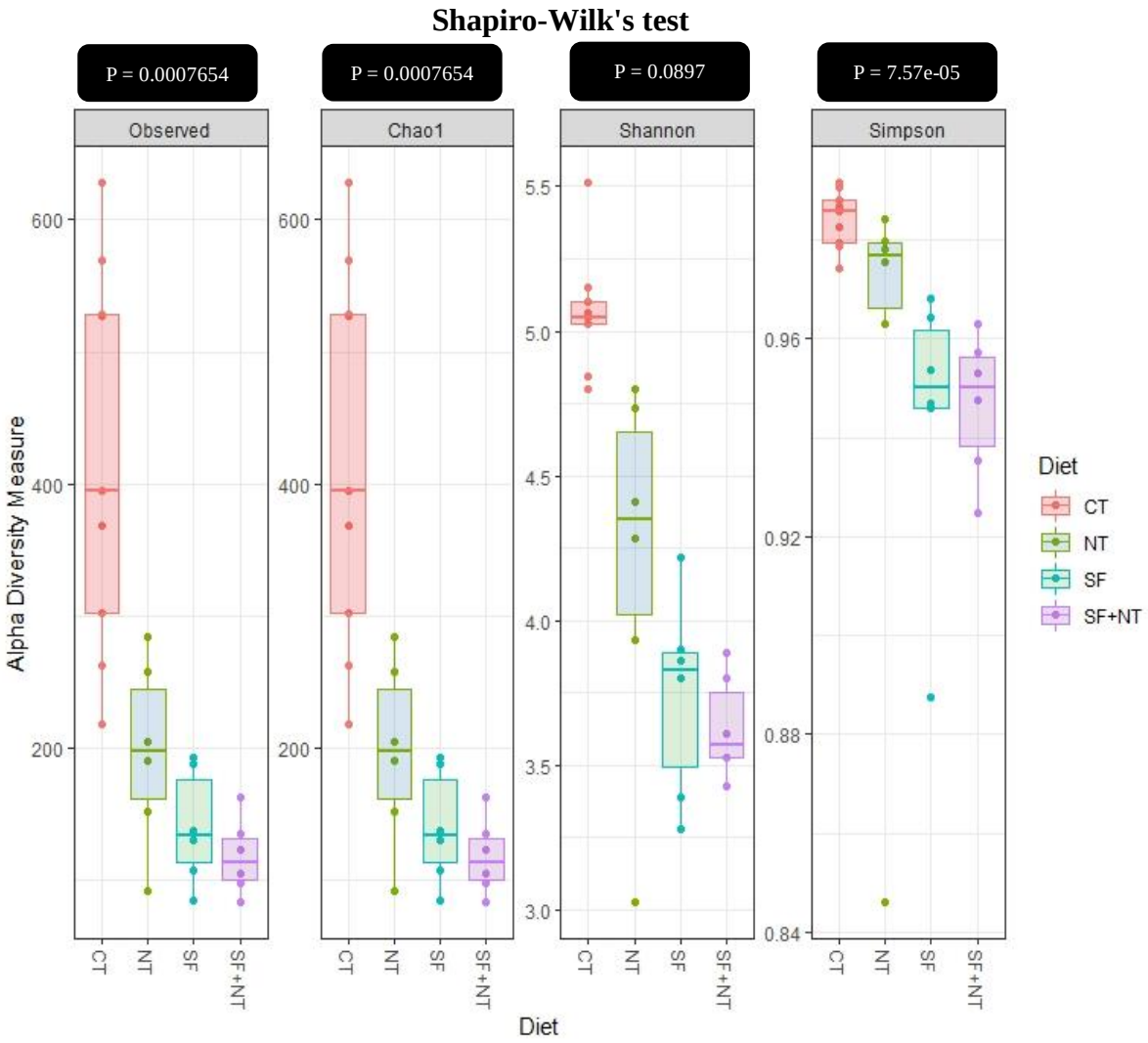


Figure 1. Bacteria Alpha Diversity Between Different Diets. The bacterial community was significantly affected by diet (P <0.01). The Observed's (P <0.01), Chao's (P <0.01), Shannon's (P = 0.08) and Simpson's (P <0.01) alpha indices were statistically significant across different diets.

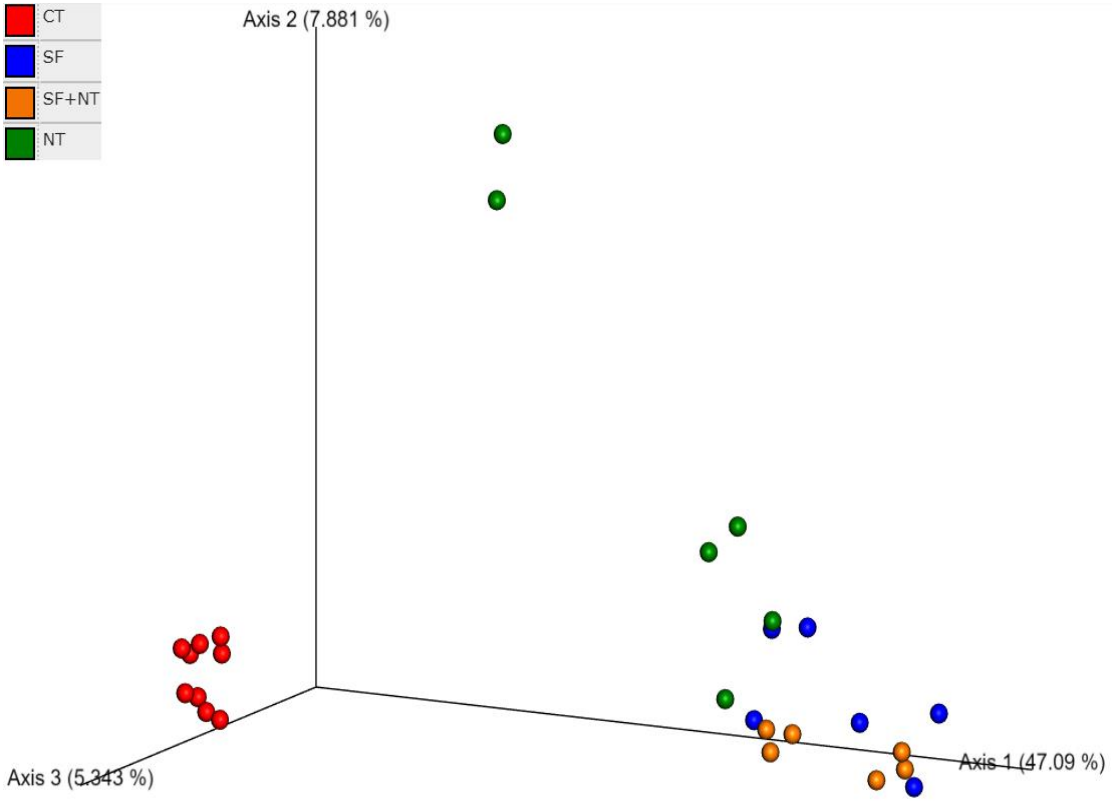


Figure 2. Bacteria PCOA of Unifrac Distances (Bray-Curtis). The PCOA plots were generated by utilizing Bray-Curtis unifrac as a measure of β-diversity. Two distinct clusters ($P < 0.01$) were observed: one cluster of common basal diet (Higher methane yield), and the other cluster of nitrate and/or sulfate treatments (Lower methane yield). Two nitrate samples with higher methane production clustered away from the other nitrates. PERMANOVA results showed statistically significant distances between common, nitrate, sulfate and COMBO diet (P=0.001). Distances between nitrate and sulfate diets (P=0.01); nitrate and COMBO (P=0.002) were significant. PERMANOVA did not show statistically significant differences between sulfate and COMBO diets (P=0.0.18).

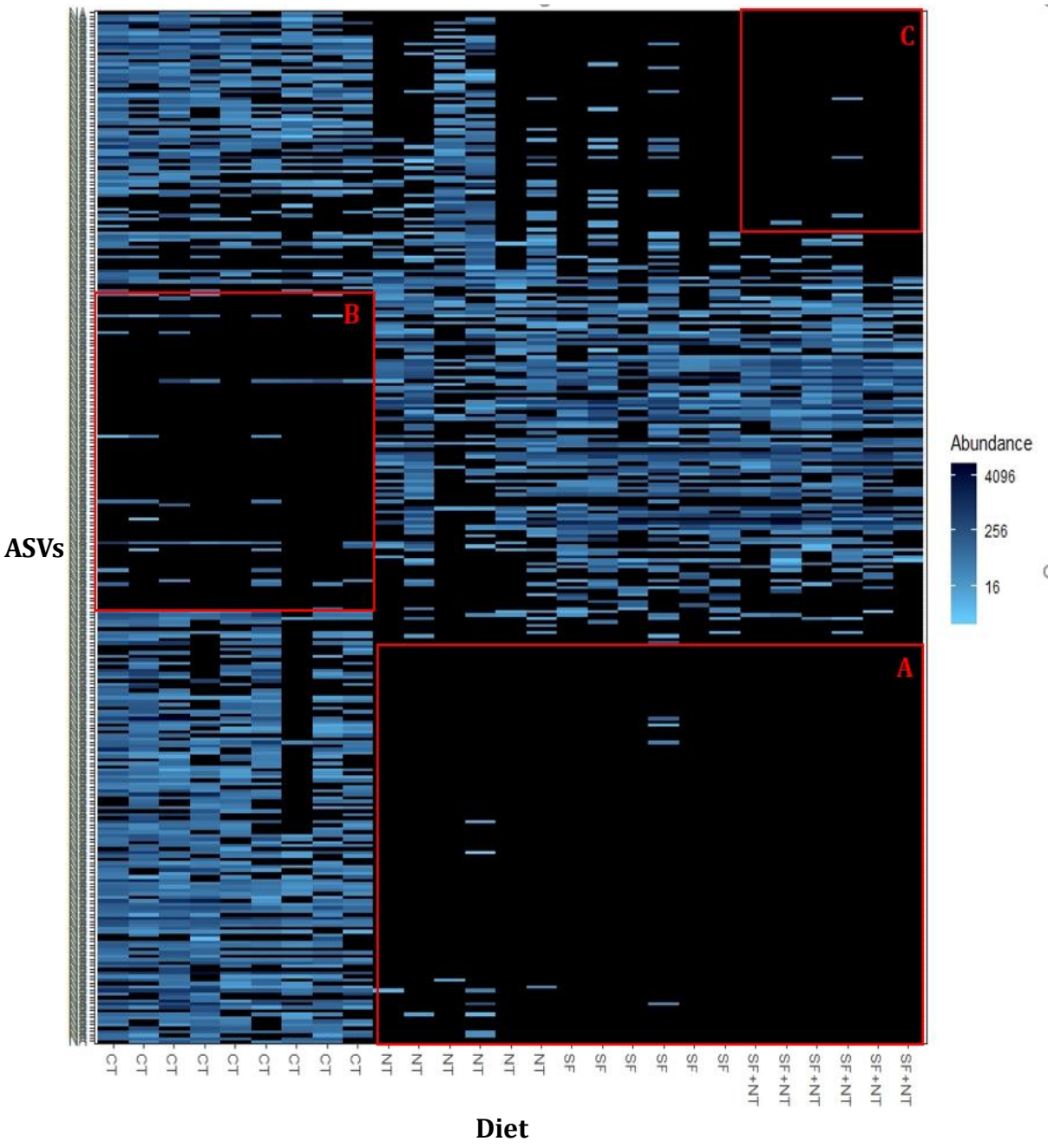


Figure 3. Heatmap of Bacterial Distribution Among the Samples of Different Diets. (A) Low abundant taxonomic ASVs in nitrate/sulfate treatment diets; (B) Low abundant taxonomic ASVs in the common basal diet; (C) Low abundant taxonomic ASVs in COMBO diet.

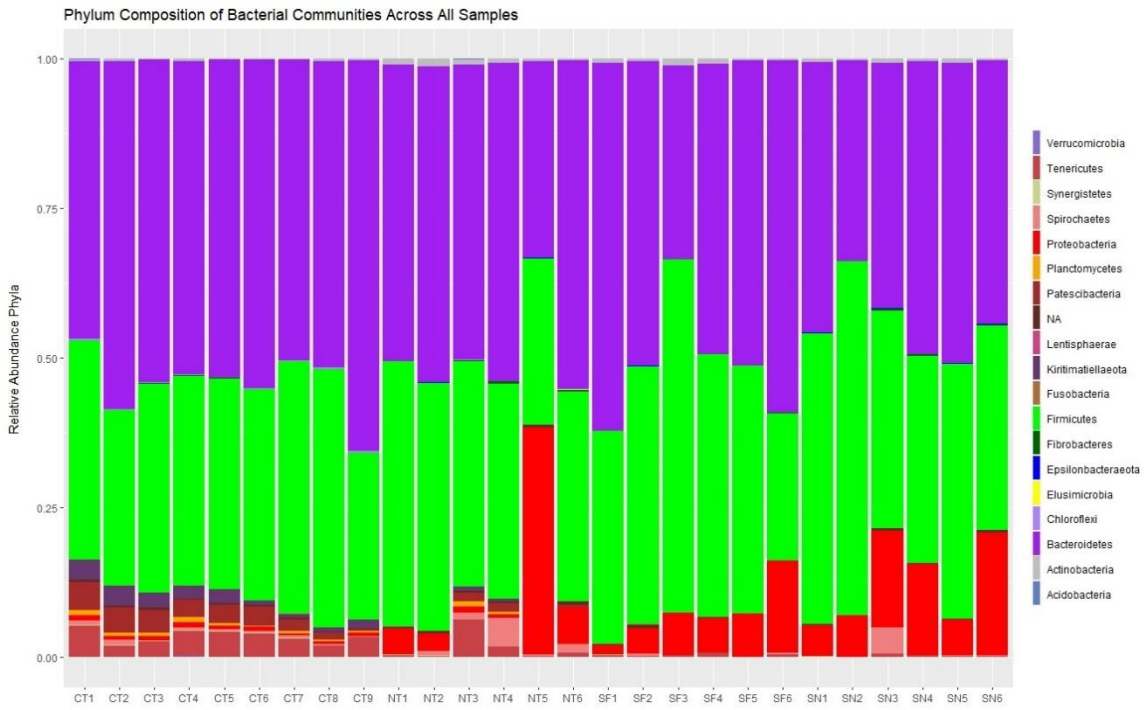


Figure 4. Bacterial Phylum Abundance Between Different Diets. Bacteroidetes and Firmicutes were highly abundant in common diet, while Proteobacteria was significantly abundant in COMBO diet

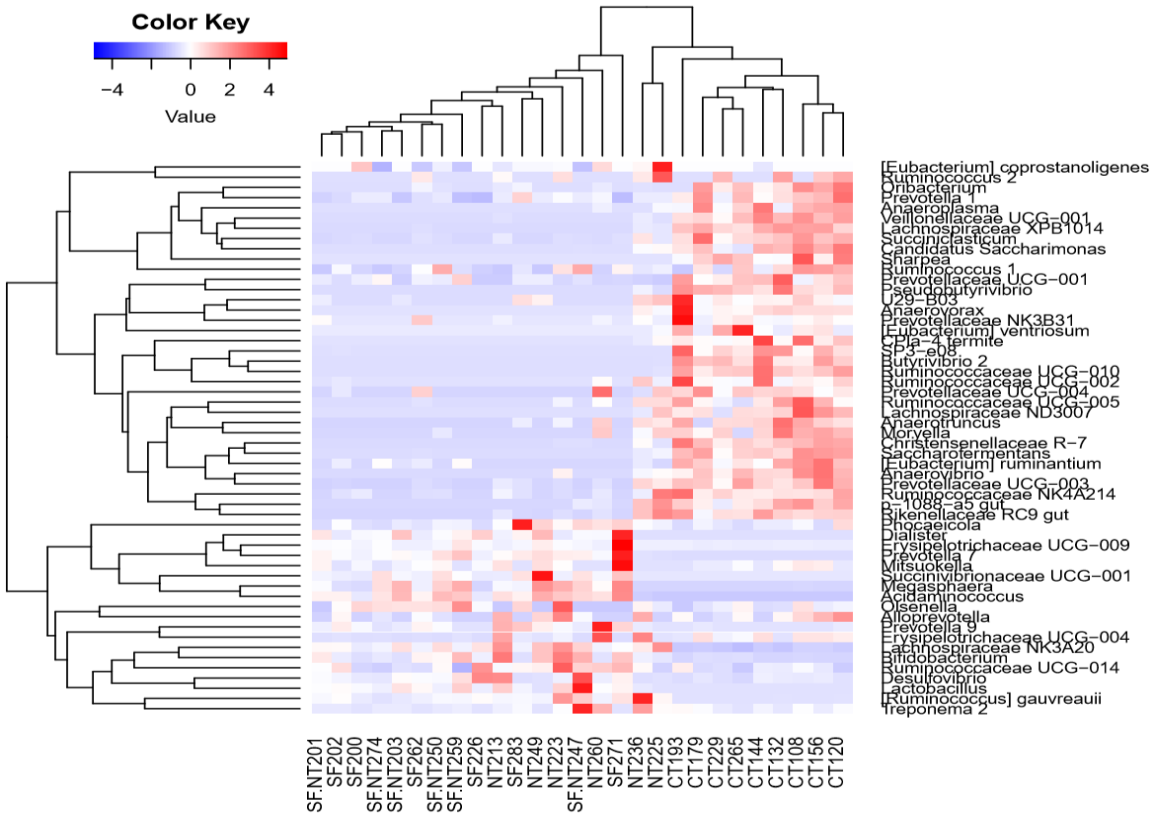


Figure 5. Heatmap of Bacterial Genera Distribution Between Different Diets. The highly abundant genera in both common and sulfate/nitrate supplementations were used to draw a heatmap to check genus abundance profile among different samples of different diets. Significant increase in some bacterial genera with H2 utilization capability e.g. *Prevotella, Megasphaera, Lactobacillus, Bifidobacterium, Desulfovibrio* was observed in COMBO diet.

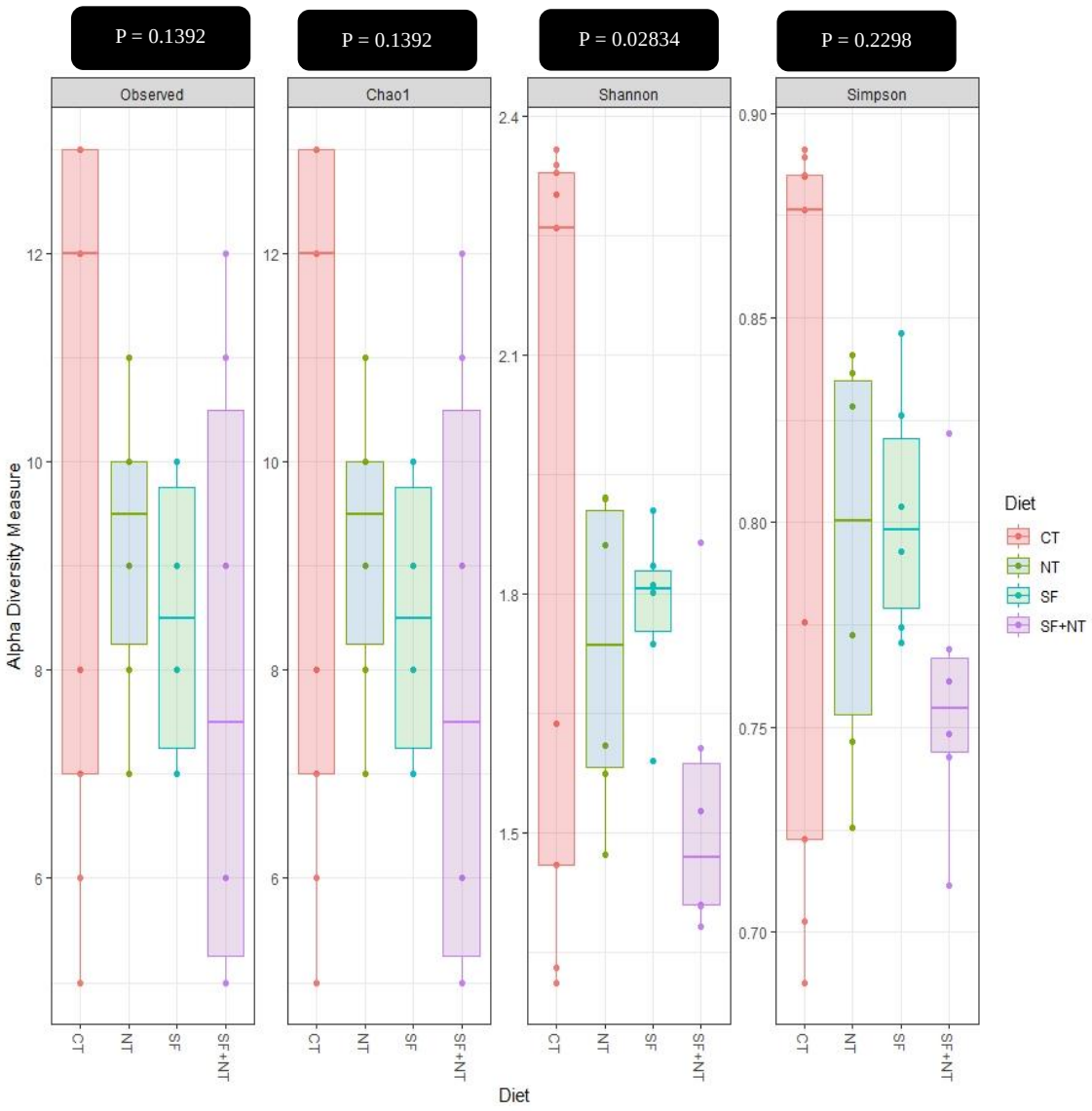


Figure 6. Archaea Alpha Diversity Between Different Diets. The Observed, Chao, Simpson indices of alpha diversity were not statistically significant between different diets. Shannon index tend to be statistically significant (P = 0.02).

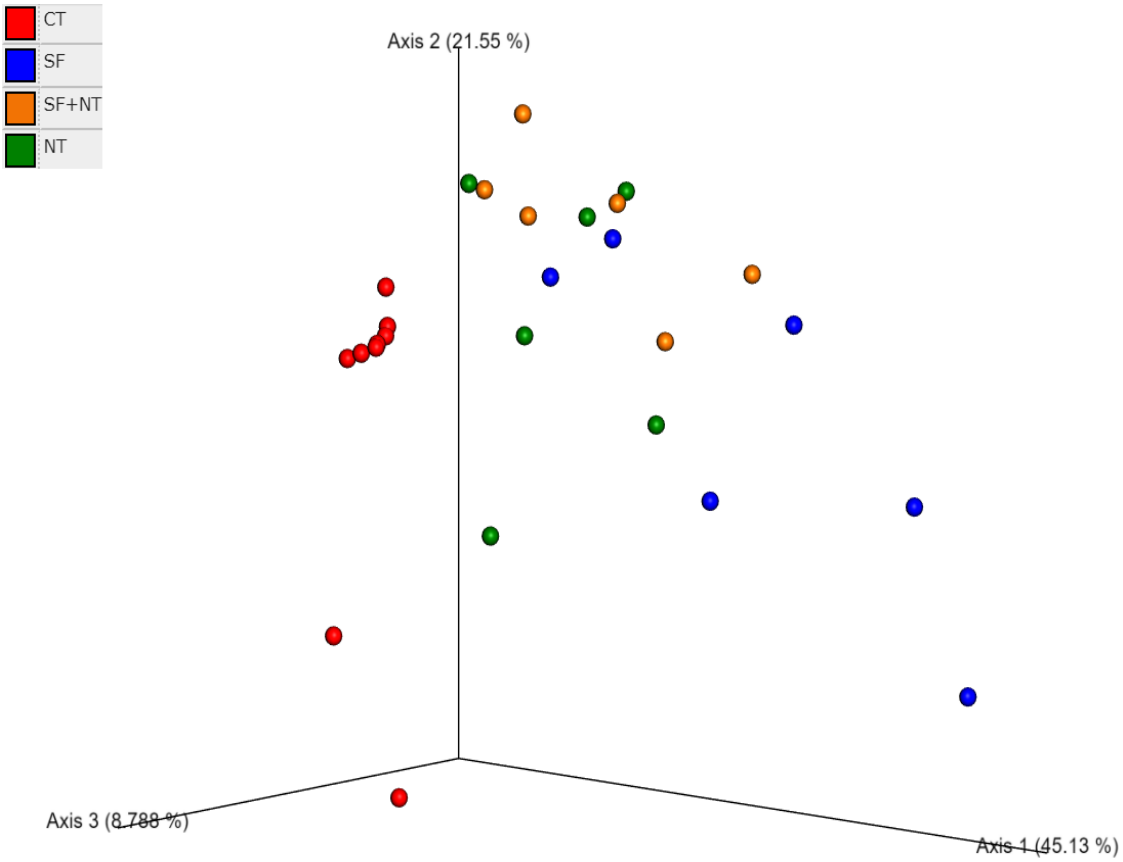


Figure 7. Archaea PCOA of Unifrac Distances (Weighted Unifrac) (P = 0.001). PERMANOVA results showed significant distances between treatment diets (P=0.001). Distances between nitrate and sulfate diets (P=0.02); nitrate and COMBO (P=0.01) tend to be statistically significant. Distances between sulfate and COMBO diets were a statistical trend (P=0.07).

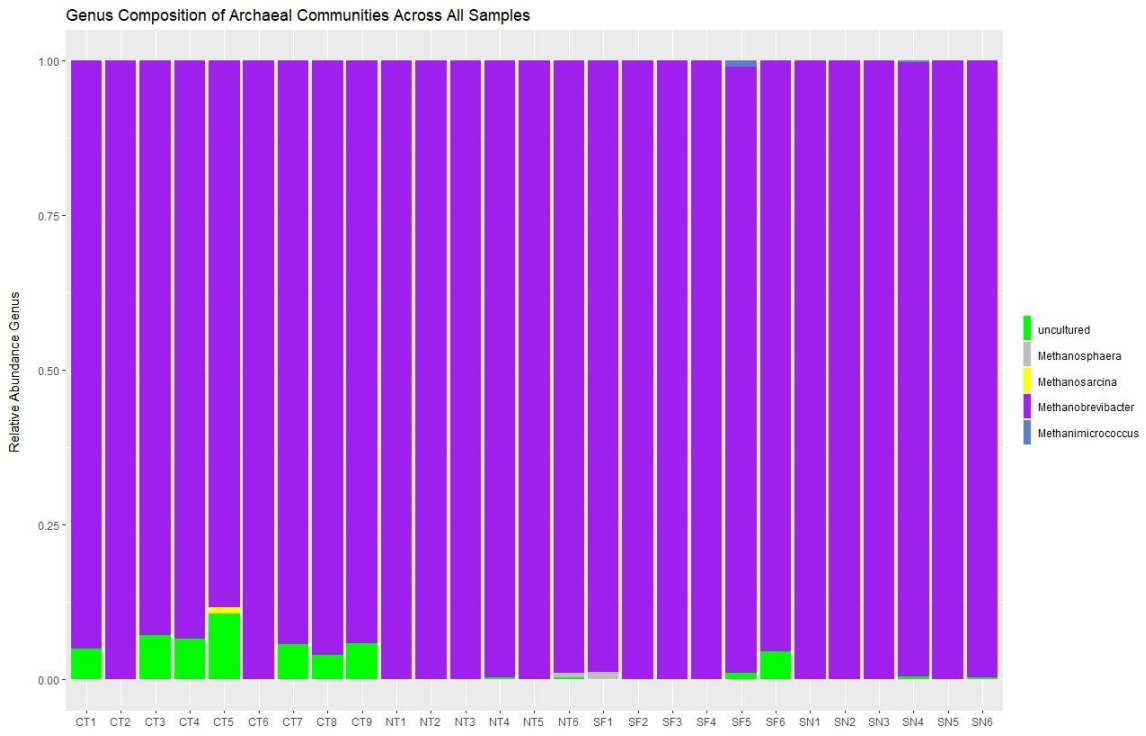


Figure 8. Archaea Taxonomic Abundance Between Different Diets. *Methanobrevibacter* was highly abundant in nitrate and COMBO diet, followed by sulfate diet and common diet. Less abundant genera e.g. *Methanimicrococcus, Methanosarcina* and *Methanosphaera* were represented.

Table 3. Metaquast Assembly Quality for Different Assemblers

Assembly	**metaSPAdes**	**MEGAHIT**	**SOAPdenovo**	**Meta Ray**
# contigs (>= 0 bp)	9342371	1344870	6906938	4475861
# contigs (>= 1000 bp)	162789	154889	96596	51336
# contigs (>= 5000 bp)	8193	10776	6929	4978
# contigs (>= 10000 bp)	1756	3017	1765	1538
# contigs (>= 25000 bp)	154	449	215	215
# contigs (>= 50000 bp)	24	84	24	20
# contigs	508583	561596	249702	126263
Largest contig	97630	210373	101386	97808
GC (%)	53.01	52.76	52.72	52.18
N50	1238	1193	1657	2021
N75	739	706	861	943
L50	113422	114080	46520	19961
L75	267654	292353	118927	55460
# N's per 100 kbp	54.23	0	5880.48	730.33
Total Predicted KOs	**4443**	**3394**	**2425**	**1828**

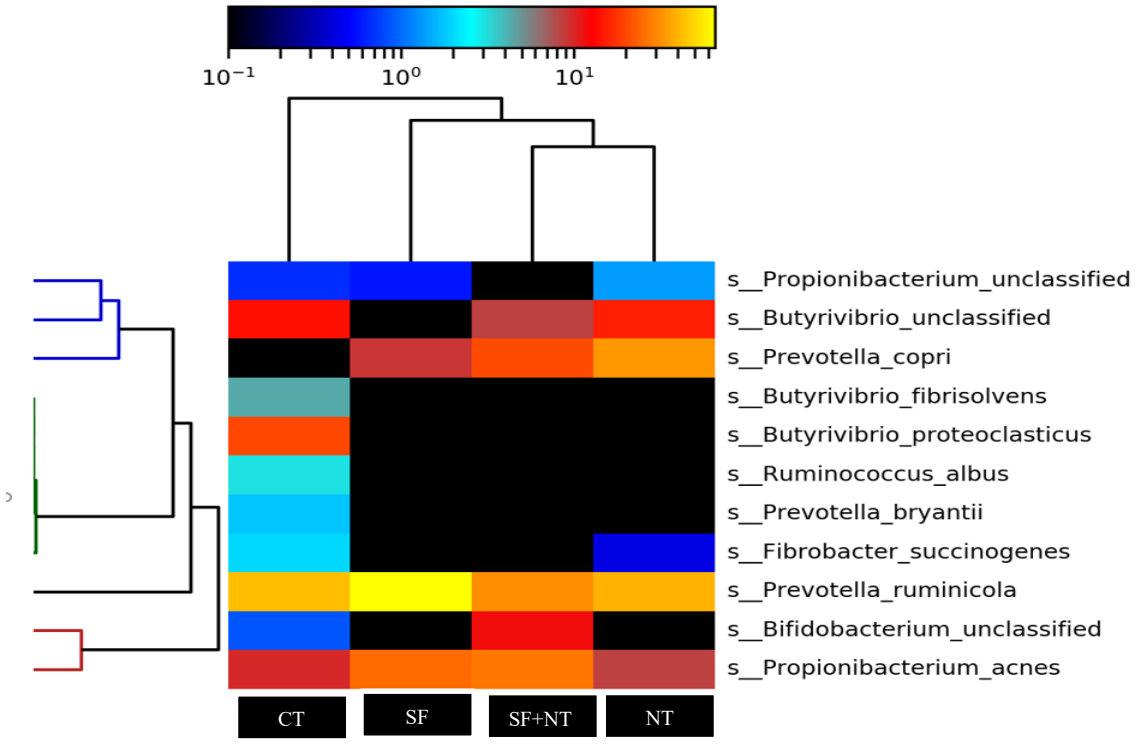


Figure 9. Metagenomic Taxonomic Abundance Profile between Different Diets. Bacterial genera with H+ utilization capability was highly abundant in COMBO diet e.g. *Butyrivibrio, Prevotella, Bifidobacterium, and Propionibacterium*.

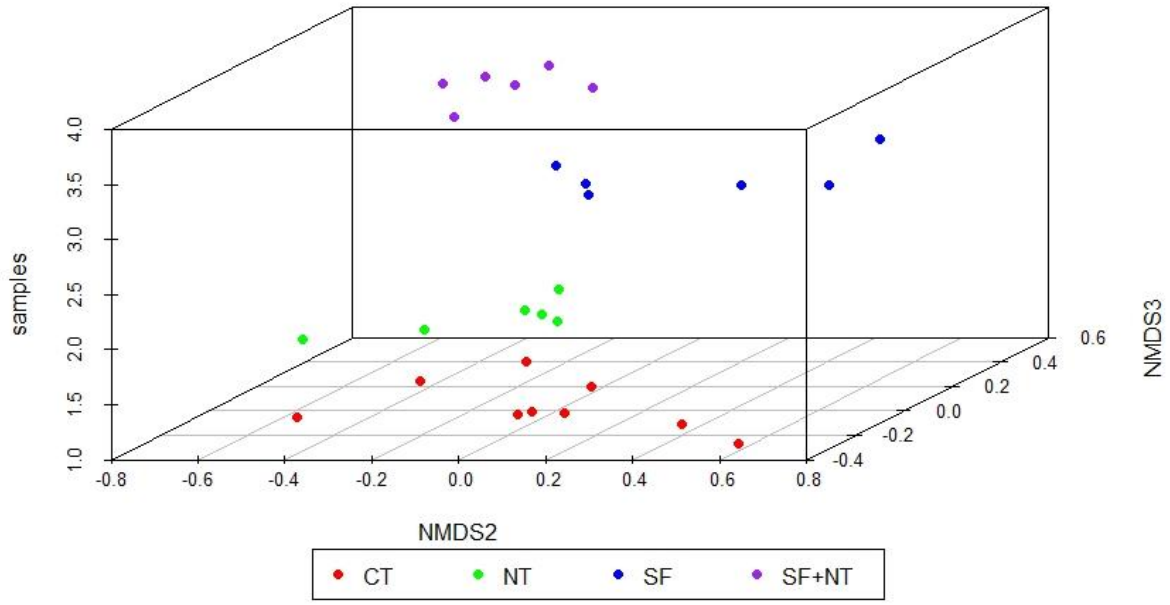


Figure 10. NMDS (Bray-Curtis) of Predicted KEGG Orthology between Different Diets. Non-metric multidimensional scaling (NMDS) was used as a measure of β-diversity to observe shifts in metabolic functions across different diets. PERMANOVA results were statistically significant between all diets (P=0.006). Distinct clusters were found between the common basal diet and nitrate/sulfate treatments. However, samples in nitrate, sulfate, COMBO diets were scattered into two separate clusters.

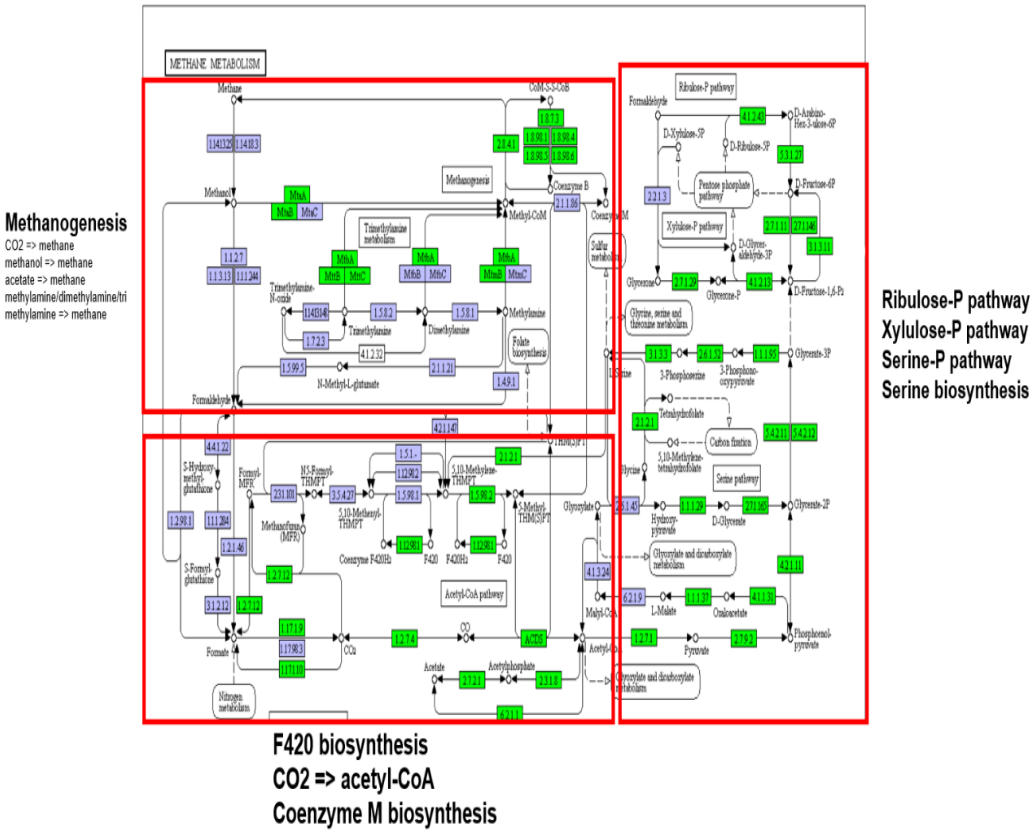


Figure 11a. KEGG Orthologs (KOs) Involved in Methane Metabolism. Green boxes represent the identified KOs enzymes and their role in methane metabolism.

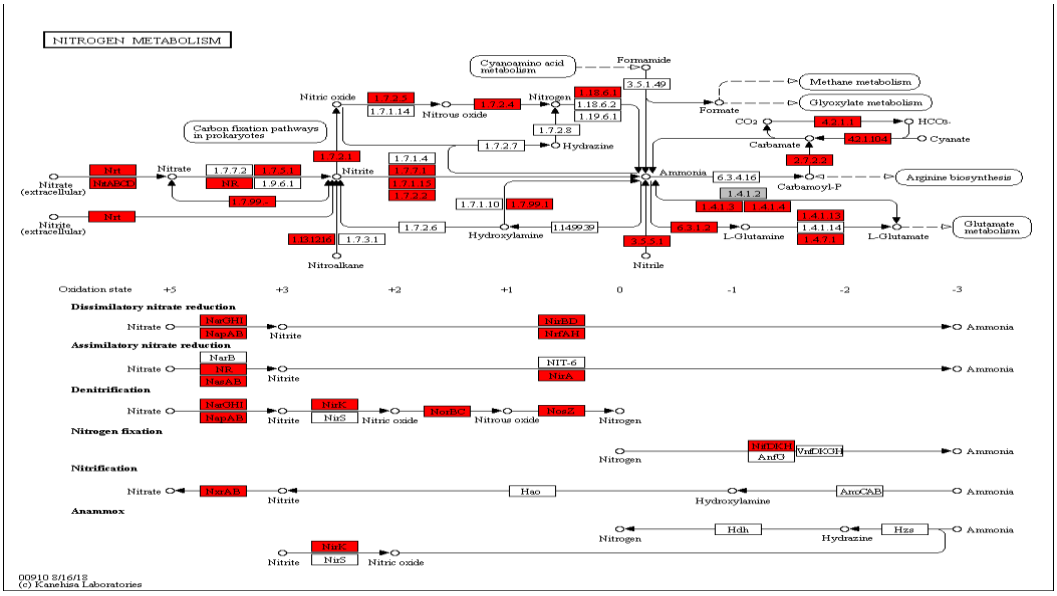


Figure 11b. KEGG Orthologs (KOs) Involved in Nitrate Metabolism. Red boxes represent the identified KOs enzymes in nitrate metabolism.

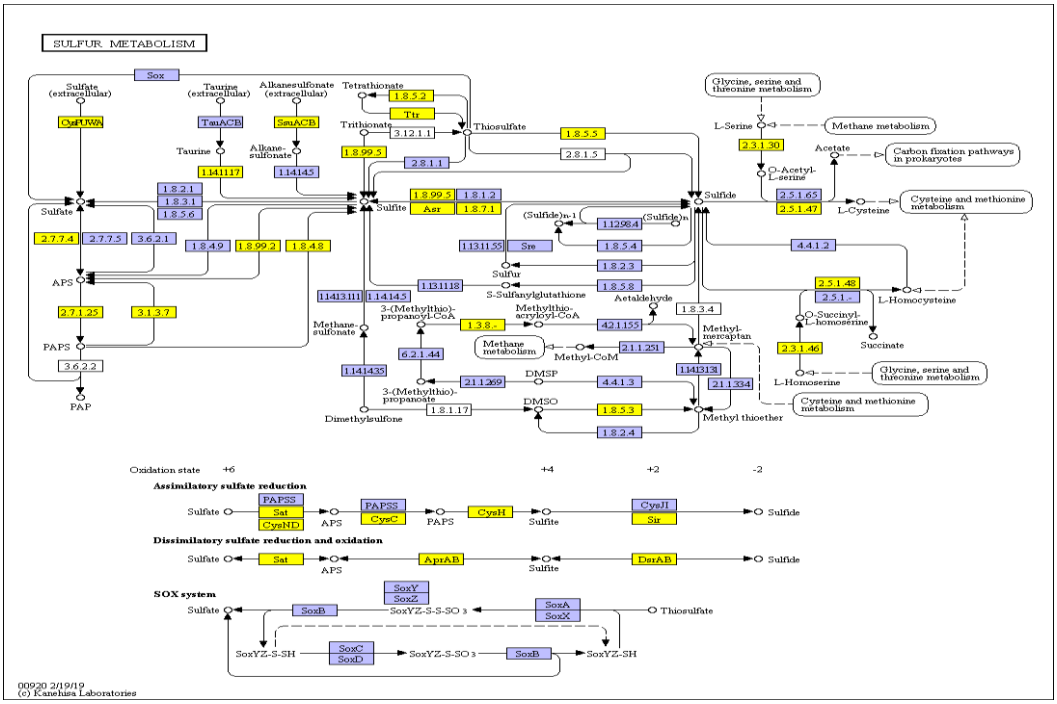


Figure 11c. KEGG Orthologs (KOs) Involved in Sulfate Metabolism. Yellow boxes represent the identified KOs enzymes in sulfate metabolism.

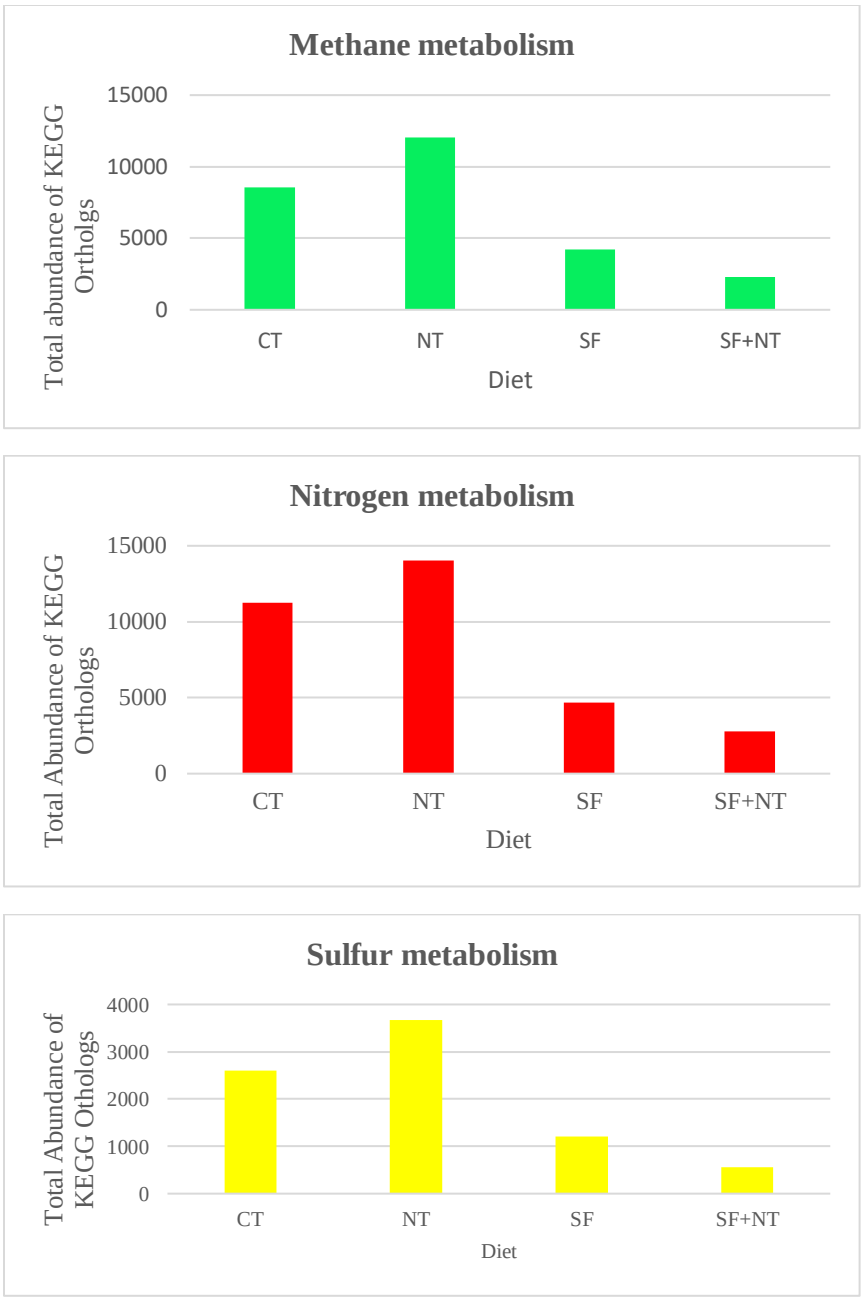


Figure 12. Total Abundance of KEGG Ortholgs involved in methane, nitrate and sulfate Metabolism. The total abundance of KOs enzymes in methane, nitrate, and sulfate metabolic pathways was higher in nitrate, followed by common basal and sulfate diets. KOs enzymes were less abundant in COMBO diet.

Table 4. Kos Enzymes Involved in Methane Metabolism

KEGG Orthology (KO)	Enzyme Family	Methanogenesis Step
K14080	mtaA; [methyl-Co(III) methanol-specific corrinoid protein):coenzyme M methyltransferase [EC:2.1.1.24]	methanol => methane
K04480	mtaB; methanol---5-hydroxybenzimidazolylcobamide Co-methyltransferase [EC:2.1.1.90]	
K14082	mtbA; [methyl-Co(III) methylamine-specific corrinoid protein):coenzyme M methyltransferase [EC:2.1.1]	methylamine/dimethylamine/trimethylamine => methane
K14083	mttB; trimethylamine---corrinoid protein Co-methyltransferase [EC:2.1.1.250]	
K14084	mttC; trimethylamine corrinoid protein	
K16176	mtmB; methylamine---corrinoid protein Co-methyltransferase [EC:2.1.1.248]	
K00925	ackA; acetate kinase [EC:2.7.2.1)]	acetate => methane
K01895	ACSS1_2, acs; acetyl-CoA synthetase [EC:6.2.1.1]	
K00625	E2.3.1.8, pta; phosphate acetyltransferase [EC:2.3.1.8]	
K00197	cdhE, acsC; acetyl-CoA decarbonylase/synthase, CODH/ACS complex subunit gamma [EC:2.1.1.245]	
K00194	cdhD, acsD; acetyl-CoA decarbonylase/synthase, CODH/ACS complex subunit delta [EC:2.1.1.245]	
K13788	pta; phosphate acetyltransferase [EC:2.3.1.8]	
K00320	mer; 5,10-methylenetetrahydromethanopterin reductase [EC:1.5.98.2]	CO2 => methane
K11261	fwdE, fmdE; formylmethanofuran dehydrogenase subunit E [EC:1.2.7.12]	
K14126	mvhA, vhuA, vhcA; F420-non-reducing hydrogenase large subunit [EC:1.12.99.-1.8.98.5]	methanol => methane; methylamine/dimethylamine/trimethylamine => methane; acetate => methane; CO2 => methane
K03388	hdrA2; heterodisulfide reductase subunit A2 [EC:1.8.7.3 1.8.98.4 1.8.98.5 1.8.98.6]	
K03389	hdrB2; heterodisulfide reductase subunit B2 [EC:1.8.7.3 1.8.98.4 1.8.98.5 1.8.98.6]	
K03390	hdrC2; heterodisulfide reductase subunit C2 [EC:1.8.7.3 1.8.98.4 1.8.98.5 1.8.98.6]	
K00399	mcrA; methyl-coenzyme M reductase alpha subunit [EC:2.8.4.1]	
K08264	hdrD; heterodisulfide reductase subunit D [EC:1.8.98.1]	
K14127	mvhD, vhuD, vhcD; F420-non-reducing hydrogenase iron-sulfur subunit [EC:1.12.99.- 1.8.98.5 1.8.98.6]	

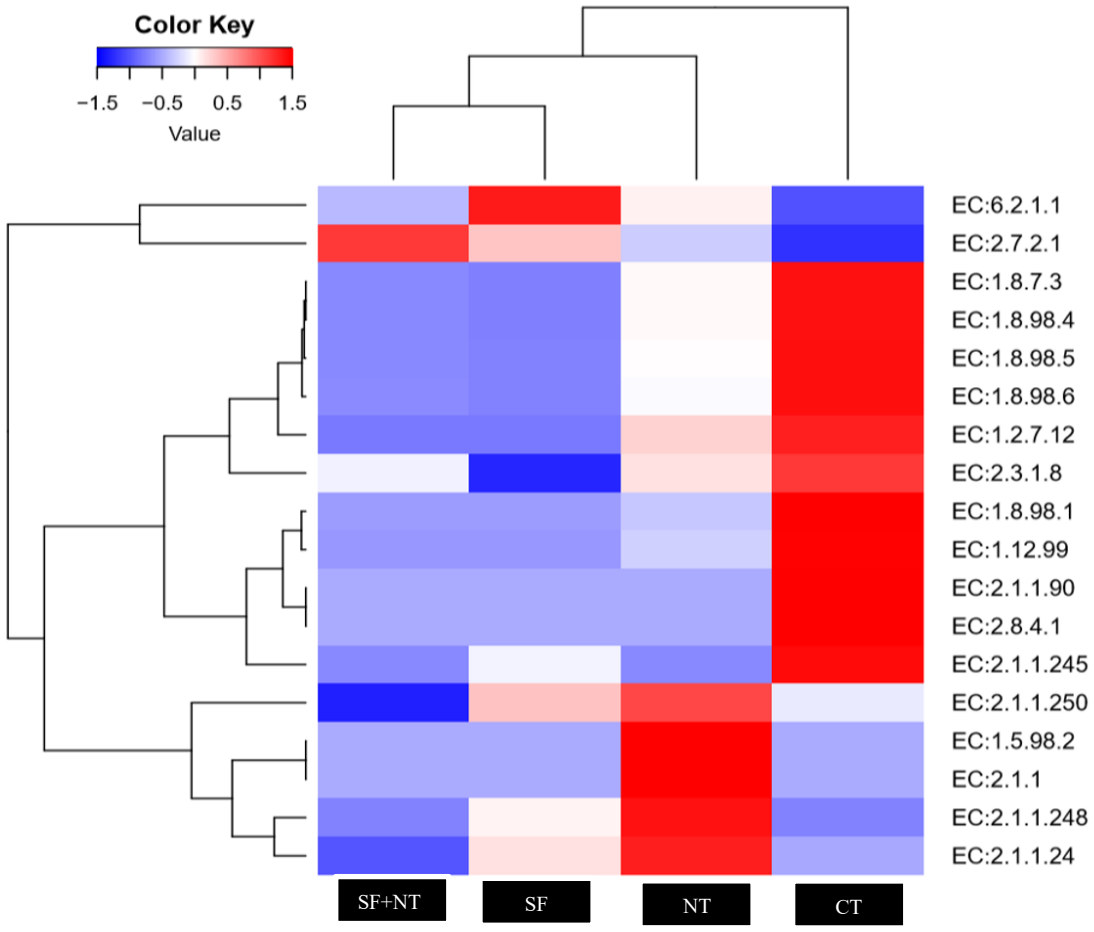


Table 13. Abundance Profile of Enzymes Involved in Methanogenesis. Enzymes which are responsible for conversion of methanol, acetate, CO2, and methylamines into methane were significantly decreased in case of sulfate and nitrate combination (COMBO) diet. However, a significant increase in acetate kinase enzyme [EC 2.7.2.1] has been observed in COMBO diet.

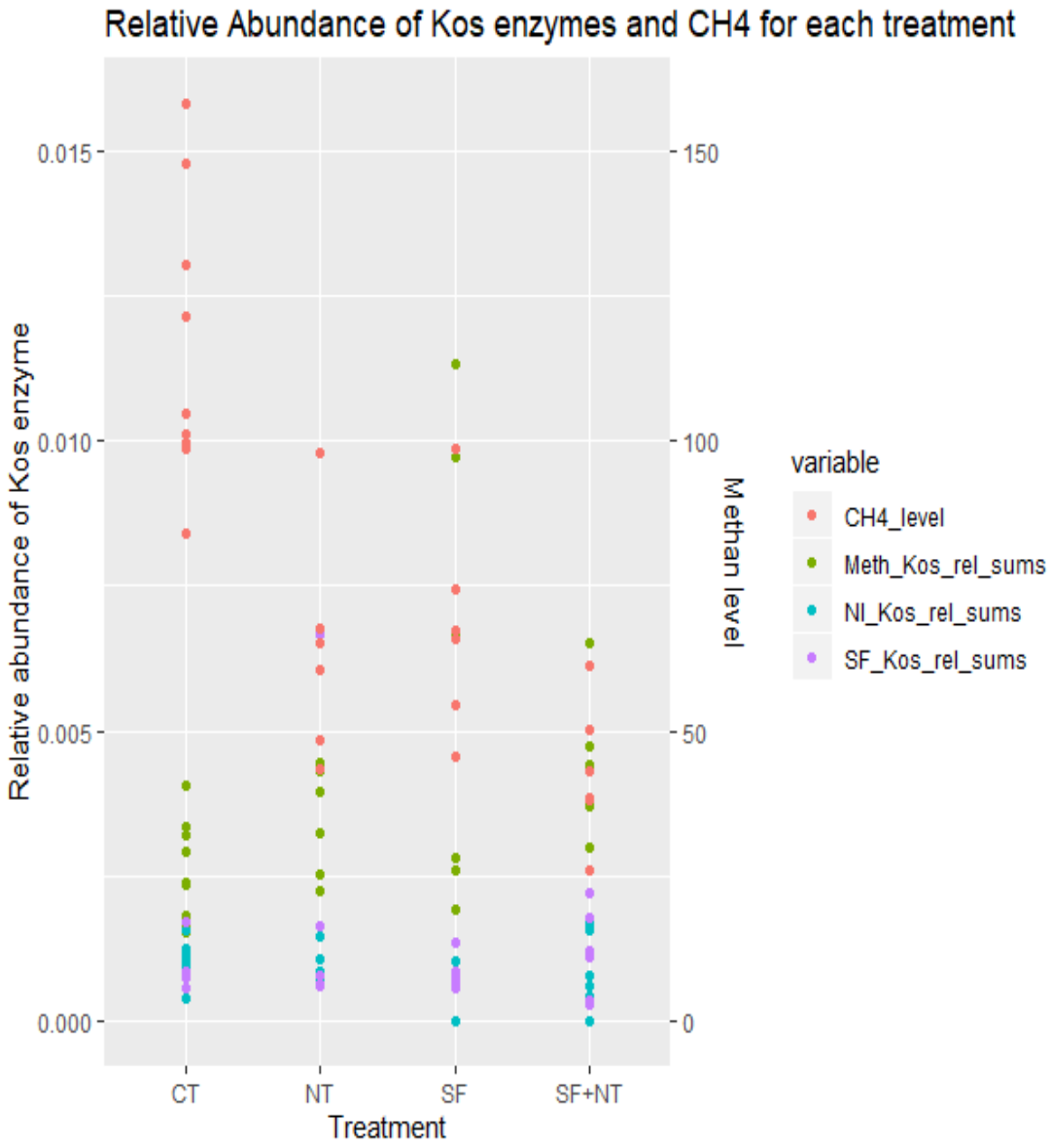


Figure 14. Relative Abundance of KOs Enzymes and Methane Yield in Different Diet Treatments. Correlation between the relative sums of KOs enzymes in methane, nitrate, and sulfate metabolism and methane yield is illustrated. The methane yield was higher in common diet, followed by sulfate/nitrate diets. Methane KOs enzymes showed different patterns between different diets, while nitrate and sulfate KOs enzymes were consistent between different diets. Increase in some methane KOs enzymes was observed in some samples of sulfate and COMBO diets because of high abundance of acetate kinase enzymes.

6.6. References

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: :http://www.bioinformatics.babraham.ac.uk/projects- /fastqc

Bankevich, A., Nurk, S., Dmitry Antipov, A., et al, 2012. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." Journal of computational biology: a journal of computational molecular cell biology. 19(5):455-77.

Belanche, A., de la Fuente, G., Newbold, C.G., 2014. Study of methanogen communities associated with different rumen protozoal populations. FEMS microbiology ecology. 90(3):663–677.

Boisvert, S., Raymond, F., dzaridis, E., et al, 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13:R122.

Cai, S., Li, J., Hu, F.Z., et al, 2010. Cellulosilyticum ruminicola, a newly describes rumen bacterium that possesses redundant fibrolytic-protein-encoding genes and degrades lignocellulose with multiple carbohydrate-borne fibrolytic enzymes. Applied and Environmental Microbiology. 76: 3818-3824.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., et al, 2016. Han, A. W., Johnson, A. J., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods. 13(7):581–583.

Caporaso, J., Kuczynski, J., Stombaugh, J., et al, 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 7:335–336.

Crutzen, P.J., Aselmann, I., Seiler, W., 1986. Methane production by domestic animals, wild ruminants, other herbivorous fauna and humans. Tellus 38B:271-284.

Danielsson, R., 2016. Methane production in dairy cows. P. 45.

Danielsson, R., Schnurer, A., Arthurson, V., Bertilsson J., 2012. Methanogenic population and CH4 production in Swedish dairy cows fed different levels of forage. Appl Environ Microbiol. 78:6172–6179.

Danielsson, R., Dicksved, J., Sun, L., et al, 2017. Methane Production in Dairy Cows Correlates with Rumen Methanogenic and Bacterial Community Structure. Frontiers in microbiology. P. 8:226.

Flint, H.J., Bayer, E.A., Rincon, M.T., et al, 2008. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nat. Rev. Microbiol. 6:121–131.

Gurevich, A., Saveliev, V., Vyahhi N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072-1075.

Hales, K.E., Cole, N.A., MacDonald, J.C., 2012. Effects of corn processing method and dietary inclusion of wet distillers grains with solubles on energy metabolism, carbon–nitrogen balance, and methane emissions of cattle. J. Anim. Sci. 90:3174–3185.

Hristov, A.N., Firkins, J.L. Dijkstra, J., et al, 2013. Special Topics—Mitigation of methane and nitrous oxide emissions from animal operations: I. A review of enteric methane mitigation options. J. Anim. Sci. 91:5045–5069.

Hyatt, D., Chen, G.L., Locascio, P.F., et al, 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 11(1):119.

J.R. Newbold, van Zijderveld, S.M., Hulshof, R.B.A., et al, 2014. The effect of incremental levels of dietary nitrate on methane emissions in Holstein steers and performance in Nelore bulls. J. Anim. Sci. 92:5032-5040.

Johnson, D.E., Ward, G.M., 1996. Estimates of animal methane emissions. Environ. Monit. Assess. 42:133–141.

Johnson, K.A., Johnson, D.E., 1995. Methane emissions form cattle. J. Anim. Sci. 73:2483–2492.

Kanehisa, M., Sato, Y., Kawashima, M., et al, 2016. KEGG as a reference resource for gene and protein annotation. Nucleic acids research. 44(D1):D457–D462.

Kittelmann, S., Pinares-Patino, C.S., Seedorf, H., et al, 2014. Two different bacterial community types are linked with the low-methane emission trait in sheep. Plos One. 9(7):e103171.

Kozich, J.J., Westcott, S.L., Baxter, N.T., et al, 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and environmental microbiology. 79(17):5112–5120.

Kuczynski, J., Lauber, C.L., Walters, W.A., et al, 2011. Experimental and analytical tools for studying the human microbiome. Nature reviews. Genetics, 13(1), 47–58.

Lahti, L., et al. (Bioconductor, 2017-2019). Tools for microbiome analysis in R. Microbiome package version . URL: (http://microbiome.github.io/microbiome)

Li, D., Liu, C.M., Luo R., et al, 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 31(10):1674–6.

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

Lozupone, C., Lladser, M.E., Knights, D., et al, 2011. UniFrac: an effective distance metric for microbial community comparison. The ISME journal. 5(2), 169–172.

Luo, R., Liu, B., Xie, Y., et al, 2012. "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." GigaScience. 1:1-18.

Mangino, J., Peterson, P., Jacobs, H., 2007. Development of an Emissions Model to Estimate Methane Fermentation in Cattle. US-Environmental Protection Agency. Accessed Aug. 16, 2015.

McMurdie, P., Holmes, S., 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 8(4):e61217

Morgavi, D.P., Martin, C., Jouany, J.P., Ranilla, M.J., 2012. Rumen protozoa and methanogenesis: not a simple cause-effect relationship. British Journal of Nutrition. 107:388-397.

Moss, A., Jouany, J., Newbold, J., 2000. Methane production by ruminants: its contribution to global warming. Annales de zootechnie, INRA/EDP Sciences. 49(3):31-253.

Myhre, G., Shindell, D., Bréon, F.M., Collins, W., et al, 2013. Anthropogenic and Natural Radiative Forcing. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Paz, H.A., Anderson, C.L., Muller, M.J., Kononoff, P.J., Fernando, S.C, 2016. Rumen Bacterial Community Composition in Holstein and Jersey Cows Is Different under Same Dietary Condition and Is Not Affected by Sampling Method. Front Microbiol., 7:1206.

Pesta, A.C., 2015. Dietary strategies for the mitigation of methane production by growing and finishing cattle. Ph.D. Diss. Univ. of Nebraska-Lincoln.

Ploner, A., 2014. Heatplus: Heatmaps with row and/or column covariates and colored clusters. R package version 2.14.0.

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 26(1):139–140.

Rognes, T., Flouri, T., Nichols, B., et al, 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 4:e2584.

Russell, J. B., 2002. Rumen Microbiology and Its Role in Ruminant Nutrition. Published by James B. Russell, Ithaca, NY.

Sarturi, J. O., Erickson, G.E., Klopfenstein, T.J., Vasconcelos, J.T., Griffin, W.A., et al, 2013. Effect of sulfur content in wet or dry distillers grains fed at several inclusions on cattle growth performance, ruminal parameters, and hydrogen sulfide. J. Anim. Sci. 91:4849-4860.

Segata N., Waldron, L., Ballarini, A., et al, 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nature Methods. 8:811–814.

Shabat, S.K., Sasson, G., Doron-Faigenboim, A., et al, 2016. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. The ISME journal. 10(12):2958–2972.

Shah, A., Mahmood, F., Maroof, Q., et al, 2014. Microbial ecology of anaerobic digesters: the key players of anaerobiosis. The Scientific World Journal. P. 183752.

Shi, W.B., Moon, C.D., Leahy, S.C., et al, 2014. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res. 24:1517–25.

Tapio, I., Snelling, T.J., Strozzi, F., Wallace, R.J., 2017. The ruminal microbiome associated with methane emissions from ruminant livestock. Journal of animal science and biotechnology. 8:7.

Troy, S., Duthie, C.A., Hyslop, J., et al, 2015. Effectiveness of nitrate addition and increased oil content as methane mitigation strategies for beef cattle fed two contrasting basal diets. Journal of Animal Science. 93(4):1815-23.

Ungerfeld, E.M., Kohn, R.A., 2006. The role of thermodynamics in the control of ruminal fermentation. In: K. Sejrsen, T. Hvelplund, and M. O. Nielsen, editors, Ruminant physiology: Digestion, metabolism, and impact of nutrition on gene expression, immunology, and stress. Wageningen Academic Publishers, Wageningen, The Netherlands. P. 55-85.

United States Environmental Protection Agency (EPA), 2020. Draft Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2018. Complete Report. P. 719.

Van Zijderveld, S.M., Gerrits, J.J., Apajalahti, J.A., et al, 2010. Nitrate and sulfate: Effective alternative hydrogen sinks for mitigation of ruminal methane production in sheep. J. Dairy Sci. 93:5856-5866.

Whiteley, A.S., Jenkins, S., Waite, I., et al, 2012. Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) platform. Journal of Microbiological Methods. 91:80-88.

Zhou, M., Chung, Y.H., Beauchemin, K.A., et al, 2011. Relationship between rumen methanogens and methane production in dairy cows fed diets supplemented with a feed enzyme additive. J Appl Microbiol. 111:1148–58.

CHAPTER 7

CONCLUSIONS

The objective of this study was: (1) to evaluate the potential allergy risks for consumption of novel foods and GE organisms including microalgae, fungi, insects and GE canola to comply with a regulatory request; and (2) to study the effects of nitrate and sulfate supplementations on ruminal archaeal and bacterial composition and functionality linked to methane mitigation in ruminants.

The proteomes of studied novel foods have been predicted through combination of whole genome sequencing, genomic, transcriptomic, and proteomic tools. However, proteins from these foods had hundreds of matches to extensively conserved proteins in different allergenic, and non-allergenic species following CODEX limits. Highly conserved proteins across diverse taxa are unlikely to pose risks. Therefore, critical evaluation of the current guidelines may provide guidance to classify some allergenic proteins as of lower risk with higher identity matches.

There are no published genomes or proteomes for some newly developed foods, and the publically available protein databases may not contain such useful information to be used for prediction of potential cross-reactivity to allergens. In this study, we presented an alternative workflow to develop reference protein databases through bioinformatics analysis of published genomic and transcriptomic raw sequencing data. We developed a protein database for House Cricket which has been validated using proteomic data. The use of this bioinformatics approach demonstrated that shrimp allergic patients may experience cross-reaction if they consume novel edible insects.

EFSA is asking food developers to evaluate the potential horizontal gene transfer from plants to microbes through comparison of the DNA inserts against the genomes of bacteria and archaea. Bioinformatics analysis raises no concerns that the inserted DNA in transgenic canola would be transferrable to bacteria or archaea. A sequence searchable celiac database has been developed to identify proteins or peptides for risk assessment of novel food proteins. The database has been updated in 2018, filtering peptides shorter than 9 AA. Bioinformatics comparisons with homologous proteins from Pooideae and from non-Pooideae monocots, dicots and animal proteins were used to predict the FASTA35 defaults. Taken together, bioinformatics tools provide useful evaluations for risk assessment of novel food sources.

In the last part, 16S sequencing and metagenomics have been used to investigate the effect of nitrate and sulfate dietary interventions on microbiome composition and function, and their impacts on finishing cattle performance and methane emissions. Sulfate and nitrate combinations helped to reduce methane emissions, but with a decrease in cattle performance data. 16S reported significant changes in the ruminal bacterial composition which are assigned to H_2 utilization in formation of fatty acids, nitrate and sulfate reduction instead of methane formation in COMBO diet. Metagenomic shotgun sequencing demonstrated a significant deacrease in enzymes linked to conversion of CO2, methanol, acetate, and methylamines into methane in case of COMBO diet. Therefore, this study provides evidence that methane production is linked to diet type, microbiome structure, and differential gene abundance in the cattle rumen microbiome. Therefore, integration of 16S and shotgun metagenomics helped to predict such a correlation between the microbiome and the functional methane attibutes.

Overall, bioinformatics tools can be used as a preliminary predictive screening for risk assessment of novel food ingredient sources; and to understand the ecological and functional insights between microbiome and dietary interventions.

APPENDIX I

FASTA Comparison of Predicted Proteins of *Fusarium sp.* to AOL V18B (*E*-Score: 10e-07). Only matches over 50% sequence identity are shown.

Fusarium proteins	AllergenOnline V18B	%Seq_Id	Align_lgth	E-score
RFSUS48114	gid|2243|transaldolase [Fusarium proliferatum]	100	323	4.40E-133
RFSUS31770	gid|543|60S acidic ribosomal protein P2 [Fusarium culmorum]	92.7	110	3.90E-36
RFSUS12296	gid|544|thioredoxin-like protein [Fusarium culmorum]	91.7	121	2.20E-60
RFSUS18429	gid|329|Enolase (2-phosphoglycerate dehydratase) (2-phospho-D-glycerate hydro-lyase) (Allergen Asp f 22) [Aspergillus fumigatus]	85.8	438	1.40E-163
RFSUS60254	gid|519|Heat shock 70 kDa protein (Allergen Cla h 4) (Cla h IV) [Davidiella tassiana]	85	652	0
RFSUS46264	gid|338|60S ribosomal protein L3 (Allergen Asp f 23) [Aspergillus fumigatus]	84.1	391	1.40E-152
RFSUS54841	gid|1033|cytochrome c [Curvularia lunata]	83.5	103	2.80E-53
RFSUS30096	gid|73|60S acidic ribosomal protein P1 (Allergen Alt a 12) (Alt a XII) [Alternaria alternata]	78.2	110	5.10E-30
RFSUS09150	gid|799|NADP-dependent mannitol dehydrogenase [Davidiella tassiana]	75	264	2.10E-85
RFSUS53964	gid|2582|alcohol dehydrogenase [Curvularia lunata]	74.2	349	1.60E-118
RFSUS24836	gid|545|helix-loop-helix protein [Fusarium culmorum]	73.5	381	1.40E-119
RFSUS64420	gid|2291|Der f 33 allergen [Dermatophagoides farinae]	73.2	451	1.90E-150
RFSUS69279	gid|1376|vacuolar serine protease [Cladosporium cladosporioides]	72.6	383	2.40E-119
RFSUS64116	gid|518|aldehyde dehydrogenase (NAD+) [Davidiella tassiana]	70.9	488	4.30E-160
RFSUS44926	gid|2301|glyceraldehyde-3-phosphate dehydrogenase [Triticum aestivum]	70.1	335	7.40E-101
RFSUS31134	gid|1926|cyclophilin [Catharanthus roseus]	69.8	169	1.30E-49
RFSUS54801	gid|2291|Der f 33 allergen [Dermatophagoides farinae]	69.6	447	1.00E-141
RFSUS39995	gid|64|Minor allergen Alt a 7 (Alt a VII) [Alternaria alternata]	68.7	201	1.50E-61
RFSUS25656	gid|1885|manganese superoxide dismutase [Alternaria alternata]	68.6	191	2.20E-56
RFSUS26854	gid|2582|alcohol dehydrogenase [Curvularia lunata]	67.9	346	9.90E-111
RFSUS24234	gid|1337|TCTP [Alternaria alternata]	67.6	170	7.10E-69
RFSUS14314	gid|2708|heat shock cognate 70 [Aedes aegypti]	66	656	1.80E-192
RFSUS41889	gid|246|elongation factor 1 beta-like [Penicillium citrinum]	65.9	232	1.40E-66
RFSUS55614	gid|2330|RecName: Full=Endo-chitosanase; Flags: Precursor [Aspergillus fumigatus]	65.8	234	6.70E-84
RFSUS64885	gid|863|cyclophilin [Aspergillus fumigatus]	65.8	161	1.90E-44
RFSUS45564	gid|336|RecName: Full=Extracellular elastinolytic metalloproteinase; Flags: Precursor [Aspergillus fumigatus]	64.7	634	2.80E-184
RFSUS39951	gid|2070|SchS21 protein, partial [Stachybotrys chartarum]	64.3	140	3.50E-41
RFSUS18445	gid|64|Minor allergen Alt a 7 (Alt a VII) [Alternaria alternata]	64.2	204	1.30E-58
RFSUS33126	gid|648|major allergenic protein Mal f4 [Malassezia furfur]	63.8	329	5.00E-113
RFSUS66716	gid|325|PPIase [Aspergillus fumigatus]	62.6	187	1.80E-48
RFSUS62635	gid|1941|cyclophilin [Daucus carota]	58.7	172	3.10E-42
RFSUS00452	gid|518|aldehyde dehydrogenase (NAD+) [Davidiella tassiana]	58.7	450	6.50E-122
RFSUS57549	gid|332|rAsp f 9 [Aspergillus fumigatus]	58.1	298	4.00E-73
RFSUS61329	gid|2591|heat shock-like protein [Tyrophagus putrescentiae]	57.9	580	8.20E-143

RFSUS45400	gid|251|peroxisomal membrane protein [Penicillium citrinum]	57.6	165	1.30E-52
RFSUS36734	gid|489|putative nuclear transport factor 2 [Davidiella tassiana]	57.4	115	3.50E-29
RFSUS26970	gid|63|Protein disulfide-isomerase (PDI) (Allergen Alt a 4) [Alternaria alternata]	56.7	379	7.70E-86
RFSUS30974	gid|925|pectate lyase [Penicillium citrinum]	55.9	295	6.20E-65
RFSUS31114	gid|2457|extracellular alkaline serine protease [Aspergillus versicolor]	55.8	419	2.00E-85
RFSUS10681	gid|1228|putative alpha/beta hydrolase superfamily protein [Davidiella tassiana]	55.3	262	8.70E-58
RFSUS48009	gid|694|allergen Ole e 5 [Olea europaea]	55.3	152	3.40E-35
RFSUS60641	gid|518|aldehyde dehydrogenase (NAD+) [Davidiella tassiana]	55	500	3.40E-117
RFSUS50720	gid|1338|ragweed homologue of Art v 1 precursor [Ambrosia artemisiifolia]	54.3	81	5.90E-16
RFSUS20874	gid|544|thioredoxin-like protein [Fusarium culmorum]	54.3	105	1.40E-20
RFSUS30736	gid|65|aldehyde dehydrogenase (NAD+) [Alternaria alternata]	54.1	492	2.50E-119
RFSUS25721	gid|648|major allergenic protein Mal f4 [Malassezia furfur]	53.8	333	1.30E-61
RFSUS19680	gid|317|Oryzin precursor (Alkaline proteinase) (ALP) (Aspergillus proteinase B) (Aspergillopeptidase B) [Aspergillus oryzae]	53.4	397	1.20E-83
RFSUS50614	gid|951|Der f Mal f 6 allergen [Dermatophagoides farinae]	53.1	145	8.80E-28
RFSUS07749	gid|317|Oryzin precursor (Alkaline proteinase) (ALP) (Aspergillus proteinase B) (Aspergillopeptidase B) [Aspergillus oryzae]	53	419	7.40E-82
RFSUS27504	gid|400|29 kDa IgE-binging protein [Candida albicans]	52.8	231	2.70E-49
RFSUS54369	gid|323|major allergen Asp F2 [Aspergillus fumigatus Af293]	52.7	264	1.70E-59
RFSUS63171	gid|324|aspergillopepsin i [Aspergillus fumigatus]	52.4	401	3.80E-82
RFSUS64496	gid|876|thioredoxin [Aspergillus fumigatus]	52.3	107	9.50E-22
RFSUS49759	gid|2271|aspartyl endopeptidase [Rhizopus oryzae]	51.7	400	5.20E-91
RFSUS55321	gid|925|pectate lyase [Penicillium citrinum]	51.6	289	3.60E-59
RFSUS70059	gid|518|aldehyde dehydrogenase (NAD+) [Davidiella tassiana]	51.5	499	4.60E-110
RFSUS24160	gid|590|superoxide dismutase (manganese) [Hevea brasiliensis]	51.1	227	1.70E-58
RFSUS67229	gid|317|Oryzin precursor (Alkaline proteinase) (ALP) (Aspergillus proteinase B) (Aspergillopeptidase B) [Aspergillus oryzae]	51	420	7.10E-71
RFSUS27814	gid|166|triosephosphat-isomerase [Triticum aestivum]	50.9	279	4.40E-53
RFSUS10384	gid|925|pectate lyase [Penicillium citrinum]	50.5	291	5.40E-55
RFSUS69615	gid|332|rAsp f 9 [Aspergillus fumigatus]	50.4	266	3.60E-55
RFSUS20171	gid|518|aldehyde dehydrogenase (NAD+) [Davidiella tassiana]	50.4	492	7.00E-113
RFSUS42495	gid|518|aldehyde dehydrogenase (NAD+) [Davidiella tassiana]	50.2	496	8.40E-108