5-16-2022

INVESTIGATION INTO THE GENETIC BASIS OF CAPSAICIN PRODUCTION IN PEPPERS USING NEXT GENERATION RNA SEQUENCING AND SYNTHETIC BIOLOGY APPROACHES

Ryan Patrick Calhoun
Rowan University

INVESTIGATION IN TO THE GENETIC BASIS OF CAPSAICIN PRODUCTION IN PEPPERS USING NEXT GENERATION RNA SEQUENCING AND SYNTHETIC BIOLOGY APPROACHES

by

Ryan Patrick Calhoun

A Thesis

Submitted to the
Department of Molecular and Cellular Biosciences
College of Science and Mathematics
In partial fulfillment of the requirement
For the degree of
Master of Science in Bioinformatics
at
Rowan University
August 13, 2021

Thesis Chair: Benjamin R. Carone, Ph.D., Professor of Molecular and Cellular Biosciences

Committee Members:
Yong Chen, Ph.D., Professor of Bioinformatics & Mathematics
Thomas Keck, Ph.D., Professor of Chemistry & Biochemistry/Molecular & Cellular Biosciences

Dedication

I would like to dedicate this thesis to my mentor Dr. Benjamin Carone for his support, guidance, and creativity in this project and pushing me to become a better scientist. He has truly shaped my academic career in both undergraduate and graduate, guiding me every step of the way and I would not be the person I am without him. I am ever so grateful for him taking me into his laboratory and teaching me everything I have come to know and love about science. He has truly influenced me to become a better student, worker, scientist and has helped pave the road to my future. I would not be where I am today if it were not for Dr. Carone.

Acknowledgements

I would like to thank my thesis chair and mentor Dr. Benjamin R Carone who taught me all of my wet lab experience and helped guide me through my masters program and thesis. I would also like to thank my committee member Dr. Yong Chen for teaching me in-depth bioinformatic techniques and analysis which were instrumental in my thesis work. I would also like to thank Dr. Thomas Keck for his guidance and feedback for my thesis and everyone's contribution in making my final project.

I would like to acknowledge members of the Carone laboratory for their contributions to the projects and specifically, Nicholas Paterna for his help and assistance in teaching me the R language for statistical computing and graphics. Members of the Carone laboratory always promoted a joyful and friendly working environment which helped encourage me and support me through my work. There was never a dull moment in the laboratory and it was always a pleasure to be surrounded by such kind friends and mentors who made my science career fun and memorable.

I would also like to thank my family members, specifically my mother and father who have made my education career possible and always supported me and my decisions. They have made me the best version of myself and I cannot thank them enough for everything they have done for me and allowing me to further pursue my educational career. I would also like to thank my girlfriend, Nicole Varrella for her unmatched support and motivation that got me through my thesis program.

Abstract

Ryan Patrick Calhoun
INVESTIGATION IN TO THE GENETIC BASIS OF CAPSAICIN PRODUCTION IN PEPPERS USING NEXT GENERATION RNA SEQUENCING AND SYNTHETIC BIOLOGY APPROACHES
2020-2021
Benjamin R Carone, Ph.D.,
Master of Science in Bioinformatics

Capsaicin, a molecule synthesized by plants in the *Capsicum* genus, is popular for its ability to produce a sensation of burning in any tissue it encounters. The synthesis of capsaicin molecules is achieved through the capsaicin biosynthesis pathway. In this dual study, our goal was to insert two crucial genes, *pun1* and *pAMT*, into a strain of *Saccharomyces cerevisiae* to allow capsaicin synthesis and perform Illumina RNA sequencing on seven pepper species of increasing pungency to identify other key or novel genes needed or related to capsaicin synthesis. We implemented a golden gate cloning strategy to insert our genes of interest into bacteria to then be cloned into yeast. We believe that successful insertion into our yeast strain was achieved for one of the genes, *pAMT*, but the other, *pun1*, appears to not be inserted. We hypothesize that correct insertion and expression of *pun1* would achieve capsaicin synthesis alongside expression of *pAMT*, as these two genes would complete the missing parts of the capsaicin pathway. We identified five possible new gene candidates with unknown functions grouped together with similar expression to known genes present in the capsaicin pathway. These novel genes were identified as follows: *CA01g11020*, *CA12g21630*, *CA09g00520*, *CA03g28900*, *CA09g15570*. We also identified five regions of interest that showed similar trends in expression patterns that could contain new promising genes that are not known to participate in the capsaicin biosynthesis pathway.

Table of Contents

Table of Contents (Continued)

Table of Contents (Continued)

List of Figures

Figure Page

List of Figures (Continued)

Figure Page

List of Tables

Table Page

Chapter 1

Introduction

Capsaicin Overview

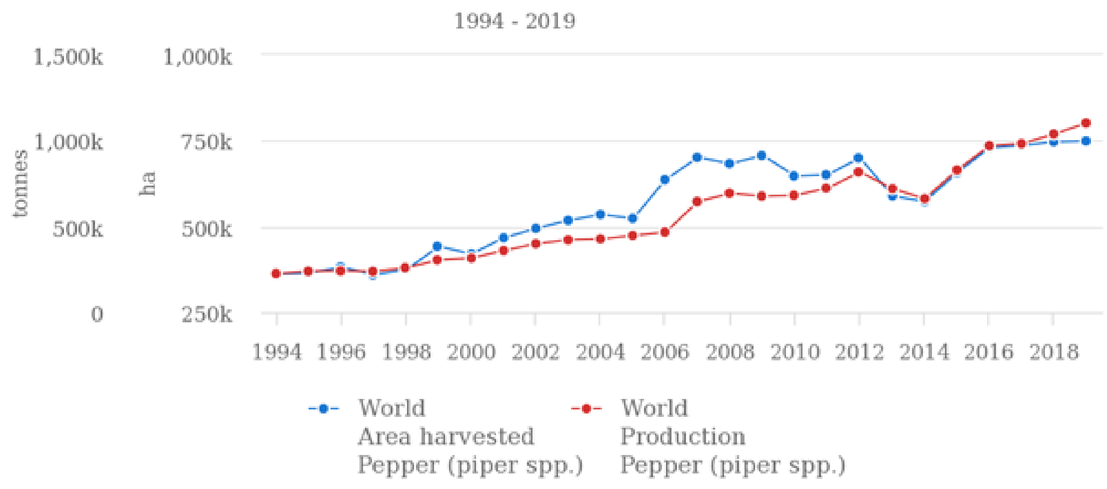
Capsaicin, 8-methyl-N-vanillyl-6-nonenamide, is a pungent molecule that is synthesized from the capsaicin biosynthesis pathway, a pathway that is unique to *Capsicum* plants. Capsaicin is a unique molecule for its ability to elicit a burning sensation in the mucous membranes of mammals (Fitzgerald, 1983). Capsaicin will bind to nociceptors, a group of sensory neurons which respond to specific sensory stimuli, which transmits information regarding tissue damage to pain-processing centers in the spinal cord and brain (Fitzgerald, 1983) (Frias & Merighi, 2016). Specifically, the transient receptor potential vanilloid type 1 (TRPV1) receptor is the receptor in which capsaicin will bind to. These ligand-gated non-selective cation channels are activated by a range of noxious stimuli such as extremes in thermal, poisonous, chemical, or mechanical stimuli (Ross, 2003) (Winter & Campbell, 1995). When the neurons are exposed to such stimuli, the pain response that is stimulated helps the individual recognize the danger and lead them away from this stimulus (Winter & Campbell, 1995). The most common sensation when consuming capscium plants is the spicy tingling burn and this burning sensation is dependent on the type of *Capsicum* plant that is ingested and your own sensitivity of your TRPV1 receptors. Different *Capsicum* plants can produce different amounts of capsaicin and an individual's pain receptors can dictate their response to the pepper stimuli (Caterina et. al., 1997). These differences in pungency, which are a result from the different amounts of capsaicin production, can be measured using the scoville scale which is recorded in scoville heat units (Sanatombi & Sharma, 2008). The scale is based on the concentration of capsaicinoids and was named after Wilbur Scoville in 1912 who originally used an organoleptic test where the exact weight of a dried pepper is dissolved in alcohol to extract capsaicinoids. They were then diluted in a solution of sugar water and given to a panel of five trained tasters until they could no longer taste the heat in the diluted sugar water solution. The heat level is then based on the dilution and rated in multiples of 100 SHU (Bosland et. al., 2008). There are problems with this type of testing such as imprecision based on human subjectivity which arises with each person having a different number of TRPV1 receptors so they will experience different associated ranges of heat. Another weakness is the desensitization that arises when these receptors are stimulated which will be expanded on in the next section. Nowadays, the pungency units for peppers are quantitatively assessed using high-performance liquid chromatography (HPLC) as this form of analysis is highly accurate, reliable, and reproducible for samples (Sanatombi & Sharma, 2008). When a pepper is consumed, the capsaicin that was synthesized and stored within the pepper is released, and the capsaicin will find and bind to TRPV1 receptors located on the tongue, esophagus, and stomach as it is ingested (Fitzgerald, 1983). The binding causes upregulation of the neurons resulting in reduced stimulation thresholds and an increase in pain perception (Ross, 2003). Although our body recognizes capsaicin as an irritant and noxious chemical, there are still many benefits to our health and medicinal uses for capsaicin along with a wide range of commercial and economic impacts it can affect.

The ability for capsaicin to interact with pain receptors has made it an interesting subject for treatment for neurological and pain conditions in patients. As capsaicin molecules interact with their respective nocireceptors, this can lead to a desensitization of the receptor. Receptor desensitization results in a decreased responsiveness due to repeated exposure to agonists (Szolcsanyi, 1997). This decrease in responsiveness results in less ion gated channels being opened, generating less electrical responses to the brain, which results in a lower pain response (Bosland et. al., 2012) (O'Neill et. al., 2012). Scientists can take advantage of this interaction to help treat different pain conditions. Studies by Nolano et. al., 1999 and Mason et. al., 2004, demonstrate that a single application of a low-dose capsaicin cream can result in an inactivation of peripheral afferent fibers and primary and secondary hyperalgesia. However, repeated application of a topical capsaicin cream over a few weeks dampens responses to both mechanical stimuli and heat (Nolano et. al., 1999) (Mason et. al., 2004). Capsaicin has been used in numerous pain studies from somatic to visceral models and can assess both primary and secondary hyperalgesia (O'Neill et. al., 2012). Capsaicin has also seen potential in dietary strategies such as managing gastrointestinal distress, effects on weight loss and weight maintenance (Singletary, 2011). Two studies conducted by Yoshioka et. al., demonstrated an increase in energy release when capsaicin is mixed with a high fat diet (Yoshioka et. al., 1995) (Yoshioka et. al., 1998). This production of heat after eating contributes to the body's resting metabolic rate and the addition of capsaicin in the diet increases the resting metabolic rate (Sharma et. al., 2013).

It also has positive effects on insulin resistance as a study done in which obese mice were fed on a high fat diet with a capsaicin supplement led to a decrease in fasting glucose and plasma triglycerice levels (Szolcsanyi, 1997) (Yoshioka et. al., 1998). New studies are even demonstrating links between capsaicin containing anti-cancer properties. A study conducted studied the effects of capsaicin on human cancer cell lines by monitoring how these cancerous cells had their cell cycle affected by capsaicin. They reported that the cell line had reduced proliferation and viability of cells, early apoptosis cell signaling, and even arrest points in the cell cycle preventing cancerous cell replication (Lin et. al., 2013). There are many other beneficial applications of capsaicin for medicine and health purposes which make it a unique and interesting molecule to pursue. The economic impacts of the *Capsicum* genus is a huge market itself. Peppers are a common culinary ingredient due to their rich diversity as a crop. According to the Food and Agriculture Organization of the United Nations, FAOSTAT, world pepper production was estimated to be 1,103,024 metric tons in 2019. In addition, the world area for land utilized for pepper production was 749,088 hectares of land. These trends can be seen in figure 1 below demonstrating pepper production and land utilization from 1994-2019 for world pepper production. For economic impact, pepper production totaled an estimated 3.8 billion U.S. dollars in 2018 in export prices (FAOSTAT, 2017). The wide range of capsaicin levels and flavors they come in make them highly desired. They are utilized in flavoring in food, manufacturing, coloring in cosmetics, and transmitting heat to medicines (Frias & Merighi, 2016).

Figure 1

World Production and World Area Harvested for Peppers According to the Food and Agriculture Organization of the United Nations



Source: FAOSTAT (Feb 10, 2021)

Note. World production of pepper is recorded in tonnes and is represented as the red dotted line. World area harvested is recorded in hectares and is represented as the blue dotted line.

The *Capsicum* genus belongs in the Solanaceae family which consists of several important crops and model plants such as the potato, tomato, eggplant, and pepper species (Mueller et. al., 2005). The main domesticated *Capsicum* species are C. annuum, C. baccatum, C. chinense, C. frutescens, and C. pubescens (Wang & Bosland, 2006) (Heiser & Pickersgill, 1969). All plants of the *Capsicum* genus produce capsaicinoids with the exception of the bell pepper, *Capsicum annuum*. The major capsaicinoids produced from this genus are capsaicin and dihydrocapsaicin with other analogs such as nordihydrocapsaicin, homocapsaicin and more produced in lower amounts (Davis et. al., 2007). Capsaicinoids can be found to be most concentrated in the placenta of the pepper as this is the major site for capsaicinoid biosynthesis (Wang & Bosland, 2006) (Frias &

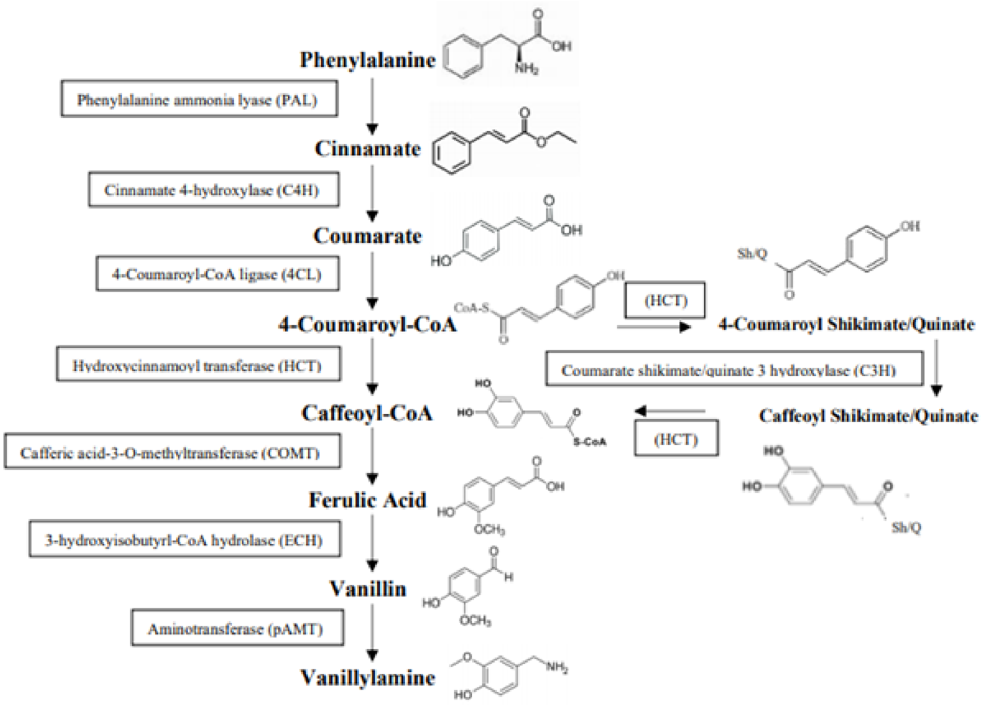
Merighi, 2016). Capsaicin is the main active component in chili peppers and the fundamental structure of it as its analogs are a branched-chain fatty acid amide of vanillylamine. The synthesis of capsaicin evolved as a defensive mechanism for the *Capsicum* genus to prevent insects or mammals from eating the plant since they eat the plant and destroy the seeds in the process or prevent the seeds from being able to be dispersed. Thereby when insects and mammals ingest *Capsicum* species, capsaicin will bind to their TRPV1 receptors eliciting a burning sensation to deter these pests (Jordt & Julius, 2002). The active component of plants belonging in the *Capsicum* genus that give them their spiciness is their capsaicinoid production. The actual synthesis of capsaicinoids is done through two molecular pathways, the phenylpropanoid and the branched chain fatty acid pathways (Davis et. al., 2007) (Zhang et al., 2016) There are many important enzymes that are involved in the synthesis of these pathways but characterization and regulation of the pathways are still being researched.

Capsaicin Biosynthesis Pathways

The phenylpropanoid pathway starts with the precursor molecule phenylalanine, which is converted to cinnamate through phenylalanine ammonia lyase (*PAL*). Cinnamate is then turned into coumarate via cinnamate 4-hydroxylase (*C4H*) and coumarate is turned into 4-coumaroyl-CoA through 4-coumaroyl-CoA ligase (*4CL*). 4-couramoyl-CoA can be converted to Caffeoyl-CoA through hydroxycinnamoyl transferase (*HCT*) or 4-couramoyl-CoA can also be converted to 4-coumaroyle Shikimate/Quinate via hydroxycinnamoyl transferase (*HCT*). Then 4-coumaroyle Shikimate/Quinate can interact with the enzyme coumarate shikimate/quinate 3 hydroxylase (*C3H*) into caffeoyl shikimate/quinate which can be converted to caffeoyl-CoA through hydroxycinnamoyl transferase again (*HCT*). Caffeoyl-CoA interacts with the cafferic acid-3-O-methyltransferase (*COMT*) enzyme to yield ferulic acid. Ferulic acid is converted to vanillin through the 3-hydroxyisobutyrl-CoA hydrolase (*ECH*). Finally, vanillin is converted to vanillylamine through an aminotransferase (*pAMT*) to complete the phenylpropanoid pathway (Wang & Bosland, 2016) (Zhang et. al., 2016) (Bennett & Kirby, 1968) (Sukrasno & Yeoman, 1993). An overview of the phenylpropanoid pathway can be seen better in figure 2 below showing the flow of molecules, structures of those molecules, and important enzymes that participate in the reaction.

Figure 2

Phenylpropanoid Pathway

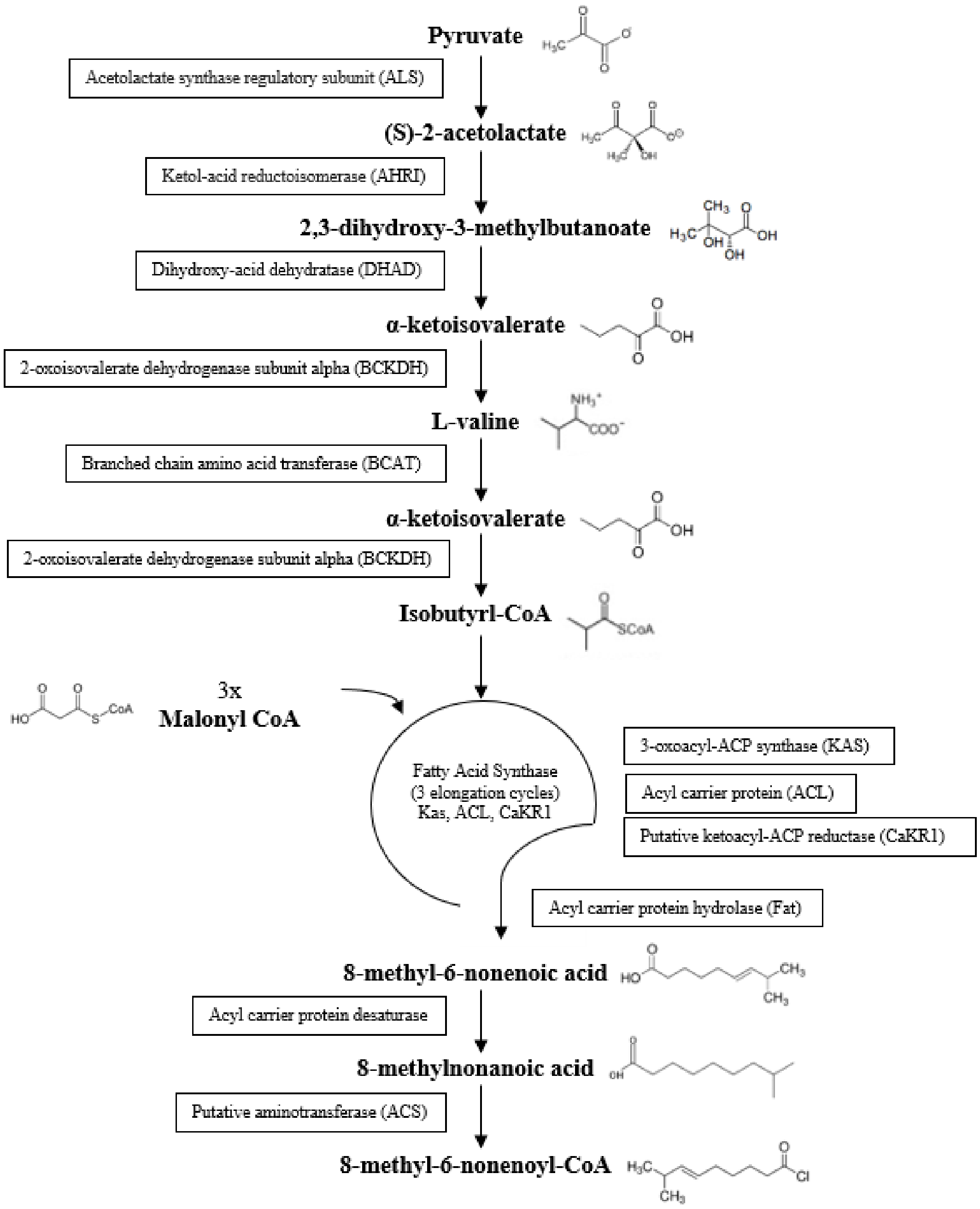

Phenylalanine
Phenylalanine ammonia lyase (PAL)
Cinnamate
Cinnamate 4-hydroxylase (C4H)
Coumarate
4-Coumaroyl-CoA ligase (4CL)
4-Coumaroyl-CoA
(HCT)
4-Coumaroyl Shikimate/Quinate
Hydroxycinnamoyl transferase (HCT)
Coumarate shikimate/quinate 3 hydroxylase (C3H)
Caffeoyl-CoA
(HCT)
Caffeoyl Shikimate/Quinate
Cafferic acid-3-O-methyltransferase (COMT)
Ferulic Acid
3-hydroxyisobutyrl-CoA hydrolase (ECH)
Vanillin
Aminotransferase (pAMT)
Vanillylamine

Note. Molecules of each reaction are indicated as bold text, structures of each molecule are listed besides, enzymes that catalyze each step are listed inside the boxes beside each arrow.

The branched fatty acid pathway starts with pyruvate molecules being converted to (S)-2-acetolactate by an acetolactate synthase regulatory subunit (*ALS*), then (S)-2-acetolactate is converted to 2,3-dihydroxy-3-methylbutanoate by the ketol-acid reductoisomerase (*AHRI*). 2,3-dihydroxy-3-methylbutanoate is converted into α-ketoisovalerate by dihydroxy-acid dehydratase (*DHAD*), α-ketoisovalerate is converted to L-valine through 2-oxoisovalerate dehydrogenase subunit alpha (*BCKDH*), and then L-valine is converted back to α-ketoisovalerate through branched chain amino acid transferase (*BCAT*). The α-ketoisovalerate molecule will turn into isobutyrl-CoA through the 2-oxoisovalerate dehydrogenase subunit alpha (*BCKHD*). Isobutyrl-CoA, along with malonyl CoA, will go into the fatty acid synthase for 3 rounds of elongation utilizing the 3-oxoacyl-ACP synthase (*KAS*), acyl carrier protein (*ACL*), and putative ketoacyle-ACP reductase (*CaKR1*) enzymes. After 3 rounds, the product interacts with the enzyme acyl carrier protein hydrolase (*FatA*) to make 8-methyl-6-nonenoic acid. 8-methyl-6-nonenoic acid is converted 8-methylnonanoic acid via an acyl carrier protein desaturase and 8-methylnonanoic acid will be converted to 8-methyl-6-nonenoyl-CoA through a putative aminotransferase (*ACS*) enzyme (Wang & Bosland, 2016) (Zhang et. al., 2016) (Bennett & Kirby, 1968) (Suzuki et. al., 1981). An overview of the branched fatty acid pathway can be seen better in figure 3 below showing the flow of molecules, structures of those molecules, and important enzymes that participate in the reaction.

Figure 3

Branched-Chain Fatty Acid Pathway

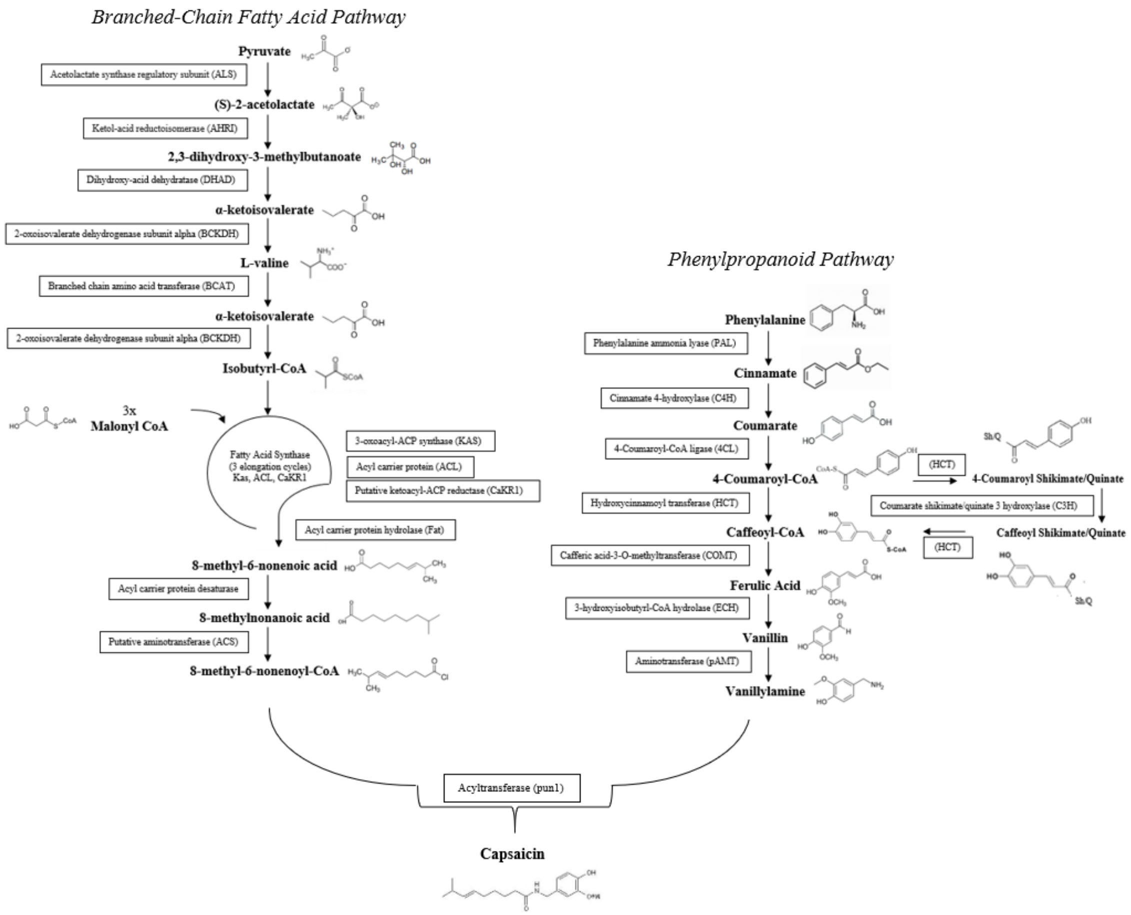

Pyruvate
Acetolactate synthase regulatory subunit (ALS)
(S)-2-acetolactate
Ketol-acid reductoisomerase (AHRI)
2,3-dihydroxy-3-methylbutanoate
Dihydroxy-acid dehydratase (DHAD)
α-ketoisovalerate
2-oxoisovalerate dehydrogenase subunit alpha (BCKDH)
L-valine
Branched chain amino acid transferase (BCAT)
α-ketoisovalerate
2-oxoisovalerate dehydrogenase subunit alpha (BCKDH)
Isobutyrl-CoA
3x
Malonyl CoA
Fatty Acid Synthase
(3 elongation cycles)
Kas, ACL, CaKR1
3-oxoacyl-ACP synthase (KAS)
Acyl carrier protein (ACL)
Putative ketoacyl-ACP reductase (CaKR1)
Acyl carrier protein hydrolase (Fat)
8-methyl-6-nonenoic acid
Acyl carrier protein desaturase
8-methylnonanoic acid
Putative aminotransferase (ACS)
8-methyl-6-nonenoyl-CoA

Note. Molecules of each reaction are indicated as bold text, structures of each molecule are listed besides, enzymes that catalyze each step are listed inside the boxes beside each arrow.

The combination of these two pathways together through a coenzyme A-dependent acyltransferase allows for the synthesis of capsaicin. The coenzyme A-dependent acyltransferase was identified by Stewart et. al. which is encoded by the *AT3* gene, namely pungent gene 1 (*pun1*) as capsaicin synthase (*CS*) (Stewart et. al., 2005). The entire phenylpropanoid pathway and branched-chain fatty acid pathway can be seen below in figure 4.

Figure 4

Capsaicin Biosynthesis Pathway


Branched-Chain Fatty Acid Pathway
Pyruvate
Acetolactate synthase regulatory subunit (ALS)
(S)-2-acetolactate
Ketol-acid reductoisomerase (AHRI)
2,3-dihydroxy-3-methylbutanoate
Dihydroxy-acid dehydratase (DHAD)
α-ketoisovalerate
2-oxoisovalerate dehydrogenase subunit alpha (BCKDH)
L-valine
Branched chain amino acid transferase (BCAT)
α-ketoisovalerate
2-oxoisovalerate dehydrogenase subunit alpha (BCKDH)
Isobutyrl-CoA
3x
Malonyl CoA
Fatty Acid Synthase
(3 elongation cycles)
Kas, ACL, CaKR1
3-oxoacyl-ACP synthase (KAS)
Acyl carrier protein (ACL)
Putative ketoacyl-ACP reductase (CaKR1)
Acyl carrier protein hydrolase (Fat)
8-methyl-6-nonenoic acid
Acyl carrier protein desaturase
8-methylnonanoic acid
Putative aminotransferase (ACS)
8-methyl-6-nonenoyl-CoA
Phenylpropanoid Pathway
Phenylalanine
Phenylalanine ammonia lyase (PAL)
Cinnamate
Cinnamate 4-hydroxylase (C4H)
Coumarate
4-Coumaroyl-CoA ligase (4CL)
4-Coumaroyl-CoA
(HCT)
4-Coumaroyl Shikimate/Quinate
Hydroxycinnamoyl transferase (HCT)
Coumarate shikimate/quinate 3 hydroxylase (C3H)
Caffeoyl-CoA
(HCT)
Caffeoyl Shikimate/Quinate
Cafferic acid-3-O-methyltransferase (COMT)
Ferulic Acid
3-hydroxyisobutyrl-CoA hydrolase (ECH)
Vanillin
Aminotransferase (pAMT)
Vanillylamine
Acyltransferase (pun1)
Capsaicin

*Note***.** Bolded text indicates molecules that are synthesized or required for synthesis of capsaicin. Molecular structures of molecules can be illustrated adjacently to the named molecule. Enzymes that catalyze each step are listed inside the boxes.

There are many important genes involved in capsaicin synthesis, however some studies showed that specifically the *pun1* gene is of high importance when it comes to capsaicin accumulation. A study done by Ogawa et. al*.,* demonstrated that the *pun1* gene and its gene product is essential for capsaicin synthesis (Ogawa et. al., 2015). This study looked at the expression levels of *pAMT* and *pun1* in various pepper cultivars and found that high expression of both *pAMT* and *pun1* lead to high levels of capsaicin production. Peppers that are pungent have high expression of both genes while non-pungent peppers have low expression of both genes. However, on comparison of the accumulated levels of vanillylamine (a precursor to capsaicin) and capsaicin itself, it was found that pungent peppers have a low level of vanillylamine present while non-pungent peppers have a high accumulation of vanillylamine. It appears that in non-pungent cultivars that have low expression of *pAMT* and *pun1* compared to pungent cultivars that have high expression of both *pAMT* and *pun1*, the vanillylamine levels were five times higher in the non-pungent cultivar than the pungent cultivar (Ogawa et. al., 2015). This correlation suggests that although *pun1* and *pAMT* both are important and expression of both are needed for production of capsaicin, it is apparent that *pun1* plays a more critical role in capsaicin production as a high or low expression of *pAMT* will yield vanillylamine but high or low expression of *pun1* determines overall capsaicin production.

Capsaicin and *Saccharomyces cerevisiae*

This was an important discovery as it demonstrated the importance of the *pun1* gene in capsaicin production and demonstrated more closely the role *pAMT* plays in the pathway. As time has gone on, the capsaicin biosynthesis pathway has been refined and amended to include the most important genes interacting in the pathway and there is a better understanding of the overall pathway given now than in previous years. Since capsaicin is synthesized through a biochemical pathway, one way to increase capsaicin production in a fast and affordable way would be to incorporate this pathway into an organism that would have a faster growth rate and require less maintenance and cost to grow it. With the high demand for peppers for their use in culinary and medicinal applications, alternative methods for capsaicin production should be evaluated to reduce costs and time. There are many expenses that go into large scale-pepper growing such as land usage, volume of water used for irrigation, and the time it takes for peppers to grow. There are a variety of additional considerations such as the specific seasonal conditions that must be met or created for the peppers, nutrients, chemicals to ensure the peppers safety or longevity, and the many hours needed for harvesting and processing the peppers all equate to a large time and financial commitment. Therefore, to meet this demand, alternative methods for capsaicin production should be evaluated that would save money and time. *Saccharomyces cerevisiae* is a model organism as it has been widely studied, is easily maintained, grows rapidly in a large quantity, has a sequenced genome and genetic manipulation of this organism has been well established (Karathia et. al., 2011). Incorporating the genes from the capsaicin biosynthesis pathway into yeast would allow for capsaicin production if the precursor molecules in the pathway are readily available.

When looking over the phenylpropanoid and branched-chain fatty acid pathways, it appeared that for yeast to synthesize capsaicin, we may only need to incorporate a few genes if the correct precursor molecules are present. The reason for this is that some of the genes in the branched-chain fatty acid pathway perform similar roles in converting pyruvate molecules to isobutyryl-CoA which then are used in the fatty acid synthase cycle to create elongated branched-chained fatty acids (Dittrich et. al., 1998). Fatty acid synthesis is the creation of fatty acids, which play an important role in the cell to create phospholipids which surround organelles (Dittrich et. al., 1998) (Tehlivets et. al., 2007). Yeast can utilize fatty acids for essential processes or as an energy source by first converting fatty acids into a usable intermediate such as acyl-CoAs. This is achieved through thioesterification of fatty acids with coenzyme A. Yeast can also utilize exogenous fatty acids and these exogenous fatty acids are activated by one of five acyl-CoA synthetases such as Faa1p, Faa2p, Faa3p, Faa4p, or Fat1p (Athenstaedt et. al., 1999) (Johnson et. al., 1994). The elongation pathway in yeast utilizes enzymes that would have similar function in the fatty acid synthesis section in the branched-chain fatty acid pathway. These similarities in yeast to synthesize long chain fatty acids make yeast an interesting organism to evaluate for capsaicin synthesis as it may already have some of the machinery present for capsaicin synthesis. Although, this is only one of the two pathways needed for capsaicin synthesis and the major gene that combines the two pathways together would still be absent. However, because yeast can be easily genetically modified, missing genes can always be incorporated and the more similarities that the yeast have in common with the capsaicin pathway, the less genes would have to be introduced.

In order to design yeast that can synthesize capsaicin, we first had to gain a better understanding of what genes *Saccharomyces cerevisiae* may share with the capsaicin biosynthesis pathway. To do this, we searched the literature and assembled the major genes and products of the capsaicin biosynthesis pathway until we were able to piece together the most complete picture of the pathway. We then utilized the U.S. National Library of Medicine's Blast tools to compare similarities in the genes of the capsaicin pathway against the *Saccharomyces cerevisiae* genome to evaluate if they share genes that have similar functions (Johnson et. al., 2008). Since shape dictates function, the more similar the sequence of the genes are to each other the more likely they may play similar roles in their respective organisms. It appears that there may be more in common in *Saccharomyces cerevisiae* with the branched-chain fatty acid pathway than the phenylpropanoid pathway. In table 1 below, there is a list of genes from the capsaicin biosynthesis pathway blasted against *Saccharomyces cerevisiae* along with some statistical output generated from blast such as E value, total score, query coverage, and identity scores of each of the genes. It can be seen that most of the genes listed in the branched-chain fatty acid pathway have a similar identity to genes found in *Saccharomyces cerevisiae* based on their small E values which is the number of expected occurrences by random chance. This value measures similarity beyond randomness for a sequence so the lower the score the more significant it is while the percent identity will simply show the percent of exact characters matching.

Table 1

Compiled Blast Results of Known Genes Involved in the Capsaicin Biosynthesis Pathway

Phenylpropanoid Pathway					
Gene	***Organism***	***E value***	***Total Score***	***Query Coverage***	***Identity***
PAL	*Capsicum annuum*	No significant similarity found			
C4H	*Capsicum annuum*	$2e^{-08}$	59.3	29 %	29.41 %
4CL	*Capsicum annuum*	$5e^{-39}$	152	86 %	26.77 %
HCT	*Capsicum annuum*	$1e^{-05}$	50.1	65 %	23.18 %
C3H	*Capsicum annuum*	No significant similarity found			
COMT	*Capsicum annuum*	No significant similarity found			
ECH	*Capsicum annuum*	No significant similarity found			
pAMT	*Capsicum annuum*	$2e^{-48}$	174	91 %	31.63 %
Pun1	*Capsicum annuum*	No significant similarity found			
Branched-Chain Fatty Acid Pathway					
Gene	***Organism***	***E value***	***Total Score***	***Query Coverage***	***Identity***
ALS	*Capsicum annuum*	$3e^{-20}$	180	72 %	27.42 %
AHRI	*Capsicum annuum*	$4e^{-38}$	212	63 %	32.17 %
DHAD	*Capsicum annuum*	0	677	90 %	58.41 %
BCAT	*Capsicum annuum*	$5e^{-47}$	168	76 %	32.25 %
BCKDH	*Capsicum annuum*	$9e^{-66}$	214	89 %	37.76 %
KAS	*Capsicum annuum*	$5e^{-65}$	218	87 %	32.06 %
ACL	*Capsicum annuum*	$1e^{-33}$	226	71 %	39.09 %
CaKR1	*Capsicum chinense*	$6e^{-19}$	135	88 %	31.84 %
FAT	*Capsicum annuum*	$1e^{-93}$	333	30 %	32.10 %
ACS	*Capsicum annuum*	$1e^{-13}$	75.5	55 %	22.22 %
Pun1	*Capsicum annuum*	No significant similarity found			

Note. Protein sequences of the genes were obtained from uniprot which listed which specific species the sequence was related to. The blastp program was used to blast the protein sequence against Saccharomyces cerevisiae using the non-redundant protein sequence database. The E value, total score, query coverage, and identity were all recorded for the best resulting search for each blast sequence.

Incorporating Capsaicin Biosynthesis in *Saccharomyces cerevisiae*

The first major experiment we will be conducting will be to genetically modify a strain of yeast to incorporate genes from the capsaicin pathway to hopefully achieve capsaicin synthesis. We decided that there are two important genes we should try to incorporate first into our yeast stain which are the *pAMT* and *pun1* genes. These two genes are essential for capsaicin synthesis and would allow us to skip several steps in the phenylpropanoid pathway. If we add the precursor molecule vanillin, the *pAMT* gene will be able to convert vanillin to vanillylamine and then the *pun1* gene would have the vanillylamine molecule ready to synthesize capsaicin. The other precursor molecule needed from the branched-chain fatty acid pathway would be 8-methyl-6-nonenoyl-CoA and this molecule may be already present in yeast as it is a product from the fatty acid synthase cycle which is found in yeast. After the addition of these two genes and testing for capsaicin production, additional genes can be added from either pathway until capsaicin synthesis can be achieved. To achieve spicy yeast, we first will start by incorporating these two genes, *pAMT* and *pun1* through a golden gate cloning assembly and then utilizing the correct plasmid that contains our promoter, coding region, and terminator sequence to then transform into a strain of *Saccharomyces cerevisiae*. If there is no capsaicin production after successful incorporation and expression of our genes of interest, then we will re-evaluate the pathway and decide on other potential genes of interest to incorporate into *Saccharomyces cerevisiae* that may be required for capsaicin synthesis.

Identifying New Capsaicin Genes

As previously mentioned, there are many important enzymes that are involved in the synthesis of these pathways but characterization and regulation of the pathways are still being researched. In order to successfully synthesize the end molecule from a pathway, a robust understanding of what genes are involved and what processes they influence are essential for successful synthesis. There are many ways for researchers to identify the influence genes have and how they relate to pathways and many of these tools are bioinformatics in nature. Gene ontology (GO) is a common form of functional analysis which provides structured, controlled vocabularies which allow it to be used across several domains to annotate gene, gene products, and sequences (Gene Ontology Consortium, 2004). The GO database is useful in determining how differentially expressed genes in a study may be related to each other, what pathways they are part of, what products the genes produce, and much more information. The DAVID database, which stands for Database for Annotation, Visualization, and Integration Discovery and is useful in interpreting a large list of genes (Huang et. al., 2007). The amount of data and information to interpret after being generated from high-throughput sequencing experiments is a daunting task. The many tools that the DAVID database comes equipped with allow for much better discovery and analysis of such large datasets through functional classification, biochemical pathway maps, conserved protein domain architectures, while all being linked to sources of biological annotation (Dennis et. al., 2003). Once a list of genes are generated from the data of high-throughput sequencing, a variety of data can be extracted from said list where different connections and conclusions can be drawn by analyzing the data in many different ways. Since the capsaicin biosynthesis pathway was a pathway that is still being researched and amended with new findings, we wanted to add to this research effort by contributing our own data analysis to hopefully draw some new conclusions.

We decided that paired-end RNA sequencing of a variety of peppers that ranged in scoville intensity would be a good place to start. There have been numerous advances in sequencing techniques since the very first sequencing of the human genome project that has made it much more affordable and possible to perform sequencing in much greater quantities and depths (Chan, 2005). RNA sequencing is a method that examines the quantity and sequences of RNA in a sample allowing us to analyse the transcriptome of said sample. The transcriptome is the total cellular contents of RNAs such as the tRNA, rRNA, mRNA, and others (Ozsolak & Milos, 2011). This is of high importance as it allows researchers to make connections between genes and their protein products. RNA sequencing can tell which genes are being upregulated, downregulated, when they are being expressed, their level of expression, and much more. It gives a more detailed and quantitative view of alternative splicing, allele-specific expression, and gene expression in general (Kukurba & Montgomery, 2015). Being able to connect the pieces between the genome and the functional proteins that are produced allow scientists to more deeply understand the biology of the cell and assess these changes (Ozsolak & Milos, 2011). The development of high throughput next generation sequencing (NGS) revolutionized transcriptomics by enabling RNA analysis through the sequencing of complementary DNA (cDNA) (Kukurba & Montgomery, 2015). A typical RNA-Seq experiment consists of isolating RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on an NGS platform (Ozsolak & Milos, 2011). A more detailed approach for RNA sequencing is as follows. RNA must first be extracted and isolated from a sample, sufficient quantity and more importantly quality is needed form the extraction as this will provide the basis for the sequencing. The RNA molecules are then reverse-transcribed to cDNA which is a much more stable product than the RNA molecule. This cDNA sample is then fragmented randomly to obtain random sized sequences that can be pieced back together in the end to form a more complete picture (Kukurba & Montgomery, 2015). These cDNA fragments are able to be utilized in the NGS workflow where adapters are added to the end of the fragments which allow the sample to attach to the flow cell for Illumina sequencing where sequencing will be performed. The adapters contain the elements which all for the start of sequencing, these elements are the amplification element and primary sequencing site. Then during the sequencing step, clusters of cDNA fragments are amplified through polymerase chain reaction in a process called cluster generation, resulting in millions of copies of cDNA (Ozsolak & Milos, 2011). The next step is to determine what these sequences are and to do this primers are attached, reversible terminators, DNA polymerase, and TCEP used to determine the base sequence through fluorescence for all of the sequences generated. The last step is the software will identify the nucleotides through fluorescence and the accuracy in identification of said nucleotides. Through generating millions of sequences and those sequences being fragmented randomly, the entire transcriptome of a single sample can be pieced together through computer programs to match different fragmented overlapping portions together to get an idea of the entire sequence (Kukurba & Montgomery, 2015). RNA sequencing is a powerful tool and utilizing the RNA sequences that are generated, we can perform the bioinformatics analyses on the sequences and go further to identify possible candidate genes that may participate in the capsaicin biosynthesis pathway which may be important to include when attempting to create genetically modified *Saccharomyces cerevisiae* that can synthesize capsaicin.

Discovery of Additional Genes in the Capsaicin Biosynthesis Pathway

The second major experiment we decided to perform was to send out pepper samples for Illumina paired-end RNA sequencing to determine gene expression levels of pungent cultivars and non-pungent cultivars to determine if there are any important genes that should be incorporated into the pathway that are highly expressed in pungent peppers. There was a study done by Zhang Z. X. et. al., published in 2016 in the journal of Nature titled “Discovery of putative capsaicin biosynthetic genes by RNA-Seq and digital gene expression analysis of pepper.” In their study, they used an Indian pepper called ‘Guijiangwang’ which is one of the world's hottest chili peppers. They harvested portions of the placenta region of this pepper at five different developmental stages and performed RNA-seq to identify assumed genes involved in capsaicin synthesis. They identified 135 genes of known function that were identified as most likely to be involved in regulating capsaicin synthesis with 20 new candidate genes that may play a role too. This was a great study done to identify new genes that may play a role in capsaicin production but we felt that we could expand on this study in various ways. One factor is that they collected one pepper and analyzed capsaicin production in different developmental stages as this pepper, ‘Guijiangwang’, as it has visual differences in its developmental stages. For their study, they only had one pepper collected for RNA expression comparison to compare their gene expression against. Pepper pungency does change as the pepper ages where capsaicin production increases in pepper growth, but they did not include other peppers in their sequencing to compare too and we believe that the study can be expanded on by doing this. In order to expand on the results from this study, we first obtained seven different peppers ranging on the Scoville scale from low

Scoville heat units (100-500 SHU) to a high level of Scoville heat units (800,000-1,000,000 SHU). These seven peppers in order from increasing pungency are as follows: cherry peppers, jalapeno peppers, hungarian wax peppers, serrano peppers, cayenne peppers, habanero peppers, and ghost peppers. Five of the seven peppers belong to the *Capsicum annuum* genus which are cherry peppers, jalapeno peppers, hungarian wax peppers, serrano peppers, and cayenne peppers while the habanero peppers and ghost peppers belong to the *Capsicum chinense* genus. Table 2 below lists the different pepper samples that we sent out for sequencing along with their scoville range and family they belong to. These peppers had their skin and placenta regions harvested and then RNA extractions of these samples were performed and sent out for Illumina-RNA sequencing through Genewiz. The RNA sequencing results that will be obtained will be processed through FastQC, bowtie, and cufflinks to evaluate gene expression between the peppers. Through analyzing different species of peppers and the varied pungency between the peppers, we hope to identify other genes that are important in capsaicin production and compare the known genes associated with capsaicin synthesis from Zhang Z. X. et. al., to our own RNA sequencing results. Using this information, we may also reevaluate and identify other important genes that we may want to incorporate into our genetically modified yeast for capsaicin production if newly identified genes appear to be present and influential in the pathway.

Table 2

List of the Seven Pepper Samples sent for Illumina RNA Sequencing

Pepper Sample	Species	Scoville Range
Cherry Pepper	*Capsicum annuum*	100-500 SHU
Jalapeno Pepper	*Capsicum annuum*	3,500-8,000 SHU
Hungarian Wax Pepper	*Capsicum annuum*	5,000-15,000 SHU
Serrano Pepper	*Capsicum annuum*	10,000-23,000 SHU
Cayenne Pepper	*Capsicum annuum*	30,000-50,000 SHU
Habanero Pepper	*Capsicum chinense*	100,000-350,000 SHU
Ghost Pepper	*Capsicum chinense*	800,000-1,000,000 SHU
Pure capsaicin		15,000,000 SHU

Pungency	SHU
Non	0-700
Mildly	700-3,000
Moderately	3,000-25,000
Highly	25,000-70,000
Very Highly	<80,000

Note. Peppers are sorted from top to bottom by scoville intensity along with pure capsaicin as a reference. Along with the peppers sent out for sequencing is a pungency table included which ranks the pungency of the peppers based on their scoville heat units.

Chapter 2

Methods

Part 1: Genetically Modifying *Saccharomyces cerevisiae* with the Capsaicin Biosynthesis Pathway

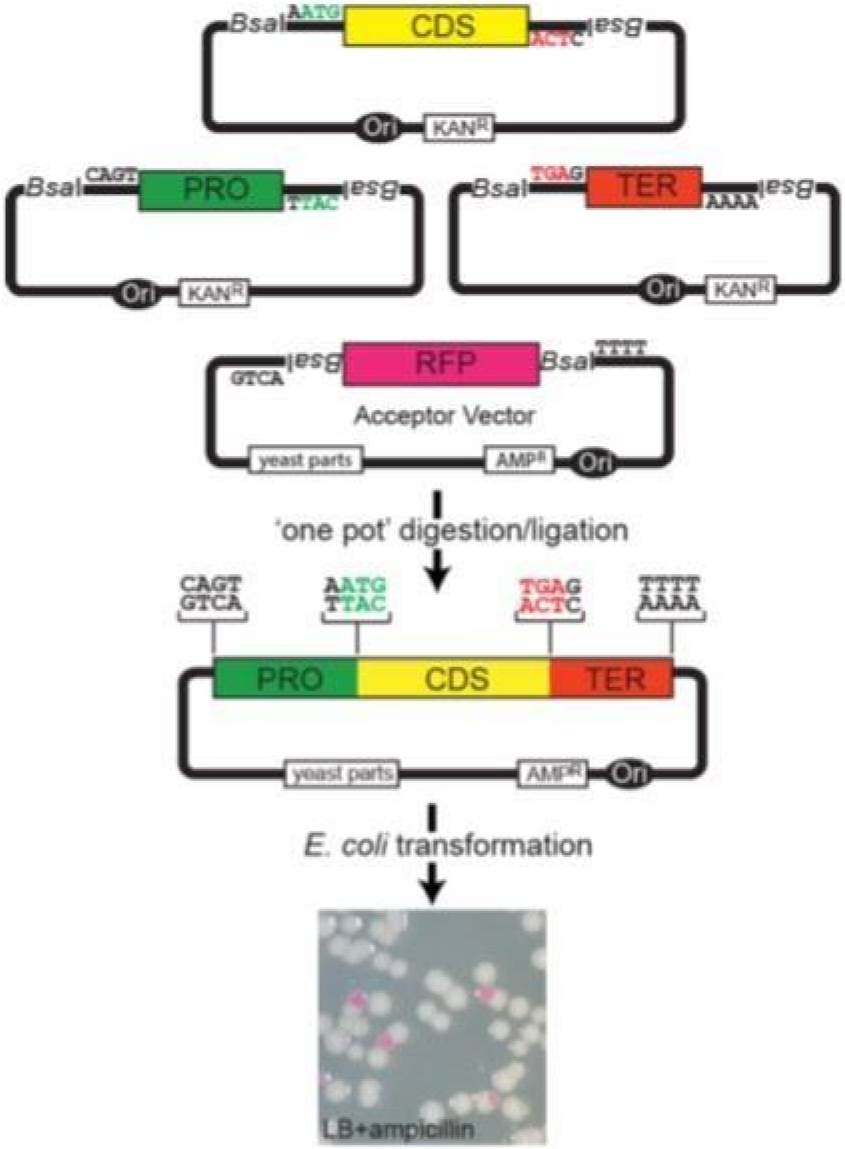
Golden Gate Cloning

The yeast golden gate DNA assembly method was designed by Agmon et. al as an easy way to incorporate genes of interest, promoters, and terminators in an efficient assembly system into *Saccharomyces cerevisiae* (Agmon et. al., 2015). Golden gate cloning utilizes type II restriction enzymes which will cleave outside of their target sequence leaving nucleotide overhangs. *BsaI*-HFV2 is the specific type II restriction enzyme that is utilized which cleaves the specific six nucleotide target sequence, 5' GGTCTC **|** $N_1N_2N_3N_4$ 3', located on the designed plasmids which then results in a small four base nucleotide overhang. The plasmids were created with the *BsaI*-HFV2 sites already located on either side of a red fluorescent protein region. We can create our transcriptional units by adding on the specific six base pair sequence that *BsaI*-HFV2 recognizes with a four base overhang that will remain after the sequence is cut. This allows us to design the correct flow of promoter, coding sequence, and terminator by pairing the overhanging fragments with its complementary DNA sequence of the next transcription unit to ensure the correct order of inserts. For example, we would take the sequence for the *pun1* gene which was one of the coding sequences we were interested in inserting. We would then add the *BsaI*-HFV2 sequence to the ends of the coding sequence and add a four base overhang sequence next to the restriction sites. This would result in when the transcriptional units are cut with the *BsaI*-HFV2 enzyme, a remaining four base overhang of our implementation would reside and we can strategically match these overhangs for each transcriptional unit to create the correct order of inserts into the plasmid itself.

To distinguish modified clones from unmodified clones, the red fluorescent protein region of the designed plasmids is the area in which the new transcriptional assembly will be inserted. For each plasmid, the *BsaI*-HFV2 restriction sites surround a red fluorescent protein so after the golden gate assembly process, colonies that grow white are ones that have been cut with the restriction enzyme and could contain the transcriptional units while red colonies are unmodified colonies. Using this technology, we were able to customize our plasmids to whatever promoter, gene, and terminator we were interested in inserting into bacteria. We chose two plasmids for our two genes, the pAV113 and pAV115 plasmids provided from Agmon et. al., 2015. The pAV113 plasmid would have the promoter PGK1, gene *pAMT*, and terminator CYC1 inserted into it. The pAV115 would have the promoter TEF1, gene *pun1*, and terminator ADH1. Both plasmids have a bacterial resistance of ampicillin and each plasmid contains a different gene for amino acid expression which will be used later for auxotrophic selection. Plasmids were combined with their respective promoters, terminators, and coding sequences in equimolar ratios, DNA ligase buffer, DNA ligase, water, and *BsaI*-HFV2 were all added and placed into a PCR machine to allow for amplification. The PCR products were then used in a bacterial transformation using chemically competent *E. coli* cells and then plated onto LB plates with ampicillin antibiotic resistance to grow overnight.

Figure 5

One-Pot Yeast Golden Gate Assembly



Note. PRO (Promoter), CDS (coding sequence), and TER (terminator) parts flanked by the appropriate prefix and suffix sequences are cloned into ampicillin resistant vectors. Cloned parts were mixed in equimolar ratio and the parental acceptor vector encodes a red fluorescent protein (RFP) gene with *E. coli* promoter and terminator sequences. Following *E. coli* transformation, white/red screening can be used to distinguish clones encoding putative transcriptional unit assemblies as compared to unmodified parental vectors. Figure obtained from Agmon et. al, 2015.

Bacterial Screening

After Golden Gate cloning, *E. coli* bacteria cells were screened utilizing restriction enzyme digestion, polymerase chain reaction, and sanger sequencing. White colonies were modified and could contain promoter, coding sequence, and terminator as these have been successfully cut with the *BsaI*-HFV2 restriction enzyme. Red colonies are colonies that have been unmodified and would not contain the correct transcriptional units which allow for a quick and easy visual screening. Using the red and white colonies, they can be digested with another restriction enzyme such as *PVUII*-HFV2, which recognizes two specific sites on a white colony and recognizes four specific sites on an unmodified red colony. Using the base pair sizes of the plasmid, promoter sequence, coding sequence, terminator sequence, and a deduction of the red fluorescent protein size, we are able to estimate the expected band sizes of colonies that would contain all the units of interest. By digesting our colonies with restriction enzymes, we can verify if our plasmid contained specific inserts based on the sizes of the remaining DNA bands after digestion. These DNA bands were evaluated on a 1% agarose gel utilizing gel electrophoresis. In addition to restriction enzyme digests, polymerase chain reaction was performed utilizing pBluescriptSK forward and pBluescriptKS reverse primers which recognize specific sequences that reside outside of the *BsaI*-HFV2 sites and amplify toward each other encompassing the red fluorescent protein region. Polymerase chain reaction can be utilized as a screening technique as the plasmid has specific sites located on it where a forward and reverse primer would amplify the DNA between these two sites creating a band size based on the distance. This distance can be determined and can span the region in which your transcriptional units may reside so it is possible to verify if the transcriptional units are inserted. Through PCR amplification utilizing the pBluescriptSK and pBluescriptKS primers, we can visualize the region where our transcriptional inserts will reside. To do this we will use gel electrophoresis and visualize this on a 1% agarose gel to see if the expected band sizes correlate to the base pair sizes of the sum of transcriptional inserts. The last step performed for successful screening of colonies was to send the isolated DNA samples of the bacterial plasmids out for sanger sequencing. Sanger sequencing is a method in determining the nucleotide sequence of DNA which utilizes fluorescently labeled dideoxynucleotides in a chain termination method which results in nucleotides being added to the DNA sequence which prevents other nucleotides from being added on. This chain termination method essentially allows for millions to billions copies of the DNA sequence of interest to be terminated at random lengths. The DNA sequence can be determined by sorting these fragments by size and using fluorescence to determine which base, A, T, C, or G has attached itself which when read out, creates the DNA sequence (Sanger et. al., 1977). The results obtained could then be used to verify if our inserts are present by comparing the sanger sequence results to our insert's sequences.

Auxotrophic Selection

Once we had confirmed the colonies had the proper pieces, we performed a yeast transformation with those samples. The vectors that we chose from the beginning, pAV113 and pAV115 have specific yeast markers they lack such as pAV113 lacks Histidine-3 and pAV115 lacks Leucine-2. This type of auxotrophic selection allows us to grow these samples on a plate that would lack one of these necessary organic compounds needed for growth, and the colony will only grow if the yeast did incorporate the plasmid that carries the gene needed for expression of the amino acid. The yeast strain we were transforming them into is a strain called By4741 which needs the following macromolecules to grow which are His3, Leu2, Met2, and Ura3. Using this knowledge, we made three different types of synthetic dextrose plates that lacked all amino acids and nucleobases except for ones we added. For the first set, we added His3, Met2, and Ura3. This plate lacks Leu2 and therefore only yeast that have the pAV115 plasmid incorporated in it will grow. The second set of plates had Leu2, Met2, and Ura3 so only pAV115 should grow on it since it lacks His3. The last set of plates only had Met2 and Ura3 meaning that if the yeast incorporated both plasmids then colonies should form. We performed the yeast transformation utilizing the correctly screened bacterial plasmid DNA to obtain yeast with our desired genes of interest.

Testing Gene Expression in Genetically Modified Yeast

Once the genes of interest were inserted into the yeast, we then wanted to evaluate gene expression levels of our *pAMT* and *pun1* genes to make sure they are being expressed. To do this, we used quantitative Reverse-Transcription Polymerase Chain Reaction or qRT-PCR for short. We extracted the RNA of our four different strains of yeast which were By4741, By4741 with *pAMT*, By4741 with *pun1*, and By4741 with *pAMT* and *pun1*. The extracted RNA from the yeast samples was used to synthesize cDNA and both were used for the polymerase chain reaction. The RNA samples were used as a control to determine if there was DNA contamination and if a DNase step was needed prior to creating the cDNA. The reason is contamination of DNA in the RNA sample would lead to inaccurate cDNA sequences. We utilized DNA primers for our *pun1* and *pAMT* genes which would recognize a specific DNA sequence from their coding region that would result in amplification of the sequence if present. We used water as a control for each of the primer sets. We also used the *Alg9* gene, which is a housekeeping gene in *Saccharomyces cerevisiae* which is responsible for the synthesis of oligosaccharide precursors for N-linked protein glycosylation (Frank & Aebi, 2005). Housekeeping genes are constituently expressed which allows it to serve as a positive control to compare our gene expression values against. Using SYBR green dye, we could visualize DNA synthesis occurring in the samples as DNA synthesis causes SYBR green dye to fluoresce brightly as it binds to the minor groove of DNA so the more DNA that is synthesized the brighter the fluorescence becomes (Noble & Fuhrman, 1998). Cycle threshold (CT) is a value is a measure of the number of cycles required for the fluorescent signal to cross a threshold or background level. The lower the CT value means that the threshold was crossed early relating to a strong positive fluorescence signal occurring. The SYBR Green dye was analyzed using the FAM fluorophore which has an excitation of 490 nm and emission of 520 nm. The polymerase chain reactions had a denaturation temperature of 95^{o}C, annealing temperature of 55^{o}C, and an elongation temperature of 72^{o}C, which repeated for 40 cycles.

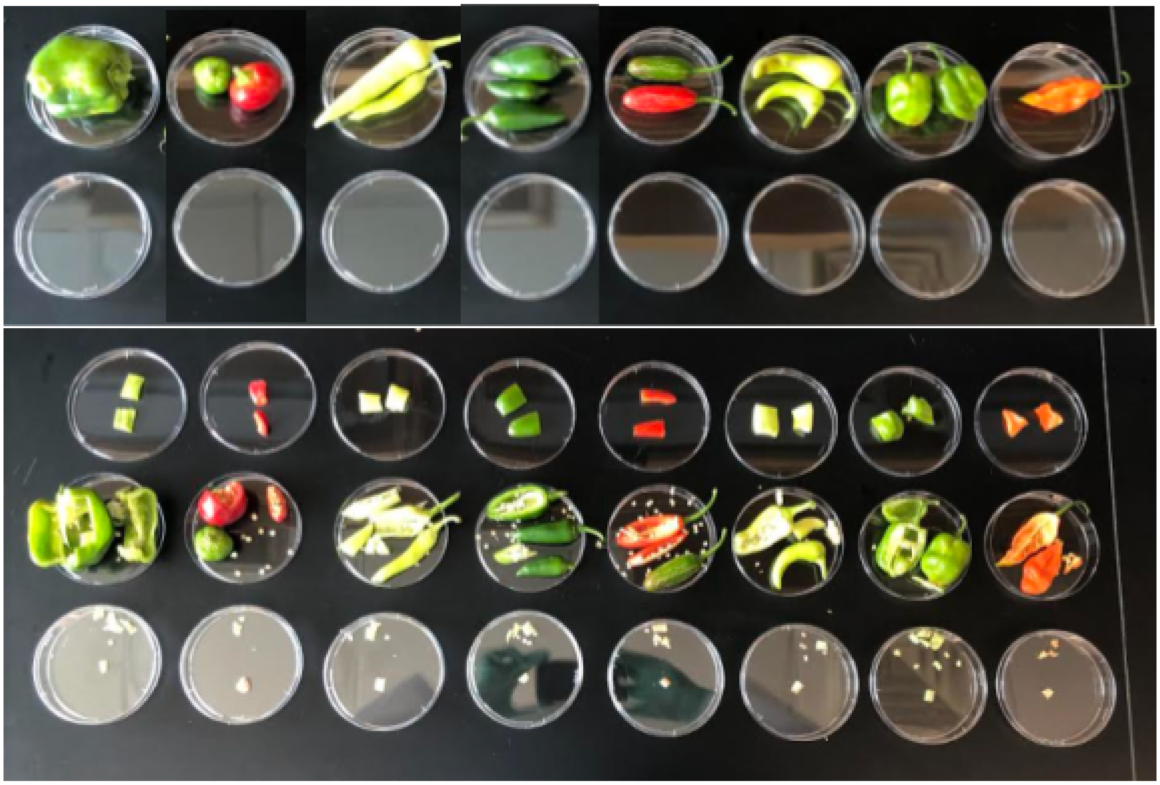
Part 2: Performing Illumina RNA Sequencing Analysis on Placenta and Skin Tissue Samples from Seven Different Peppers of Varying Scoville Intensity to Identify Novel Genes for Capsaicin Synthesis

Collection of Peppers for RNA Sequencing

A variety of peppers species were collected from the surrounding area and dissected for RNA processing. We collected the following pepper samples bell pepper, jalapeno pepper, cayenne pepper, hungarian wax pepper, serrano pepper, cherry pepper, habanero pepper, carolina reaper pepper, and ghost pepper. We attempted to obtain the highest RNA quality of pepper samples by immediately processing them after they were harvested to prevent RNA degradation. When pepper samples were picked from the plant, they were immediately placed in an ice cooler where they were brought back to Rowan University. Sterile dissecting tools such as exacto knives, razor blades, and tweezers were used to dissect the pepper and harvest three distinct portions, the placental region, the skin, and the seeds which can be seen in figure 6 below. Each of these pieces were placed into a 1.5mL tube and stored in the -80°C freezer until RNA extraction of placenta and skin regions were performed. The peppers were collected from a variety of places and consistency in this extraction process was done to maintain evenness among samples. Carolina reaper peppers as well as habanero peppers were collected and harvested from Rowan University in Glassboro, New Jersey. Ghost peppers, jalapeno peppers, and cayenne peppers were cultivated and grown by Ryan Calhoun in Turnersville, New Jersey. Hungarian wax peppers, serrano peppers, cherry peppers, and bell peppers were acquired by Dr. Benjamin Carone from Visalli's Farm Market located in Mullica Hill, New Jersey. Multiples of the same pepper were collected from each site allowing us to collect multiple pepper extractions for a broader range of results. This allowed us to obtain multiple skin and placenta sections from each pepper so we could perform multiple experiments in the future if needed.

Figure 6

Six of the Nine Pepper Samples Dissected for RNA Extraction



Note. The peppers are listed starting from left to right as the following: bell pepper, serrano pepper, hungarian wax pepper, cherry pepper, serrano pepper, hungarian wax pepper, habanero pepper, and ghost pepper. Not pictured are the jalapeno, carolina reaper, and cayenne peppers. The top picture illustrates the peppers prior to dissection. The bottom picture illustrates the peppers after extraction top to bottom where the top petri dish houses the skin extracts, the middle petri dish houses the seeds and remaining pepper, while the bottom petri dish houses the placenta regions. These regions of the dissected peppers were labeled and placed into 1.5mL tubes and stored in the -80°C until ready for RNA extraction.

Extraction of RNA from Peppers

RNA samples were processed by first taking 1 cm by 1 cm incisions of either the placenta or skin regions and mixing them with ribozol. Early samples were processed utilizing a bead beater where the sample would undergo 10 minutes of bead beating followed by the Ribozol RNA Extraction Reagent protocol for extracting RNA from plant tissue provided by VWR Life Science. Although bead beating was effective, a VWR micro homogenizer was purchased to provide better cell disruption as the plant cell wall proved to be a difficult material to homogenize. The VWR standard Micro-Homogenizer allowed for processing of small sample sizes in small microcentrifuge tubes. After homogenization, these samples were also followed by the same RNA extraction protocol as stated above. To assess RNA integrity, the samples were quantitatively analyzed using a Quibit Fluorometric Quantitation and were analyzed using gel electrophoresis on a 1.5% agarose gel to visualize integrity of the bands. As previously stated, we obtained multiple samples of skin and placenta regions for each pepper so we could perform multiple RNA extractions on the samples if needed. In total, we performed a total of 48 RNA extractions on different skin and placenta regions of the peppers and assessed their RNA quality and concentrations.

FastQC

Using the results from the RNA extraction process, we selected the best pepper samples that had high RNA integrity and were highly concentrated to send out for sequencing. We also attempted to include samples that would result in a wide range of the Scoville scale meaning we did not want to send out multiple non-pungent peppers or pungent peppers, we wanted a wide distribution from non-pungent to highly pungent. We decided to send out the following peppers for paired-end Illumina RNA Sequencing: cherry peppers, jalapeno peppers, hungarian wax peppers, serrano peppers, cayenne peppers, habanero peppers, and ghost peppers. Pepper samples were placed in a container, covered in dry ice, then sealed in a styrofoam box which was then sent out to GENEWIZ to perform standard paired-end RNA-Seq analysis. Genewiz processed the samples and evaluated many factors of the samples such as RNA integrity, average nucleotide size, region molarity, RNA concentration, and more before performing sequencing on the Illumina platform. After sequencing, the bioinformatics pipeline will be utilized to analyze the sequencing results to generate gene expression values. To do this, we are first given the FASTQ files and we utilize FASTQC to determine the quality of our data. Using the fastq files, we received a variety of information on our data such as the per base sequence quality, base N content (could not identify nucleotides), if there are any overrepresented sequences, adapter content, and much more. This step is necessary as it allows the user to determine the quality of their data and if the raw files need to be processed prior to continuing through the pipeline (Andrews, 2017). Alterations to the fastq files can be done at this step such as trimming the overall reads if there is a drop in quality, checking for adapter contamination, if there are overrepresented sequences, and many other variations to the raw files to generate the best sequence reads for further analysis.

Bowtie

After utilizing FASTQC, we used Bowtie to align our fastq files to a reference sequence. Bowtie is an ultrafast short read aligner which is aimed to quickly align large sets of short DNA sequences to a genome. This produces a SAM output file which shows the overall alignment score obtained between the sample sequence and the reference sequence. It also shows how many reads in total there are in the sample sequence, how many of those are aligned 0 times, aligned 1 time, and aligned more than 1 time (Langmead & Salzburg). For our experiment, the reference genome that we will be using will be the *Capsicum annuum* reference genome. The reference genome was obtained from Sol Genomics Network and it is *Capsicum annuum* cv CM334 Genome CDS (release 1.55) (Qin et. al., 2014).

Cufflinks

After the bowtie alignment, we then can take the Sam file and convert it to a Bam file utilizing samtools. The Bam file was also sorted in order of chromosomes to match how the reference genome is set up. After sorting, cufflinks can be used to evaluate gene expression on our pepper samples. Cufflinks is a program that assembles transcripts, estimates their abundances, and can test for differential expression and regulation in RNA-seq samples (Ghosh & Chan, 2016). It produces several output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes. It also tracks changes in the relative abundance of transcripts sharing a common transcription start site, and in the relative abundances of the primary transcripts of each gene. Tracking the former allows one to see changes in splicing, and the latter lets one see changes in relative promoter use within a gene (Trapnell et. al., 2012).

Analyzing Cufflinks Output

After running cufflinks, we analyzed the results utilizing the FPKM output. FPKM stands for fragments per kilobase of transcript per million mapped reads. The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. Utilizing these FPKM output values and the free software environment R, we processed the data creating various graphical plots and images to interrogate the data (Team, 2013). The most important visual was the generation of a heatmap to visualize the differences between gene expression values for the pepper samples to make connections and comparisons between expression levels and genes.

Chapter 3

Results

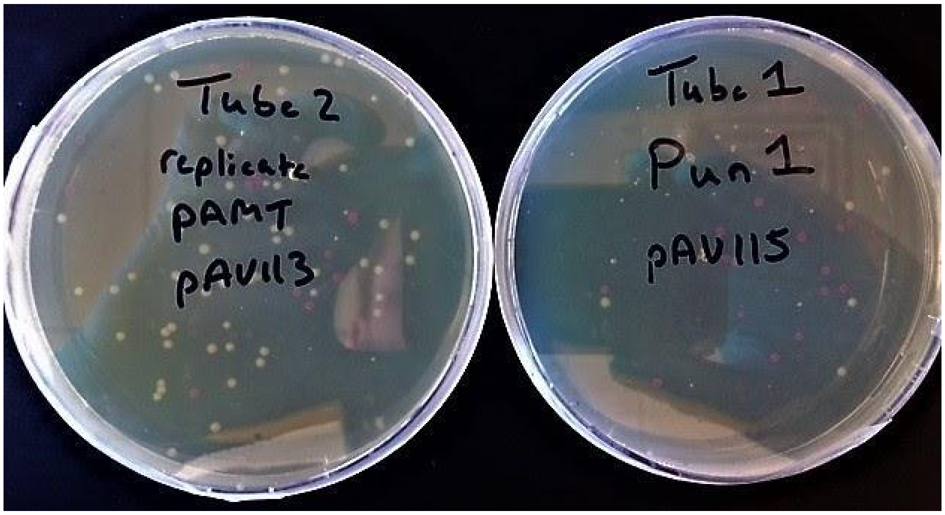
Part 1: Genetically Modifying *Saccharomyces cerevisiae* with the Capsaicin Biosynthesis Pathway

Golden Gate Cloning

Utilizing the yeast Golden Gate DNA assembly method as described previously, we incorporated the PGK1 promoter, *pAMT* gene, and CYC1 terminator into the pAV113 plasmid and incorporated the TEF1 promoter, *pun1* gene, and ADH1 terminator into the pAV115 plasmid provided from Agmon et. al., 2015. This concoction, after PCR amplification, was used in a bacterial transformation into chemically competent *E. coli* cells which were grown overnight at 37°C. The resulting colonies that were grown were a mixture of single red bacteria colonies and single white bacteria colonies which can be seen in figure 7 below. Red colonies are bacteria cells that still have the original red fluorescent protein portion remaining in their plasmid. White colonies are bacteria cells that have had their red fluorescent protein portions successfully cut out by the type II restriction enzyme digest and could contain the correct transcriptional inserts from the golden gate cloning. The red colonies however will not have the correct transcriptional inserts as these plasmids are unmodified since they were not cut by the *BsaI*-HFV2 restriction enzyme. Using both the red and white colonies, we can perform plasmid DNA extractions and perform various screening assays to verify if they do or do not contain our transcriptional units of interest. As seen in figure 7, we were able to grow both red and white colonies. It appears that there were more white colonies that grew in the pAV113 plasmid compared to the pAV115 plasmid which grew more red colonies. There also appears to be more colonies overall in general on the pAV113 plate. Being able to grow both red and white colonies is beneficial since we can test in the screening process an unmodified vector against a modified vector to see more substantial differences.

Figure 7

E. coli Colonies after Bacterial Transformation with Golden Gate Assembly



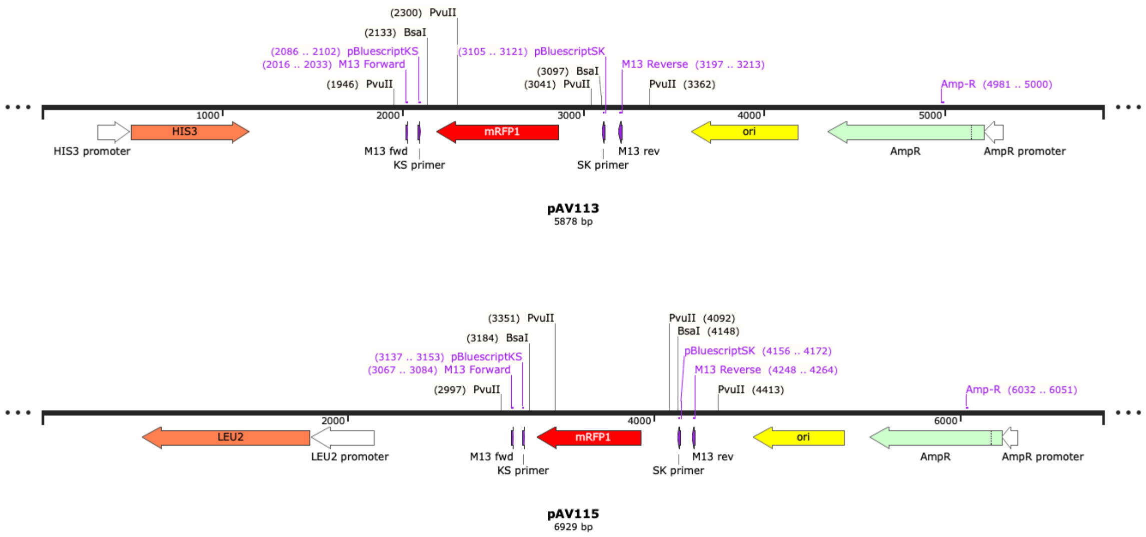
Note. The red colonies are unmodified parental vectors whose red fluorescent protein coding sequence region remained unchanged in the vector. The white colonies are bacterial colonies that have been modified in the golden gate transformation process and may contain the inserted CDS region.

Bacterial Screening

After Golden Gate cloning, *E. coli* bacteria cells were screened utilizing restriction enzyme digestion, polymerase chain reaction, and sanger sequencing. Included below in figure 8 are the linear plasmid maps of pAV113 and pAV115. Both of these maps include the important features such as restriction enzyme site information, red fluorescent protein regions, or primer site information that we exploited in the screening process. One form of screening we utilized was taking our purified plasmid DNA from red and white colonies performing a restriction enzyme digest with the *PVUII*-HFV2 enzyme. This specific restriction enzyme is useful as it recognizes a different number of sites on a plasmid cut with *BsaI*-HFV2, two, compared to an unmodified sample, four, as seen in figure 8's plasmid maps. Using this information we would ideally be looking for a different number of bands to show up on the agarose gel for a white colony compared to a red colony. In addition, since we are replacing the red fluorescent region with our transcriptional units of interest, and we know the base pair lengths of these promoters, coding sequences, and terminators, we can estimate the band size lengths that are expected to show up if the plasmid does successfully contain each of these units. To calculate the estimated band size, we would take the total size of our plasmid and first subtract the region we cut out with the *BsaI*-HFV2 enzyme as this region is no longer present in the plasmid. Since there are two *PVUII*-HFV2 sites left in the plasmid, one band size will be large consisting of the distance from one site to the other and a simple calculation of this distance only needs to be determined. To determine the other band size, a little more math will be required. This second band size will be the remaining region between the *PVUII*-HFV2 sites and *BsaI*-HFV2 sites with the addition of the total size of our transcriptional units we are inserting. After performing this calculation, we determined that the pAV115 plasmid would have two expected band sizes of 2732 bp and 5965 bp while the pAV113 plasmid would have 3228 and 4462 expected band sizes. Based on these expected band lengths, we ran multiple restriction enzyme digests with the *PVUII*-HFV2 enzyme and visualized the results on agarose gels at a 1% concentration. After multiple gel electrophoresis experiments, we did obtain samples that had demonstrated two unique bands that did resemble the expected band sizes of interest for both the pAV113 and pAV115 plasmid. Typically for the samples that did show distinct banding, the top which consisted of the larger of the two DNA segments was a thicker and fuller band while the lower band representing the smaller fragment was fainter and less concentrated. It did appear that the samples we had matched our expected results but to be sure, we also conducted polymerase chain reactions on our samples and performed sanger sequencing as another form of validation.

Figure 8

Linear Plasmid Maps for Plasmids pAV113 and pAV115

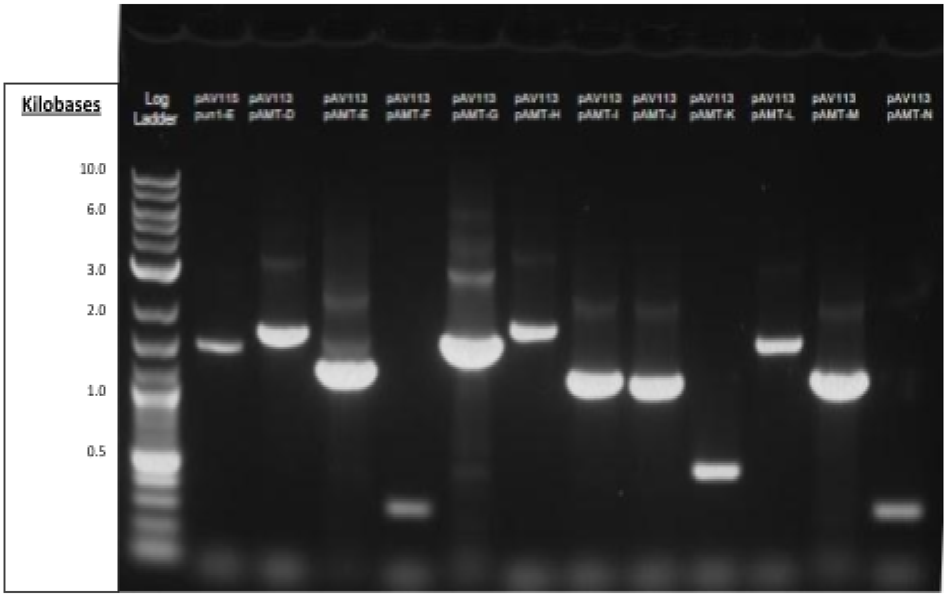


Note. These plasmid maps highlight the key features that were manipulated during the screening process such as the polymerase chain reaction sites and the restriction enzyme digest sites as well as the red fluorescent protein region of interest.

Polymerase chain reaction (PCR) was performed utilizing pBluescriptSK and pBluescriptKS forward and reverse primers which are specific sequences that reside outside of the *BsaI*-HFV2 sites. These primers amplify toward each other encompassing the red fluorescent protein region. Through PCR amplification utilizing these SK and KS primers, we will have an expected band length size estimation of our transcriptional unit inserts. These polymerase chain reaction and restriction enzyme digest results were visualized on a 1% agarose gel to see if the expected band sizes correlate to the base pair sizes of the sum of transcriptional inserts. The results from the SK and KS polymerase chain reaction band sizes can be seen in figure 9 below. Each plasmid had an expected band size based on the size of DNA that the PCR would amplify if the transcriptional units were present. To calculate this expected band length, the region size that will be amplified between the primers is first determined which we will call the primer region. Then, the region between the two *BsaI*-HFV2 sites is determined which will be the restriction enzyme digest region. This is the region that has been cut out and replaced with our transcriptional units. We then will determine the length of our transcription units which is the sequence lengths of the promoter, terminator, and coding sequence which will be our transcriptional unit region. This allows us to determine the final band size we expect to see as it will be the size of our transcriptional insert added on to the remaining bp size of the primer region after subtracting the primer region by the restriction enzyme digest region. It was expected that the pAV115 plasmid would have a band size around 2500 and the pAV113 plasmid would have a band size around 2100. The PCR band sizes that we visualized on the agarose gel vary widely from sample to sample. We had our expected band sizes for each plasmid but it appeared that the most prominent band sizes for our samples typically did not line up with our expected band size reference on the DNA ladder. Only the samples pAV115 *pun1*-E and pAV113 *pAMT*-L appeared to reside around the expected band length sizes of our modified plasmids. Many samples were located a bit too far out of this expected region and these two samples appear to be the closest to the 2500 band size for pAV115 and 2100 bp size for pAV113. There were other samples that potentially were located near the expected band lengths, but none were located as close to the expected size as samples pAV115 *pun1*-E and pAV113 *pAMT*-L.

Figure 9

Gel Electrophoresis Results for Bacterial Samples that Underwent Polymerase Chain Reaction with SK and KS Primers



Expected band length size = Transcriptional unit region + (Primer Region - Restriction Enzyme Digest Region)
expected band size for pAV113 = PGK1+CYC1+*pAMT* + (primer region - RED region) expected band size for pAV113 = 983 + 190 + 1380 + (1035 - 964) = **2553**
expected band size for pAV115 = ADH1+TEF1+*pun1* + (primer region - RED region) expected band size for pAV115 = 292 + 422 + 1323 + (1035 - 964) = **2108**

Note. Agarose gel was a 1% gel, 4 uL of 2 Log Ladder was loaded, 5 uL of sample added with 3 uL of Ficoll orange loading dye. Gel electrophoresis was performed at 90 volts for 45 minutes. Expected band size for the pAV113 vector with correct transcriptional inserts is 2553 bp and 2108 bp for the pAV115 vector. The expected band length calculations can be seen below the figure.

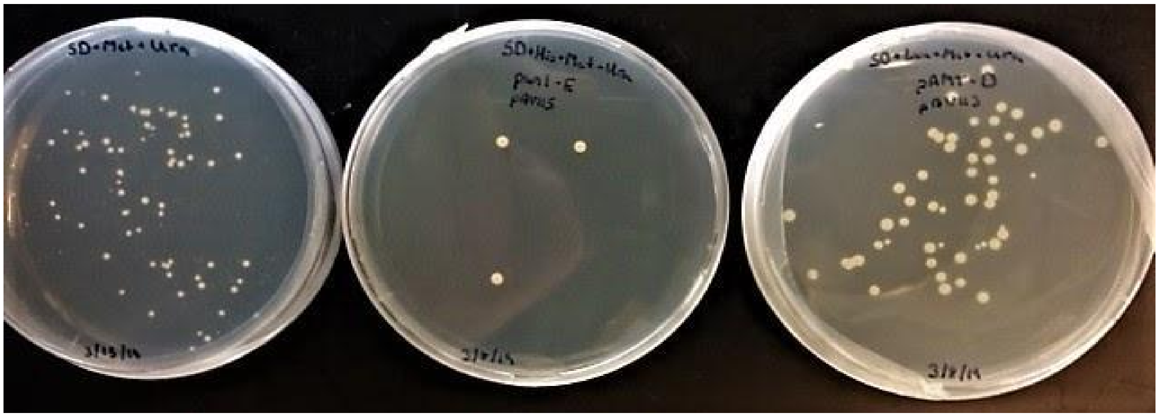
The final screening experiment we performed was to send out our purified plasmid DNA sequences out for sanger sequencing analysis to confirm that our transcriptional units were present in the plasmid. The DNA sequencing results for sanger sequencing can be useful as we can determine if the sequences of our transcriptional units match those found in the sequencing of our samples. We sent out a total of five samples for sanger sequencing, one sample was a possible *pun1* candidate, and the other four were possible *pAMT* candidates. We were fairly confident through previous experimental tests that we did indeed have a positive match for our *pun1*-pAV115 plasmid and therefore only sent that specific one out for sequencing. Compared to the *pun1*-pAV115 plasmid, we were unsure of our *pAMT*-pAV113 samples and sent out multiple to be sequenced as a precautionary measure. Using the blast program, we were able to blast our sanger sequence results against the *Capsicum annuum* genome to determine if our *pun1* and *pAMT* genes of interest were present based on the similarity result. We also directly compared the sequences of our sanger sequence results to the known promoter sequences, coding sequences, and terminator sequences we were using to determine if they were present or not. We were able to identify that four of the five sequences had promising sanger sequencing results which demonstrated to include part of our coding sequence and terminator sequence of interest. We were not able to determine if the promoter region was present in the samples because it was outside the range of sanger sequencing capabilities as sanger sequencing is able to sequence about 800 base pairs from its start site. Alternative primers would have to be designed to further explore if the entire coding region was present and promoter region.

Auxotrophic Selection

After we identified which plasmid sequences contained the proper transcriptional units, we would then be able to perform a yeast transformation to insert our plasmid DNA into a strain of *Saccharomyces cerevisiae.* The strain of yeast we will be using is By4741 which is a yeast strain with the following genotype: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0. This yeast strain requires the following nutrients for growth which are histidine, leucine, methionine, and uracil. The plasmids that we are inserting into this yeast strain both code for a specific amino acid where pAV113 has a histidine promoter region and pAV115 has a leucine promoter region. Using this information, we can specifically select which yeast are able to grow on specialized plates that lack the required nutrients. These plates are made to contain yeast nitrogen without amino acids, water, and sugar. Then the specific amino acids and nucleobases were added to create a variety of unique plates. The first one contained just methionine and uracil, the second contained histidine, methionine, and uracil, the third with leucine, methionine and uracil, and the fourth contained no added amino acids or uracil. These series of plates will allow yeast to grow if they have both pAV113 and pAV115 plasmids for the first plate, only pAV115 to grow on the second plate, only pAV113 to grow on the third plate, and nothing to grow on the fourth plate. Once the plates were created, we performed a yeast transformation using the plasmid DNA that we believed to contain the correct transcriptional inserts and plated them onto our SD plates which can be seen in figure 10 below. The plates were grown for one day in a 32ºC incubator. The plate located on the far left consists of SD media, methionine and uracil. The plate located in the middle consists of SD media, histidine, methionine, and uracil. The plate on the far right consists of SD media, leucine, methionine, and uracil. Not pictured is the plate containing just SD media and no other added nutrients. We have growth on all three of the plates with added nutrients suggesting that each of the yeast cells grown on the plate contains the correct plasmid or plasmids. There was no growth on the plate lacking any type of added chemicals. There appears to be many colonies that grew on both the SD+Met+Ura and SD+Leu+Met+Ura plates. Although the SD+His+Met+Ura plate grew far fewer colonies, it still grew colonies nonetheless. The colonies also on the combination plasmid plate appear to have grown smaller in size compared to the other two plates. All together, the colonies growing on these plates are expected to have the correct plasmids inserted into the yeast which is the reason they were able to grow.

Figure 10

Saccharomyces cerevisiae Strain By4741 Colonies Grown on Synthetic Dextrose



Note. The leftmost plate contains methionine and uracil. The middle plate contains histidine, methionine, and uracil. The right most plate contains leucine, methionine, and uracil. Yeast colonies were grown for 24 hours in a 32ºC incubator. Plasmids pAV113 and pAV115 were used in the yeast transformation.

Testing Gene Expression in Genetically Modified Yeast

Following auxotrophic selection, we then performed a follow up analysis experiment to determine if our genetically modified yeast was expressing their newly added genes. If these genes were not being expressed then capsaicin would not be synthesized according to our hypothesis. The samples that underwent qRT-PCR were the genetically modified yeast samples and their RNA was extracted, isolated, and used to create their cDNA strands. The genes we were analyzing were *pun1*, *pAMT*, and the *Alg9* housekeeping gene as a control. Each of the yeast were combined with their respective primers to visualize if fluorescence occurred after 40 cycles of polymerase chain reaction. Using the fluorescence readings from the polymerase chain reaction machine, we can see how many cycles it took for the sample to start fluorescing if there was any fluorescence at all. The CT results for our different cDNA and RNA yeast samples can be seen below in table 3. The results demonstrate that the yeast samples that contained the *pAMT* gene had low CT values (CT score < 20) for both the RNA and cDNA sets when tested with the pAMT primer. On average, the CT scores of the *pAMT* samples when treated with the pAMT primer were in the 16-18 CT range. Samples that did not contain the *pAMT* gene, but were treated with the pAMT primer had a CT score in the 32-33 CT range with the highest at 38 which was water. Yeast samples that contained the *pun1* gene and the pun1 primer however did not result in low CT values. These samples had a CT score typically in the range of 36-37. The water standard had a CT score of 38 and there were two samples resulting in 0 CT scores which means that no fluorescence was ever recorded for them so they are essentially beyond the 40 cycle mark. Lastly, the *Alg9* samples when combined with the Alg9 primer, all resulted in significant CT values in the range of 17-19 CT for both the RNA and cDNA samples besides the water sample with a CT score of 34.

Table 3

qRT-PCR Count Threshold Results from By4741 Yeast Samples

Primers:	*FAM CT:*	*Samples:*								
pun1 primer		By4741 cDNA	By4741 pun1 cDNA	By4741 pAMT cDNA	By4741 pun1+ pAMT cDNA	By4741 RNA	By4741 pun1 RNA	By4741 pAMT RNA	By4741 pun1+ pAMT RNA	H20
	CT Score	**38.05**	**37.91**	**37.51**	**36.36**	**0**	**0**	**37.44**	**36.48**	**38.26**
pAMT primer		By4741 cDNA	By4741 pun1 cDNA	By4741 pAMT cDNA	By4741 pun1+ pAMT cDNA	By4741 RNA	By4741 pun1 RNA	By4741 pAMT RNA	By4741 pun1+ pAMT RNA	H20
	CT Score	**33.95**	**32.73**	**16.04**	**16.13**	**32.65**	**33.95**	**17.34**	**18.83**	**38.39**
Alg9 primer		By4741 cDNA	By4741 pun1 cDNA	By4741 pAMT cDNA	By4741 pun1+ pAMT cDNA	By4741 RNA	By4741 pun1 RNA	By4741 pAMT RNA	By4741 pun1+ pAMT RNA	H20
	CT Score	**19.17**	**19.32**	**19.03**	**19.22**	**18.00**	**18.14**	**17.83**	**18.68**	**34.23**

Note. There are four different types of samples which are By4741, By4741+pun1, By4741+pAMT, and By4741+pun1+pAMT. Each of these four samples had their cDNA and RNA variations tested against three different sets of primers which were pun1, pAMT, and Alg9. Count threshold values were recorded and samples with a strong positive fluorescence were highlighted in green (CT <=20).

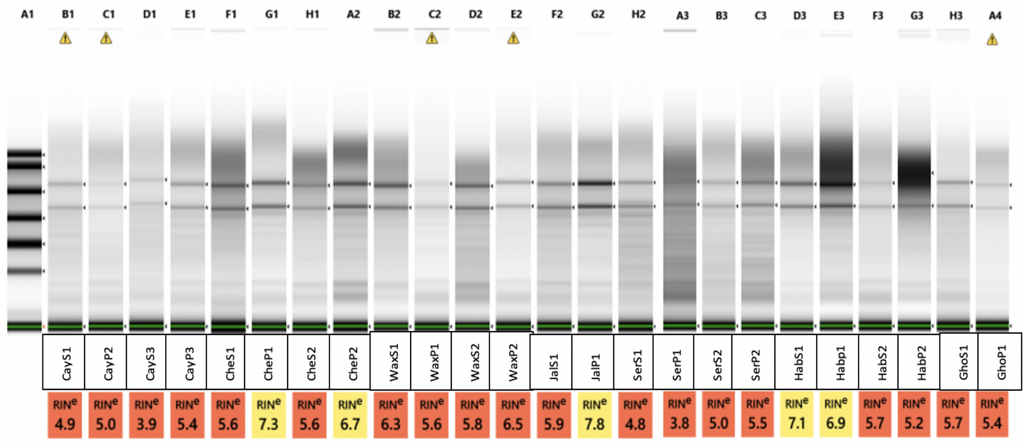
Part 2: Performing Illumina RNA Sequencing Analysis on Placenta and Skin Tissue Samples from Seven Different Peppers of Varying Scoville Intensity to Identify Novel Genes for Capsaicin Synthesis

Collection of Peppers for RNA Sequencing

After collecting a large variety of pepper, we had many samples we could send out for RNA sequencing. We selected the pepper samples that appeared to have the highest RNA integrity from our agarose gel electrophoresis experiments, samples that had high RNA concentrations, and samples that would result in a generous range and reflection of the Scoville scale. As previously mentioned, the following pepper samples were sent out for paired-end Illumina RNA sequencing: cherry peppers, jalapeno peppers, hungarian wax peppers, serrano peppers, cayenne peppers, habanero peppers, and ghost peppers. GENEWIZ performed standard RNA-Seq analysis and processed each of the pepper samples, evaluating RNA integrity, average nucleotide size, region molarity, RNA concentration, and then performed sequencing utilizing the Illumina platform. RIN^{e} is an abbreviation for RNA integrity number equivalent which is an algorithm that is calculated to assign an integrity score to an RNA sample. The score is calculated on a scale from 1-10 with 10 being the least degraded or having the highest integrity. The RNA Agilent TapeStation system produces the electropherogram which is used to calculate the RIN^{e} score for each sample and agarose gel electrophoresis images were also created from Genewiz which can be seen in figure 11 below. In addition to the RIN^{e} scores, Genewiz generated a table of information regarding our pepper samples consisting of concentration, average size, region molarity, and more which can be seen in table 4 below. RIN^{e} scores that are lower than 6 are highlighted as cautionary samples according to a threshold determined by GeneWiz. Most of the pepper samples appear to have RIN^{e} scores below 6.0, but many of them are close to the threshold of 6 besides a select few that fall below 5.0 such as CayS1, CayS3, SerS1, and SerP1. The average RIN^{e} score was 5.73 with a maximum score of 7.8 from sample Jalapeno Placenta 1 and a minimum score of 3.8 from sample Serrano Placenta 1. When visualizing the RNA bands for the pepper samples, for most of the samples we see two distinct bands at the 28S and 18S locations on the gel with a few samples having either faint or blurry bands. DV 200 is a measurement that represents the percentage of RNA fragments that are larger than 200 nucleotides in size. GeneWiz marks samples with a DV 200 percentage that falls below 70 for the samples and of the 24 samples, four samples had a DV 200 score that fell below this threshold, which was CayP2, WaxP1, WaxP2, and SerS2. Detailed values for these scores such as the RIN^{e} and DV 200 can be found in table 4 below along with other outputs such as RNA concentration values and average size of nucleotides. Most samples obtained yielded nucleotide sizes in the ranges of 4000-5000 nucleotides long and had an overall average nucleotide size of 5025. The sample with the longest nucleotide length was Cherry Placenta 2 with 5625 nt and the sample with the shortest nucleotide length was Hungarian Wax Skin 2 with 3913. When analyzing the nucleic acid concentrations for the samples, the average concentration for the pepper samples was 80.82 ng/uL. A few samples had a relatively low RNA concentration which were Habanero Skin 2 at 22.4 ng/uL, Cayenne Skin 3 at 18.96 ng/uL, Hungarian Wax Placenta 1 at 23.2, Ghost Skin 1 at 34 ng/uL, and Cayenne Placenta 3 at 27.6 ng/uL. The DV 200 score has an average value of 75.67 with a maximum of 90.32 for sample Habanero Placenta 2 and a minimum value of 61.92 for sample Hungarian Wax Placenta 1.

Figure 11

Gel Electrophoresis Results for Sequenced Peppers Provided from GeneWiz Agilent TapeStation Analysis Technologies



Note. Twenty-four of our pepper samples were analyzed and had their RIN^{e} scores calculated along with a visual representation of the RNA bands. Samples colored yellow are samples with a RIN^{e} score above 6.5 while samples colored orange are those with RIN^{e} scores below 6.5.

Table 4

Summary of TapeStation Results

Sample	Sample ID	Nucleic Acid Conc. (ng/uL)	RIN^e Score	DV 200	Average size (nt)
Cayenne Skin 1	CayS1	51.20	4.9	73.17	4883
Cayenne Placenta 2	CayP2	47.60	5.0	69.82	5591
Cayenne Skin 3	CayS3	18.96	3.9	70.92	5237
Cayenne Placenta 3	CayP3	27.60	5.4	74.55	5557
Cherry Skin 1	CheS1	62.00	5.6	82.20	5023
Cherry Placenta 1	CheP1	55.60	7.3	70.84	5094
Cherry Skin 2	CheS2	58.80	5.6	84.75	4622
Cherry Placenta 2	CheP2	107.60	6.7	78.79	5625
Hungarian Wax Skin 1	WaxS1	172.40	6.3	83.12	4965
Hungarian Wax Placenta 1	WaxP1	23.20	5.6	61.92	4323
Hungarian Wax Skin 2	WaxS2	180.40	5.8	76.51	3913
Hungarian Wax Placenta 2	WaxP2	48.00	6.5	66.79	5084
Jalapeno Skin 1	JalS1	96.40	5.9	74.42	5091
Jalapeno Placenta 1	JalP1	87.60	7.8	80.62	5002
Serrano Skin 1	SerS1	78.00	4.8	72.37	5244
Serrano Placenta 1	SerP1	139.60	3.8	70.66	4701
Serrano Skin 2	SerS2	111.20	5.0	69.28	5486
Serrano Placenta 2	SerP2	200.00	5.5	74.54	5235
Habanero Skin 1	HabS1	66.40	7.1	81.39	5300
Habanero Placenta 1	HabP1	108.00	6.9	83.47	5335
Habanero Skin 2	HabS2	22.40	5.7	70.68	5449
Habanero Placenta 2	HabP2	82.80	5.2	90.32	4477
Ghost Skin 1	GhoS1	34.00	5.7	74.03	4343
Ghost Placenta 1	GhoP1	60.00	5.4	81.01	5042

Note. Twenty-four pepper samples were processed and given unique RIN^e scores, calculated RNA concentrations, average sizes, and calculated DV 200 scores. Boxes that were highlighted yellow indicate warnings for samples as they either demonstrated a cautionary RIN^e score or DV 200 value according to GENEWIZ criteria.

FastQC

The first step when analyzing RNA sequencing data is to perform a quick qualitative control test on the raw sequence data. FastQC outputs a list of quality control checks and gives the user a quick impression of the overall status of the sample. We had a total of 24 samples we sent out for RNA sequencing and because we did paired-end sequencing, which is sequencing from both ends of a DNA fragment, we ended up with two output files from one sample. This means we had a total of 48 sample files output from RNA sequencing. When running FastQC, there are eleven outputs that will return as a green check mark, yellow exclamation point, or a red x mark, each indicating the quality of that check going from good to bad respectively. These eleven outputs are as follows: basic statistics, per base sequence quality, per tile sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base N content (could not identify nucleotide), sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content. Using this information, we quickly gained an overview of our data quality and table 5 below lists each sample and their results for each test. After running the FastQC program on our samples, we can see a lot of information quickly on our data such as many of the data had good scores for their base sequence quality, tile sequence quality, sequence quality score, base N content, and base sequence length distribution. All the samples returned a green check mark for those specific parameters. In addition to good quality data, we can see where the data is fairly poor such as in the base sequence content and sequence duplication columns where all the samples were flagged with a red x mark. The last couple columns such as sequence GC content, overrepresented sequences, and adapter content all have a mix of green, red and mainly yellow suggesting these samples are of lower quality. A majority of these outputs suggest that before proceeding onto the next steps in the bioinformatics pipeline, we should consider addressing these problems that the raw data is showing. To do this, there will be two things performed to manipulate the raw data. The first will be trimming the raw data sequence because the per base sequence quality becomes worse as the position in the reads gets larger. Through trimming the per base sequence length from 151 bp, our reads will be shorter but will be more accurate. When reviewing the per base sequence quality scores of all the data, although they all came back with a good value, their phred scores for some appeared to start to fall below the 28 mark and a few below the 20 mark for some of the reads around the 135-150 region. Therefore, by trimming the data by 30 base pairs from the 5' end, we will end up with a higher quality and more accurate read for each of our samples. The second thing to implement is to trim the adapter sequences which appear to be leading to poor adapter content and affecting the overrepresented sequences scores too. When illumina performs RNA sequencing, they utilize adapters for the cDNA from the RNA to bind to the flow cell which has complementary sequences present on it. Amplification will occur off of the cDNA and the adapters that are present can become part of the newly sequenced strand which will lead to a high adapter content value and will influence the overrepresented sequence. To account for this, we can use a program called trimmomatic which allows us to specify the illumina adapters used for sequencing and trim out those adapter sequences if they find them in the raw sequence therefore eliminating the adapter sequence. In figure 12 below, there are images of the raw data FastQC status checks before and after being processed by trimmomatic and shortening the per base sequence length. This generated three sets of fastq files where the first is the unaltered raw data which will be referred to as raw fastq files, the second being the raw data will have its total sequence length trimmed which we will refer to as the trimmed fastq files, and the third will be the fastq file that we cut the adapters using trimmomatic and trimmed the total sequence length which will be referred to as the cut adapters fastq files. After both of these processing steps were performed to the raw data, it can be seen that the columns for adapter content, overrepresented sequences, per base sequence content, and per sequence GC content all show an improvement in the green status check marks from their previous statuses. However, it is also important to note that the per tile sequence quality and sequence length distribution both go from the green status to the cautionary status as a result of these manipulations to the data

Table 5

FastQC Output for all 24 Pepper Samples

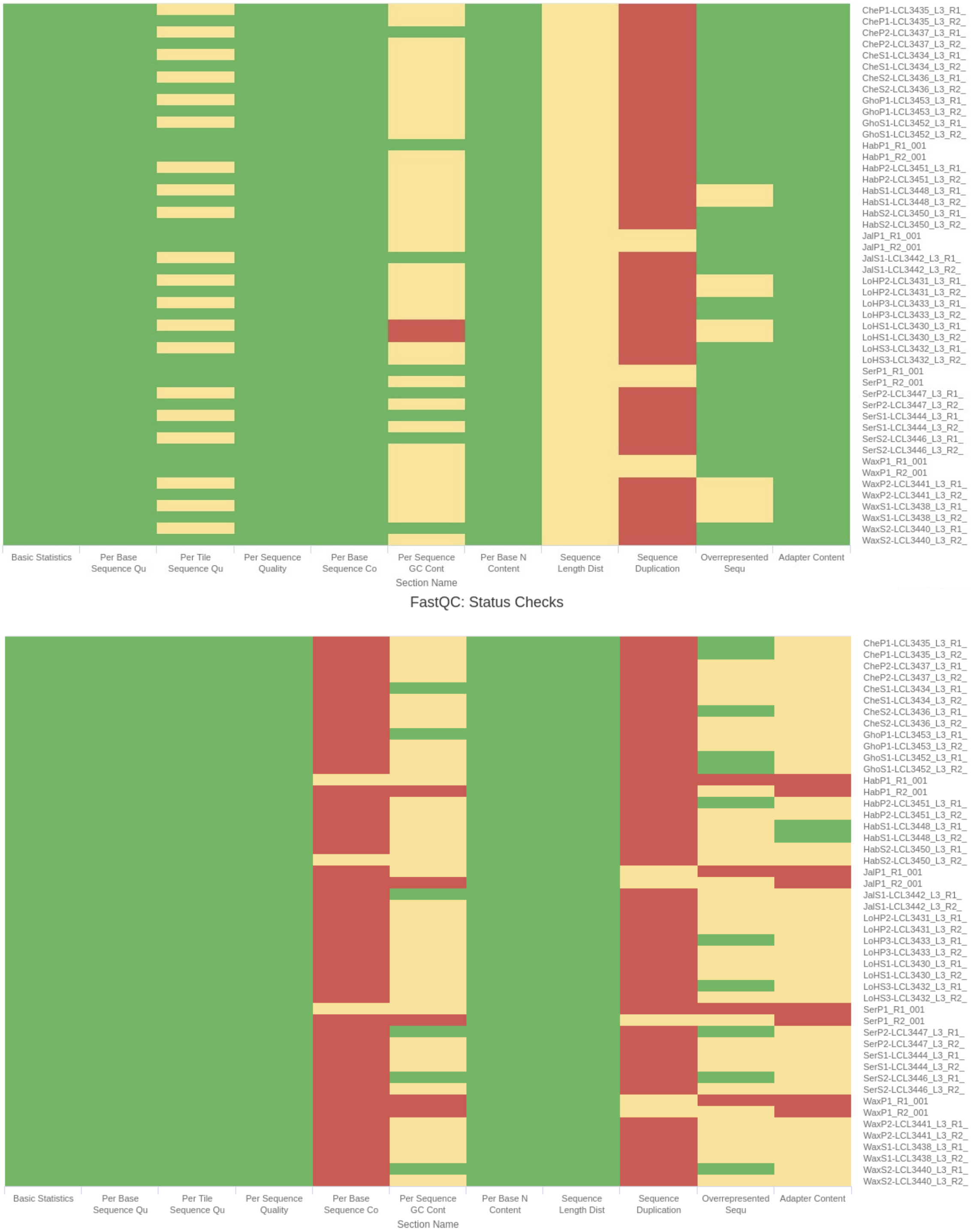
Pepper Sample	Total Sequences	Sequence Length	Statistics	Base Seq Quality	Tile Seq Quality	Sequence Quality Score	Base Seq Content	Sequence GC Content	Base N content	Seq Length Distr	Seq Duplication	Overrepresented Seq	Adapter Content
Che P1 Run1	27194587	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Che P1 Run2	27194587	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Che P2 Run1	24310395	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Che P2 Run2	24310395	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Che S1 Run1	24469899	151	Good	Good	Good	Good	Bad	Good	Good	Good	Bad	Okay	Okay
Che S1 Run2	24469899	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Che S2 Run1	29804880	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Che S2 Run2	29804880	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Gho P1 Run1	26863525	151	Good	Good	Good	Good	Bad	Good	Good	Good	Bad	Okay	Okay
Gho P1 Run2	26863525	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Gho S1 Run1	26850176	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Gho S1 Run2	26850176	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Hab P1 Run1	24222612	150	Good	Good	Good	Good	Okay	Okay	Good	Good	Bad	Bad	Bad
Hab P1 Run2	24222612	150	Good	Good	Good	Good	Bad	Bad	Good	Good	Bad	Okay	Bad
Hab P2 Run1	32608724	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Hab P2 Run2	32608724	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Hab S1 Run1	47424533	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Good
Hab S1 Run2	47424533	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Good
Hab S2 Run1	28928231	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Hab S2 Run2	28928231	151	Good	Good	Good	Good	Okay	Okay	Good	Good	Bad	Okay	Okay
Jal P1 Run1	17295942	150	Good	Good	Good	Good	Bad	Okay	Good	Good	Okay	Bad	Bad
Jal P1 Run2	17295942	150	Good	Good	Good	Good	Bad	Bad	Good	Good	Okay	Okay	Bad
Jal S1 Run1	33886935	151	Good	Good	Good	Good	Bad	Good	Good	Good	Bad	Okay	Okay
Jal S1 Run2	33886935	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Cay P2 Run1	42014619	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Cay P2 Run2	42014619	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Cay P3 Run1	27563358	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Cay P3 Run2	27563358	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Cay S1 Run1	43595331	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Cay S1 Run2	43595331	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Cay S3 Run1	31214834	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Good	Okay
Cay S3 Run2	31214834	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay

Pepper Sample	Total Sequences	Sequence Length	Statistics	Base Seq Quality	Tile Seq Quality	Sequence Quality Score	Base Seq Content	Sequence GC Content	Base N content	Seq Length Distr	Seq Duplication	Overrepresented Seq	Adapter Content
Ser P1 Run1	19020160	150	Good	Good	Good	Good	Okay	Okay	Good	Good	Bad	Bad	Bad
Ser P1 Run2	19020160	150	Good	Good	Good	Good	Bad	Bad	Good	Good	Okay	Okay	Bad
Ser P2 Run1	28339488	151	Good	Good	Good	Good	Bad	Good	Good	Good	Bad	Good	Okay
Ser P2 Run2	28339488	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Ser S1 Run1	26052653	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Ser S1 Run2	26052653	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Ser S2 Run1	25816590	151	Good	Good	Good	Good	Bad	Good	Good	Good	Bad	Good	Okay
Ser S2 Run2	25816590	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Wax P1 Run1	20019634	150	Good	Good	Good	Good	Bad	Bad	Good	Good	Okay	Bad	Bad
Wax P1 Run2	20019634	150	Good	Good	Good	Good	Bad	Bad	Good	Good	Okay	Okay	Bad
Wax P2 Run1	42021312	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Wax P2 Run2	42021312	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Wax S1 Run1	39400448	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Wax S1 Run2	39400448	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay
Wax S2 Run1	27711941	151	Good	Good	Good	Good	Bad	Good	Good	Good	Bad	Good	Okay
Wax S2 Run2	27711941	151	Good	Good	Good	Good	Bad	Okay	Good	Good	Bad	Okay	Okay

Note. Pepper samples are abbreviated and are as follows: Che = Cherry pepper, Gho = Ghost pepper, Hab = Habanero pepper, Jal = Jalapeno Pepper, Cay = Cayenne pepper, Ser = Serrano pepper, Wax = Hungarian wax pepper. Pepper samples are differentiated with either a S or P indicating if they are a skin or placenta sample respectively. Run1 and Run2 differentiate between the paired end run results. Green highlighted boxes with the “good” text indicate good quality data, yellow highlighted boxes with the “okay” text indicate cautionary data, and red highlighted boxes with the “bad” text indicate poor quality data.

Figure 12

FastQC Results Before and After Data Processing



Note. The top figure represents the raw data files obtained from FastQC. The bottom figure represents the altered data after processing with trimmomatic and cutting down on the per base sequence length by 30 base pairs. On the y-axis are the listed pepper samples and the text on the x-axis lists each of the eleven FastQC outputs. Data are highlighted as green, yellow, or red to indicate good quality, cautionary quality, or poor quality data respectively.

Bowtie

After utilizing FASTQC, the Bowtie software program was used to align the fastq files to our reference sequence. There were three iterations of fastq files that were generated from the previous steps which were the raw fastq files, cut adapters fastq files, and the trimmed fastq files. Each of these file types were mapped to our reference genome and their respective overall alignment scores were generated and compared to one another to see how the differences after each data manipulation step affected the overall bowtie scores. For our experiment, the reference genome that we will be using will be the *Capsicum annuum* reference genome. The reference genome was obtained from Sol Genomics Network and it is *Capsicum annuum* cv CM334 Genome CDS (release 1.55). The first bowtie alignment we ran on the raw fastq file resulted in an average alignment score of 0 times at 35.7%, an average alignment score of 1 time at 41.54%, an average alignment score of more than 1 times at 22.76% with an overall alignment score of 64.30%. The second bowtie alignment we ran on the trimmed fastq file, which only had its per base sequence length shortened, this resulted in an average alignment score of 0 times at 29.32%, an average alignment score of 1 time at 45.77%, an average alignment score of more than 1 times at 24.92% with an overall alignment score of 70.69%. The third bowtie alignment we ran on the cut adapters fastq file, which had its per base sequence length shortened and the adapters cut out, this resulted in an average alignment score of 0 times at 23.26%, an average alignment score of 1 time at 48.67%, an average alignment score of more than 1 times at 28.07% with an overall alignment score of 76.74%. Table 6 below demonstrates the results for each of the bowtie alignments that we generated using each of the pepper sample fastq files. The most noticeable difference we can see after processing the data is in the “Overall Score Improvement (Cut Adapters-Raw)” column which is the difference between the cut adapters bowtie scores and the raw alignment scores for each of the pepper samples. In this section, values that are highlighted green are samples that had the highest score improvement while samples that are highlighted in red are ones that had the lowest score improvement after being processed. The main column to look at is the overall cut adapters alignment column. This column lists the overall bowtie score for the pepper samples after having their sequences trimmed and adapters cut. This column contains the highest bowtie alignment scores for the pepper samples when we compare the same pepper samples against their raw alignment scores and trimmed alignment scores. This score and output will be used for the cufflinks assessment as these pepper samples demonstrated the best alignment to their reference genome.

Table 6

Bowtie Alignment Output for all 24 Pepper Samples

Scoville Heat Units	**Genus**	**Reads**	**Pepper**	**Overall Raw Alignment**	**Overall Trimmed Alignment**	**Overall Cut Adapters Alignment**	**Overall Score Improvement (Cut Adapters - Raw)**
Cherry Peppers: 100-500 SHU	*Capsicum annuum*	29804880	CheS2-R2	66.29%	71.84%	76.65%	10.36%
	Capsicum annuum	29804880	CheS2-R1	67.61%	73.29%	76.76%	9.15%
	Capsicum annuum	24469899	CheS1-R2	61.24%	67.13%	72.54%	11.30%
	Capsicum annuum	24469899	CheS1-R1	62.72%	68.77%	72.69%	9.97%
	Capsicum annuum	24310395	CheP2-R2	62.30%	68.18%	73.42%	11.12%
	Capsicum annuum	24310395	CheP2-R1	63.77%	69.81%	73.57%	9.80%
	Capsicum annuum	27194587	CheP1-R2	66.72%	72.59%	76.85%	10.13%
	Capsicum annuum	27194587	CheP1-R1	67.86%	73.81%	76.99%	9.13%
Jalapeno: 3,500-8,000 SHU	*Capsicum annuum*	33886935	JalS1-R2	61.78%	67.56%	72.72%	10.94%
	Capsicum annuum	33886935	JalS1-R1	63.11%	69.05%	72.88%	9.77%
	Capsicum annuum	17295942	JalP1-R2	50.63%	61.72%	73.93%	23.30%
	Capsicum annuum	17295942	JalP1-R1	51.27%	62.33%	73.89%	22.62%
Hungarian Wax: 5,000-15,000 SHU	*Capsicum annuum*	27711941	WaxS2 -R2	61.92%	67.60%	73.01%	11.09%
	Capsicum annuum	27711941	WaxS2 -R1	63.54%	69.41%	73.22%	9.68%
	Capsicum annuum	39400448	WaxS1 -R2	71.09%	75.96%	80.50%	9.41%
	Capsicum annuum	39400448	WaxS1 -R2	72.27%	77.21%	80.61%	8.34%
	Capsicum annuum	42021312	WaxP2-R2	70.04%	74.97%	79.81%	9.77%
	Capsicum annuum	42021312	WaxP2-R1	71.61%	76.65%	79.98%	8.37%
	Capsicum annuum	20019634	WaxP1-R2	62.73%	72.23%	92.01%	29.28%
	Capsicum annuum	20019634	WaxP1-R1	63.86%	73.25%	92.35%	28.49%
Serrano: 10,000-23,000 SHU	*Capsicum annuum*	25816590	SerS2-R2	61.13%	67.09%	72.32%	11.19%
	Capsicum annuum	25816590	SerS2-R1	62.62%	68.77%	72.56%	9.94%
	Capsicum annuum	26052653	SerS1-R2	62.60%	68.43%	73.77%	11.17%
	Capsicum annuum	26052653	SerS1-R1	64.15%	70.17%	73.96%	9.81%
	Capsicum annuum	28339488	SerP2-R2	62.45%	68.39%	73.36%	10.91%
	Capsicum annuum	28339488	SerP2-R1	64.00%	70.14%	73.56%	9.56%
	Capsicum annuum	19020160	SerP1-R2	49.25%	61.02%	78.00%	28.75%
	Capsicum annuum	19020160	SerP1-R1	50.09%	61.82%	78.12%	28.03%
Cayenne: 30,000-50,000 SHU	*Capsicum annuum*	31214834	CayS3-R2	64.70%	70.14%	75.69%	10.99%
	Capsicum annuum	31214834	CayS3-R1	66.40%	72.01%	75.94%	9.54%
	Capsicum annuum	43595331	CayS1-R2	75.27%	79.49%	83.48%	8.21%
	Capsicum annuum	43595331	CayS1-R1	76.33%	80.61%	83.57%	7.24%
	Capsicum annuum	27563358	CayP3-R2	63.51%	68.98%	74.19%	10.68%
	Capsicum annuum	27563358	CayP3-R1	65.02%	70.66%	74.37%	9.35%
	Capsicum annuum	42014619	CayP2-R2	74.36%	79.01%	83.55%	9.19%
	Capsicum annuum	42014619	CayP2-R1	75.71%	80.44%	83.67%	7.96%

Scoville Heat Units	**Genus**	**Reads**	**Pepper**	**Overall Raw Alignment**	**Overall Trimmed Alignment**	**Overall Cut Adapters Alignment**	**Overall Score Improvement (Cut Adapters - Raw)**
Habanero: 100,000-350,000 SHU	*Capsicum chinense*	28928231	HabS1-R2	63.94%	69.53%	74.64%	10.70%
	Capsicum chinense	28928231	HabS1-R1	65.60%	71.35%	74.85%	9.25%
	Capsicum chinense	47424533	HabS1-R2	75.86%	79.65%	82.55%	6.69%
	Capsicum chinense	47424533	HabS1-R1	76.39%	80.22%	82.64%	6.25%
	Capsicum chinense	32608724	HabP2-R2	64.36%	69.94%	74.87%	10.51%
	Capsicum chinense	32608724	HabP2-R1	65.80%	71.51%	75.02%	9.22%
	Capsicum chinense	24222612	HabP1-R2	47.80%	58.97%	73.50%	25.70%
	Capsicum chinense	24222612	HabP1-R1	48.53%	59.68%	73.57%	25.04%
Ghost: 800,000-1,000,000 SHU	*Capsicum chinense*	26850176	GhoS1-R2	66.18%	71.54%	75.71%	9.53%
	Capsicum chinense	26850176	GhoS1-R1	67.20%	72.63%	75.78%	8.58%
	Capsicum chinense	26863525	GhoP1-R2	61.82%	68.01%	72.98%	11.16%
	Capsicum chinense	26863525	GhoP1-R1	63.00%	69.32%	73.12%	10.12%

Color Legend:
Pepper Species - *Capsicum annuum*
Pepper Species - *Capsicum chinense*
Pepper Tissue - Placenta
Pepper Tissue - Skin
Bowtie Alignment Score/Improvement/Number of Reads - High Alignment
Bowtie Alignment Score/Improvement/Number of Reads - Moderate Alignment
Bowtie Alignment Score/Improvement/Number of Reads - Low Alignment

Note. Pepper samples are abbreviated and are as follows : Che = Cherry pepper, Gho = Ghost pepper, Hab = Habanero pepper, Jal = Jalapeno Pepper, Cay = Cayenne pepper, Ser = Serrano pepper, Wax = Hungarian wax pepper. Pepper samples are differentiated with either a S or P indicating if they are a skin or placenta sample respectively and are highlighted either blue or purple accordingly. The number following S or P will either be a 1, 2, or a 3 which denotes which pepper the tissue sample originated from. R1 and R2 explain if the pepper sample is Run 1 or Run 2 to differentiate between the paired end results. Pepper samples are listed based on scoville heat units from lowly pungent to highly pungent. Pepper species are color coated to differentiate between Capsicum annuum and Capsicum chinense. Green highlighted boxes indicate high bowtie alignment scores, high score improvement, or high number of reads depending on the column. Yellow highlighted boxes indicate moderate bowtie alignment scores, moderate score improvement, or moderate number of reads depending on the column. Red highlighted boxes indicate low bowtie alignment scores, low score improvement, or low number of reads depending on the column.

Cufflinks

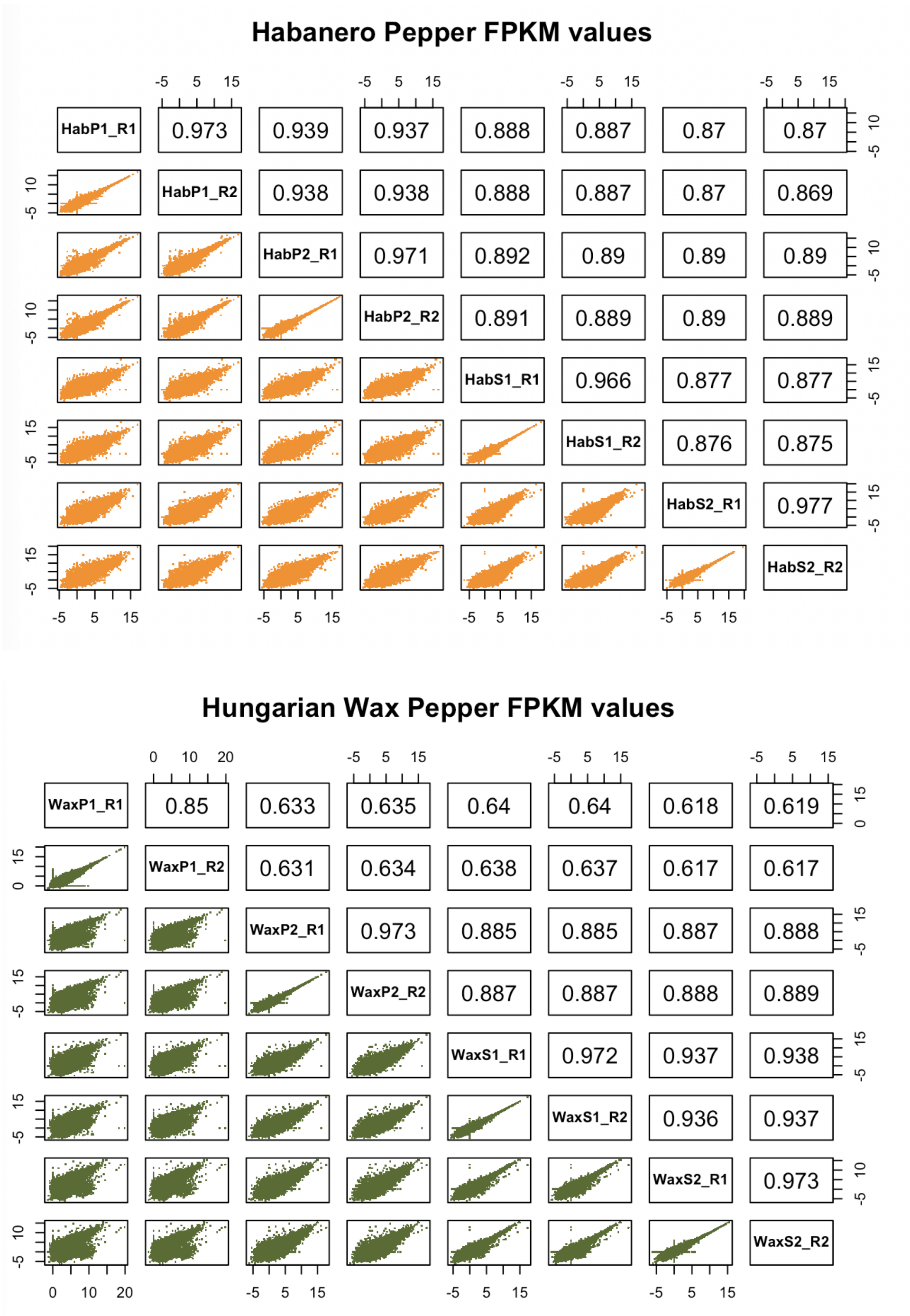
After the bowtie alignment, we then took our mapped reads and ran them through a program called cufflinks. First, bowtie outputs the files in a SAM format which must be converted to a BAM format. SAM files are files that the computer creates which stores the sequence data from the bowtie assessment in a series of tab delimited ASCII columns. These files are generated as they are "human readable" compared to its sister file which is the BAM (Binary Alignment Map) file that will store the same data in a compressed, indexed, binary format (Trapnell et. al., 2012). We first convert the SAM file to a BAM so the computer can then use the store sequence data to perform the analysis. After conversion of the SAM file to its BAM counterpart, we sort the file by chromosomes which is how the reference genome is set up. The reference genome that we will be utilizing will be the same file from the bowtie alignment which is the Capsicum annuum cv CM334 Genome CDS (release 1.55). After sorting, cufflinks can be used to evaluate gene expression on our pepper samples. Cufflinks is a program that assembles transcripts, estimates their abundances, and can test for differential expression and regulation in RNA-seq samples. It produces several output files that contain test results for changes in expression at the level of transcripts, primary transcripts, and genes (Trapnell et. al., 2012) (Ghosh & Chan, 2016). For our cufflinks output, we specifically obtained a file containing a list of 30,242 genes along with their relative expression levels (FPKM) for each of our 48 samples. The FPKM output stands for Fragments per kilobase of transcript per million mapped reads. The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. The FPKM output file can be combined with the reference genome file to obtain a compiled list of the genes, along with their location on the chromosome, and their predicted protein function if available.

Data Visualization in R

Since we have the forward and reverse runs for each of our samples due to pair-end sequencing, we can evaluate the forward against the reverse runs to identify any strong differences in FPKM values. Ideally, the forward and reverse runs should demonstrate similar FPKM results for the same genes as they would be expected to have similar gene expression levels as they are duplicate runs essentially. To validate this, we created scatter plot matrices of the different forward and reverse runs for our pepper samples and their FPKM outputs. The more linear the distribution, the more similar the samples are to one another meaning there is a stronger correlation between the FPKM values of the forward run with the reverse run. The more randomly distributed or non-linear in general, the weaker the correlation is between the forward and the reverse runs meaning that the sample output is less reliable and that sample set should be regarded with caution. We used the splom function in R which can be downloaded from the lattice package (Sarkar, 2008). Splom, is shortened for scatterplot matrix and it is a tool that uses multiple scatterplots to determine the correlation between a series of variables where the scatterplots are organized into a matrix to allow for easy visualization for any set of variables. Using the splom function, we created a series of scatterplots interrogating the relationship between skin and placenta FPKM values among the same pepper types. In total, we generated seven scatter plots, one for each of the pepper samples where we plotted their skin and placenta FPKM output for their forward and reverse runs. This allowed us to gather quick and efficient data on the correlations between the forward and reverse runs of multiple skin or placenta samples and identify specific cautionary samples that did not meet an expected R^2 value. In figure 13 below, there are two scatter plots, one demonstrating normal data and the other containing an example of samples that did not meet our expectations. The first figure, the Habanero Pepper FPKM values, did have normal data distribution and we can see strong correlation lines for our forward and reverse runs in the lower panels (R1 and R2) when they are plotted against each other and their respective R^2 values plotted on the upper panels. The placenta samples will have a strong linear relationship and the skin samples have a strong linear relationship and the forward and reverse runs of that sample will have the strongest correlation in the graph. There is not as strong of a correlation between the placenta and skin samples when plotted against one another but still resulted in a linear relationship. Now when we look at the bottom figure, the Hungarian Wax Pepper FPKM values, this scatterplot contains values that did not meet our expectations, we see a much different relationship between the forward and reverse runs for the same sample. The strength of the correlation is weaker when looking at some of the placenta forward and reverse runs when those ideally should be the strongest relationships. Out of a total of forty-eight samples, we only had two samples that demonstrated a noticeable weak linear relationship between its forward and reverse run for that same sample and across the same tissue type for that sample. The two samples were the Hungarian Wax pepper samples of WaxP1_R1 and WaxP1_R2 which had a low R^2 value at 0.85 when compared to each other and an even weaker R^2 value when compared to the other wax placenta samples of WaxP2_R1 and WaxP2_R2 of 0.631 and 0.634. This was the lowest R^2 value recorded by far out of all samples. Out of all of the other samples when comparing their forward and reverse runs against each other, none of them had a R^2 value below 0.96 while hungarian wax had a value of 0.85. In addition, out of all the samples when comparing the same pepper's tissue types, none of the samples had a lower R^2 value below 0.87 while the hungarian wax peppers had the lowest score across their placenta samples at a R^2 value of 0.63. These two placenta samples for the hungarian wax data appear to be outliers in the data set as their FPKM output varies greatly compared to each other and compared to the other hungarian wax placenta samples.

Figure 13

Scatterplot Matrix of FPKM Values Between Peppers

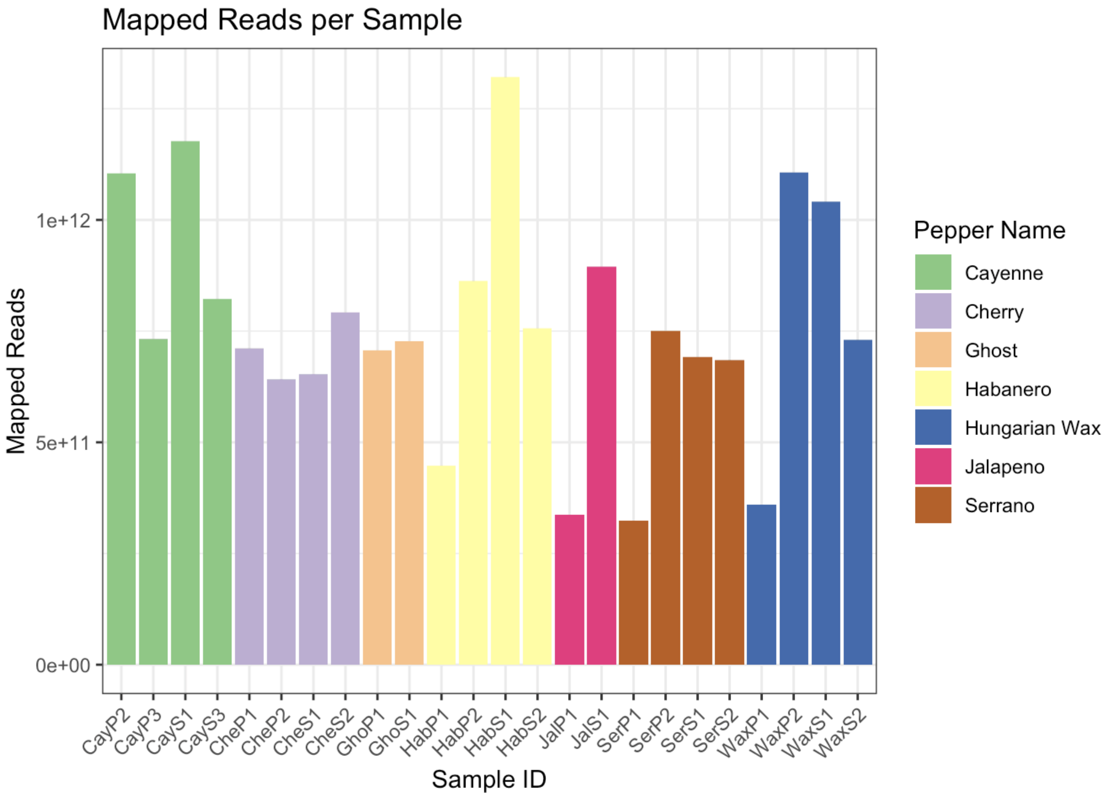


Note. Scatterplot matrix describing the correlation among the Log2 FPKM values for the pepper samples of the same pepper type. The first graph demonstrates good data and all the possibilities of the habanero pepper FPKM values and their correlation strength against each other. The second graph demonstrates bad data and all the possibilities of the hungarian wax pepper FPKM values and their correlation strength against each other. The sample ID for each pepper can be found in the diagonal panels in the scatterplot matrix. The scatterplot of the FPKM values is listed on the lower panels and their respective R^2 values are listed on the upper panels.

This first step in creating these series of scatter plots matrices allows us to easily visualize if certain samples may have poor or inaccurate data from the FPKM output provided from cufflinks. Performing this step allowed us to acknowledge those samples and treat them with caution before generating heat maps or other more involved data analysis steps. Besides evaluating the data with the scatter plot matrices alone, we also generated a bar graph visualizing the mapped library sizes for each of our pepper samples as well as a box plot of the FPKM output for each pepper sample. The tidyverse package was installed to format the data correctly for further analysis (Wickham et al., 2019). The library sizes were generated from bowtie and it is a count of the total number of mapped reads to the reference genome. This value was calculated using the processed bowtie output where the base sequence length was shortened and the adapters were also cut out for each of the pepper samples. Ultimately, this pre-processing resulted in shorter mapped reads as we had shortened the overall length of the reads from these steps. Using this processed data for the bowtie alignment, we collected the alignment scores and bowtie gave the value of how many reads were mapped back to the reference genome for each pepper sample. Using this information, we created a bar graph to visualize the differences in the sizes of the mapped library reads for the different pepper species which can be seen in figure 14 below. There are a total of 24 pepper samples plotted on the bar graph below and they represent the forward and reverse runs for each pepper as the forward and reverse library sizes were identical. The pepper species are grouped together based on pepper name through color coding and the pepper sample IDs are located on the x-axis with their abbreviations and tissue identifiers. The average mapped reads for the library sizes was 25,297,134 bp and a majority of the mapped reads of the pepper samples fell in this range. The largest mapped library size was 43,670,172 which was the HabS1 samples and the smallest mapped library size was 10,726,756 which was the SerP1 sample with the WaxP1 and JalP1 samples not far behind. Some of the pepper samples had rather large library sizes such as CayP2, CayS1, HabS1, WaxP1, and WaxS1 and some of the pepper samples had rather small library sizes such as HabP1, JalP1, SerP1, and WaxP1 samples. The majority of the other samples had their average library sizes around 20,000,000 - 25,000,000 base reads. A boxplot was also created utilizing the log2 FPKM values generated from cufflinks and we plotted the different pepper samples to visualize the differences in distribution of their FPKM values across pepper samples which can be seen in figure 15 below. The boxplot is useful to visualize where the average is among samples and outliers are easily identifiable as they will reside outside the Q1 and Q3 whisker portion of the boxplot. When comparing the averages among samples and their IQR ranges, it appears that many of the samples are similar in terms of their IQR sizes. As for the averages, many of the samples have a similar distribution ranging from 0 to 1, skewing in the negative direction except for the WaxP1 forward and reverse run samples which are strongly skewed toward the positive direction. In addition, the whisker length distributions are very similar among many of the pepper samples besides the WaxP1 forward and reverse samples which have a very positively skewed distribution. There are a very large number of outliers that reside with a FPKM value of 10 or higher for all the pepper samples which are seen as plotted points. The plotted points can also be seen as identical in the forward and reverse runs for the same pepper sample which is supportive that the FPKM output is consistent.

Figure 14

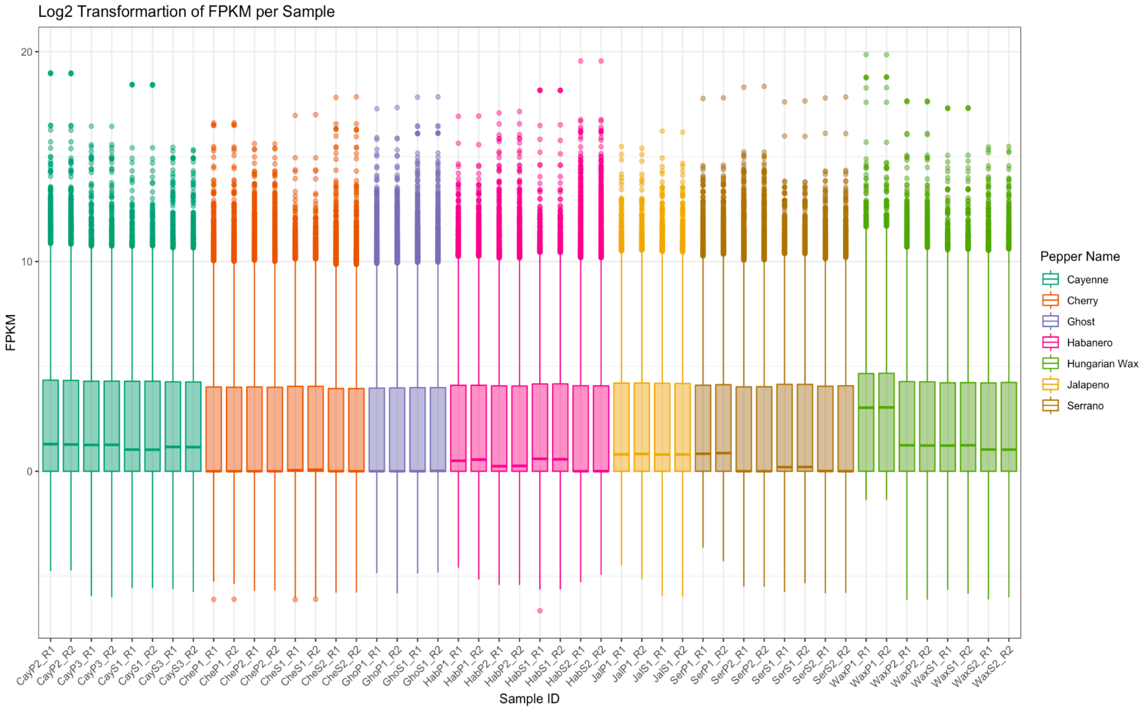
Total Number of Mapped Reads per Sample



Note. The data has been processed by having its adapter sequences cut and per base sequence length shortened. Data is represented as the total number of mapped reads and the samples are color coated by groupings based on the scientific names of the peppers. There are a total of 24 individual pepper samples plotted and their three-letter abbreviation as the sample ID on the x-axis are as follows: Cay = Cayenne pepper, Che = Cherry pepper, Gho = Ghost pepper, Hab = Habanero pepper, Jal = Jalapeno Pepper, Ser = Serrano pepper, Wax = Hungarian wax pepper. Pepper samples are differentiated with either a S or P after their three-letter abbreviation indicating if they are a skin or placenta sample. The number following S or P will either be a 1,2, or a 3 which denotes which pepper the tissue sample originated from. Forward and reverse runs for each pepper were not specified and instead replaced with a single sample ID as the forward and reverse runs had identical total number of mapped reads.

Figure 15

FPKM Values per Pepper Sample after Log2 Transformation

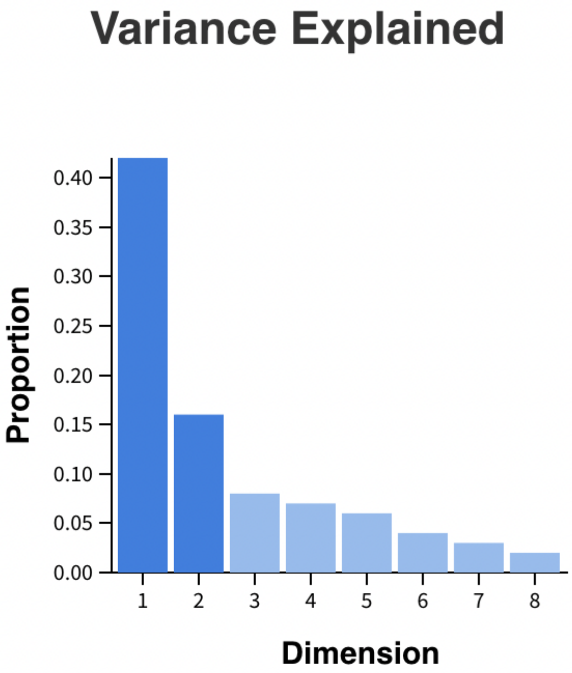


Note. A total of 30,242 FPKM values were generated for each pepper sample and 24 of the peppers were plotted with their forward and reverse runs for a total of 48 pepper samples. The horizontal darkened line in the box represents the mean for that pepper sample and the shaded box region itself represents the IQR. The whiskers represent the minimum and maximum values while any plotted points fall outside the whisker range and are considered outliers. A total of 48 individual pepper samples are plotted and their three-letter abbreviation as the sample ID on the x-axis and are as follows: Cay = Cayenne pepper, Che = Cherry pepper, Gho = Ghost pepper, Hab = Habanero pepper, Jal = Jalapeno Pepper, Ser = Serrano pepper, Wax = Hungarian wax pepper. Pepper samples are differentiated with either a S or P after their three-letter abbreviation indicating if they are a skin or placenta sample. The number following S or P will either be a 1,2, or a 3 which denotes which pepper the tissue sample originated from. Forward runs are identified with the suffix R1 and reverse runs are identified with the suffix as R2 for each of the pepper samples.

After generating the boxplot, bar graph, and scatterplot, we decided that a principal component analysis and a correlation matrix would be beneficial visuals to evaluate the relationship amongst our samples. The glimma package and limma packages were both installed to utilize the glMDSPlot and plotMDS functions provided to generate the pieces of the principal components of analysis (Su et. al., 2017) (Smyth, 2005). The two pieces of the principal components of analysis we generated are the scree plot which determines the number of statistically significant factors and the principal component analysis graph itself plotting our first and second components against each other. The scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. The value of a scree plot is that it tells you the number of factors that are present in your data and what principal components you should evaluate. Typically, scree plots demonstrate a sharp reduction in size of their eigenvalues and when this occurs, any factors that fall below this sharp reduction will add relatively little to no information to the graph as they contribute significantly less to any variation observed in the data. Therefore, any values that fall before this reduction in eigenvalues is important to evaluate as these components account for the majority of variation in the data. According to the scree plot, there are two strong principal components which are components 1 and 2 which can be seen in figure 16 below. Principal component 1 makes up for over 40% proportion of the variation in the data while principal component 2 only accounts for a little over 15%. Principal components 3, 4, and 5 are relatively low in the 6% - 8% proportion range and the rest of the components after 5, all fall below 5% proportion. Using this information from the scree plot, we then made one principal component of analysis graph plotting component 1 on the x-axis and component 2 on the y-axis as seen in figure 17 below. The principal component analysis figure demonstrates some relationships in the data. For the samples, their forward and reverse runs are all plotted right over each other for each sample and generally, most of the samples are grouped in the same relative area according to their pepper name. There appears to be three different clusterings of the data, one group in the top right corner, one in the middle bottom, and the last in the top left. These groups contain the following pepper samples; in the top right corner the cherry and serrano peppers, in the middle bottom the habanero and ghost peppers reside, and in the top left the cayenne, hungarian wax, and jalapeno peppers reside. Most of the clustering of the placenta and skin samples are near each other with a stronger clustering occurring between placenta samples with placenta samples and the skin samples with the skin samples of the same pepper type. Again, compared to most other samples, the WaxP1 samples are located a further distance away from the WaxP2 sample set and even further from the WasS1 and WaxS2 samples. Most samples may be located some distance away between the skin and placenta samples of the same pepper type but rarely are the samples located far away of the same tissue type and same pepper type.

Figure 16

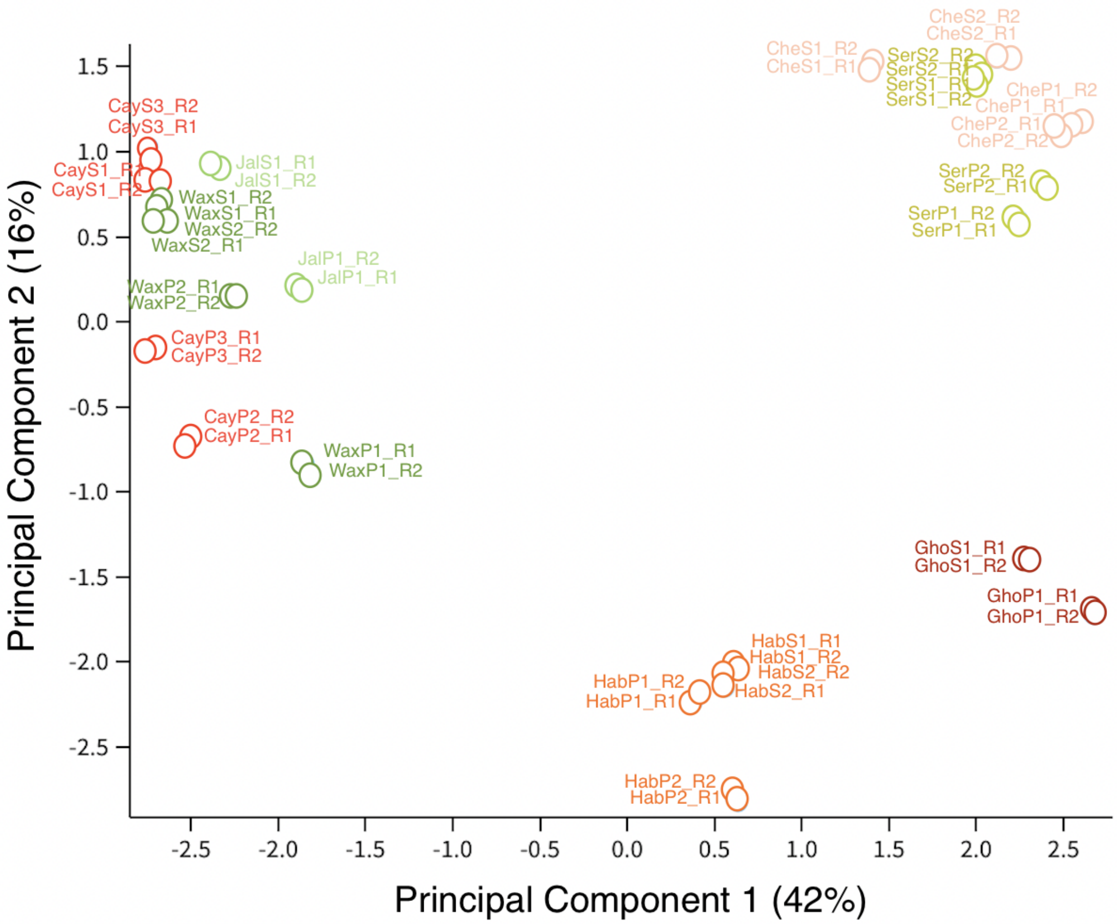
Scree Plot Demonstrating the Variation Between Principal Components



Note. Scree plot demonstrating the eigenvalues of the principal components identified across the pepper samples utilizing their Log2 FPKM values. Eigenvalues are displayed on the y-axis which explains the degree of variation in each sample. Each principal component is displayed on the x-axis. The bar graphs are ordered from largest to the smallest according to their eigenvalues.

Figure 17

Principal Component Analysis for the Log2 FPKM Values of the Pepper Samples

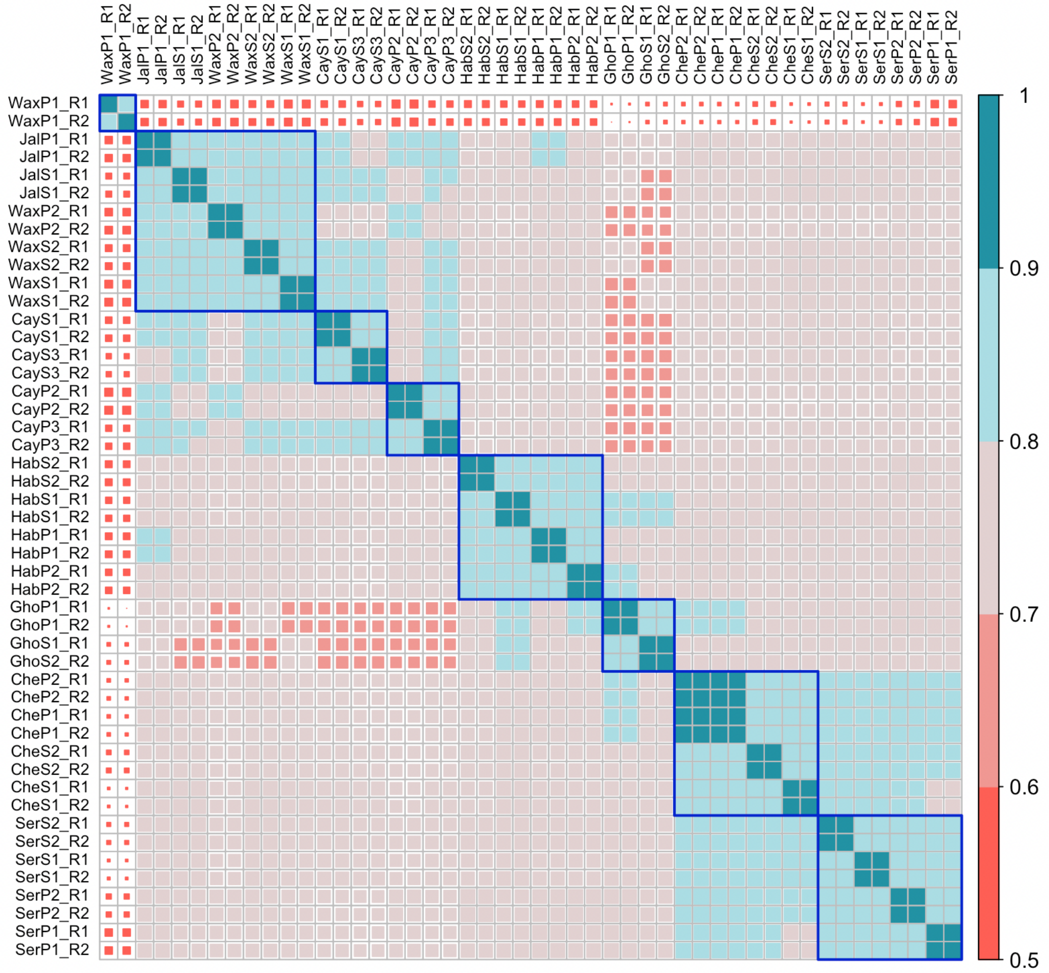


Note. Principal components 1 and 2 were plotted with their proportion of variation contributing to the data displayed as a percentage. Pepper samples are color coded according to the pepper name. A total of 48 individual pepper samples are plotted and their three-letter abbreviation are as follows: Cay = Cayenne pepper, Che = Cherry pepper, Gho = Ghost pepper, Hab = Habanero pepper, Jal = Jalapeno Pepper, Ser = Serrano pepper, Wax = Hungarian wax pepper. Pepper samples are differentiated with either a S or P after their three-letter abbreviation indicating if they are a skin or placenta sample. The number following S or P will either be a 1,2, or a 3 which denotes which pepper the tissue sample originated from. Forward runs are identified with the suffix R1 and reverse runs are identified with the suffix R2 for each of the pepper samples.

Another visual representation of the data that we created was a correlation matrix that was plotted using hierarchical clustering. The correlation matrix plot was created in R using the corrplot package (Wei, 2021). Correlation matrix is a visual table that displays the correlation among the data and by plotting it in a hierarchical fashion, the graph will place samples that have a stronger correlation to each other, next to each other. There were two correlation matrices we made, the first one had the WaxP1 samples in the matrix and the second graph we removed them to see how it affected the relationships in the graph. Figure 18 below shows the correlation matrix we created, plotting the relationships of all pepper samples against each other with no exclusions. The strongest correlation a sample could have is a value of 1, which is colored coded on the y-axis as dark turquoise shaded squares, and this suggests that this sample demonstrates a perfectly positive linear relationship between two variables. As we can see in the figure, all the forward and reverse runs of the same pepper have a perfect linear relationship with each other except for the WaxP1_R1 and WaxP1_R2 samples. They have a strong linear relationship when WaxP1_R1 is plotted against WaxP1_R1, but that is to be expected as they are the same sample. In addition, when any sample is plotted against the WaxP1 sample set, even the same WaxP2 samples, the correlation is very low at about 50 - 60 %. When the forward run is plotted against the reverse run, they are only about 80 - 90% correlated. They are also not clustered next to the other hungarian wax pepper samples, and the jalapeno peppers appear to be more similar to the hungarian wax samples than the hungarian wax placenta samples are to the hungarian wax samples. All of the other pepper samples are clustered next to their unique pepper types and even further all of those unique pepper types have their skin samples next to their skin samples and their placenta samples next to their placenta samples. Clearly, the hungarian wax placenta 1 forward and reverse run remains an outlier in the data so we removed this sample set and remade the matrix which can be seen in figure 19 below. After removing the wax placenta 1 sample set, we see an overall better correlation matrix with a majority of the samples having a stronger correlation with each other. Each of the forward and reverse runs for each pepper type again has a strong correlation with each other indicated by the dark turquoise shaded squares. The cherry pepper placenta samples are all strongly correlated with each other even with the placenta samples being taken from two different cherry peppers. None of the other samples demonstrate a linear relationship of 90 - 100 % when comparing the placenta samples or skin samples of the same pepper but different extractions. When selecting the option for the program to identify samples with strong clustering properties, it identified samples by drawing a blue square around them. We can see that the program distinctly identifies the same pepper types such as saying all the jalapeno samples are closely related and all the wax samples are closely related. It identified a total of 8 clusters with only the unusual one being two separate clusters for the cayenne samples separating the tissue types between placenta and skin. There also appears to be a stronger correlation in the cayenne placenta samples with each other compared to any other skin or placenta samples of the same pepper type. The cayenne placenta samples are all shaded a dark turquoise demonstrating their strong correlation of 90 - 100 % with each other. None of the other samples have all their placenta samples and skin samples of the same pepper type shaded with the dark turquoise together, only the forward and reverse samples have this in the other pepper samples. After interrogating the data with these graphics for outliers or abnormalities in the data, we decided to remove the hungarian wax placenta 1 samples from future analyses as this was a poor data set and an outlier. Using this information, we would then create a heatmap to visualize the differences in gene expression among the pungent and non-pungent peppers to identify any other genes that may play a role in pepper pungency.

Figure 18

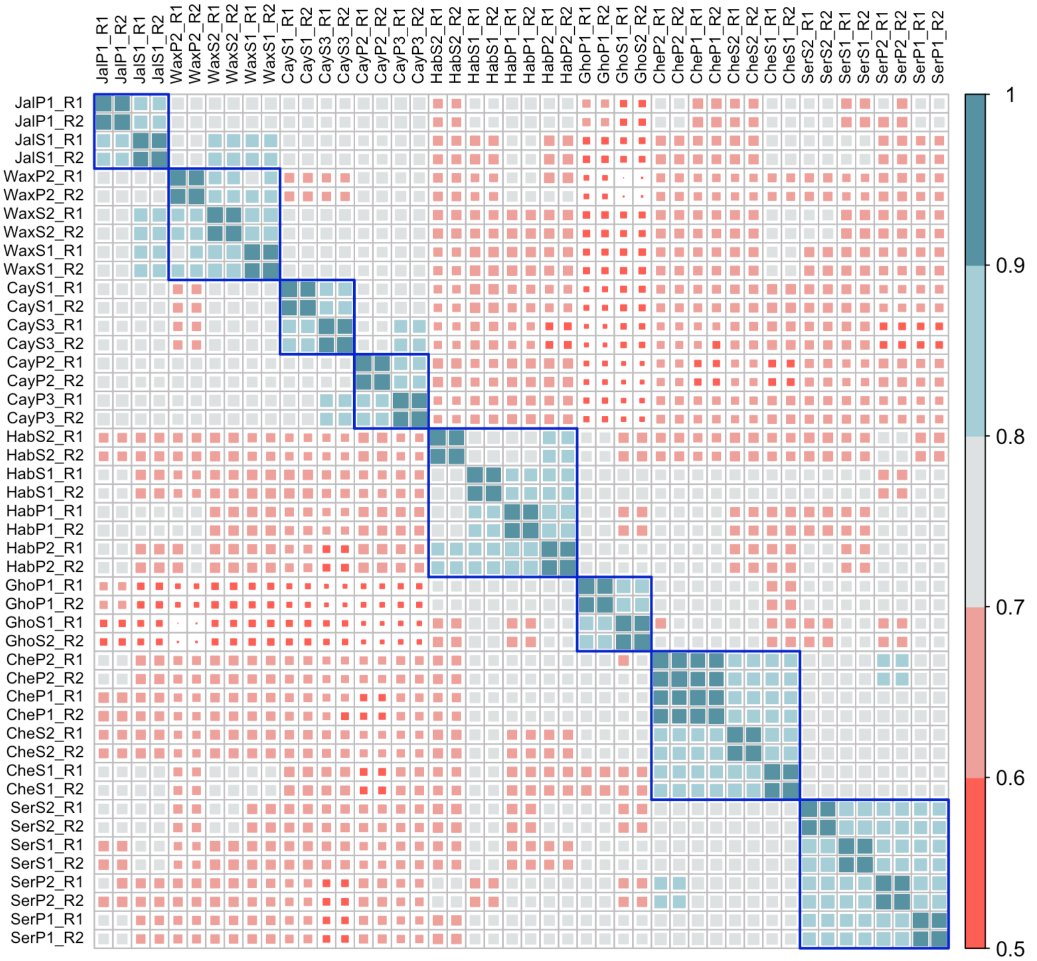
Correlation Matrix Ordered by Hierarchical Clustering Prior to Outlier Removal



Note. The 48 pepper samples are listed on the x and y axis and are organized by their similarity through hierarchical clustering. The blue squares highlight groups in the data that are highly similar and there are a total of 8. The matrix is color coded according to correlation strength where strongly correlated samples are color coded by dark turquoise with a correlation value of 1 while weaker correlated samples are colored as dark red with a value of 0.5. The matrix is also oriented by square size to demonstrate correlation strength where large squares indicate a strong correlation and small colored squares a weak correlation.

Figure 19

Correlation Matrix Ordered by Hierarchical Clustering Post Outlier Removal



Log2 FPKM values for the 46 pepper samples after removing the WaxP1 sample set. The 46 pepper samples are listed on the x and y axis and are organized by their similarity through hierarchical clustering. The blue squares highlight groups in the data that are highly similar and there are a total of 8. The matrix is color coded according to correlation strength where strongly correlated samples are color coded by dark turquoise with a correlation value of 1 while weaker correlated samples are colored as dark red with a value of 0.5. The matrix is also oriented by square size to demonstrate correlation

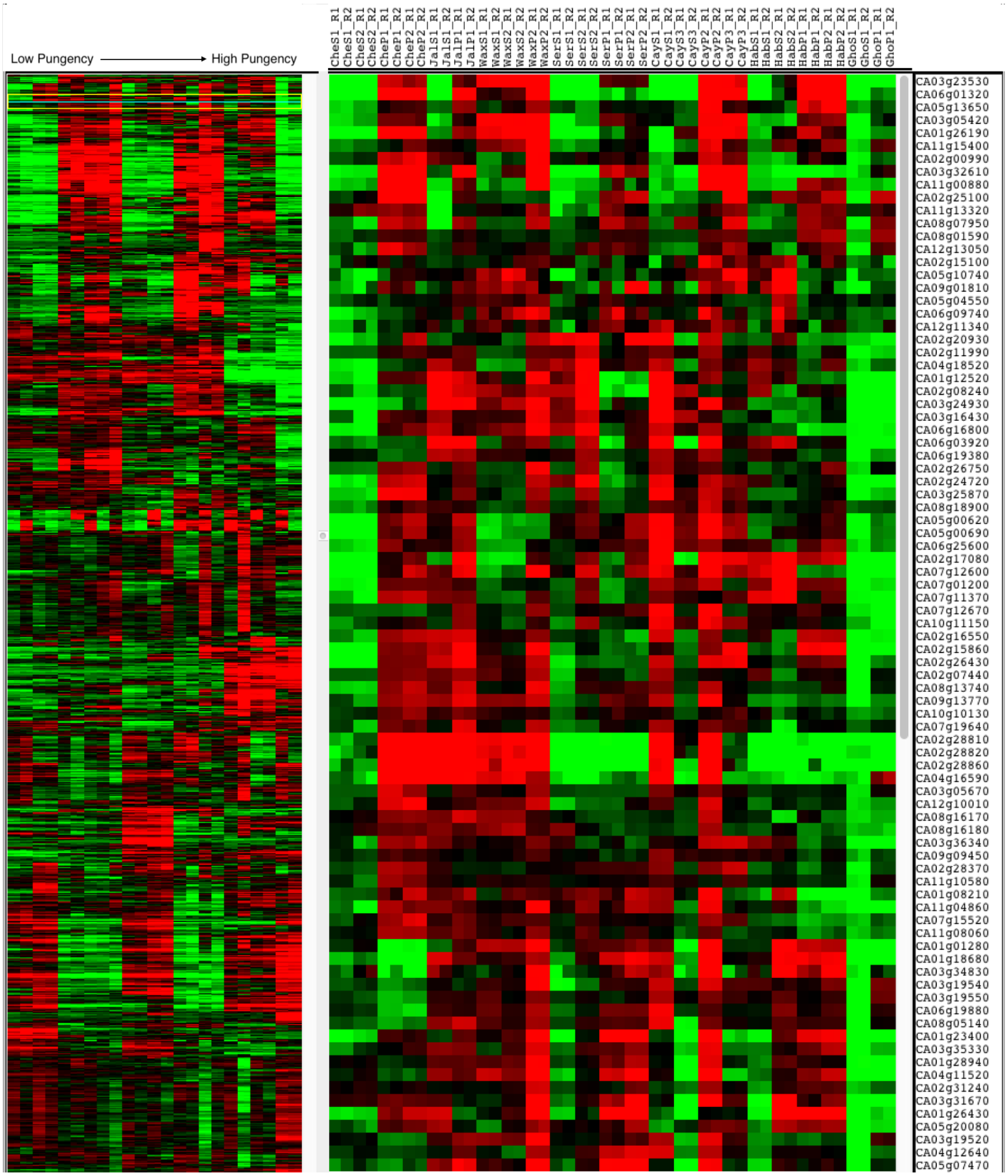
strength where large squares indicate a strong correlation and small colored squares a weak correlation.

Heatmap

In order to create the visual heat map, we utilized Gene Cluster 3.0, a program that has been improved by M.J.K. de Hoon, S. Imoto, J. Nolan, and S. Miyano which is an updated version of Michael Eisen's Cluster program of Berkeley Lab (De Hoon et. al., 2004). The Cluster 3.0 program can be used for gene expression clustering which is useful in finding patterns in the data and identifying outliers, incorrectly annotated samples and more. Our goal for gene expression clustering is to look at our pungent pepper samples and evaluate genes that are highly expressed in the pungent peppers compared to the less-pungent peppers and see if they are associated with capsaicin production. We can evaluate trends in the data set or even search for specific genes that are related to capsaicin production and look at other genes that cluster around them. The other genes that cluster around these searched genes are ones that the cluster program determined to be highly similar to each other so it is possible they could play a part in the capsaicin biosynthesis pathway. When using the Gene Cluster 3.0 program, we first applied some filters to the data. Besides removing the wax placenta 1 sample set, we log2 transformed the data set to bring everything into a more manageable scale. The data was filtered to remove lowly expressed genes, reducing the noise these genes would produce. To remove lowly expressed genes, we removed any that did not have a sufficient read depth by setting the number of observations to one with an absolute value being greater than five. The second filtering action we performed was to set the maximum value subtracted from the minimum value to be greater or equal to 1 which removes samples that do not have much variation in them. These filtering actions resulted in a total of 11,052 genes passing out of a total of 30,242 total genes. In the appendix section, figure A6 demonstrates the heatmap output when no filtering actions are performed. That heatmap contains all 30,242 genes and has many spots that are gray or black in the heatmap. These areas are ones that had low gene expression and the information in these sections was not of much interest to us. Then, using these 11,052 genes that passed the filtering criteria, we then centered the genes which is done because each gene represents a vector of values and by subtracting the average values of the gene from each experiment we get a better idea of the relative expression of each gene compared to each other. Lastly, we performed hierarchical clustering for just the genes of the samples and not the arrays while performing an average linkage as the clustering method. We did not cluster the arrays because we had previously sorted the array from least pungent to most pungent to easily visualize genes across samples based on pungency so only the genes need to be clustered. To visualize the microarray that was generated from the cluster program, we utilized the Java Treeview program created by Alok J. Saldanha (Saldanha, 2004). Java Treeview generates several interactive views of the gene expression data which allows the user to easily navigate through the image and examine samples. The samples will be colored based on their gene expression level, highly expressed genes will be red in our case and lowly expressed genes will be green. This allows easy visualization of trends in the data. The columns will be representative of the samples while the rows are representative of the genes. An example of what the output image looks like can be seen in figure 20 below. This image is an overview of the entire 11,052 genes that were sorted by hierarchical clustering for the pepper samples. In the file that we used for the cluster program, pepper samples were organized left to right from least pungent to most pungent and ordered with skin samples coming first and placenta samples coming second for each pepper. The most interesting sections for us will be any that have a large number of samples with high expression in the highly pungent peppers and low expression in the lowly pungent peppers. These sections will be useful as genes that may play a role in capsaicin production could possibly reside here. Another useful tool to utilize would be the search gene function. Using this tool to search out genes that are known to be highly expressed in pungent peppers and non-pungent peppers is useful too, as other genes that have similar expression patterns will be clustered together with the known gene.

Figure 20

Gene Expression Data Generated by Cluster 3.0 and Visualized on Java Treeview

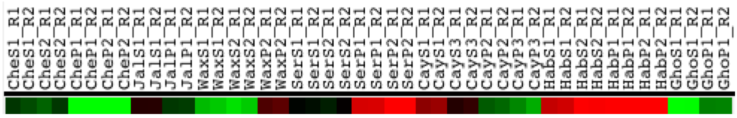
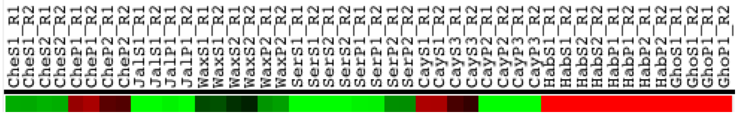
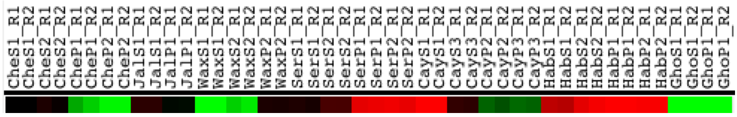
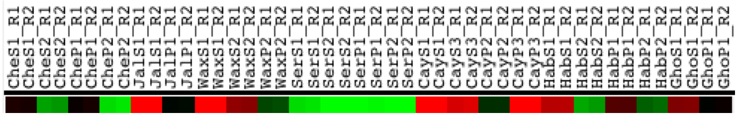
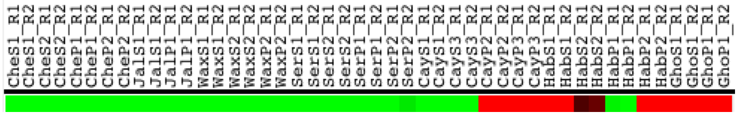
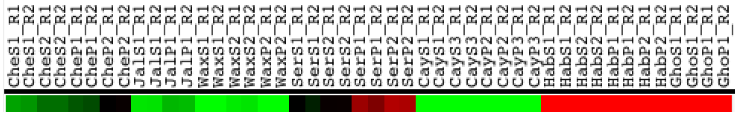
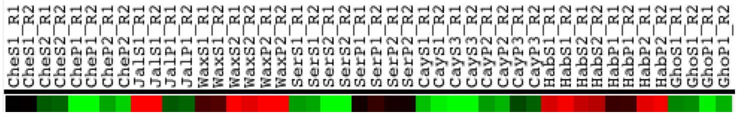
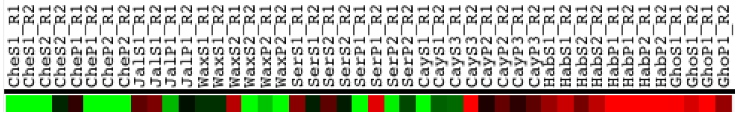
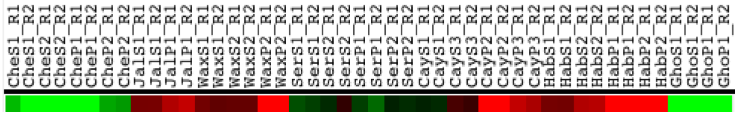


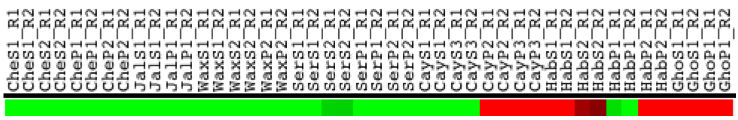
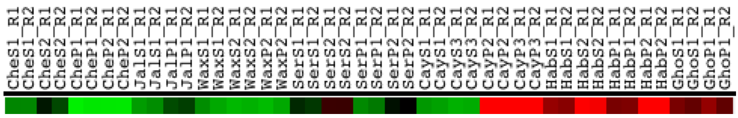
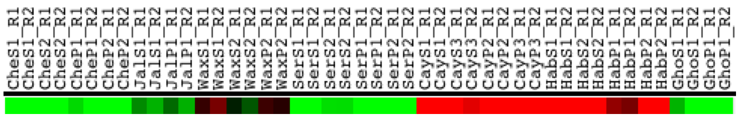
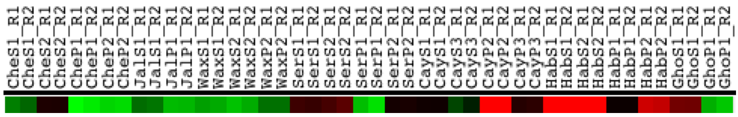
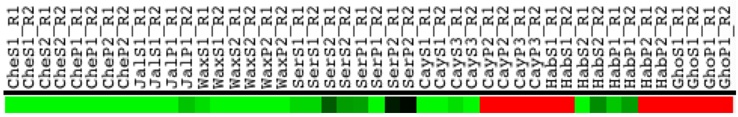
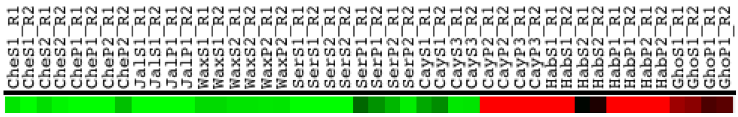
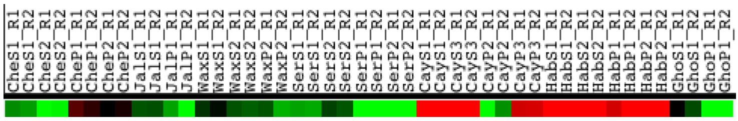
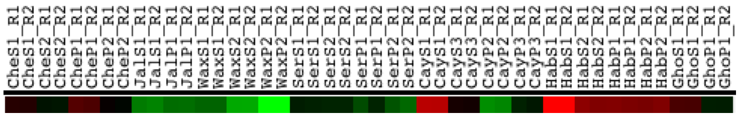
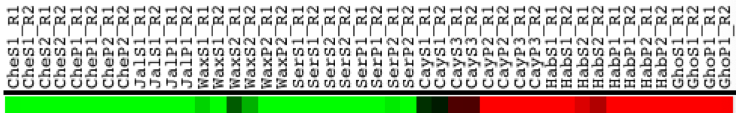
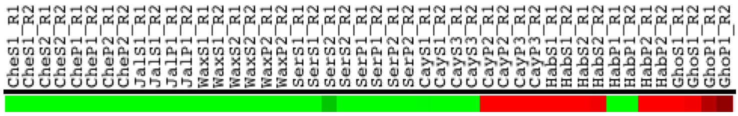
Note. There are a total of 11,052 genes plotted for the 46 pepper samples and they are color coded based on gene expression level where red is high expression or upregulation of a gene, green is low expression or downregulation of a gene, and black is no change in gene regulation. The first window shows all genes and all samples and are listed left to right from low pugency to high pungency. The second window is a blown up view of selected samples with the pepper samples representative as the columns and gene IDs are representative of the rows.

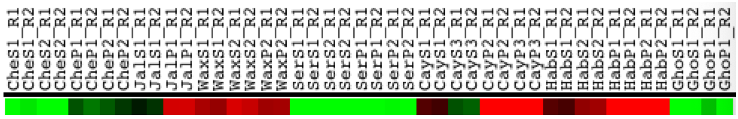
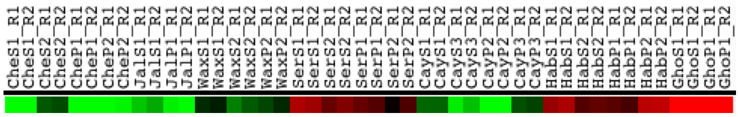
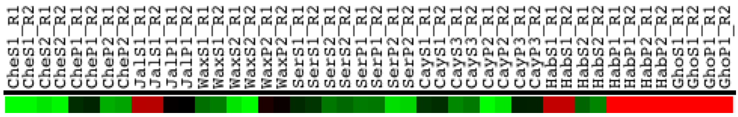
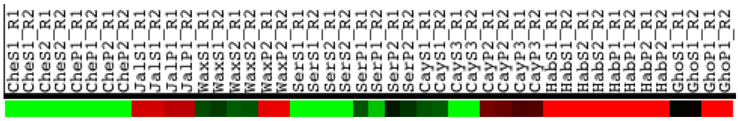
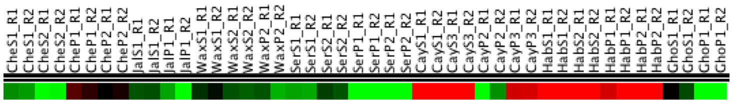
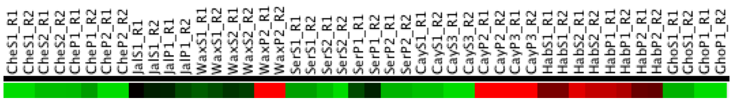
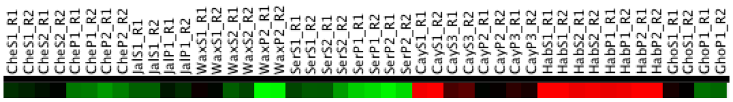
To begin the analysis of the heatmap, we will begin by searching known genes that are present in the capsaicin pathway and then analyze genes that are located near these known genes. The reason for this is because the clustering program organized the genes based on hierarchical ordering where similar things are located near each other. For this reason, we would predict other genes that share a similar expression to known genes in the capsaicin pathway may also play an important role in capsaicin synthesis. Using the paper “Discovery of putative capsaicin biosynthetic genes by RNA-Seq and digital gene expression of analysis of pepper” by Zhang et. al., they have listed known genes that are associated with the capsaicin biosynthesis pathway and novel candidate genes that may be associated with capsaicin production (Zhang et. al., 2016). Using the identified genes from their paper, we searched every gene they listed, a total of one hundred and thirty five, and compared them to our Java Treeview output file. We then identified if we saw any expression or not which can be seen in table A1 in the appendix section as some genes were removed after filtering our data. The genes may not be present in the output because if it was lowly expressed or had little variation, it would have been filtered out and removed. If the gene was present in our gene expression output, we would then visualize its expression across our generated heatmap and we recorded any genes if they had high expression for the pungent peppers and low expression for non-pungent peppers which can be seen in table 7. There were a total of twenty-six genes that had an expression profile similar to the one we are interested in which was high expression in pungent peppers and low expression in less-pungent peppers. Although not all of these genes in our heatmap exhibited this trend perfectly, we included a majority of them as they had high expression for pungent peppers as these areas were vibrant red. A few of the genes that did fit the expression trend very well were BCKDH E1a, KASI, FatA, pAMT, and AT. These samples typically had red expression levels for the cayenne, habanero, and ghost pepper samples which are all in the upper regions of our scoville list. Conversely, these samples also had either a green or black coloring in the less pungent peppers such as cherry, jalapeno, hungarian wax, and serrano pepper samples. Some of the genes from the twenty-six we identified exhibited more variation in their expression levels which resulted in not only the pungent peppers being highly expressed, some of the non-pungent peppers had high expression too. These genes were included in the list however because typically only a few non-pungent pepper samples were expressed and these samples still had very vibrant red, high expression in the pungent samples. Some of these genes were CCoAOMT, NADH-GOGAT, C4H, CCR, and HCT where they all have a little more variation in their expression but still fit the relative trend we were expecting. These trends can all be seen as the small section of the heatmap listed in table 7 which demonstrates the expression across all forty-six of our pepper samples. Using these samples, we will then broaden our search to identify other genes that may participate in the capsaicin biosynthesis pathway.

Table 7

Comparing Known Capsaicin Genes to our Heatmap and Identifying Genes that Demonstrated Upregulation in Pungent Peppers Compared to Non-Pungent Peppers

Protein	Gene ID	JavaTree Gene Expression Output
C4H - Cinnamate 4-hydroxylase	CA06g25930	
PAL - Phe ammonia-lyase	CA05g20790	
C4H - Cinnamate 4-hydroxylase	CA06g25940	
NADH-GOGAT - NADH-dependent Glu synthase	CA03g19580	
AT - Acyltransferase 2	CA01g32880	
BCKDH E2 - Dihydrolipoamide transacylase	CA01g18360	
CCoAOMT - Caffeoyl-CoA 3-O-methyltransferase	CA02g14470	
KasI - Ketoacyl-ACP synthase I	CA07g11150	
CM1 - Chorismate mutase	CA02g27850	

Protein	Gene ID	JavaTree Gene Expression Output
AT - acyltransferase 2	CA01g32920	
FatA - Acyl-ACP thioesterase	CA06g26640	
ENRa - Enoyl-ACP reductase	CA10g20920	
DH - hydroxyacyl-ACP dehydratase	CA08g15600	
KASIII - Ketoacyl-ACP synthase III	CA01g28560	
KASI - Ketoacyl-ACP synthase I	CA01g00840	
BCCP - Biotin carboxyl carrier protein	CA06g18470	
PDH E1a - Pyruvate dehydrogenase E1a	CA07g07490	
BCKDH E1a - a-Ketoacid decarboxylase E1a	CA06g10910	
pAMT - putative aminotransferase	CA03g08530	

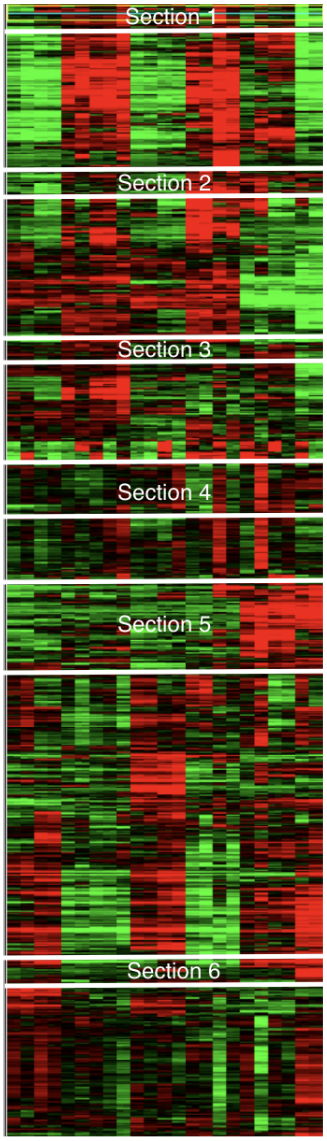
Protein	Gene ID	JavaTree Gene Expression Output
CAD - Cinnamyl alcohol dehydrogenase	CA06g10220	
CCR - Cinnamoyl-CoA reductase	CA08g13650	
SAMSyn - S-Adenosylmethionine synthetase	CA10g15500	
HCT - Hydroxycinnamoyl transferase	CA03g30250	
BCCP - Biotin carboxyl carrier protein	CA06g18470	
CAD - Cinnamyl alcohol dehydrogenase	CA02g00320	
BC - Biotin carboxylase	CA11g09810	

Note. The one hundred and thirty five genes identified by Zhang et. al. were searched in our heatmap and genes that demonstrated the trend of high expression in pungent peppers and low expression in less pungent peppers were recorded. A total of twenty-six were identified to exhibit this trend and those samples were recorded in the table. The respective gene IDs were recorded as well as the protein they produce in the annotation column. A visual representation of the expression profile for the sample was also included. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation.

The next steps will be to evaluate the twenty-six samples and identify other samples that were clustered next to or near them because these samples will share a strong similarity to these known genes that are present and participate in the capsaicin biosynthesis pathway. Using the genes from the list, we looked up each gene and identified where they fell on the heatmap and recorded the sections where a majority of the genes were located. These sections would be interesting places to dig through and pick out other possible genes of interest. There were roughly six different sections in which the twenty-six genes were falling in which can be seen as the white boxes in figure 21 below. A majority of the genes appeared to fall into sections 2, 5, and 6 as these sections were similar to the trend we were looking for. Sections 1, 3, and 4 did have spots that were similar to the trend we were looking for but more of the genes appeared to fall in those first three sections described. Using mainly the sections of 2, 5, and 6 since many of the genes were present in these sections, we generated some additional heatmap figures which can be seen in the appendix sections in figures A8 through A10. These heatmaps demonstrate high expression in the pungent peppers and low expression in the non-pungent peppers for many genes that were not identified as part of the capsaicin biosynthesis pathway. Each of these heatmaps also have a table corresponding to them listing the gene ID and the predicted protein function if available which can also be found in the appendix section in tables A2 through A5. We included them in the appendix section as these heatmaps represent genes that fit the trend in the data and could contain important information. These tables and figures contained a large list of genes, 273 in total, to analyze that we wanted to focus on one specific heatmap that looks the most promising for this report.

Figure 21

Sectioning of Heatmap for areas which Contained Higher Gene Expression in Pungent Peppers vs Non-Pungent Peppers

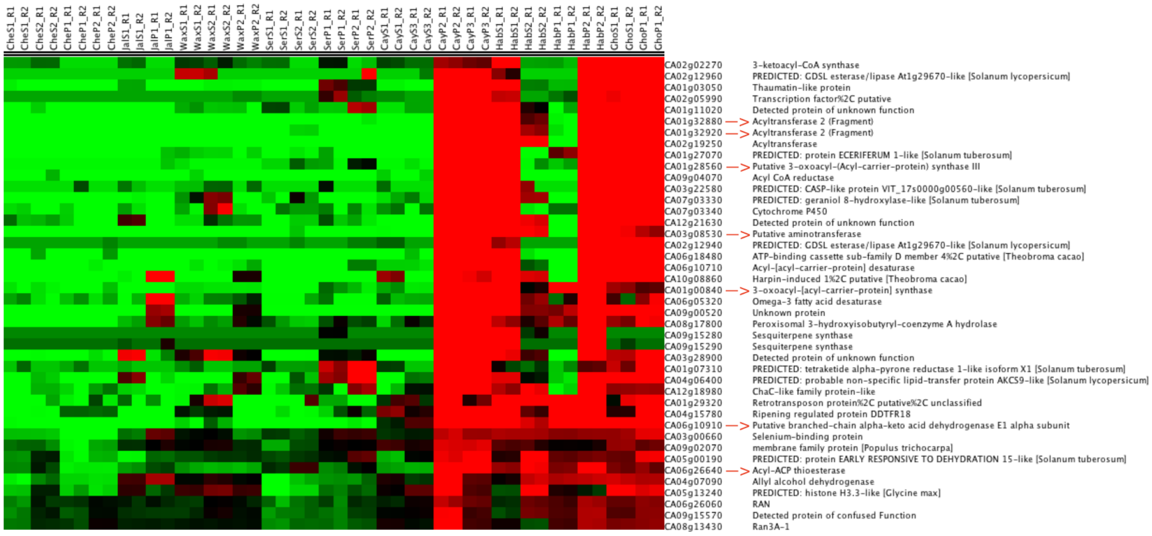


Note. The heatmap is broken down into six different sections which are the major regions where the twenty-six identified genes that had high expression for pungent peppers and low expression for less pungent peppers were found. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation.

The section of the heatmap that is of high interest to us and the section we wanted to focus on is the section below in figure 22. This section is a piece from the heatmap that contains seven of the twenty-six genes that are known to participate in the capsaicin biosynthesis pathway all clustered near each other. The genes that are present from the capsaicin pathway are as follows: CA01g32880, CA01g32920, CA01g28560, CA03g08530, CA01g00840, CA06g10910, and CA06g26640. The proteins that these genes encode are listed in the same order as their gene IDs which are as follows: AT, AT, KasII, pAMT, KasI, BCKDH E1a, and FatA. In figure 22, we can see the gene expression levels of the different peppers listed from low pungency on the left to high pungency on the right. Then, to the right of the expression level visual, listed are the gene IDs and their respective protein function listed next to them. Out of the 11,052 genes present in the heatmap, this small grouping has seven of the genes that are known to participate in the capsaicin biosynthesis pathway. There are a total of forty-two genes shown in figure 22 but these seven all fall within a thirty-one gene span of each other. With these many genes part of the capsaicin biosynthesis pathway being clustered together in such a small section, there could be other genes that may not have yet been identified in the pathway that could reside in this location too.

Figure 22

Heatmap Containing Seven of the Twenty-Six Genes Known to Participate in Capsaicin Biosynthesis



Note. Visual heatmap containing seven of the twenty-six genes known to participate in the capsaicin biosynthesis pathway with similar expression to our trend. The columns are representative of the 46 pepper samples while the rows are representative of the gene ID and protein functions associated with each ID. Pepper samples are ordered from left to right in increasing pungency. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation. The seven gene IDs that were present out of the twenty-six are identified with a red arrow pointing from the gene ID to the protein function.

Identifying Protein Functions from Heatmap

After identifying this specific section of the heatmap which contained a high clustering of known genes part of the capsaicin biosynthesis pathway grouped with other genes that have unknown functions or are not associated with capsaicin biosynthesis, we wanted to identify if these genes have any associations with capsaicin biosynthesis and possibly identify any functions of the unknown genes too. To start, we researched the literature of these gene IDs to evaluate if there is any known protein functionality for these genes that may have not been associated with the database. After assigning some functionality for a few of these genes based on the literature, we wanted to get a better understanding of these genes by evaluating their sequences and comparing them to other sequences. To resolve this, we looked up the gene IDs in the Sol Genomics Network which provided the annotation file which we used for mapping. Matching the gene IDs from our heatmap to the ones from their database, we were able to obtain the protein sequences for said genes. Using the protein sequences, we performed manual blast protein searches to identify the top hits each protein sequence had and then analyzed those top hits to determine what function our genes may relate to. Using these top hits, we researched them to identify what functions they may be related to and then used these results to determine what most likely our genes would be related to in function. Using the idea of shape dictates function and the shape of the protein is influenced from its sequence of amino acids, by matching our proteins against others with very similar sequences, we can gain a good understanding of what processes our proteins may play a part in. We compiled a table of these genes, their protein functions, and the proposed functions they are related to. We grouped the genes based on factors that are interesting to our project by color coding them six different ways. They are grouped according to their relation to the fatty acid biosynthesis pathway, the phenylpropanoid pathway, both pathways, transcriptional regulation, unknown relationship, and unknown functions.

Table 8

List of Genes from Figure 22 Heatmap with their Protein and Proposed Functions

Gene ID	Protein Function	Proposed Function
CA01g00840	3-oxoacyl-[acyl-carrier-protein] synthase	capsiconiate biosynthesis
CA09g04070	Acyl CoA reductase	long chain fatty acid synthesis
CA08g17800	Peroxisomal 3-hydroxyisobutyryl-coenzyme A hydrolase	fatty acid &beta;-oxidation II (peroxisome) - enoyl-CoA hydratase
CA01g27070	Predicted: protein ECERIFERUM 1-like	cuticular wax biosynthesis - fatty aldehyde decarbonylase
CA01g28560	Putative 3-oxoacyl-(Acyl-carrier-protein) synthase III	fatty acid biosynthesis initiation I - beta;-ketoacyl-ACP synthase
CA02g02270	3-ketoacyl-CoA synthase	very long chain fatty acid biosynthesis I
CA02g12940	TGL1 - GDSL esterase/lipase At1g29670-like	triacylglycerol degradation
CA02g12960	TGL1 - GDSL esterase/lipase At1g29670-like	triacylglycerol degradation
CA04g06400	Predicted: probable non-specific lipid-transfer protein AKCS9-like	Fatty acid synthesis pathway
CA06g05320	Omega-3 fatty acid desaturase	alpha;-linolenate biosynthesis I - linolenoyl-lipid 15-desaturase
CA06g10710	Acyl-[acyl-carrier-protein] desaturase	oleate biosynthesis I - stearoyl-[acyl-carrier-protein] 9-desaturase
CA06g10910	Putative branched-chain alpha-keto acid dehydrogenase E1 alpha subunit	2-oxoisovalerate decarboxylation to isobutanoyl-CoA - 3-methyl-2-oxobutanoate dehydrogenase
CA06g22640	Acyl-Acp thioesterase	thioesterase
CA03g08530	Putative aminotransferase	4-aminobutanoate degradation I - 4-aminobutyrate aminotransferase
CA07g03330	Predicted: geraniol 8-hydroxylase-like [Solanum tuberosum]	omega;- hydroxylation of laurate
CA09g15280	Sesquiterpene synthase	germacrene biosynthesis - germacrene D synthase
CA09g15290	Sesquiterpene synthase	germacrene biosynthesis - (+)-germacrene D synthase
CA07g03340	Cytochrome P450	Oxidation of steroids, fatty acids, and xenobiotics
CA01g32880	Acyltransferase 2 (fragment)	volatile benzenoid biosynthesis I (ester formation) - benzoyl-CoA:benzyl alcohol benzoyltransferase
CA01g32920	Acyltransferase 2 (fragment)	volatile benzenoid biosynthesis I (ester formation) - benzoyl-CoA:benzyl alcohol benzoyltransferase
CA02g19250	Acyltransferase	Phenylpropanoid pathway

Gene ID	Protein Function	Proposed Function
CA03g22580	CASP 1F1	Casparian strip membrane proteins - transmembrane
CA01g03050	Thaumatin-like protein	Natural Sweetener found in Katefe
CA01g07310	Predicted: tetraketide alpha-pyrone reductase 1-like isoform X1	Linked to Sporopollenin formation
CA03g00660	Selenium-binding protein	Binding to Selenium and detoxification under stress conditions
CA04g07090	Allyl alcohol dehydrogenase	conversion of ethanol to acetaldehyde
CA04g15780	Ripening regulated protein DDTRF18	Transmembrane Protein involved in clearing toxic compounds from cells - Protein detoxification 27
CA05g00190	EARLY RESPONIVE TO DEHYDRATION 15-like	Dihydroneopterin aldolase 1 - Biosynthesis of pteridine
CA08g13430	RAan3A-1	Ras GTP binding protein - cell signaling protein involved many cellular processes
CA06g18480	ATP-binding cassette sub-family D member 4	protein transporters
CA10g08860	Harpin-induced 1	membrane protein
CA12g18980	ChaC-like family protein-like	gamma;-glutamyl cycle - Glutathione levels strongly elevated in pungent peppers
CA09g02070	membrane family protein [Populus trichocarpa]	membrane protein
CA09g00520	Unknown protein	No matches
CA09g15570	Detected protein of confused function	No matches
CA12g21630	Detected protein of unknown function	No matches
CA01g11020	Detected protein of unknown function	No matches
CA01g29320	Retrotransposon protein%2C putative%2C unclassified	Maybe Rnase H protein from repeat
CA03g28900	Detected protein of unknown function	No matches

Color Code:
Branched Chain Fatty Acid Pathway
Phenylpropanoid Biosynthesis Pathway
Both Pathways
Transcriptional Regulation
Unknown Relationship
Unknown Function
Known Genes Part of Capsaicin Pathway

Note. These genes from the heatmap section had a strong clustering with the known genes part of the capsaicin biosynthesis pathway. The genes in this list have their respective protein functions and proposed functions listed. Genes known to be present in the capsaicin biosynthesis pathway are highlighted yellow and grouped into different categories based on their protein functions.

Chapter 4

Discussion

Part 1: Genetically Modifying *Saccharomyces cerevisiae* with the Capsaicin Biosynthesis Pathway

Golden Gate Cloning/Screening Process

After Golden Gate cloning, *E. coli* bacteria cells were screened utilizing restriction enzyme digest, PCR, and sanger sequencing to verify if the correct transcriptional units were present. As mentioned previously, the white bacteria colonies that were grown had the potential to contain the correct transcriptional units that we were trying to incorporate since these colonies have been modified. They have been modified in the terms that since they did not grow red, they no longer contained the red fluorescent protein region in the original plasmid meaning they were successfully cut with the *BsaI*-HFV2 restriction enzyme. One form of screening we utilized was taking our purified plasmid DNA from red and white colonies of the pAV113 and pAV115 plasmids and performing a restriction enzyme digest with the *PVUII*-HFV2 enzyme. This specific restriction enzyme is useful as it recognizes two specific sites on a white colony and recognizes four specific sites on an unmodified red colony. The reason for this is that the sequence that it recognizes is found two times in the red fluorescent region of the plasmids and two times outside the red fluorescent region so a red colony which still has its red region would be cut four times in total while a white colony would only be cut twice. In addition, since we are replacing the red fluorescent region with our transcriptional units of interest, and we know the base pair lengths of these promoters, coding sequences, and terminators, we can estimate the band size lengths that are expected to show up if the plasmid does successfully contain each of these units. It was determined that we did indeed have two possible samples that yielded the expected band length sizes that were expected. The band size expectations we were expecting was 5965 and 2732 bp for the pAV115 plasmid and 4462 and 3228 bp size for the pAV113 plasmid. After performing a series of three restriction enzyme digests with *PVUII*-HFV2 on a total of 30 bacterial samples, there were a total of 10 samples that appeared to have a similar banding pattern that was expected for the plasmid. Of the 10 bacterial samples, 6 of them were for the pAV113 plasmid and 4 of them were for the pAV115 plasmid. We continued to interrogate these colonies and other bacterial colonies utilizing polymerase chain reaction and sanger sequencing.

Polymerase chain reaction was then performed utilizing pBluescriptSK and pBluescriptKS forward and reverse primers which are specific sequences that reside outside of the *BsaI*-HFV2 sites and amplify toward each other encompassing the red fluorescent protein region. Since these primers reside outside of the red fluorescent region and amplify towards each other, we can utilize them for both white and red colonies to gain information from the plasmid. Red colonies still contain their red fluorescent protein regions and by looking at the plasmid map, we know the distance between these two primer regions and what size DNA fragment we should expect in an unmodified plasmid. On the other hand, in a modified plasmid, the red fluorescent protein region should be absent since it was cut with *BsaI*-HFV2 and replaced with our transcriptional units. We would expect our DNA fragment size to be the size of our transcriptional units plus the difference between the primer sites and the *BsaI*-HFV2 restriction enzyme sites. The reason for this is that *BsaI*-HFV2 recognized and cut a specific region of the plasmid so that portion is no longer there. However, the region between the primer site and *BsaI*-HFV2 sites are still present and will be present in PCR amplification. Therefore, we add the sizes of these remaining regions with the overall size of our transcriptional insert we implemented to calculate our new overall fragment size. It was expected that the pAV115 plasmid would have a band size around 2500 and the pAV113 plasmid would have a band size around 2100. Of the multiple polymerase chain reactions we performed and gel electrophoresis experiments to visualize the results, we determined that two possible samples appeared highly likely to contain the correct band sizes. These samples were pAV115 *pun1*-E and pAV113 *pAMT*-L. Although there were other samples that could have possibly been in the range of our expected band sizes, these two specific samples did appear most likely to contain the correct transcriptional units.

The final screening experiment we performed was to send out our purified plasmid DNA sequences for sanger sequencing analysis to confirm that our transcriptional units were present in the plasmid. The DNA sequencing results for sanger sequencing are very useful as we can determine if the sequences of our transcriptional units match those found in the sequencing of our samples. One unfortunate downside to sanger sequencing is that it is only effective for sequencing of about 800 or so base pairs as the sequence quality deteriorates beyond that. The base pairs and the length of our transcriptional units of promoter, coding sequence, and terminator all together are about 2322 base pairs for pAV113 plasmid and 1422 for the pAV115 plasmid. Therefore, when we sent out our samples for sequencing, we primarily were interested in the coding sequence to observe if it was present or not and therefore used specific plasmid primers that amplify the region right outside of the *BsaI*-HFV2 cut sites allowing for inclusion of a promoter or terminator and our coding sequence. We sent out a total of five samples for sanger sequencing, one sample was a possible *pun1* candidate, and the other four were possible *pAMT* candidates. Using the blast program, we were able to blast our sanger sequence results against the *Capsicum* genome to determine if our genes of interest were present based on the similarity result. We also directly compared the sequences of our sanger results to the known promoter sequences, coding sequences, and terminator sequences we were using to determine if they were present or not. We were able to identify that four of the five sequences had promising sanger sequencing results where they contained part of our coding sequence and terminator sequence of interest. We were unable to determine if the promoter region was present in the samples because it was outside the range of sanger sequencing capabilities as sanger sequencing is able to sequence about 800 base pairs from its start site. Alternative primers would have to be designed to further explore if the entire coding region was present and promoter region.

Auxotrophic Selection

Using the results from our screening processes, we were able to identify two samples, *pun1*-E pAV115 and *pAMT*-D pAV113 which should contain our genes of interest. We would use these two bacterial samples to perform a yeast transformation to insert our plasmid DNA into a strain of *Saccharomyces cerevisiae.* When selecting the yeast strain to utilize, we wanted to pick a strain that would allow for auxotrophic selection. Each plasmid encodes a specific amino acid, pAV113 encodes a histidine promoter region and pAV115 encodes a leucine promoter region. Therefore, the yeast strain we would like to transform these plasmids into should require histidine and leucine for growth. The yeast strain B4741 was determined to be the appropriate choice as it has the following genotype: MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0. This yeast strain requires the following amino acids and ribonucleic acids for growth which are histidine, leucine, methionine, and uracil. Using this information, we can specifically select which transformed plasmid would be able to grow on what specialized plates if it lacked certain amino acids or ribonucleic acids. The specialized plates would be composed of yeast nitrogen without amino acids, water, sugar, and then added any specific amino acids or ribonucleic acids to each plate.

We successfully grew colonies on each of our experimental plates which can be seen in figure 10. The plate located on the far left consists of SD media, methionine, and uracil and therefore only yeast that contain a combination of both plasmids, pAV113 and pAV115 would be able to grow. The plate located in the middle consists of SD media, histidine, methionine, and uracil and therefore only yeast containing the pAV115 plasmid would be able to grow. The plate on the far right consists of SD media, leucine, methionine, and uracil and therefore only yeast with the pAV113 plasmid would be able to grow. Each of the colonies that grew on the plate suggest that those yeast colonies contain the correct plasmids or plasmid of interest. There is a noticeable difference in colony growth for the different plasmids as some grew better than others. For example, there are few colonies that grew on the *pun1*-E pAV115 plate compared to the *pAMT*-D pAV113 plate and the combination of the two plasmids plate. The colonies also on the combination plasmid plate appear to have grown smaller in size compared to the other two plates. The growth of these colonies did meet our expectations as we predicted there would be no growth on the SD plate that contained no additional macromolecules since the specific strain we were utilizing required four essential macromolecules for growth, and at most the yeast containing both plasmids could only synthesize two macromolecules of the four. We did grow colonies on our expected plate and these colonies should have the correct plasmid inserts. Using these yeast colonies, we can then determine if the genes that we inserted into the plasmids are being expressed or not, which is essential to produce the enzymes needed to complete the capsaicin biosynthesis pathway.

Testing Gene Expression in Genetically Modified Yeast

Following auxotrophic selection, we then performed a follow up analysis experiment to determine if our genetically modified yeast was expressing their newly added genes. This needed to be done to validate that our genes were being expressed because if the yeast are not expressing the genes we inserted, then the proteins these genes encode will not be present and will not complete the capsaicin biosynthesis pathway. To do this, we performed a quantitative reverse transcription polymerase chain reaction experiment. We did this utilizing the RNA of our three different genetically modified yeast which were By4741 with *pAMT*, By4741 with *pun1*, and By4741 with *pAMT* and *pun1*. The extracted RNA from the yeast samples was used to synthesize cDNA and both were used for the polymerase chain reaction. The RNA samples were used as a control since the primers we were utilizing are DNA based primers and no amplification of the RNA samples should occur. This is the same result we would expect for the water samples as they would lack any DNA for amplification to occur. Only the cDNA samples should experience amplification as the primers would recognize the DNA bases during the process. We utilize specific primers for coding sequences which are the *pun1* and *pAMT* genes. These primers are around 20 base pairs long which makes them highly selective and they recognize a specific sequence located on the *pun1* and *pAMT* genes. We also used the *Alg9* housekeeping gene since housekeeping genes are constituently expressed as a positive control to compare our gene expression values against.

The results suggest that the yeast samples that contained the *pAMT* gene had low CT values when tested with the pAMT primer for both the RNA and cDNA samples. These low CT values which are below 20 suggest that the *pAMT* is present in the sample but we are unable to determine its expression levels. Since there is a low CT value for both the RNA and cDNA sample, it means that there is DNA contamination in the RNA strand as the primers we used are DNA sequence primers. There should be no amplification of the RNA as the primers would not recognize a sequence present in that sample. There must be some sort of DNA contamination occurring in the sample which is resulting in primer amplification for the RNA samples. We are able to verify that the *pAMT* gene is present in the sample though because we had amplification of the cDNA samples meaning the gene is present. We ideally would have liked to have seen there to be CT values of 0 or a very high CT value indicating no significance in the RNA and water samples since the primers would not recognize any DNA in these samples.

The results for the *pun1* gene tell us that it does not appear to be present in the genetically modified yeast strain. The yeast samples that contained the *pun1* gene did not result in low CT values, they are relatively high with most of them resulting in a CT reading above 36 for both the RNA and cDNA sample sets. This suggests that the *pun1* gene is not present in the yeast as the primers were not able to identify the *pun1* sequence in the yeast. The important samples to look at are the cDNA samples where we combine the gene with that gene's specific primer sequence. These samples we would expect to see a low CT value because during polymerase chain reaction, we would expect high amplification to occur as we are combining samples that should contain our gene of interest and primers that recognize that gene. This would result in amplification or the synthesis of more DNA which then should result in a higher fluorescence reading due to SYBR dye binding to the DNA leading to a lower cycle threshold as a result. When we look at the samples, we only see that the samples with the pAMT primer and the *pAMT* gene resulted in low CT values or our control of the Alg9 primer in the yeast. The pun1 primer does not appear to amplify the *pun1* gene in the pun1 samples and as a result does not yield a low CT value. This can also be clearly seen in the combination plasmid samples which contain both *pun1* and *pAMT* genes. When these samples are treated with both pun1 and pAMT primers, only the pAMT primer sample resulted in a significant low CT value while the pun1 primer samples were not. We can also evaluate how our samples compare to the housekeeping gene control samples. Each sample of the *Alg9* when treated with the Alg9 primer had low CT values all around the 18-19 range. This was expected as these genes are always being expressed and are necessary for cellular functioning. When we compare the CT value of our *pAMT* gene to these housekeeping genes, the *pAMT* samples obtained a similar but slightly lower CT score compared to the

Alg9 samples. For the *pun1* samples, they are nowhere close to our *Alg9* or *pAMT* samples CT count.

The outcome of this experiment informs us that we have successfully inserted the *pAMT* plasmid into our genetically modified yeast but it appears that the *pun1* plasmid is not present at all. We will have to reinsert the *pun1* plasmid into the strain of yeast and perform the qRT-PCR experiment over again. We would also need to add a DNase step into the experiment to remove any contaminating DNA prior to synthesizing the cDNA from the extracted RNA. This step will help us create more accurate results and get a better sense of how the expression levels compare between our inserted genes and housekeeping genes. This DNase step will help remove any remaining DNA before the RNA is converted to cDNA and we would need a follow up step to remove the DNase before we start synthesizing cDNA as the DNase would also digest newly synthesized cDNA (Añez-Lingerfelt et. al., 2009). The resulting CT values for both the *pAMT* and *Alg9* cDNA samples for this experiment are highly similar which may mean that they are both being expressed to a similar degree. Even though their CT values are similar, we are unable to comment on how the expression of our *pAMT* gene may compare to the *Alg9* gene. The reason for this is the PCR reaction is unable to distinguish between cDNA that was synthesized from reverse transcription and that of contaminating genomic DNA that was not removed in the previous steps. Because of this, genomic DNA contamination in the cDNA samples will result in an overestimation of the amount of RNA present if the primers recognize this contaminating sequence. This amplification will lead to an overestimation which will affect the CT results of this experiment and although they appear to have similar CT values and would possibly possess similar expression levels, this may not be the truth based on the amount of contaminating genomic DNA from the start (Lingerfelt et. al., 2009). What we can say for certain is that the *pAMT* gene appears to be present in our genetically modified yeast and the *pun1* gene does not appear to be present.

Part 2: Performing Illumina RNA Sequencing Analysis on Placenta and Skin Tissue Samples from Seven Different Peppers of Varying Scoville Intensity to Identify Novel Genes for Capsaicin Synthesis

Collection of Peppers for RNA Sequencing

A large number of pepper species were dissected, extracting their placenta and tissue regions, and the RNA from these regions was isolated. Pepper samples were sent out for Illumina RNA sequencing if they had high RNA integrity, high RNA concentrations, and a good balance of non-pungent to pungent peppers according to the Scoville Scale. The following peppers were sent to GENEWIZ for Illumina RNA sequencing: cherry peppers, jalapeno peppers, hungarian wax peppers, serrano peppers, cayenne peppers, habanero peppers, and ghost peppers. GENEWIZ evaluates the RNA samples prior to RNA sequencing analysis and yields a variety of information for the customer about the RNA. RIN^e is an abbreviation for RNA integrity number equivalent which is an algorithm that is calculated to assign an integrity score to an RNA sample. The score is calculated on a scale from 1-10 with 10 being the least degraded or having the highest integrity. Messenger RNA only comprises a small percent of total RNA so it is not readily detectable and instead, ribosomal RNA is measured as it accounts for more than 80% of the RNA with a majority of that comprised by the 28S and 18S rRNA species in mammalian systems (Palmer & Prediger, 2004). Plant tissues are composed of three types of ribosomal RNAs which are cytosolic, chloroplastic, and mitochondrial which all vary in size from 5S to 25S. In addition, greener plant tissues can contain additional ribosomal RNAs in contrast to non-green tissues (Babu & Gassmann., 2016). To account for these differences in RIN^e results, a Bioanalyzer can be used for plants which is able to differentiate these complex types of plant tissues. However, for our samples Genewiz utilized a standard eukaryotic RNA analysis using a Agilent TapeStation system. The RNA Agilent TapeStation system produces the electropherogram which is used to calculate the RIN^e score for each sample and agarose gel electrophoresis images were also created from Genewiz. Genewiz also generated a table of information regarding our pepper samples consisting of concentration, average size, region molarity, and more which can be seen in table 4 below. RIN^e scores that are lower than 6 are highlighted as cautionary samples according to a threshold determined by GeneWiz. These samples have higher RNA degradation than would be desired for RNA sequencing and can result in poor sequencing results. As seen in the results, a majority of the pepper samples appear to have RIN^e scores below 6.0 but many of them fall close to this threshold of 6 with the average score being 5.73. RIN^e is an important value when determining the reliability of the sample to send out for sequencing. Poor quality RNA samples can lead to uneven gene coverage or a sample that is highly degraded may not truly represent gene expression at that time. Therefore, high-quality samples are needed for RNA sequencing as a degraded sample sent out for RNA sequencing will yield inaccurate sequence results. Although our scores are lower than the desired threshold for RNA sequencing, we are sequencing plant tissue, and as previously mentioned, there are discrepancies in RIN^e score for plant tissue samples if a device like the bioanalyzer is not utilized. Since plant tissues are composed of three different types of rRNAs, cytosolic, chloroplastic, and mitochondrial, they are a rather complex type of tissue for the RIN^e values to be determined from. They are more variable in rRNA size (5S, 8S, 16S, 18S, 23S, and 25S) which is important as the RIN^e algorithm calculates the area under these peaks. With this addition of chloroplastic RNA into the mix, instead of the typical two distinct bands, there are multiple RNA bands which will interfere with the algorithm used to calculate the RIN^e score and ultimately, an inaccurate RIN^e score as a result (Kim & Haj-Ahmod, 2016). Ultimately, although our scores are lower than the desired threshold, we went ahead with the processing aspect of the data as we hope that the RIN^e scores are not a true reflection of our sample integrity and that the RNA quality is actually higher than what was recorded which will result in more accurate sequencing results.

Another factor that is typically generated in RNA sequencing to evaluate sample quality is the DV 200 score. DV 200 is a way for researchers to reliably classify degraded RNA by size and remove suitable samples from unsuitable ones. Essentially, DV 200 will evaluate the percentage of fragments that are larger than 200 nucleotides in length and return a numerical percentage value (Matsubara et. al., 2020). GeneWiz marks samples with a DV 200 percentage that falls below 70 for the samples and of the 24 samples, four samples had a DV 200 score that fell below this threshold which was CayP2, WaxP1, WaxP2, and SerS2. The DV 200 value is another way to assess RNA integrity compared to the RIN^e score. According to a study done by Masubara et. al., 2019, they evaluated the accuracy in the DV 200 index for assessing RNA integrity in next-generation sequencing compared to the RIN^e score. Typically, the RNA integrity number equivalent

RIN^e is the widely used method for analyzing RNA integrity and the DV 200 as a quality assessment standard. Masaburas lab compared the RIN^e and DV 200 RNA quality indexes to determine the most suitable RNA index for next generation sequencing. They first assessed the RNA quality by using both previously stated methods, prepared two kinds of sequencing libraries, and then calculated the correlation between each of the RNA quality indexes and the amount of library product. It was determined that the DV 200 calculated value showed a stronger correlation with the amount of library product produced and was a better marker for predicting library production (Matsubara et. al., 2020). Our average DV 200 score for our pepper samples was calculated to be 75.67 on average and only three of the twenty-four samples were flagged for being below the threshold determined by GENEWIZ. Therefore, using this information and the outcome of our DV 200 values, we believe that our library will not suffer from severe RNA degradation and will be a reliable representation of the data. However, we still will keep in mind the samples that had severely low RIN^e scores or low DV 200 scores when processing the data as a few may yield accurate sequencing results, but overall the majority of the samples appear to be fine to proceed with sequencing.

FastQC

The first step when analyzing RNA sequencing data is to perform a qualitative control test on the raw sequence data. The qualitative control test will identify any possible problems that may need addressing in the data as these problems will result in poor results if not processed. When running FastQC, there are eleven outputs that will return as a green check mark, yellow exclamation point, or a red x mark, each indicating the quality of that check going from good to bad respectively. These eleven outputs are as follows: basic statistics, per base sequence quality, per tile sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base N content (could not identify nucleotide), sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content. Two of the more important statistical outputs to evaluate from the FastQC data are the per base sequence quality and the overrepresented sequences. The overrepresented sequences display the sequences (at least 20bp) that occur in more than 0.1% of the total number of sequences. This is important as this output will inform you if there is any form of contamination in the data such as remaining adapter sequences. The per base sequence quality provides the distribution of quality scores across all bases at each position in the read. These quality scores are a representation of the probability that each of the corresponding nucleotides are called correctly. These quality scores are called Phred quality scores and exist in the range of 0-40. A Phred quality score of 10 means there is a 1 in 10 chance the base was called incorrectly and a score of 20 means there was a 1 in 100 chance for an incorrect base call. For Phred scores of 30 and 40, there is a 1 in 1000 and a 1 in 10000 chance a base call is incorrect respectively. Using this information, we can determine the accuracy of our reads and trim the data so that it will include only high Phred scores to yield highly accurate data.

After running the FastQC program on our samples, we can see a lot of information quickly such as many of the samples had good scores for their tile sequence quality, sequence quality score, base N content, and base sequence length distribution. All the samples returned a green check mark for those specific parameters. In addition to good quality data, we can see where the data is fairly poor such as in the base sequence content and sequence duplication columns where all the samples were flagged with a red x mark. The last couple columns such as sequence GC content, overrepresented sequences, and adapter content all have a mix of green, red and mainly yellow suggesting these samples are of lower quality. Due to the possible caution in the overrepresented samples column and after evaluating each of the per base sequence quality Phred scores of the samples, some processing of the raw data should be implemented to get better results. Most of the per base sequence quality statistics came back as a green check mark but it was evident that as the sequencing progressed, the Phred scores started to drop more and more in quality. This is a typical problem when sequencing which is a drop in quality as the sequencing length increases. Generally, all of the samples obtained a fairly high Phred score in the range of 30 and above with only a few select samples having Phred scores that dropped below the 30 threshold when reaching around the 130-150 base pair position in the read. Therefore, we determined that trimming the tail end of the read by 30 base pairs would be a sufficient amount of bases to remove for better quality sequences. In addition to removing poor base pairs at the end of the sequencing position, we determined that we had contamination in our samples based on the overrepresented sequences output. For a majority of the samples, their overrepresented sequence output according to FastQC was of cautionary status, 37 out of 48 samples were labeled either yellow or red. According to the output, we had contamination with our illumina adapters which must still have been present during the sequencing and leading to the overrepresented sequence. When illumina performs RNA sequencing, the RNA strands are randomly fragmented and cDNA is synthesized through reverse transcription of the randomly fragmented RNA strands. Then sequence adapters ligated to the cDNA strand and amplification will occur off of the cDNA. The problem is that the adapters that are present to allow for amplification are sometimes read back and incorporated in the RNA sequencing read which will influence the overrepresented sequences as these are not truly present. To account for this, we can use a program called trimmomatic which allows us to specify the illumina adapters used for sequencing and trim out those adapter sequences if they are found in the raw sequence and remove them. It was apparent that after performing both forms of trimming for all of the samples, trimming the adapters and trimming the base pairs, our overall FastQC output files for the overrepresented sequences and per base sequence quality dramatically improved. In figure 12, there are images of the raw data FastQC status check and then the results after manipulating the raw data with the trimmomatic tool and shortening the per base sequence length. On comparison of the raw data with the processed data, it can be seen that the processed data had overall score improvements in the columns for adapter content, overrepresented sequences, per base sequence content, and per sequence GC content. Specifically, these changes we made to the data yielded better results and this can be seen true as we take the data further and perform mapping utilizing bowtie and comparing how the overall alignment scores are changed based on each of the processing techniques to the raw data. In addition, we performed pair-end RNA sequencing instead of single-end RNA sequencing but we did not combine the forward and reverse runs of the paired-end reads together. Typically these paired-end reads are combined together and are a way to create more accurate reads during the mapping process at the extra cost of time and money. In our experiment, we decided to not combine these two reads together but treat them as two distinct reads. Although combining them together will increase the accuracy of mapping, allowing for better detection of splice junctions and such, we are not too interested in this study detecting splice junctions, insertions, deletions, or mutations. By not combining the paired-end reads, they will now serve as another set of essential replicates and the sequencing of both ends of these fragments increases the coverage of our experiment. The reason for this is that when the DNA is fragmented for sequencing, in single-end the fragmented DNA is only sequenced from one side and only for a certain number of base pairs depending on the sequencing depth. Therefore, the entire fragmented DNA piece is not sequenced as a result and by performing paired-end, more information is gathered from the fragments as it is sequenced from both sides instead of just one. By not combining them and still mapping these paired-end reads separately, we can increase the range and coverage to detect more gene expression information for our experiment.

Bowtie

After utilizing FASTQC, the Bowtie software program was used to align the fastq files to our reference sequence. As previously mentioned, there were three different output files we created after analyzing the data in FastQC. The first was the raw unaltered data, the second was a fastq file which only had its per base sequence length shortened, and the last had its adapters cut and per base sequence length shortened. Each of these file types were mapped to our reference genome and their respective overall alignment scores were generated. Using these three different sets of data, we were able to compare how each processing step affected the overall mapping score compared to the raw data. We saw that after each of the processing steps were performed, the overall bowtie alignment score was improved for all the samples. The raw fastq file resulted in an average alignment score of 0 times at 35.7%, an average alignment score of 1 time at 41.54%, an average alignment score of more than 1 times at 22.76% with an overall alignment score of 64.30%. When we compare this raw alignment score with the alignment score of the trimmed fastq file, it has an average alignment score of 0 times at 29.32%, an average alignment score of 1 time at 45.77%, an average alignment score of more than 1 times at 24.92% with an overall alignment score of 70.69%. We can see that the average overall alignment score increases from 64% to 71% which is a fair increase considering as an average for a total of 48 samples. We see that the average alignment scores of 0 times decreases from the raw run to the trimmed run which is a good sign because this value is a calculation of the total number of reads that were unable to map at all to the reference sequence so a decrease in this value means more reads were able to map after being processed. In addition, the average alignment score of 1 time and more than one time also increases from the raw data to the processed data which is again a positive sign as this is the number of times a read is mapped to the reference genome. Now, when we compare the trimmed fastq file to the cut adapters file which had its per base sequence length shortened and adapters cut out, this resulted in an average alignment score of 0 times at 23.26%, an average alignment score of 1 time at 48.67%, an average alignment score of more than 1 times at 28.07% with an overall alignment score of 76.74%. These values resulted in an even better alignment score than the previous raw data and trimmed processing data. Overall, it is evident that after processing the raw data, the alignment scores increased significantly and the cut adapters fastq file showed the most significant increase in alignment score as seen in table 6. The most noticeable difference we can see after processing the data is in the “Overall Score Improvement (Cut Adapters-Raw)” column which is the difference between the cut adapters bowtie scores and the raw alignment scores for each of the pepper samples. In this section, values that are highlighted green are samples that had the highest score improvement while samples that are highlighted in red are ones that had the lowest score improvement after being processed. This column shows how important the processing of the raw data can be to the RNA-sequencing data because most peppers had a 10% increase in alignment score and a select few pepper samples had double that with a 20% increase in alignment score. Although some cells in this column are colored red, that does not necessarily mean they are bad samples. These red samples in this section mean that those samples were less impacted by the before and after processing techniques as their alignment score is relatively similar before and after. The column that contains the data that we will be using for our next analysis will be the data in the overall cut adapters alignment column. This is a significant column as we can see the overall alignment scores for each of our 48 pepper samples and their respective bowtie scores and this column has the highest scores. The samples in this column contain the darkest green - yellowish samples which was a color scale applied to the following three columns (Overall Raw Alignment, Overall Trimmed Alignment, Overall Cut Adapters Alignment). This color scale would highlight the largest values green, the lowest values red, and values that reside in the middle yellow. Using this gradient of high to low to color code this section, it can be visually seen that this section contains the most green to greenish yellow colored cells. This column of data has been processed by having their adapters cut and having their sequence lengths trimmed and this group of data will be used for the next step since this column contains the samples with the highest bowtie alignment scores. This score and output will be used for the cufflinks assessment as these pepper samples demonstrated the best alignment to their reference genome.

Cufflinks

Cufflinks is a program that assembles transcripts, estimates their abundances, and can test for differential expression and regulation in RNA-seq samples. For our cufflinks output, we specifically obtained a file containing a list of 30,242 genes along with their relative expression levels (FPKM) for each of our 48 samples (Trapnell et. al., 2012). Since we had our reference genome which was labeled with their respective genes and the gene's predicted protein functions, we were able to combine this information to our file to allow for quick identification of highly expressed genes and their functions they may perform. Using this FPKM output file, we would then interrogate the data utilizing R studio to create visualizations to identify important genes or examine the quality of the data in general.

Data Visualization in R

One of the first graphs we decided to create was a scatter plot for our pepper samples. Before we started to create more elaborate graphs identifying highly expressed genes and relationships between our samples, we wanted to make sure that our FPKM output from the cufflinks program seemed accurate. To do this, we decided that a scatter plot matrix that would plot the log2 FPKM values of the forward and reverse runs for that sample would be a valuable way to evaluate the data. We log2 transformed the data as this makes it easier to evaluate the FPKM output of samples on a simpler scale. The reason being is since we performed pair-end sequencing, the forward and the reverse run are essentially duplicates of each other. They were sequenced in the forward direction and in the reverse direction and should have the same output in general and their FPKM values for the forward and reverse direction should be the similar. Since we are using a scatter plot matrix, we can also compare multiple samples against each other at the same time and what we would expect is strong linear relationships between samples that are of the same pepper and tissue type such as between peppers CheS2-R1 and CheS1-R1 and an even stronger linear relationship between pepper samples CheS2-R1 and CheS2-R2 as these are the same tissues of the same type of pepper. The more linear the distribution, the more similar the samples are to one another meaning there is a stronger correlation between the FPKM values of the forward run with the reverse run. The more randomly distributed or non-linear in general, the weaker the correlation is between the forward and the reverse runs meaning that the sample output is less reliable and that sample set should be regarded with caution. We decided to create a total of 7 different scatter plots matrices, one for each of the pepper samples where we plotted their skin and placenta log2 FPKM output for their forward and reverse runs as seen in figure 15. We would then evaluate the R^2 values generated from the graphs and any samples that had a low R^2 value when graphed against their own tissue type, especially if it is the same exact pepper, would be regarded with caution. As we can see from the results section, there were two samples that appear to be outliers compared to the other data. These samples were the WaxP1_R1 and the WaxP1_R2 samples as they had a very low R^2 value of 0.85 when plotted against each other. They had an even lower R^2 value when these samples were plotted against the other wax placenta samples such as WaxP2_R1 and WaxP2_R2 with a score of 0.631 and 0.634 respectively. Out of all of the other samples when comparing their forward and reverse runs against each other, none of them had a R^2 value below 0.96 while hungarian wax had a value of 0.85. In addition, out of all the samples when comparing the same pepper's tissue types, none of the samples had a lower R^2 value below 0.87 while the hungarian wax peppers had the lowest score across their placenta samples at a R^2 value of 0.63. The reason for such a low R^2 value, which can also be visually seen in the scatterplot matrix of these samples plotted against each other, would have to be due to the fact that the WaxP1 data is poor quality data and is different compared to the other hungarian wax samples. This could be an error that extended back to the beginning of the experiments such as the RNA quality of the data set being poor in general. When looking back, the WaxP1 sample set had a poor Rin^e score and a poor DV 200 value. If the RNA sequencing was done with poor quality data, then the mapping of that sequence and anysteps after would result in poor data. Although fastqc can be used to visualize the results of the RNA sequencing data and fix some early warning signs in the data that can be seen such as we did with trimming or removing adapters. However, it cannot always change a bad sample to a good sample and you can only do so much to the sample to pre-process it before continuing down the pipeline. We continued using the WaxP1 data set for creation of other figures and graphs in the beginning to further interrogate the sample set before fully removing it from further forms of data analysis.

Another way to interrogate the data besides the scatterplot matrix was creating a visual bar graph demonstrating the mapped library sizes for each sample and also generating a boxplot of the log2 FPKM values to visualize the distributions across the different samples. The box plot was a useful graphic as it is easy to visualize the averages of FPKM across all samples which is denoted by the line dividing the box into two, the type of skew in the data depending on how the box is divided, and outliers are easily seen as they reside outside the whisker portion. This is a great way to compare differences in the data sets and when we generated the boxplots for all of the pepper samples, we found a similar outcome to the scatter plot graphic. It was seen that again the WaxP1 sample set for the forward and reverse run behaved abnormally compared to the rest of the data sets. For the most part, the average log2 FPKM value for the sample sets ranged anywhere from a value of 0 - 1 while the WaxP1 dataset had a large average value around 4. Many of the samples have the same inter quartile range which is denoted by the shaded box for each sample set, and then their minimum and maximum values are identified by the length of the whisker. Outliers are any plotted points that reside outside of the whisker portion. The WaxP1 sample set has an abnormal whisker distribution as their whisker length in the negative direction is very short while their whisker length going to the positive direction is long and similar to the other sample sets. No other sample set has such a short negative direction whisker and this is suggestive that the WaxP1 sample set has a lot more highly expressed genes as their average FPKM value is a lot higher. They also have such a small negative whisker indicative that their minimum values are not that low so many of their FPKM values are larger compared to the other data sets. All of the peppers have a large number of positive outliers, too many to count as the dotted points are plotted right on top of each other. This is not unexpected as these FPKM values are related to highly expressed genes in the pepper samples which is to be expected in any sample set. There will be genes that are more highly expressed compared to other genes in the data set and our goal is to identify these highly expressed genes and see if there is a correlation between these genes and producing highly pungent peppers.

The library sizes that we utilized to generate the bar graph was collected from the bowtie output of our processed fastq files. The fastq files were processed so that their adapters were trimmed and their per base sequence length was reduced by 30 base pairs resulting in shorter sequence lengths, but higher quality data. Since the data was processed, the overall number of mapped reads output from bowtie were shortened compared to the raw data since we had to trim our samples before mapping them so base pairs were lost in the processing section. The alignments scores were used for each sample which was a number associated with the number of reads mapped back to the reference genome for each pepper sample. The bowtie output for mappeds reads is very important because essentially when we sequence a set of samples, we get a collection of sequences that have no genomic context. They are a long list of nucleotides that at face value do not mean much and we do not know what part of the genome these sequences would belong to. Mapping the reads to a reference genome, the reads are assigned to a specific location in the genome. It is important to note that a larger sequencing depth will generate more reads and this will increase the power to detect differential expression among genes. Using the bowtie output, we generated a bar graph showing the number of mapped reads for each pepper sample set, their forward and reverse reads, because the forward and reverse were identical for each sample. As you can see from figure 14, a majority of the sample sets have a similar number of mapped reads with a few exceptions in the data. The four following samples had a very small number of mapped reads compared to the other samples which were HabP1, JalP1, SerP1, and WaxP1. Again, we can see the WaxP1 sample is falling under poor quality data with having such a small number of mapped reads. The other samples in this small number of mapped reads section are interesting too as when compared with their P2 sample sets, if they have one, those sample sets fall more around the average number of mapped reads size. Again, something like this could be connected back to the poor quality of RNA at the start before performing RNA sequencing as these samples had a much smaller sequence length compared to the others. Since they had such a small sequence length collected from illumina sequencing, then their mapped sample reads were going to be small too, less than 20,000,000 base pairs, since they had less overall reads to map since their RNA sequencing lengths were already so short. The total sequence lengths collected from RNA sequencing can be seen below in table 9 and these short reads are all highlighted red supporting their small sequence depth from the illumina sequencing. On the other hand, the opposite is true for the samples that had a large sequence depth from illumina sequencing. The samples that have very large mapped library sizes were the CayP2, CayS1, HabS1, WaxP2, and WaxS1 which were greater than 40,000,000 base pairs. These sample sets also had some of the largest sequencing depths and are all highlighted a dark green in table 9. The majority of the other samples are highlighted yellow indicating that they fell between 20,000,000 and 40,000,000 base pairs long for their sequencing lengths. The overall sequencing depth of the samples is good, even the samples that have a low sequencing depth below 20,000,000 reads. We should still be able to use the cufflinks output to evaluate the gene expression levels of the samples and come to a conclusion as these are still substantial read lengths. They are just shorter compared to the rest of the data and the statistical power generated from the cufflinks output will be less powerful, but still informative. The opposite goes for the larger datasets where the information gathered from them will be more informative and we would be more confident that the gene expression data is a truer representation of the data. Since there are so many reads, the power to map back to the reference genome is much greater and we should get a solid understanding of these sample sets and how they compare against each other.

Table 9

Illumina RNA-sequencing Read Depth of the 48 Pepper Samples

Pepper Sample	Total Sequence		Pepper Sample	Total Sequence
Che P1 Run1	27194587		Cay P2 Run1	42014619
CheP1 Run2	27194587		Cay P2 Run2	42014619
Che P2 Run1	24310395		Cay P3 Run1	27563358
Che P2 Run2	24310395		Cay P3 Run2	27563358
Che S1 Run1	24469899		Cay S1 Run1	43595331
Che S1 Run2	24469899		Cay S1 Run2	43595331
Che S2 Run1	29804880		Cay S3 Run1	31214834
Che S2 Run2	29804880		Cay S3 Run2	31214834
Gho P1 Run1	26863525		Ser P1 Run1	19020160
Gho P1 Run2	26863525		Ser P1 Run2	19020160
Gho S1 Run1	26850176		Ser P2 Run1	28339488
Gho S1 Run2	26850176		Ser P2 Run2	28339488
Hab P1 Run1	24222612		Ser S1 Run1	26052653
Hab P1 Run2	24222612		Ser S1 Run2	26052653
Hab P2 Run1	32608724		Ser S2 Run1	25816590
Hab P2 Run2	32608724		Ser S2 Run2	25816590
Hab S1 Run1	47424533		Wax P1 Run1	20019634
Hab S1 Run2	47424533		Wax P1 Run2	20019634
Hab S2 Run1	28928231		Wax P2 Run1	42021312
Hab S2 Run2	28928231		Wax P2 Run2	42021312
Jal P1 Run1	17295942		Wax S1 Run1	39400448
Jal P1 Run2	17295942		Wax S1 Run2	39400448
Jal S1 Run1	33886935		Wax S2 Run1	27711941
Jal S1 Run2	33886935		Wax S2 Run2	27711941

Note. These are the sequence lengths for the pepper samples after being processed. Samples are highlighted according to sequence length where large sequences are highlighted green, small sequences are highlighted red, and yellow samples are in between. Under pepper sample, their three-letter abbreviation as the sample ID are as follows: Cay = Cayenne pepper, Che = Cherry pepper, Gho = Ghost pepper, Hab = Habanero pepper, Jal = Jalapeno Pepper, Ser = Serrano pepper, Wax = Hungarian wax pepper. Pepper samples are differentiated with either a S or P after their three-letter abbreviation indicating if they are a skin or placenta sample. The number following S or P will either be a 1,2, or a 3 which denotes which pepper the tissue sample originated from. Forward and reverse runs for each pepper are identified as Run1 and Run2 respectively.

After interrogating the data sets a bit, we decided a principal component analysis and a correlation matrix ordered by hierarchical clustering would be beneficial to visualize more trends in the data. For the principal component analysis, we first generated a scree plot to decide how many statistically significant factors are present in the data. The scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. The value of a scree plot is that it tells you the number of factors that are present in your data and what principal components you should evaluate. Typically, Scree plots demonstrate a sharp reduction in size of their eigenvalues and when this occurs, any factors that fall below this sharp reduction will add relatively little to no information to the graph as they contribute significantly less to any variation observed in the data. Therefore, any values that reside before this reduction in eigenvalues is important to evaluate as these components account for the majority of variation in the data. When we generated our scree plot, there were two main components, three weaker components, and forty-three very weak components. In figure 16 the two main components account for 42% and 16% of the main variation in the data set. The three weaker components account for roughly 8%, 7%, and 6% of the variation in the data set. All in all, these first five components account for a total of 79% of the variation in the samples while the first two components account for 58% of the total variation. We focused on these first two components as they accounted for the majority of the variation in the data. When analyzing the principal components of analysis graph, plotting component 1 against component 2, we can see how the data is grouped of these two components to contribute to over 58% of the variation. The sample points for the forward and reverse runs for that sample are plotted directly on top of each other which was to be expected as these samples are essentially duplicates so they should reside in the same area. Another thing to notice is how the skin samples reside near the skin samples of the same pepper type and placenta samples reside near placenta samples of the same pepper type. There appears to be three different forms of clustering generally speaking. These groups contain the following pepper samples; in the top right corner the cherry and serrano peppers reside, in the middle bottom the habanero and ghost peppers reside, and in the top left the cayenne, hungarian wax, and jalapeno peppers reside. The clustering of the samples together essentially demonstrates how similar those samples are to each other. We see this as the forward and reverse samples are plotted directly on top of each other. For most samples, the placenta and skin samples are clustered relatively close to each other with exception to the WaxP1 samples to the hungarian wax samples and the cayenne placenta samples to the cayenne skin samples. Again the WaxP1 sample set is a bit abnormal as they are located further away from the other hungarian wax placenta samples and even further away from the hungarian wax skin samples. These two sample sets have the greatest difference between all other peppers when comparing the same type of tissue of the same pepper. Since these two samples are so distant from each other, and not clustered with the other hungarian wax samples, the PCA is essentially saying that they are not similar to the other hungarian wax samples and are identified as something different. Another set of samples that are located a decent distance away from each other are the cayenne placenta samples compared to the cayenne skin samples. A majority of the other samples are located fairly close to each other and to verify the clustering of the samples, we generated a correlation matrix using hierarchical clustering method to see how the samples would arrange and their correlation between each other. There were two correlation matrices that we created, one utilizing all of the pepper samples and the other utilizing all of the pepper samples besides the WaxP1 data set. When looking at figure 16 which consisted of all of the pepper samples, we can see that all the forward and reverse samples have a very strong linear correlation except for the WaxP1 samples. This is seen as each of the samples have a dark turquoise shaded color when plotting their forward and reverse runs against each other. The only sample set that has a more unique correlation pattern compared to the other samples is the cherry placenta data set. These samples are the only ones that have a dark turquoise shading for all of their placenta samples. All the other samples only have dark turquoise shading for their forward and reverse runs of the same pepper set while these samples have it for their forward and reverse runs for both their placenta sets meaning that the cherry placenta peppers are all extremely similar to each other. We ideally would have liked to have seen this trend for all of the skin and placenta samples of the same pepper type as this would reinforce the strength of our analysis because it would suggest that the placenta extractions of the same type of pepper but different extraction resulted in highly similar results. In our experiment, the data does not suggest that but it does order the tissue samples right next to each other and also pepper samples are grouped together so we are confident enough in the groupings of our correlation matrix to go on and utilize this data.

Heatmap

Using the data sets, we went ahead and created a visual heatmap of our log2 FPKM values for the 46 pepper samples (excluded the wax placenta 1 data set) to attempt to determine if we could identify any genes that were related to the capsaicin pathway by searching for ones that are highly expressed in the pungent peppers and lowly expressed in the less pungent peppers. When creating our visual heatmap, we used a program called Cluster 3.0 which allowed us to filter and sort and determine what form we wanted our data to be used in the final heatmap creation (Hoon et. al., 2004). It was determined that filtering the data would be ideal as lowly expressed reads would be filtered out as these reads would not give us much information. They were removed by filtering by read depth and then we filtered out samples that did not have much variation to them. The filtering process reduced our total of 30,242 genes down to 11,052 genes that passed the criteria. These genes were then cented to easily visualize the relative expression of each gene compared to one another. As for the method we used to map the genes, we wanted them ordered by hierarchical clustering and wanted them clustered by their average linkage. The output file that was generated was visualized on the Java Treeview program (Saldanha, 2004). The heatmap that was produced held a large amount of information that could have been evaluated in a number of different ways.

We decided to first look for any obvious trends in the data and identify sections that exhibited specific trends or patterns we were looking for. We ideally were interested in genes that had high expression in our pungent peppers and low expression in our less pungent peppers. Any genes that met this criteria were of interest to us and we located a few sections that appeared to meet these trends. There was a large list of genes that could be found in these sections and listing them one by one would take a while and may or may not be meaningful. Although we are searching for genes that are upregulated in pungent peppers, most may not be associated at all with pungency for those peppers. There were a total of 30,242 genes that were mapped from the annotation file and they all serve a specific purpose for a pepper. Just because a gene is upregulated in a spicy pepper vs a less spicy pepper, it does not mean that it will be related to capsaicin production. There were plenty of genes that were upregulated in the less pungent peppers and downregulated in the pungent peppers which is the exact opposite of the trend we are looking for. Then there are many other sections with some more diversified trends such as certain peppers upregulated when others are downregulated and some sections where the trend we were looking for appeared in small pockets. As mentioned previously, when we searched through the entire heatmap, we saw six different sections where the twenty-six genes known to participate in capsaicin biosynthesis fall. A few of these sections were large and easily identifiable by eye with the trend we were looking for while some of the other smaller sections had small parts that exhibited the trend but were not easily visualized without combing through in a more magnified lens. Therefore, looking through the sections that exhibited the trend we were looking for and trying to make a connection between capsaicin production and the gene itself would be difficult. Therefore, we decided the better method to explore our data was to take known genes that are associated with capsaicin production and locate where they fell on our heatmap which could give rise to interesting sections since the heatmap is organized by hierarchical clustering. Although the capsaicin pathway is still not well established, there was a paper released in 2016 by Zhang et. al., called "Discovery of putative capsaicin biosynthetic genes by RNA-Seq and digital expression of analysis of pepper”. This paper was discussed previously and for our study, these authors generated a list of a one hundred and thirty-five genes that are known to be present in the capsaicin biosynthesis pathway (Zhang et. al., 2016). Using this list of specific genes, we searched each of the genes up in our heatmap and identified if they were present or not because many of the genes were removed in the filtering step. All of the genes should have been present using the unfiltered 30,242 heatmap but if those genes did not have much variation or a significant read depth, then they would not tell us much information. Therefore, we decided to filter the data and then search to see if the genes were still present or not from the list of one hundred and thirty-five.

After identifying if the genes were present or not in our heatmap, we then recorded the samples that had a trend similar to our expression we were looking for. Many of the genes from the list were still present and many did have strong variations in expression, but we narrowed down the list to include a total of twenty-six genes that exhibited the relative trend of high expression in pungent peppers and low expression in non pungent peppers. The list of these twenty-six genes can be found in table 7 where they exhibit the trend we were looking for to some extent. Some of them such as BCKDH E1a, KASI, FatA, pAMT, and AT all exhibited relatively strong expression in the pungent peppers and low expression in the less-pungent peppers. Other samples such as CCoAOMT, NADH-GOGAT, C4H, CCR, and HCT all have a little more variation in their expression where there are some high expression levels in a few of the less pungent peppers but were still included as their pungent pepper expression levels were very high too. Using this information, we then constructed multiple heatmaps to evaluate the data.

Four of the five heatmaps are listed in the appendix section and these heatmaps were included as they are ones that show strong trends of high expression in pungent peppers and low expression in less pungent peppers. These are heatmap figures A8 through A10 with their corresponding tables also listed in the appendix section as tables A2 through A5. These heatmaps could contain useful information as there are a large number of genes that exhibit this trend and included after each heatmap figure is a table listing the genes corresponding protein functions if available. Even though we narrowed down the list to twenty-six known genes and obtained the sections they resided in, there was still such a large number of genes to scan through these specific sections. Since the heatmaps were generated through hierarchical clustering, the other genes that were clustered together according to the program must exhibit strong similarities to the genes around them. Therefore, by searching the genes that are known to be part of the capsaicin biosynthesis pathway, we will find genes surrounding them that would exhibit similar expression patterns and may also partake in the pathway.

When we searched all twenty-six genes that appeared to have similarities to the trend we were looking for and are known to be part of the capsaicin biosynthesis pathway, we were looking for the regions these genes resided. To keep things simple, we took sections of the heatmap that spanned eleven genes in total, five above and five below the gene we searched. This means that we would search each of the twenty-six genes in our heatmap, then record the five genes that fell above and five genes that fell below the gene we searched, and compile screenshots of the heatmap for this section. We decided arbitrarily on recording five genes above and below for a total span of eleven for each search, this could have been changed to a broader search of ten, twenty, or whatever number of genes above and below but we chose five to keep the search range relatively small increasing the significance of the results. We then tried to determine if there was any overlap between the regions we recorded by pieces together the sections if they had genes in the same order. The reason we attempted to piece together the sections was to see if a majority of the genes we searched were clustering in specific regions and how many were in that region. The more of these searched genes are located together in the same general region means the more similar those genes are related and the other genes listed in this region have strong similarities to these searched genes. What we found was a section of the heatmap where seven of the twenty-six genes appeared to cluster together in a span of thirty-one genes within each other. This section of the heatmap which can be seen in figure 22, is of high interest to us and may contribute to new identified genes that could relate to capsaicin synthesis. The reason for this is that out of the 11,052 genes that were present in the heatmap, we identified twenty-six that have similar expression trends that are known to contribute to capsaicin synthesis. Of these twenty-six, seven of them all are clustered in one small section, a section that is thirty-one genes long. So, out of the 11,052 possible genes we have a large percent of genes responsible for capsaicin synthesis that exhibit the trend we are looking for, all clustered together in a very small section.

This means that these seven genes out of the twenty-six are highly similar to each other and they were clustered here because of that, but more importantly it means that the other genes listed in this section are highly similar to these known genes and may play a role in capsaicin synthesis. Some important genes that would be interesting to look into are the ones that do not have a known protein function such as *CA01g11020*, *CA12g21630*, *CA09g00520*, *CA03g28900*, *CA09g15570*. These genes all resulted in the literature saying they detected proteins of unknown function with the exception of CA09g15570 which resulted in confused function. All of these genes could use more research into them to identify the role they play and they are extra important as they all resided in this section where many genes from the capsaicin pathway resided. Identifying these gene functions could result in understanding more about the capsaicin pathway and these genes may have a connection to the pathway when more information of their protein function is uncovered. Another thing to consider is genes that are located closely to the known capsaicin genes and seeing the similarities in their protein functions. For example, the gene *CA02g19250* encodes an acyltransferase and it resides below the other two genes *CA01g32880* and *CA01g32920* which both encode for an acyltransferase 2. However, the gene *CA02g19250* that encodes the acyltransferase was not listed in the paper by Zhang et. al. as one of the known genes to participate in the capsaicin biosynthesis pathway. This gene has an extremely similar expression profile as both the acyltransferase 2 genes and is located right next to them but through literature searches and the published paper by Zhang et. al., we do not see it being incorporated as a gene that is part of the capsaicin biosynthesis pathway. This could be true, that it does not partake in the capsaicin biosynthesis pathway or it could also be an oversight as there are a total of twenty-one known acyltransferases listed by Zhang et. al., two are acyltransferase 3s, eleven are acyltransferase 2s, four are acyltransferase 1s, three are acyltransferases, and one is an acyl transferase-like protein. This is a large list of acyltransferases that have been discovered thus far and based on the proximity and similarity that the *CA02g19250* gene exhibits to the other acyltransferases, but its absence in the contribution to the capsaicin synthesis pathway could be an oversight. This specific section of the heatmap could contain promising data that could lead to the identification of other important genes being part of the capsaicin pathway. In addition to this heatmap, there are also the four other heatmaps that had very high expression for pungent peppers and low expression for non-pungent peppers which are all included in the appendix section. There is a lot of data that was generated from the heatmap and we picked out what we thought were some of the most important sections.

Identifying Protein Functions from Heatmap

After generating our heatmap which contained the genes of high interest, we decided to further evaluate these genes identifying their protein and molecular functions to determine if other genes in these sections perform similar roles. The idea was to identify other important genes that were clustered to these known genes that may also play a role in capsaicin biosynthesis. The three main groups we grouped our genes into were if they were related to the branched-chain fatty acid pathway, phenylpropanoid pathway, and capsaicin biosynthesis pathway. The three other groups we clustered our genes into if they did not fit the first groups were if they played a part in transcriptional regulation, had an unknown relationship compared to any of the other groups, or an unknown function. What we found from this further investigation was that the fatty acid biosynthesis pathway is the most prominent feature in pungent peppers. We come to this conclusion because when identifying these functions for the proteins and grouping genes together based on pathways they contribute to, we found a large majority of the genes produce proteins that participate in the fatty acid biosynthesis pathway. Out of the forty-two total genes we investigated from our heatmap section, thirteen of those total genes participate in the fatty acid biosynthesis pathway. We had four genes of those thirteen that are known to participate in the capsaicin biosynthesis that were grouped into the fatty acid biosynthesis pathway. On the other hand, there are only four genes that are listed that participate in the phenylpropanoid biosynthesis pathway, one of those four that is known to participate in capsaicin biosynthesis. Then there are four that are listed to be associated with both pathways, where only two of those four are known to participate in capsaicin biosynthesis. Although this is a small section of the entire heatmap, this section was a very interesting sub suction with the large clustering of known capsaicin related genes so it makes it interesting that there are so many genes with high expression in pungent peppers that produce proteins whose functions are related to the branched-chain fatty acid pathway. This makes the branched-chain fatty acid pathway a particularly interesting pathway to look more closely into, as much of the early research and experiments for our study focused on the phenylpropanoid pathway. It may be that peppers that are highly pungent, producing a large amount of capsaicin may have higher expression of genes in the branched-chain fatty acid pathway. Studies we originally investigated looked into the two important genes *pun1* and *pAMT* which reside in the phenylpropanoid pathway and how in highly pungent peppers *pun1* expression is more of a driving force for capsaicin production while *pAMT* is still needed. Maybe for highly pungent peppers too, high expression of the genes in the branched chain fatty acid pathway also contribute to the overall capsaicin production and limiting the expression of one of these genes could contribute to a reduction in overall capsaicin production like we see in *pun1*. Another interesting feature in this heatmap section is that many of the genes that are upregulated in pungent peppers are transmembrane and stress response genes. Since capsaicin is a noxious stimuli, at least when ingested by mammals, maybe high production of capsaicin molecules in turn also leads to high expression of stress response genes to cope with the production of this molecule. The stress response genes exhibited similar expression of upregulation in pungent peppers so as a response to producing capsaicin, peppers may also increase expression of stress response genes to compensate.

Chapter 5

Conclusion and Future Directions

Part 1: Genetically Modifying *Saccharomyces cerevisiae* with the Capsaicin Biosynthesis Pathway

The overall experiment to induce a strain of *Saccharomyces cerevisiae* to synthesize capsaicin was not achieved but progress has been made that may allow for proper synthesis in the future. It would appear that of the two crucial genes we were inserting into the yeast, *pAMT* and *pun1*, we achieved successful insertion pAMT but not successful insertion or expression of *pun1*. Although it appeared that the *pun1* may have been present at some point due to the series of screening processes we performed, it is possible for the yeast to lose the plasmid after a series of replications. Plasmid stability in yeast is an issue where the yeast can lose or alter the foreign plasmid. There was a large time span between the creation of the yeast and the testing for expression experiments due to closures that arose from covid-19. This large time span resulted in many colonies being inoculated over multiple months to keep the yeast cell line going, so through these multiple generations the *pun1* plasmid could have been lost even though we believed we had a successful colony with plasmid at one point. We believe that the pAMT plasmid has been successfully inserted into our yeast strain but we have to repeat the qRT-PCR experiment to determine the expression level due to contaminating genomic DNA. Contaminating genomic DNA can result in an overrepresentation of the DNA sequence that would result in inaccurate amplification readings compared to a truly isolated RNA sequence.

What we would plan on doing next is to repeat the golden gate transformation experiment for the *pun1* gene and conduct another series of screening experiments to confirm if the transcriptional units are present or not. After confirming this, we would again transform the plasmid into the By4741 strain of yeast and test for expression again. We would make sure to utilize DNase to remove any contaminating genomic DNA from the RNA samples prior to creating their complementary DNA sequences. In the end, we would hope to obtain a strain of yeast that would have both plasmids present in it with expression of said genes. This would allow us to see if the addition of the precursor molecule vanillin would result in synthesis of capsaicin. If this strain of yeast did not synthesize capsaicin then we would have to reevaluate the capsaicin pathway and possibly add other genes that may not be highly expressed in our strain of yeast in the branched-chain fatty acid pathway. The reason is that the phenylpropanoid pathway would be essentially complete as the last steps are vanillin converts to vanillylamine through pAMT which then is combined with 8-methyl-6-nonenoyl-CoA, the last step in the branched-chain fatty acid pathway to synthesize capsaicin. We would predict that there would be missing genes or low expression of previous genes in the branched-chain fatty acid pathway that would result in the failure to synthesize capsaicin if we did have *pun1* and *pAMT* present and expressing. A closer look at the regulation and expression of the genes in this pathway will show insight into what cellular conditions may influence the expression of the genes or what else we would have to add. This is what the future plan would be to further pursue the spicy yeast project and, in addition, any findings from our RNA sequencing experiments of important genes that may not have been identified can also influence what we would want to insert into our yeast.

Part 2: Performing Illumina RNA Sequencing Analysis on Placenta and Skin Tissue Samples from Seven Different Peppers of Varying Scoville Intensity to Identify Novel Genes for Capsaicin Synthesis

The results that we obtained from our RNA-sequencing data appears to be promising as we generated several heatmaps with large lists of genes that appear to have trends of high expression in highly pungent peppers and low expression in lowly pungent peppers. More importantly, one specific heatmap had a large clustering of known associated capsaicin genes present and in this section there may be more possible capsaicin candidate genes that exist. Specifically, there were five genes with unidentified functions present in these sections which are as follows: *CA01g11020*, *CA12g21630*, *CA09g00520*, *CA03g28900*, *CA09g15570*. The identification of the protein functions of these genes would be a valuable asset as since they are clustered closely to known genes in the capsaicin pathway, they have very similar expression patterns and could possibly be related back to the pathway. There are also many genes that were listed in the heatmaps we generated that also have unidentified functions but similar trends in which we were analyzing. There were a total of fifty other genes that had proteins with unidentified, confused, or unknown functions in the several other heatmaps listed figures A8-A10 and tables A2-A5 of the appendix section. Further analysis of these unidentified protein products could shed light onto more genes that may be associated with the capsaicin biosynthesis pathway.

After further analyzing the genes from the interesting heatmap in figure 22, we were able to assign some protein functions and proposed functions for said genes and what parts of the pathways these genes may be associated with. We discovered that out of the forty-two genes from the heatmap, there appears to be a large number of genes that are associated with the branched-chain fatty acid pathway. Out of the seven known genes to be associated with the capsaicin biosynthesis pathway, five of those genes are associated with the branched-chain fatty acid pathway, one of those genes are associated with the phenylpropanoid pathway, and lastly two of those genes are associated with both pathways. Discovering that in this section of the heatmap, there is such a large number of genes that contribute to the branched-chain fatty acid pathway is interesting to our next direction for research as we would like to incorporate a fatty acid biosynthesis regulating protein in to our *S. cerevisiae* to help induce long chain fatty acid biosynthesis. The heatmap tells us that apparently for very pungent peppers, high expression of the branched-chain fatty acid pathway occurs and since these peppers produce high levels of capsaicin, then we should attempt to incorporate a regulating protein for the fatty acid biosynthesis into our yeast as high expression of these genes help with capsaicin biosynthesis. Furthermore, we would like to evaluate the several other heatmaps we produced, evaluating the genes and identifying their protein functions, molecular functions, and grouping them into clusters to identify what pathways they are most likely associated with. Then taking this larger list into consideration, we can gain a better understanding of the distribution of genes and their relatedness to the different pathways for highly pungent peppers overall.

References

Fitzgerald, M. (1983). Capsaicin and sensory neurones—a review. Pain, 15(1-4), 109-130.

Ross, R. A. (2003). Anandamide and vanilloid TRPV1 receptors. British journal of pharmacology, 140(5), 790-801.

Winter, J., Bevan, S., & Campbell, E. A. (1995). Capsaicin and pain mechanisms. British journal of anaesthesia, 75(2), 157-168.

Caterina, M. J., Schumacher, M. A., Tominaga, M., Rosen, T. A. (1997). The capsaicin receptor: A heat activated-ion channel in the pain pathway. Nature. 389(6653): 816-824.

Sanatombi, K., & Sharma, G. J. (2008). Capsaicin content and pungency of different Capsicum spp. cultivars. Notulae Botanicae Horti Agrobotanici Cluj-Napoca, 36(2), 89-90.

Frias, B., Merighi, D. (2016). Capsaicin, nociception and pain. Molecules. 21(6): 797; doi:10.3390/molecules21060797

Bosland, P. W., Votava, E. J., & Votava, E. M. (2012). Peppers: vegetable and spice capsicums (Vol. Cabi.).

O'Neill, J., Brock, C., Olesen, A. E., Andresen, T., Nilsson, M., & Dickenson, A. H. (2012). Unravelling the mystery of capsaicin: a tool to understand and treat pain. Pharmacological reviews, 64(4), 939–971. https://doi.org/10.1124/pr.112.006163

Singletary, K. (2011). Red pepper: overview of potential health benefits. Nutrition Today, 46(1), 33-47.

Nolano, M., Simone, D. A., Wendelschafer-Crabb, G., Johnson, T., Hazen, E., & Kennedy, W. R. (1999). Topical capsaicin in humans: parallel loss of epidermal nerve fibers and pain sensation. Pain, 81(1-2), 135-145.

Mason, L., Moore, R. A., Derry, S., Edwards, J. E., & McQuay, H. J. (2004). Systematic review of topical capsaicin for the treatment of chronic pain. Bmj, 328(7446), 991.

Szolcsanyi, J. (1977). A pharmacological approach to elucidation of the role of different nerve fibres and receptor endings in mediation of pain. Journal de physiologie, 73(3), 251-259.

Sharma, S. K., Vij, A. S., & Sharma, M. (2013). Mechanisms and clinical uses of capsaicin. European journal of pharmacology, 720(1-3), 55-62.

Yoshioka, M., Lim, K., Kikuzato, S., Kiyonaga, A., Tanaka, H., Shindo, M., & Suzuki, M. (1995). Effects of red-pepper diet on the energy metabolism in men. Journal of nutritional science and vitaminology, 41(6), 647-656.

Yoshioka, M., St-Pierre, S., Suzuki, M., & Tremblay, A. (1998). Effects of red pepper added to high-fat and high-carbohydrate meals on energy metabolism and substrate utilization in Japanese women. British Journal of Nutrition, 80(6), 503-510.

Lin, C. H., Lu, W. C., Wang, C. W., Chan, Y. C., & Chen, M. K. (2013). Capsaicin induces cell cycle arrest and apoptosis in human KB cancer cells. BMC complementary and alternative medicine, 13(1), 1-9.

Wang, D., & Bosland, P. W. (2006). The genes of Capsicum. HortScience, 41(5), 1169-1187.

Heiser Jr, C.B.; Pickersgill, B. (1969). "Names for the Cultivated Capsicum Species (Solanaceae)". Taxon. 18 (3): 277–283. doi:10.2307/1218828. JSTOR 1218828.

Davis, C. B., Markey, C. E., Busch, M. A., & Busch, K. W. (2007). Determination of capsaicinoids in habanero peppers by chemometric analysis of UV spectral data. Journal of agricultural and food chemistry, 55(15), 5925-5933.

Zhang, Z. X., Zhao, S. N., Liu, G. F., Huang, Z. M., Cao, Z. M., Cheng, S. H., & Lin, S. S. (2016). Discovery of putative capsaicin biosynthetic genes by RNA-Seq and digital gene expression analysis of pepper. Scientific reports, 6(1), 1-14.

Ogawa, K., Murota, K., Shimura, H., Furuya, M., Togawa, Y., Matsumura, T., & Masuta, C. (2015). Evidence of capsaicin synthase activity of the Pun1-encoded protein and its role as a determinant of capsaicinoid accumulation in pepper. BMC plant biology, 15(1), 1-10.

FAOSTAT, R. (2017). FAOSTAT database. Food Agric. Organ. UN.

Mueller, L. A., Solow, T. H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., ... & Tanksley, S. D. (2005). The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond. Plant physiology, 138(3), 1310-1317.

Stewart Jr, C., Kang, B. C., Liu, K., Mazourek, M., Moore, S. L., Yoo, E. Y., ... & Jahn, M. M. (2005). The Pun1 gene for pungency in pepper encodes a putative acyltransferase. The Plant Journal, 42(5), 675-688.

Dittrich, F., Zajonc, D., Hühne, K., Hoja, U., Ekici, A., Greiner, E., ... & Schweizer, E. (1998). Fatty acid elongation in yeast: Biochemical characteristics of the enzyme system and isolation of elongation‐defective mutants. European Journal of Biochemistry, 252(3), 477-485.

Tehlivets, O., Scheuringer, K., & Kohlwein, S. D. (2007). Fatty acid synthesis and elongation in yeast. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1771(3), 255-270.

Athenstaedt, K., Zweytick, D., Jandrositz, A., Kohlwein, S. D., & Daum, G. (1999). Identification and characterization of major lipid particle proteins of the yeast Saccharomyces cerevisiae. Journal of bacteriology, 181(20), 6441-6448.

Johnson, D. R., Knoll, L. J., Levin, D. E., & Gordon, J. I. (1994). Saccharomyces cerevisiae contains four fatty acid activation (FAA) genes: an assessment of their role in regulating protein N-myristoylation and cellular lipid metabolism. Journal of Cell Biology, 127(3), 751-762.

Agmon, N., Mitchell, L. A., Cai, Y., Ikushima, S., Chuang, J., Zheng, A., ... & Boeke, J. D. (2015). Yeast Golden Gate (yGG) for the efficient assembly of S. cerevisiae transcription units. ACS synthetic biology, 4(7), 853-859.

Palmer, M., & Prediger, E. (2004). Assessing RNA quality. Ambion TechNotes, 11(1).

Babu, C. V. S., & Gassmann, M. (2016). Assessing integrity of plant RNA with the Agilent 2100 Bioanalyzer System. Waldbronn: Agilent Technologies.

Noble, R. T., & Fuhrman, J. A. (1998). Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquatic Microbial Ecology, 14(2), 113-118.

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

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 7(3), 562.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences, 74(12), 5463-5467.

Kim, W. S., & Haj-Ahmod, Y. (2016). Evaluation of Plant RNA Integrity Number (RIN) Generated Using an Agilent BioAnalyzer 2100. Application Note 80 Plant/Fungi RNA Sample Preparation-Norgen Biotek Corp.

Matsubara, T., Soh, J., Morita, M., Uwabo, T., Tomida, S., Fujiwara, T., ... & Hirasawa, A. (2020). DV200 index for assessing RNA integrity in next-generation sequencing. BioMed research international, 2020.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. Nucleic acids research, 36(suppl_2), W5-W9.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., ... & Lempicki, R. A. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic acids research, 35(suppl_2), W169-W175.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl_1), D258-D261.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: database for annotation, visualization, and integrated discovery. Genome biology, 4(9), 1-11.

Chan, E. Y. (2005). Advances in sequencing technology. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 573(1-2), 13-40.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews genetics, 12(2), 87-98.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. Cold Spring Harbor Protocols, 2015(11), pdb-top084970.

Jordt, S. E., & Julius, D. (2002). Molecular basis for species-specific sensitivity to “hot” chili peppers. Cell, 108(3), 421-430.

Bennett, D. J., & Kirby, G. W. (1968). Constitution and biosynthesis of capsaicin. Journal of the Chemical Society C: Organic, 442-446.

Sukrasno, N., & Yeoman, M. M. (1993). Phenylpropanoid metabolism during growth and development of Capsicum frutescens fruits. Phytochemistry, 32(4), 839-844.

Suzuki, T., Kawada, T., & Iwai, K. (1981). Biosynthesis of acyl moieties of capsaicin and its analogues from valine and leucine in Capsicum fruits. Plant and Cell Physiology, 22(1), 23-32.

Karathia, H., Vilaprinyo, E., Sorribas, A., & Alves, R. (2011). Saccharomyces cerevisiae as a model organism: a comparative study. PloS one, 6(2), e16015.

Su S, Law CW, Ah-Cann C, Asselin-Labat M, Blewitt ME, Ritchie ME (2017). “Glimma: interactive graphics for gene expression analysis.” Bioinformatics, 33(13), 2050-2052.

Wei T, Simko V (2021). R package "corrplot": Visualization of a Correlation Matrix. (Version 0.89), https://github.com/taiyun/corrplot.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Team, R. C. (2013). R: A language and environment for statistical computing.

Sarkar, D. (2008). Lattice: multivariate data visualization with R. Springer Science & Business Media.

Smyth, G. K. (2005). Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor (pp. 397-420). Springer, New York, NY.

De Hoon, M. J., Imoto, S., Nolan, J., & Miyano, S. (2004). Open source clustering software. Bioinformatics, 20(9), 1453-1454.

Saldanha, A. J. (2004). Java Treeview—extensible visualization of microarray data. Bioinformatics, 20(17), 3246-3248.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., ... & Zhang, Z. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. Proceedings of the National Academy of Sciences, 111(14), 5135-5140.

Añez-Lingerfelt, M., Fox, G. E., & Willson, R. C. (2009). Reduction of DNA Contamination in RNA Samples for RT-PCR using Selective Precipitation by Compaction Agents. Analytical biochemistry, 384(1), 79.

Andrews, S. (2017). FastQC: a quality control tool for high throughput sequence data. 2010.

Ghosh, S., & Chan, C. K. K. (2016). Analysis of RNA-Seq data using TopHat and Cufflinks. In Plant Bioinformatics (pp. 339-361). Humana Press, New York, NY.

Frank, C. G., & Aebi, M. (2005). ALG9 mannosyltransferase is involved in two different steps of lipid-linked oligosaccharide biosynthesis. Glycobiology, 15(11), 1156-1163.

Appendix

Supplemental Material

Figure A1

Scatterplot Matrix of the Serrano Pepper Log2 FPKM Values

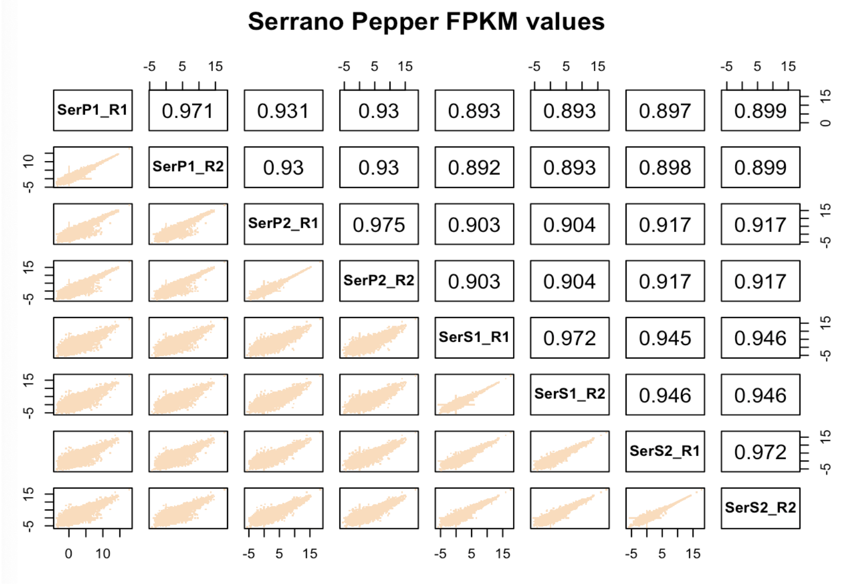


Figure A2

Scatterplot Matrix of the Cayenne Pepper Log2 FPKM Values

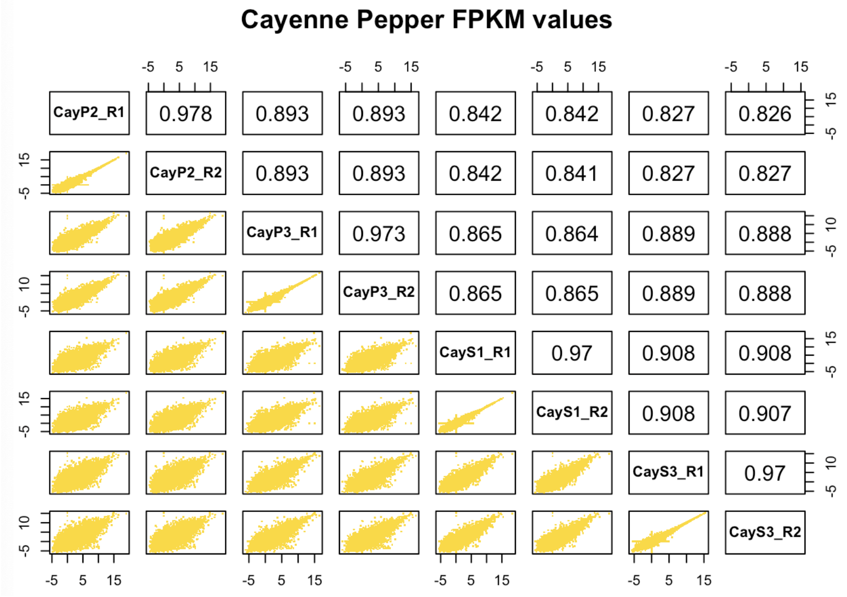


Figure A3

Scatterplot Matrix of the Ghost Pepper Log2 FPKM Values

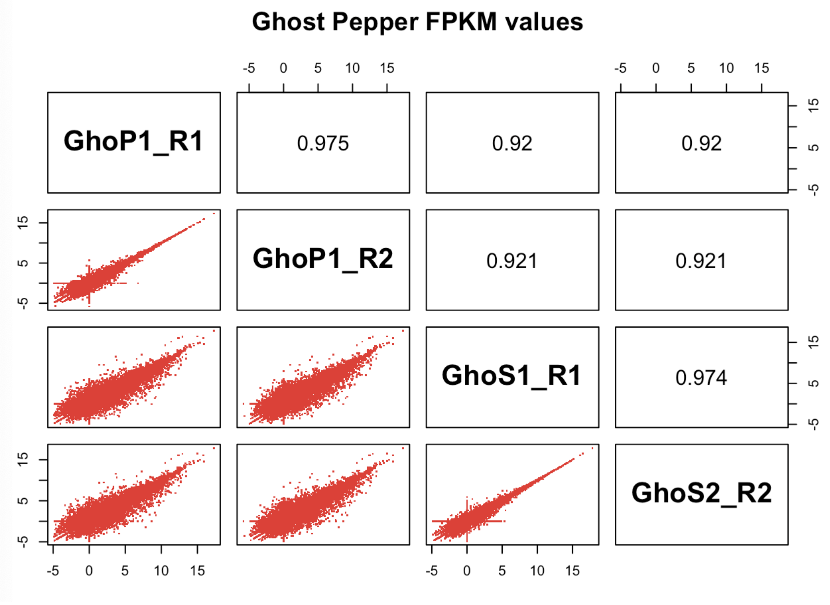


Figure A4

Scatterplot Matrix of the CherryPepper Log2 FPKM Values

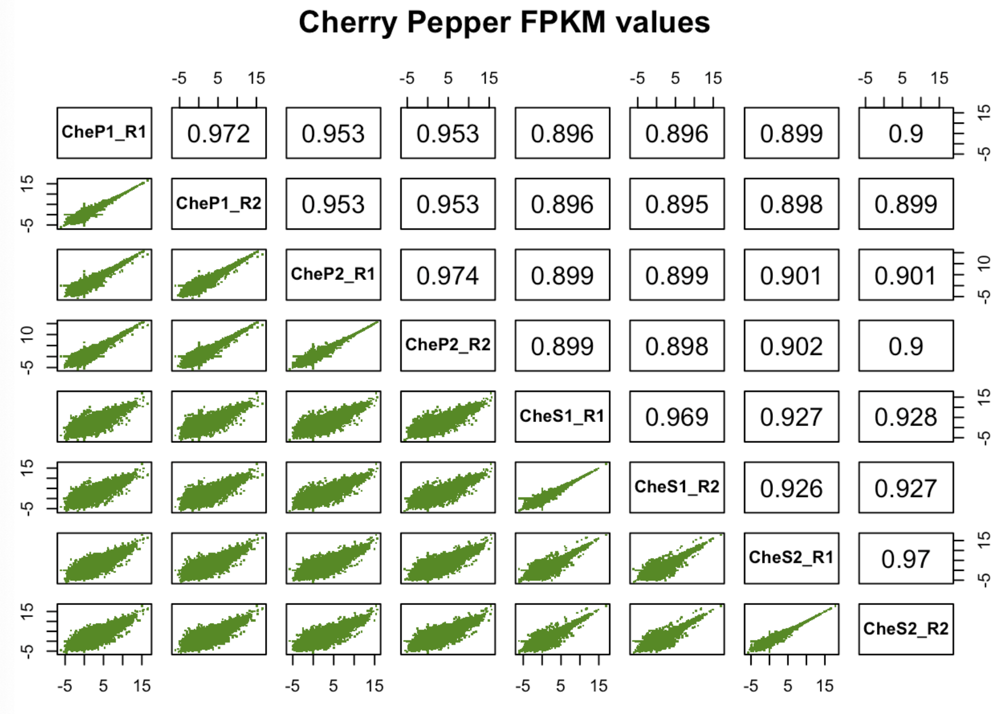


Figure A5

Scatterplot Matrix of the JalapenoPepper Log2 FPKM Values

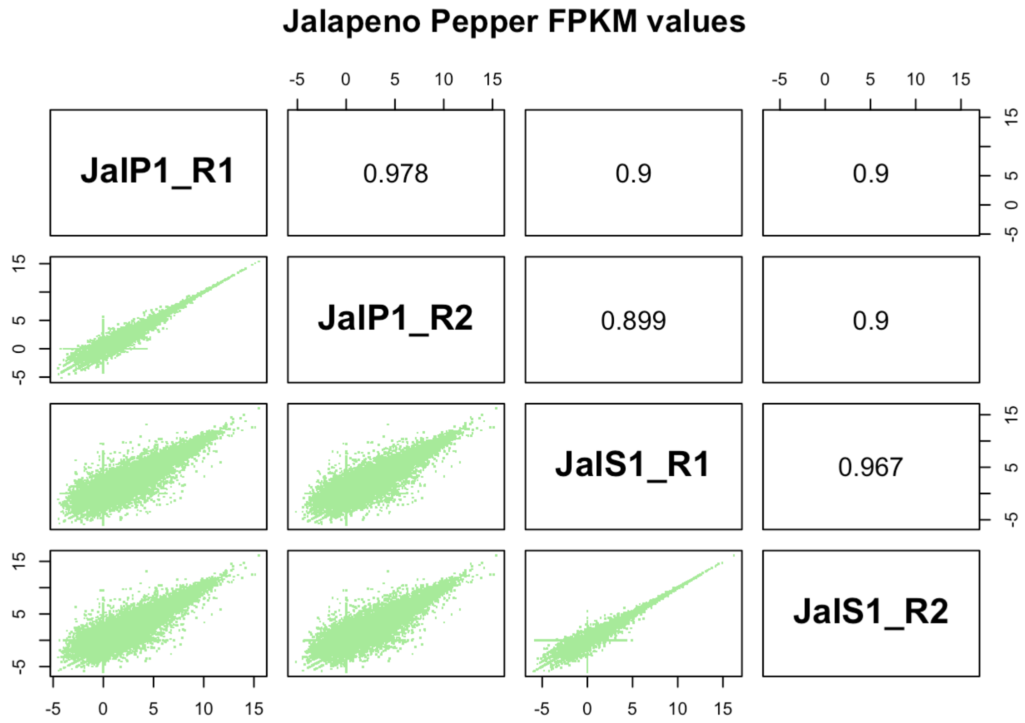


Figure A6

Java Treeview Heatmap for the Unfiltered Data of the Log2 FPKM Values from the 46 Pepper Samples

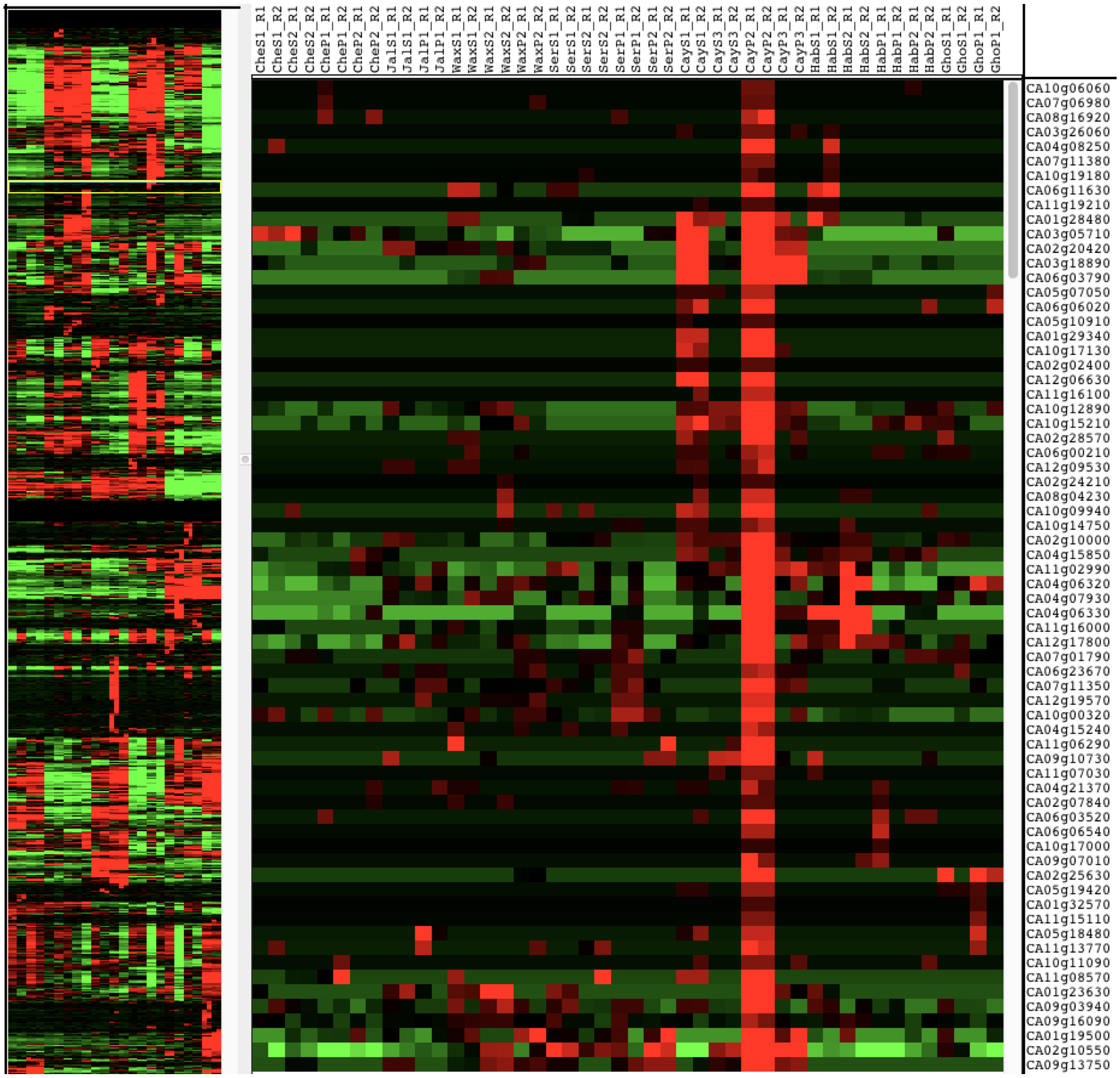


Table A1

List of 135 Genes that were Identified to be Associated with the Capsaicin Biosynthesis Pathway from Zhang Z. X. et. al.

Enzyme	Abbreviation	GeneID	Present?
Chorismate mutase	*CM1*	CA02g27850	Yes
Prephenate aminotransferase	*PAT*	CA12g09590	Yes
Arogenate dehydratase	*ADT*	CA11g14180	No
Arogenate dehydratase	*ADT*	CA02g18350	No
Arogenate dehydratase	*ADT*	CA06g23200	Yes
Phe ammonia-lyase	*PAL*	CA00g95510	No
Phe ammonia-lyase	*PAL*	CA05g20790	Yes
Phe ammonia-lyase	*PAL*	CA09g02420	Yes
Phe ammonia-lyase	*PAL*	CA10g12380	Yes
Phe ammonia-lyase	*PAL*	CA09g02410	Yes
Gln synthetase	*GS2*	CA01g24340	Yes
NADH-dependent Glu synthase	*NADH-GOGAT*	CA02g10690	No
NADH-dependent Glu synthase	*NADH-GOGAT*	CA03g19580	Yes
Cinnamate 4-hydroxylase	*C4H*	CA06g25930	Yes
Cinnamate 4-hydroxylase	*C4H*	CA06g25940	Yes
4-Coumaroyl-CoA ligase	*4CL*	CA03g30500	Yes
Hydroxycinnamoyl transferase	*HCT*	CA03g30250	Yes
p-Coumaroyl shikimate/quinate 3-hydroxylase	*C3H*	CA08g09680	Yes
Cytochrome P450 reductase	*CPR*	CA04g12460	Yes
Cytochrome P450 reductase	*CPR*	CA04g16280	Yes
Caffeoyl-CoA 3-O-methyltransferase	*CCoAOMT*	CA00g52190	No
Caffeoyl-CoA 3-O-methyltransferase	*CCoAOMT*	CA02g14450	Yes
Caffeoyl-CoA 3-O-methyltransferase	*CCoAOMT*	CA02g14470	Yes
Caffeoyl-CoA 3-O-methyltransferase	*CCoAOMT*	CA02g14460	Yes
Caffeoyl-CoA 3-O-methyltransferase	*CCoAOMT*	CA00g18340	No

Enzyme	Abbreviation	GeneID	Present?
S-Adenosylmethionine synthetase	*SAMSyn*	CA09g01970	Yes
S-Adenosylmethionine synthetase	*SAMSyn*	CA10g15500	Yes
Cinnamoyl-CoA reductase	*CCR*	CA04g23420	Yes
Cinnamoyl-CoA reductase	*CCR*	CA03g32090	Yes
Cinnamoyl-CoA reductase	*CCR*	CA08g13650	Yes
Cinnamoyl-CoA reductase	*CCR*	CA03g32100	No
Cinnamoyl-CoA reductase	*CCR*	CA00g69270	No
Cinnamoyl-CoA reductase	*CCR*	CA02g29680	No
Cinnamoyl-CoA reductase	*CCR*	CA01g16660	No
Cinnamyl alcohol dehydrogenase	*CAD*	CA08g04910	Yes
Cinnamyl alcohol dehydrogenase	*CAD*	CA08g04950	No
Cinnamyl alcohol dehydrogenase	*CAD*	CA00g51870	No
Cinnamyl alcohol dehydrogenase	*CAD*	CA06g10220	Yes
Cinnamyl alcohol dehydrogenase	*CAD*	CA02g03280	Yes
Cinnamyl alcohol dehydrogenase	*CAD*	CA12g14850	Yes
Cinnamyl alcohol dehydrogenase	*CAD*	CA00g84360	No
Cinnamyl alcohol dehydrogenase	*CAD*	CA02g00320	Yes
Putative aminotransferase	*pAMT*	CA12g10090	Yes
Putative aminotransferase	*pAMT*	CA03g08530	Yes
Thr deaminase1	*TD*	CA00g84990	No
Acetolactate synthase	*ALS*	CA04g12110	Yes
Acetohydroxyacid reductoisomerase	*AHRI*	CA07g14810	Yes
Dihydroxyacid dehydratase	*DHAD*	CA05g17070	No
Isopropylmalate synthase	*IPMS*	CA01g29060	Yes
Isopropylmalate synthase	*IPMS*	CA06g08090	Yes
Isopropylmalate dehydrogenase	*IPMDH*	CA02g23730	Yes
Isopropylmalate dehydrogenase	*IPMDH*	CA02g23740	Yes
Isopropylmalate dehydrogenase	*IPMDH*	CA11g00650	Yes
Branched-chain amino acid aminotransferase	*BCAT*	CA04g16630	Yes

Enzyme	Abbreviation	GeneID	Present?
Branched-chain amino acid aminotransferase	*BCAT*	CA04g16660	Yes
Branched-chain amino acid aminotransferase	*BCAT*	CA12g15820	No
a-Ketoacid decarboxylase E1a	*BCKDH E1a*	CA06g10910	Yes
a-Ketoacid decarboxylase E1b	*BCKDH E1b*	CA01g17970	Yes
Dihydrolipoamide transacylase	*BCKDH E2*	CA01g18360	Yes
Dihydrolipoamide dehydrogenase	*BCKDH E3*	CA12g21080	Yes
Pyruvate dehydrogenase E1a	*PDH E1a*	CA07g07490	Yes
Pyruvate dehydrogenase E1b	*PDH E1a*	CA05g07180	Yes
Pyruvate dehydrogenase E1b	*PDH E1a*	CA03g22370	Yes
Dihydrolipoamide acetyltransferase	*PDH E1b*	CA04g11620	Yes
Dihydrolipoamide dehydrogenase	*PDH E1b*	CA00g80920	No
Pyruvate dehydrogenase E1a	*PDH E1b*	CA05g08380	Yes
Pyruvate dehydrogenase E1b	*PDH E1b*	CA08g03630	Yes
Pyruvate dehydrogenase E1b	*PDH E1b*	CA06g21080	No
Pyruvate dehydrogenase E1a	*PDH E2*	CA09g17310	No
Pyruvate dehydrogenase E1b	*PDH E3*	CA11g14500	Yes
a-Carboxyltransferase	*α-CT*	CA09g06200	Yes
Biotin carboxylase	*BC*	CA08g00580	Yes
Biotin carboxyl carrier protein	*BCCP*	CA00g87420	No
Biotin carboxyl carrier protein	*BCCP*	CA06g18470	Yes
b-Carboxyltransferase	*β-CT*	CA02g01350	No
b-Carboxyltransferase	*β-CT*	CA04g12500	No
b-Carboxyltransferase	*β-CT*	CA00g79900	No
Malonyl-CoA:ACP transacylase	*MCAT*	CA00g33160	No
Malonyl-CoA:ACP transacylase	*MCAT*	CA00g33150	No
Ketoacyl-ACP synthase I	*KasI*	CA02g11910	Yes
Ketoacyl-ACP synthase I	*KasI*	CA01g00840	Yes
Ketoacyl-ACP synthase I	*KasI*	CA06g02660	No
Ketoacyl-ACP synthase I	*KasI*	CA07g11150	Yes

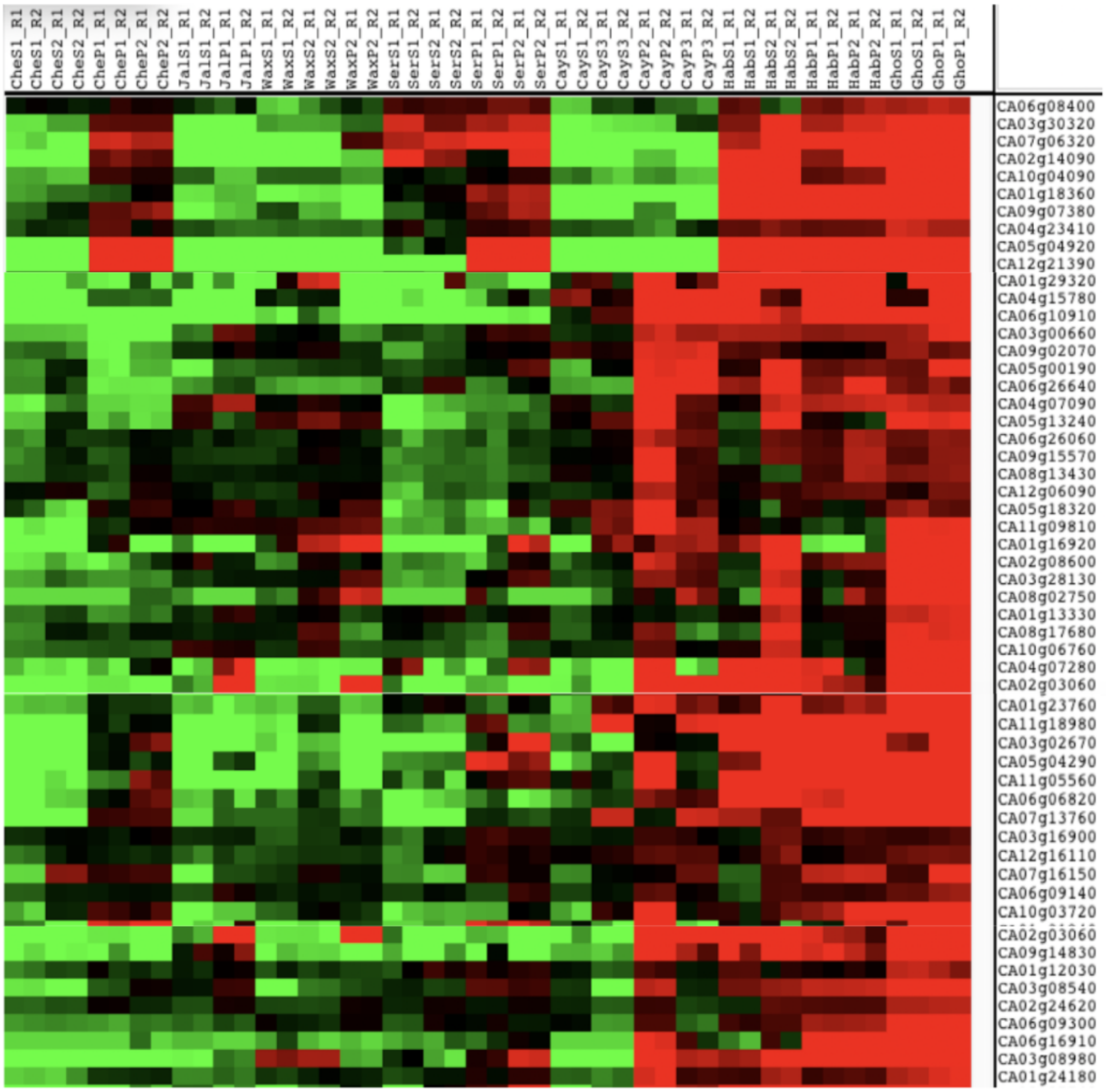
Enzyme	Abbreviation	GeneID	Present?
Ketoacyl-ACP synthase II	*KasII*	CA07g05890	Yes
Ketoacyl-ACP synthase II	*KasII*	CA03g35350	Yes
Ketoacyl-ACP synthase II	*KasII*	CA03g12190	Yes
Ketoacyl-ACP synthase II	*KasII*	CA09g14750	No
Ketoacyl-ACP synthase III	*KasIII*	CA01g28560	Yes
Ketoacyl-ACP synthase III	*KasIII*	CA10g11450	No
Ketoacyl-ACP synthase III	*KasIII*	CA10g11480	No
Ketoacyl-ACP reductase	*KR*	CA08g14900	Yes
Ketoacyl-ACP reductase	*KR*	CA10g21800	No
Ketoacyl-ACP reductase	*KR*	CA01g24640	Yes
Hydroxyacyl-ACP dehydratase	*DH*	CA08g15600	Yes
Enoyl-ACP reductase	*ENRa*	CA10g20920	Yes
Enoyl-ACP reductase	*ENRa*	CA01g27690	No
Acyl carrier protein	*ACLd*	CA01g27220	Yes
Acyl carrier protein	*ACLd*	CA05g00980	Yes
Acyl carrier protein	*ACLd*	CA06g11050	Yes
Acyl carrier protein	*ACLd*	CA12g01130	Yes
Acyl carrier protein	*ACLd*	CA03g31150	Yes
Acyl-CoA synthetase	*ACS*	CA01g01120	Yes
Acyl-CoA synthetase 9	*ACS*	CA04g10340	No
Acyl-CoA synthetase	*ACS*	CA07g08100	Yes
Acyl-CoA synthetase	*ACS*	CA03g18160	Yes
Acyl-CoA synthetase 7	*ACS*	CA08g12160	No
Acyl-CoA synthetase 8	*ACS*	CA00g74260	No
Acyl-CoA synthetase 2	*ACS*	CA08g18140	Yes
Acyl-CoA synthetase 1	*ACS*	CA01g22440	No
Acyl-CoA synthetase 4	*ACS*	CA08g08360	No
Acyl-CoA synthetase 2	*ACS*	CA08g18150	No
Acyl-CoA synthetase 4	*ACS*	CA01g34500	No

Enzyme	Abbreviation	GeneID	Present?
Acyl-ACP thioesterase	*FatB*	CA09g15720	Yes
Acyl-ACP thioesterase	*FatA*	CA06g26640	Yes
Acyltransferase 3	*AT*	CA02g31200	No
Acyltransferase 3	*AT*	CA02g19260	Yes
acyltransferase 1	*AT*	CA11g05940	Yes
acyltransferase 2	*AT*	CA01g32810	No
acyltransferase 2	*AT*	CA01g32940	No
acyltransferase 2	*AT*	CA01g32950	No
acyltransferase 2	*AT*	CA01g32880	Yes
acyltransferase 2	*AT*	CA01g32860	No
acyltransferase 2	*AT*	CA01g32960	No
acyltransferase 1	*AT*	CA03g34400	No
acyltransferase	*AT*	CA02g19270	No
acyltransferase 2	*AT*	CA01g32820	Yes
acyltransferase 2	*AT*	CA01g09180	No
acyltransferase 2	*AT*	CA00g98730	No
acyltransferase 1	*AT*	CA05g20230	No
acyltransferase 2	*AT*	CA01g32920	Yes
acyltransferase 1	*AT*	CA05g20240	No
acyltransferase 2	*AT*	CA01g32970	No
acyltransferase	*AT*	CA02g19280	No
acyltransferase	*AT*	CA02g19300	No
acyltransferase-like	*AT*	CA12g22900	Yes

Note. Each of these 135 genes were searched against our generated heatmap of 11,052 genes and we identified if the genes were present or not from the list of 135. The genes that were present then were recorded in table 7 if they had high expression in pungent peppers and low expression in less pungent peppers.

Figure A7

Heatmap 1 of the Filtered Data



Note. Columns are representative of the pepper samples for a total of 46 while rows are representative of the gene identification associated with each pepper with a total of 11,052. This is the first of four total heatmaps generated that have a strong trend of high expression in pungent peppers and low expression in less pungent peppers. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation. Peppers are ordered left to right from low pungency to high pungency.

Table A2

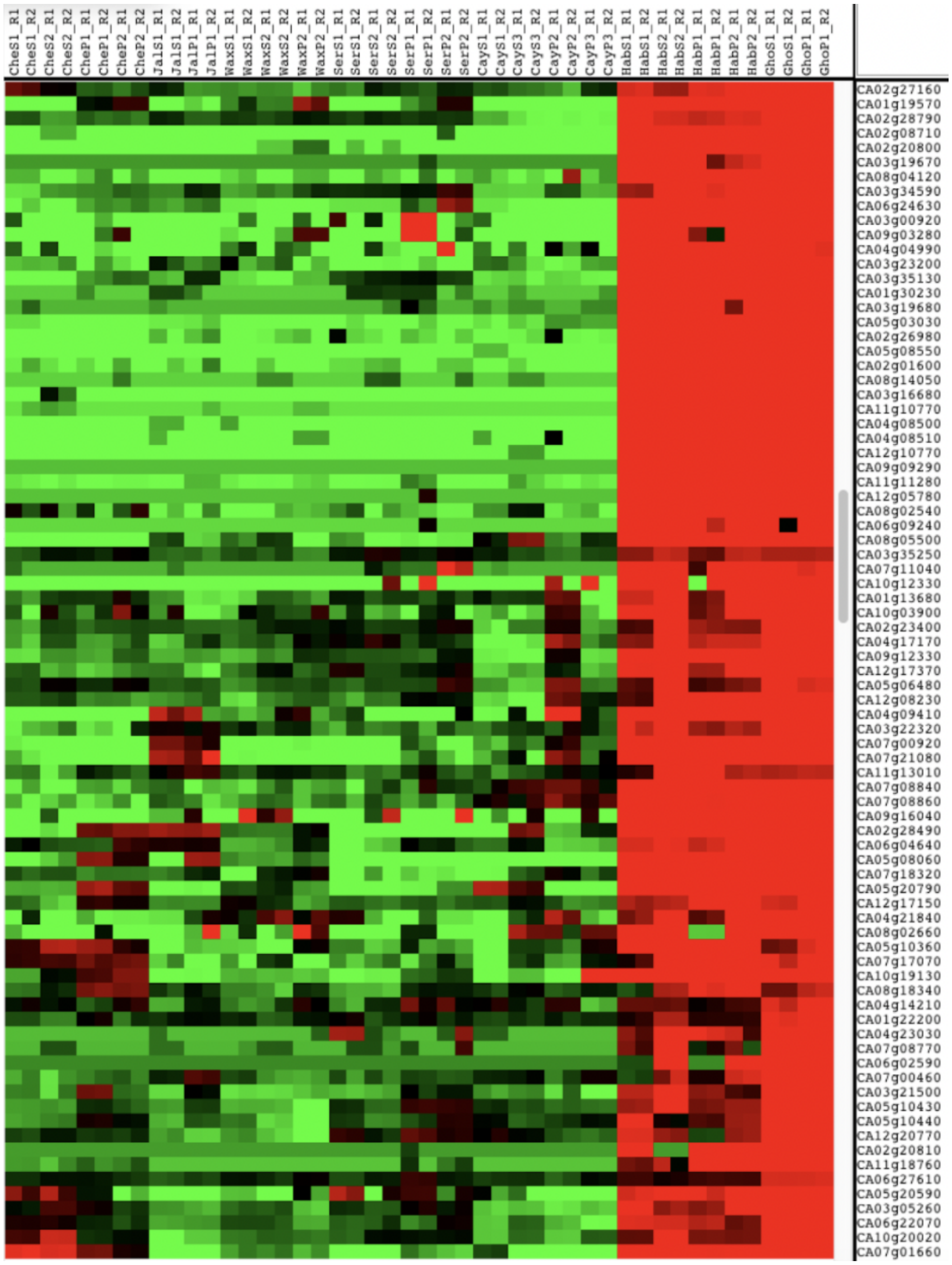
List of Genes Taken from Heatmap 1 with their Functional Annotations

Heatmap 1	
Gene ID	**Annotation**
CA06g08400	RNA-binding protein with multiple splicing
CA03g30320	Detected protein of unknown function
CA07g06320	Glycosyltransferase
CA02g14090	PREDICTED: abscisic acid receptor PYL12-like [Glycine max]
CA10g04090	Eukaryotic translation initiation factor 3 subunit%2C putative
CA01g18360	Lipoamide acyltransferase component of branched-chain alpha-keto acid dehydrogenase%2C putative
CA09g07380	Transparent testa 12 protein
CA04g23410	Chromatin remodeling complex subunit
CA05g04920	Protein kinase APK1B%2C chloroplast%2C putative
CA12g21390	Epoxide hydrolase 1
CA01g29320	Retrotransposon protein%2C putative%2C unclassified
CA04g15780	Ripening regulated protein DDTFR18
CA06g10910	Putative branched-chain alpha-keto acid dehydrogenase E1 alpha subunit
CA03g00660	Selenium-binding protein
CA09g02070	membrane family protein [Populus trichocarpa]
CA05g00190	PREDICTED: protein EARLY RESPONSIVE TO DEHYDRATION 15-like [Solanum tuberosum]
CA06g26640	Acyl-ACP thioesterase
CA04g07090	Allyl alcohol dehydrogenase
CA05g13240	PREDICTED: histone H3.3-like [Glycine max]
CA06g26060	RAN
CA09g15570	Detected protein of confused Function
CA08g13430	Ran3A-1
CA12g06090	Detected protein of unknown function
CA05g18320	PREDICTED: putative GPI-anchor transamidase-like isoform 1 [Solanum lycopersicum]
CA11g09810	MADS box protein
CA01g16920	Unknown protein
CA02g08600	Nucleosome/chromatin assembly factor group (Fragment)

Gene ID	Annotation
CA03g28130	PREDICTED: ribonuclease 3-like [Solanum lycopersicum]
CA08g02750	Detected protein of unknown function
CA01g13330	PREDICTED: E3 ubiquitin-protein ligase RING1-like isoform 1 [Vitis vinifera]
CA08g17680	Serine/threonine-protein kinase SAPK10%2C putative
CA10g06760	Detected protein of unknown function
CA04g07280	PREDICTED: ocs element-binding factor 1-like [Solanum tuberosum]
CA02g03060	NAC domain protein NAC6
CA01g23760	Hypersensitive-induced response protein
CA11g18980	Unknown protein
CA03g02670	Myb domain protein 79 isoform 1 [Theobroma cacao]
CA05g04290	Detected protein of confused Function
CA11g05560	Putative sterol desaturase
CA06g06820	Detected protein of unknown function
CA07g13760	zinc finger family protein [Populus trichocarpa]
CA03g16900	Hypersensitive induced reaction protein 4
CA12g16110	Detected protein of unknown function
CA07g16150	Detected protein of confused Function
CA06g09140	Phosphatidylcholine transfer protein%2C putative
CA10g03720	PREDICTED: nudix hydrolase 15%2C mitochondrial-like [Solanum tuberosum]
CA02g03060	NAC domain protein NAC6
CA01g23760	Hypersensitive-induced response protein
CA10g03720	PREDICTED: nudix hydrolase 15%2C mitochondrial-like [Solanum tuberosum]
CA09g14830	Pectinesterase
CA01g12030	Serine/threonine-protein phosphatase
CA03g08540	Beta-glucosidase%2C putative
CA02g24620	Cysteine desulfurase
CA06g09300	5'-nucleotidase surE
CA06g16910	2-oxoglutarate-dependent dioxygenase
CA03g08980	Flavin monooxygenase-like protein
CA01g24180	Hydroxymethylglutaryl-CoA lyase%2C putative

Figure A8

Heatmap 2 of the Filtered Data



Note. Columns are representative of the pepper samples for a total of 46 while rows are representative of the gene identification associated with each pepper with a total of 11,052. This is the second of four total heatmaps generated that have a strong trend of high expression in pungent peppers and low expression in less pungent peppers. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation. Peppers are ordered left to right from low pungency to high pungency.

Table A3

List of Genes Taken from Heatmap 2 with their Functional Annotations

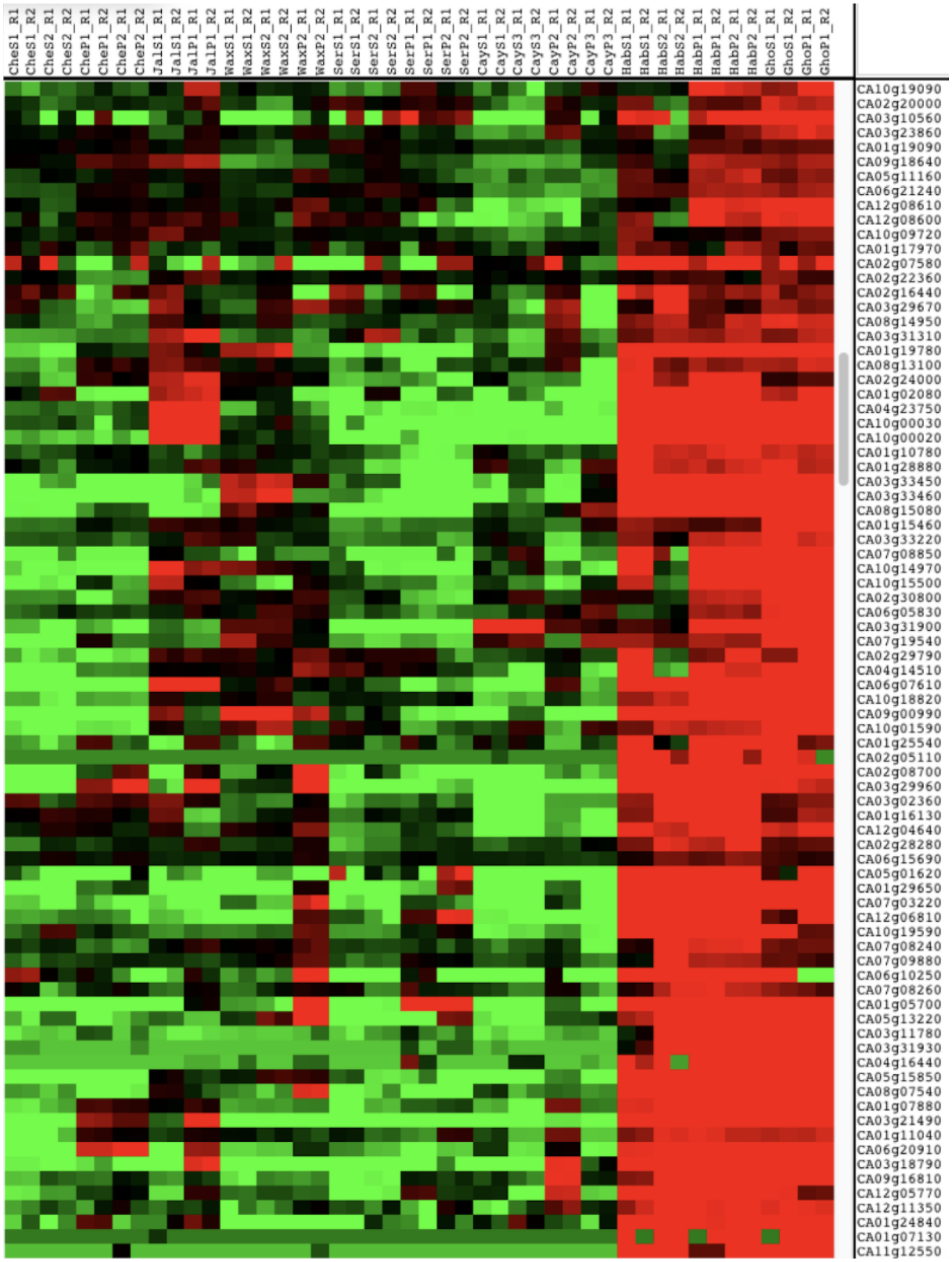
Heatmap 2	
Gene ID	**Annotation**
CA02g27160	Mitochondrial carnitine/acylcarnitine carrier protein%2C putative
CA01g19570	Papain-like cysteine proteinase isoform I
CA02g28790	Acetylglucosaminyltransferase%2C putative
CA02g08710	Detected protein of unknown function
CA02g20800	Detected protein of confused Function
CA03g19670	PREDICTED: pleiotropic drug resistance protein 1-like [Solanum lycopersicum]
CA08g04120	PREDICTED: probable phosphoinositide phosphatase SAC9-like isoform X1 [Solanum tuberosum]
CA03g34590	Aspartate aminotransferase%2C putative
CA06g24630	Putative cytochrome P450 monooxygenase
CA03g00920	Protein phosphatase 2c%2C putative
CA09g03280	ORF64c [Pinus koraiensis]
CA04g04990	PREDICTED: LRR receptor-like serine/threonine-protein kinase GSO1-like [Solanum lycopersicum]
CA03g23200	DNA polymerase epsilon subunit B%2C putative
CA03g35130	Ethylene receptor
CA01g30230	PREDICTED: probable receptor protein kinase TMK1-like [Solanum lycopersicum]
CA03g19680	PREDICTED: pleiotropic drug resistance protein 1-like [Solanum lycopersicum]
CA05g03030	Cytochrome P450 CYP736A54
CA02g26980	Protein disulfide oxidoreductase%2C putative
CA05g08550	Beta-1%2C3-glucanase 24 (Precursor)
CA02g01600	PREDICTED: type II inositol 1%2C4%2C5-trisphosphate 5-phosphatase FRA3-like isoform X1 [Solanum tuberosum]
CA08g14050	PREDICTED: CBS domain-containing protein CBSX5-like [Solanum lycopersicum]
CA03g16680	Peptidoglycan-binding LysM domain-containing protein [Theobroma cacao]
CA11g10770	Calcium-dependent lipid-binding family protein [Theobroma cacao]
CA04g08500	Dicel-like 2 (Fragment)

Gene ID	**Annotation**
CA04g08510	Dicel-like 2 (Fragment)
CA12g10770	Detected protein of unknown function
CA09g09290	Transposon MuDR mudrA-like protein%2C putative
CA11g11280	Alpha-L-fucosidase 2%2C putative
CA12g05780	T24M8.8 protein
CA08g02540	PPR repeat domain-containing protein
CA06g09240	PREDICTED: 2-hydroxyisoflavanone dehydratase-like [Glycine max]
CA08g05500	Ornithine aminotransferase
CA03g35250	Aldehyde dehydrogenase family 7 member A1
CA07g11040	PREDICTED: calcium-dependent protein kinase 3-like [Solanum lycopersicum]
CA10g12330	Serine/threonine-protein kinase
CA01g13680	Beta-D-glucosidase (Precursor)
CA10g03900	PREDICTED: transcription elongation factor 1 homolog isoform 1 [Solanum lycopersicum]
CA02g23400	Detected protein of unknown function
CA04g17170	Catalytic%2C putative
CA09g12330	PREDICTED: clathrin light chain 2-like [Solanum tuberosum]
CA12g17370	PREDICTED: pre-mRNA-splicing factor ATP-dependent RNA helicase PRP16-like [Solanum tuberosum]
CA05g06480	Aig1%2C putative
CA12g08230	PREDICTED: 3-oxo-5-alpha-steroid 4-dehydrogenase 2-like [Vitis vinifera]
CA04g09410	Pectinesterase
CA03g22320	Cathepsin B-like cysteine proteinase
CA07g00920	Pectinesterase
CA07g21080	Pectinesterase
CA11g13010	ALY protein
CA07g08840	Isoamylase isoform 1
CA07g08860	Isoamylase isoform 1
CA09g16040	Putative Ty3-gypsy-like retroelement pol polyprotein
CA02g28490	Protein C10orf22%2C putative
CA06g04640	NbPCL1 protein
CA05g08060	Bet v I allergen family protein
CA07g18320	PREDICTED: mediator of RNA polymerase II transcription subunit 18-like [Solanum

	lycopersicum]
Gene ID	**Annotation**
CA05g20790	Phenylalanine ammonia-lyase
CA12g17150	2-hydroxyacid dehydrogenase%2C putative
CA04g21840	Detected protein of unknown function
CA08g02660	PREDICTED: probable xyloglucan endotransglucosylase/hydrolase protein 8-like [Solanum lycopersicum]
CA05g10360	PREDICTED: dentin sialophosphoprotein-like [Citrus sinensis]
CA07g17070	CYP72A58
CA10g19130	F-box family protein [Theobroma cacao]
CA08g18340	1-acyl-sn-glycerol-3-phosphate acyltransferase
CA04g14210	Unknown protein
CA01g22200	Serologically defined colon cancer antigen-like protein
CA04g23030	Unknown protein
CA07g08770	Alpha tubulin (Fragment)
CA06g02590	Detected protein of confused Function
CA07g00460	Mitochondrial bifunctional diaminopelargonate synthetase
CA03g21500	Cytochrome P450
CA05g10430	4-hydroxyphenylpyruvate dioxygenase
CA05g10440	4-hydroxyphenylpyruvate dioxygenase
CA12g20770	15-cis-zeta-carotene isomerase [Theobroma cacao]
CA02g20810	Retrotransposon protein%2C putative%2C Ty1-copia subclass
CA11g18760	Ripening-related protein grip22
CA06g27610	Detected protein of unknown function
CA05g20590	PREDICTED: ankyrin repeat-containing protein At3g12360-like [Solanum tuberosum]
CA03g05260	Casein kinase%2C putative
CA06g22070	Detected protein of unknown function
CA10g20020	Thioredoxin fold
CA07g01660	PREDICTED: QWRF motif-containing protein 8-like isoform X1 [Solanum tuberosum]

Figure A9

Heatmap 3 of the Filtered Data



Note. Columns are representative of the pepper samples for a total of 46 while rows are representative of the gene identification associated with each pepper with a total of

11,052. This is the third of four total heatmaps generated that have a strong trend of high expression in pungent peppers and low expression in less pungent peppers. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation.Peppers are ordered left to right from low pungency to high pungency.

Table A4

List of Genes Taken from Heatmap 3 with their Functional Annotations

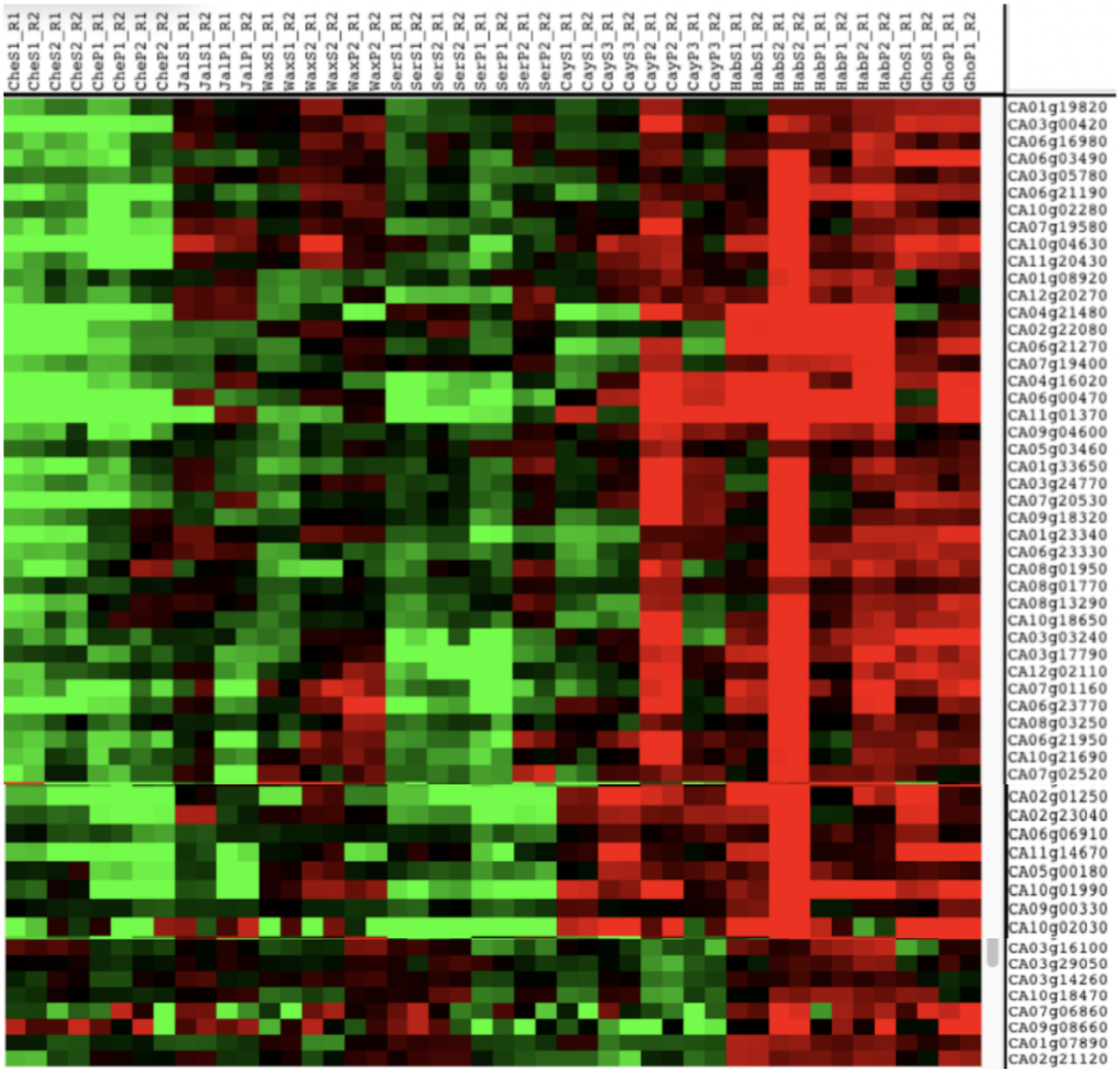
Heatmap 3	
Gene ID	**Annotation**
CA10g19090	PREDICTED: eukaryotic translation initiation factor NCBP-like isoform X1 [Glycine max]
CA02g20000	PREDICTED: SRSF protein kinase 1-like isoform X1 [Solanum tuberosum]
CA03g10560	Alcohol dehydrogenase
CA03g23860	Detected protein of confused Function
CA01g19090	Cop11 protein
CA09g18640	PREDICTED: eukaryotic peptide chain release factor subunit 1-3-like isoform 1 [Solanum lycopersicum]
CA05g11160	Protein kinase Ck2 regulatory subunit 2
CA06g21240	Nucleic acid binding protein%2C putative
CA12g08610	PREDICTED: TSL-kinase interacting protein 1-like [Vitis vinifera]
CA12g08600	PREDICTED: TSL-kinase interacting protein 1-like [Vitis vinifera]
CA10g09720	CDPK11
CA01g17970	PREDICTED: 2-oxoisovalerate dehydrogenase subunit beta 2%2C mitochondrial-like [Solanum tuberosum]
CA02g07580	HGWP repeat containing protein-like protein
CA02g22360	PREDICTED: transmembrane protein 53-like [Vitis vinifera]
CA02g16440	Cytokinin riboside 5'-monophosphate phosphoribohydrolase-like
CA03g29670	Cis-prenyltransferase 3
CA08g14950	At1g10280
CA03g31310	Detected protein of unknown function
CA01g19780	Salt-tolerance protein%2C putative
CA08g13100	Detected protein of unknown function
CA02g24000	Auxin:hydrogen symporter%2C putative
CA01g02080	PREDICTED: F-box protein At2g26160-like isoform X1 [Solanum tuberosum]

Gene ID	**Annotation**
CA04g23750	Aminotransferase family protein
CA10g00030	Aminotransferase family protein
CA10g00020	Aminotransferase family protein
CA01g10780	Sucrose-phosphatase
CA01g28880	Eukaryotic translation initiation factor 3 subunit%2C putative
CA03g33450	Putative receptor-like protein kinase
CA03g33460	Receptor-like protein kinase
CA08g15080	Detected protein of unknown function
CA01g15460	ADP%2CATP carrier protein-like
CA03g33220	PREDICTED: protein IFH1-like [Solanum tuberosum]
CA07g08850	Detected protein of unknown function
CA10g14970	Os02g0480100 [Oryza sativa] Japonica Group
CA10g15500	Detected protein of confused Function
CA02g30800	B2 protein%2C putative
CA06g05830	SGT1
CA03g31900	Detected protein of unknown function
CA07g19540	Ubiquitin2
CA02g29790	Detected protein of unknown function
CA04g14510	Disease resistance protein RGH2
CA06g07610	PREDICTED: F-box/FBD/LRR-repeat protein At1g13570-like [Solanum tuberosum]
CA10g18820	PREDICTED: probable mediator of RNA polymerase II transcription subunit 26b-like [Solanum tuberosum]
CA09g00990	Unknown protein
CA10g01590	Os01g0698300 protein
CA01g25540	Dehydroquinate dehydratase/shikimate:NADP oxidoreductase
CA02g05110	BSD domain-containing family protein [Populus trichocarpa]
CA02g08700	Detected protein of unknown function
CA03g29960	Detected protein of unknown function
CA03g02360	AG-motif binding protein-2
CA01g16130	Actin cross-linking protein%2C putative [Theobroma cacao]
CA12g04640	Serine carboxypeptidase precursor family protein [Populus trichocarpa]
CA02g28280	Delta(3%2C5)%2Cdelta(2%2C4)-dienoyl-CoA isomerase 1
CA06g15690	26S protease regulatory subunit%2C putative
CA05g01620	Unknown protein

Gene ID	**Annotation**
CA01g29650	UDP-glucose:glucosyltransferase
CA07g03220	Ty3/gypsy retrotransposon protein
CA12g06810	UDP-glucose:glucosyltransferase
CA10g19590	Somatic embryogenesis zinc finger 2
CA07g08240	PREDICTED: F-box/kelch-repeat protein At1g16250-like isoform X1 [Solanum tuberosum]
CA07g09880	Eukaryotic translation initiation factor 3 subunit 11 family protein [Populus trichocarpa]
CA06g10250	PREDICTED: photosystem II 5 kDa protein%2C chloroplastic-like [Solanum tuberosum]
CA07g08260	Guanine nucleotide exchange factor P532%2C putative
CA01g05700	AP2/ERF domain-containing transcription factor
CA05g13220	PREDICTED: protein MIZU-KUSSEI 1-like [Solanum tuberosum]
CA03g11780	Sulfate/bicarbonate/oxalate exchanger and transporter sat-1
CA03g31930	F9L1.26
CA04g16440	PREDICTED: zeatin O-glucosyltransferase-like [Solanum tuberosum]
CA05g15850	CBL-interacting protein kinase
CA08g07540	Mitogen-activated protein kinase 4
CA01g07880	Gamma-tocopherol methyltransferase
CA03g21490	Cytochrome P450
CA01g11040	Os08g0117900 protein
CA06g20910	Bcl-2-associated athanogene-like protein
CA03g18790	Defective in meristem silencing 3 [Theobroma cacao]
CA09g16810	Unknown protein
CA12g05770	PREDICTED: OTU domain-containing protein At3g57810-like [Cucumis sativus]
CA12g11350	Detected protein of unknown function
CA01g24840	Receptor serine/threonine kinase%2C putative
CA01g07130	PREDICTED: serine/threonine-protein phosphatase 7 long form homolog [Solanum tuberosum]
CA11g12550	Kinesin light chain%2C putative

Figure A10

Heatmap 4 of the Filtered Data



Note. Columns are representative of the pepper samples for a total of 46 while rows are representative of the gene identification associated with each pepper with a total of 11,052. This is the fourth of four total heatmaps generated that have a strong trend of high expression in pungent peppers and low expression in less pungent peppers. Red sections represent upregulated genes, green sections represent genes that are downregulated, and black sections represent no changes in gene regulation. Peppers are ordered left to right from low pungency to high pungency.

Table A5

List of Genes Taken from Heatmap 4 with their Functional Annotations

Heatmap 4	
Gene ID	**Annotation**
CA01g19820	PREDICTED: V-type proton ATPase 16 kDa proteolipid subunit c2-like [Solanum tuberosum]
CA03g00420	PREDICTED: alkaline ceramidase 3-like [Solanum tuberosum]
CA06g16980	PREDICTED: cytochrome c oxidase subunit 5b-1%2C mitochondrial-like [Solanum tuberosum]
CA06g03490	Calcineurin B-like 10
CA03g05780	PREDICTED: activating signal cointegrator 1-like isoform 2 [Solanum lycopersicum]
CA06g21190	PREDICTED: mitotic-spindle organizing protein 1B-like [Solanum lycopersicum]
CA10g02280	Ribose-phosphate pyrophosphokinase 4
CA07g19580	PREDICTED: GPI-anchored protein LORELEI-like [Solanum tuberosum]
CA10g04630	Nucleoporin
CA11g20430	50S ribosomal protein L27 [Medicago truncatula]
CA01g08920	Putative auxin-induced protein
CA12g20270	60S ribosomal protein L13a-like protein
CA04g21480	Unknown protein
CA02g22080	Mitogen-activated protein kinase 8
CA06g21270	PREDICTED: inositol-tetrakisphosphate 1-kinase 1-like [Solanum lycopersicum]
CA07g19400	PREDICTED: probable protein phosphatase 2C 55-like isoform X4 [Solanum tuberosum]
CA04g16020	PREDICTED: purine permease 1-like [Solanum tuberosum]
CA06g00470	Unknown protein
CA11g01370	Catalytic%2C putative
CA09g04600	PREDICTED: MOSC domain-containing protein 2%2C mitochondrial-like isoform X1 [Solanum tuberosum]
CA05g03460	GTP-binding protein
CA01g33650	Unknown protein
CA03g24770	small nuclear ribonucleoprotein D2 [Arabidopsis thaliana]
CA07g20530	SEC14 cytosolic factor%2C putative
CA09g18320	eukaryotic translation initiation factor 5A-4 [Solanum lycopersicum]
CA01g23340	Unknown protein

Gene ID	**Annotation**
CA06g23330	RNA binding protein-like protein
CA08g01950	Detected protein of unknown function
CA08g01770	PREDICTED: rab GDP dissociation inhibitor alpha-like [Solanum lycopersicum]
CA08g13290	Detected protein of unknown function
CA10g18650	Syntaxin-52
CA03g03240	Mitotic spindle assembly checkpoint protein MAD2B
CA03g17790	PREDICTED: microtubule-associated protein 1B-like [Solanum tuberosum]
CA12g02110	PREDICTED: centromere protein V-like [Solanum lycopersicum]
CA07g01160	Unknown protein
CA06g23770	Unknown protein
CA08g03250	Unknown protein
CA06g21950	Unknown protein
CA10g21690	Unknown protein
CA07g02520	Oxidoreductase%2C putative
CA02g01250	UDP-glucuronosyltransferase%2C putative
CA02g23040	Chloroplast geranylgeranyl diphosphate synthase
CA06g06910	PREDICTED: actin-related protein 5-like [Solanum tuberosum]
CA11g14670	Oxidoreductase
CA05g00180	Unknown protein
CA10g01990	Glutathione S-transferase/peroxidase
CA09g00330	Detected protein of unknown function
CA10g02030	Unknown protein
CA03g16100	Fructokinase
CA03g29050	JHL10I11.9 protein
CA03g14260	GTP-binding-like protein
CA10g18470	GrpE protein homolog
CA07g06860	Chromatin remodeling complex subunit
CA09g08660	Detected protein of unknown function
CA01g07890	Protein disulfide isomerase L-2
CA02g21120	Isocitrate dehydrogenase%2C putative