

WATSON, AMANDA L., Ph.D. Evaluation of Analytical Tools for Studying the Host and Gut Microbe Relationships in Type 2 Diabetes. (2016)  
Directed by Dr. Norman H L Chiu 97 pp.

Type 2 Diabetes (T2D) is becoming one of the most prevalent diseases in the world. With this comes an increased burden in healthcare and a lower quality of life for those affected by diabetes. Further, certain populations are more susceptible to T2D than others. For example, in the US, the Hispanic population has the second highest rate of T2D while those of Caucasian heritage have the lowest rate of T2D. Studies have shown that different environment and socioeconomic standing is not enough to explain the increased prevalence of diabetes in the Hispanic population. There must be other factors involved in disease susceptibility. Specifically, there is evidence that genetic variation and the gut microbiome play a role in human health and disease development including in T2D. Furthermore, the Hispanic population is the least studied of all populations for diabetes. Therefore, we set out to examine these factors that may increase the susceptibility for T2D in Hispanic population. To facilitate the studies on these factors, a number of new analytical tools were evaluated.

First, by using a new beadarray chip that was developed by a consortium of experts in T2D and other related diseases, a pilot study on the genetic variations of Hispanic population with T2D was completed. To determine the adequate size of cohort for this study, a reference database of a larger diabetes genetic study was used. Using the method of principle components analysis (PCA), it has shown that the genetic information obtained from more than 20 participants would have sufficient resolving power to determine whether a particular participant was healthy or diseased, providing

sufficient SNP genotypes are included in the PCA analysis. Following this, a genetic study was completed on an adult Hispanic population and 26 new SNPs were found to be associated with T2D through comparison with reference populations and PCA analysis. Future work will involve increasing the size of cohort to validate the identified SNPs and to further evaluate the use of SNP genotypes to define the host for studying the host - gut microbe relationships.

Second, recent studies have indicated the gut microbiome that lives in symbiosis with the human host can influence our health. More recently, there are evidences that gut microbes can also increase the susceptibility for T2D. Extensive work has been done to identify the microbes in the gut, but studying the activity of gut microbes is still under development. Therefore, we set out to build a simple model of gut microbes aiming to explore new ways to measure the activity. *Lactobacillus helveticus*, a probiotic gram-positive bacteria was used to build the model. The enzymatic activity of beta-galactosidase ( $\beta$ -gal) was assayed with a fluorescent substrate called 4-methylumbelliferyl  $\beta$ -D-galactopyranoside. To ensure the  $\beta$ -gal assay is compatible with subsequent studies, a whole-cell format was adopted. Since no protocol for the selected gram-positive microbe was available, the assay was developed by reducing the background noise and brought the assay time to one day with a linear dynamic range over two orders of magnitude. To validate the results, the  $\beta$ -gal assay was repeated on a different bacterial strain. With the developed  $\beta$ -gal assay, building the gut microbe model can continue towards evaluating microbe activity in relationship with the host in health and disease.

EVALUATION OF ANALYTICAL TOOLS FOR STUDYING THE HOST AND GUT  
MICROBE RELATIONSHIPS IN TYPE 2 DIABETES

by

Amanda L. Watson

A Dissertation Submitted to  
the Faculty of The Graduate School at  
The University of North Carolina at Greensboro  
in Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

Greensboro  
2016

Approved by

---

Committee Chair

APPROVAL PAGE

This dissertation written by Amanda L. Watson has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair \_\_\_\_\_

Committee Members \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

\_\_\_\_\_  
Date of Acceptance by Committee

\_\_\_\_\_  
Date of Final Oral Examination

## TABLE OF CONTENTS

	Page
LIST OF TABLES.....	vi
LIST OF FIGURES .....	vii
CHAPTER	
I. INTRODUCTION .....	1
1.1 What is Type 2 Diabetes? .....	1
1.2 Environment vs. Genetics .....	3
1.2.1 Genetics and Susceptibility for Type 2 Diabetes .....	3
1.2.2 Gut Microbiome.....	6
1.3 A Survey of Current Detection Methods .....	7
1.3.1 High-throughput Detection of Genetic Biomarkers.....	7
1.3.2 “Omics” for Microbiome Measurement .....	10
1.4 Specific Aims .....	13
1.4.1 Identifying the Role the Host Plays in Susceptibility .....	13
1.4.2 Building a Simple Model for the Gut Microbiome.....	14
II. PRINCIPLE COMPONENTS ANALYSIS OF A REFERENCE DATASET .....	15
2.1 Introduction .....	15
2.1.1 Sample Size for a Pilot Study .....	15
2.1.2 What is Principle Components Analysis?.....	16
2.1.3 Bioinformatics Tools .....	17
2.2 Method.....	18
2.2.1 Data Acquisition and Processing .....	18
2.2.2 Isolation of Subsets of Individuals.....	18
2.2.3 Data Pruning .....	18
2.2.4 PCA Analysis and Plotting .....	19
2.2.5 Reduction of Sample Size.....	19
2.2.6 Further Reduction of SNPs .....	20
2.3 Results/Discussion.....	20
2.3.1 Data Processing.....	20
2.3.2 Isolating Small Sample Groups .....	20
2.3.3 Pruning Methods.....	21
2.3.4 Testing the Reduction of the Data .....	23
2.4 Conclusion.....	30

III. GENETIC BIOMARKERS FOR TYPE 2 DIABETES IN A HISPANIC POPULATION.....	32
3.1 Introduction .....	32
3.2 Method.....	34
3.2.1 Sample Selection.....	34
3.2.2 DNA Isolation and Characterization.....	34
3.2.3 Genotyping.....	36
3.2.4 Data Processing.....	37
3.2.5 Quality Control .....	37
3.2.6 Data Analysis .....	38
3.2.7 Comparison to Reference Populations.....	38
3.2.8 PCA.....	39
3.3 Results/Discussion.....	39
3.3.1 Participant Selection .....	39
3.3.2 Genomic DNA Isolation .....	41
3.3.3 Data Preparation for Analysis.....	43
3.3.4 Quality Control .....	45
3.3.5 Association Analysis.....	46
3.3.6 Comparison to Reference Populations.....	49
3.3.7 PCA.....	52
3.4 Conclusion.....	54
IV. OPTIMIZATION OF A WHOLE CELL BETA-GALACTOSIDASE ASSAY IN GRAM-POSITIVE BACTERIA.....	55
4.1 Introduction .....	55
4.1.1 Identifying the Role the Gut Microbes Play in Susceptibility .....	55
4.1.2 Modelling the Gut Microbiome .....	55
4.1.3 Microbe Activity.....	56
4.2 Method.....	58
4.2.1 Bacteria Stock .....	58
4.2.2 Growth Curve.....	59
4.2.3 Initial Assay Conditions.....	59
4.2.4 Testing for Background Noise .....	60
4.2.5 MUG Concentration.....	60
4.2.6 Activation.....	60
4.2.7 Linear Dynamic Range .....	61
4.2.8 Reproducibility of the Assay .....	61
4.3 Results/Discussion.....	61
4.3.1 Initial Growth of <i>L. helveticus</i> .....	61
4.3.2 Results of Initial Assay .....	63

4.3.3 Background Noise.....	64
4.3.4 MUG Concentration.....	66
4.3.5 Activation of Beta-Galactosidase .....	68
4.3.6 Linear Dynamic Range .....	71
4.3.7 Reproducibility of the Assay in Another Gram-Positive Bacterial Strain .....	72
4.4 Conclusion.....	74
V. CONCLUSIONS.....	76
REFERENCES .....	78
APPENDIX A. PLINK COMMANDS.....	84
APPENDIX B. PCA BILOTS FOR SAMPLE AND SNP REDUCTION .....	86

## LIST OF TABLES

	Page
Table 1. The Cardio-Metabochip.....	36
Table 2. Participant Demographics.....	40
Table 3. Genomic DNA Concentration from Human Blood Samples.....	43
Table 4. Quality Control Results .....	44
Table 5. Top 26 SNPs .....	48
Table 6. Comparison to Reference Populations.....	51

## LIST OF FIGURES

	Page
Figure 1. PCA Biplots of Two Pruning Approaches .....	22
Figure 2. Reducing Sample Size.....	24
Figure 3. Set of 100 Samples with Number of SNPs Reduced.....	26
Figure 4. Set of 50 Samples with Number of SNPs Reduced.....	28
Figure 5. Set of 20 Samples with Number of SNPs Reduced.....	29
Figure 6. Number of Diabetes Studies by Population.....	33
Figure 7. PicoGreen Structure.....	42
Figure 8. Manhattan Plot of Genotyping Data.....	47
Figure 9. PCA of Hispanic Population Data.....	53
Figure 10. Reaction of MUG with Beta-galactosidase Enzyme .....	58
Figure 11. <i>L. helveticus</i> Growth Curve.....	62
Figure 12. Initial Assay Results .....	64
Figure 13. Sources of Background Noise .....	65
Figure 14. Changing MUG Concentration to Increase Signal-to-noise Ratio .....	67
Figure 15. Beta-galactosidase Activation .....	69
Figure 16. Results of Activation Study.....	70
Figure 17. Linear Dynamic Range of Assay for <i>L. helveticus</i> .....	72
Figure 18. Growth Curve for <i>L. reuteri</i> .....	73
Figure 19. Linear Dynamic Range for <i>L. reuteri</i> .....	74

# CHAPTER I

## INTRODUCTION

### 1.1 What is Type 2 Diabetes?

Type 2 Diabetes is a common and debilitating disease with the number of those affected increasing drastically each year (CDC Health, 2010). The Center for Disease Control (CDC) predicts that without preventative steps, one in three individuals in the United States will develop the disease (CDC Health, 2010). Thus, it is becoming increasingly more important to develop early detection methods and interventions to help prevent the rapid rise of this disease.

Type 2 Diabetes is also called non-insulin dependent or adult-onset diabetes because it usually does not require insulin for treatment and is almost always diagnosed in adults older than twenty years old. In fact, less than one percent of children are diagnosed with Type 2 Diabetes. Further, it accounts for nearly ninety percent of all diabetes cases (CDC Diabetes Statistics Report, 2014).

The disease symptoms and effects are progressive. It can start with consistently high levels of glucose introduced into the blood stream leading to slight insensitivity of the glucose receptors in the target tissues. This leads to hyperglycemia and increased insulin secretion. There can also be  $\beta$ -cell dysfunction that leads to a decrease in insulin levels leading to hyperglycemia and insulin resistance. B-cell dysfunction can be genetic, be caused by cell death due to cellular stress inflicted by obesity induced inflammation,

or Reactive Oxygen Species caused by glucolipotoxicity. Once this has begun, a viscous cycle occurs where increased insulin levels leads to insulin resistance and thus consistent hyperglycemia. This further exacerbates the  $\beta$ -cell dysfunction leading to further insulin secretion, proliferation of the  $\beta$ -cells to keep up with the demand, and imbalance in other metabolic hormones. This leads to further  $\beta$ -cell dysfunction and worsening of the disease into full blown Type 2 Diabetes. Steps can be taken to reduce the hyperglycemia and thus reduce the work load on the  $\beta$ -cells and potentially preventing progression of the disease. (Cerf, 2013)

Because there is increased glucose in the blood stream, there are symptoms of increased thirst, weight loss, and longer healing time. Often people go for some time ignoring these symptoms before getting diagnosed. It is believed that there are around seven million people in the US that have not been diagnosed, but have the disease. Late diagnosis as well as poor management can lead to worsening of the disease and an increased chance for peripheral effects. These include heart disease, stroke, blindness, poor circulation leading to limb amputation, and death. Type 2 Diabetes is currently the seventh leading cause of death in America (CDC Diabetes Statistics Report, 2014).

Nearly a third of the adult US population and half of the elderly population over 65 has prediabetes. This involves slight insulin resistance with the consequential slight increase in blood glucose levels. For most prediabetic individuals, preventative measures such as diet, exercise, and medication can protect them from fully developing Diabetes. Although Type 2 Diabetes is manageable, it is better to catch it early and to take preventative measures to avoid the disease and the many potential side effects it causes.

## **1.2 Environment vs. Genetics**

Just as the mode of how the disease starts varies, Type 2 Diabetes is a multifactorial disease that has multiple potential sources for susceptibility and development. The most commonly associated risk factor for Type 2 Diabetes is environmental causes like obesity and poor diet and exercise. However, more recent research has indicated that environment alone is not enough to cause diabetes (Lusis *et al.*, 2008). Specifically, genetics can change susceptibility for Type 2 Diabetes and the gut microbiome, which interacts with the diet and environment, can influence the host towards obesity and insulin resistance.

### **1.2.1 Genetics and Susceptibility for Type 2 Diabetes**

There has been considerable evidence that supports the fact that innate genetics plays a role in the development of diabetes. The most common genetic biomarkers when studying this are single nucleotide polymorphisms (SNPs). These are single base mutations in the genome that have to occur in at least 1% of the population to be categorized as a SNP and not a random mutation. The impact of these single base changes range from basic appearance changes to significant likelihood of disease development (Broad Institute, 2013). To date, there have been at least 88 SNPs associated with Type 2 Diabetes susceptibility (Mohlke & Boehnke, 2015) with more discovered regularly. Thus it has been shown that there is not a single mutation that causes the disease, but rather a large number of small changes that have an additive effect that leads to increased susceptibility for diabetes (Bonnetfond *et al.*, 2010). Further, the number of SNPs that have been shown to play a role in diabetes indicates that not only is Type 2 Diabetes a

complex disease, there may in fact be sub-categories of the disease based on genetics (Murea *et al.*, 2012). Therefore it is important to continue studying these genetic factors to identify all possible biomarkers and how they group in different people to cause disease.

One might doubt that there is a genetic factor if the effect is so small. To address this question, Goldfine *et al.* set out to examine if family history, and therefore genetics, really played a role in increasing the susceptibility for Type 2 Diabetes. Their study included healthy participants with and without family history of diabetes. Those participants with family history meant that both their parents had Type 2 Diabetes. The insulin sensitivity and glucose response were measured at the start of the study and then the participants were followed for an average of 25 years to determine if they would develop diabetes. By the end of the study, those with family history had a significantly greater chance of developing diabetes compared to those without family history. In addition, at the start of the study, the family history group had a greater percentage of individuals with slight insulin resistance and low glucose sensitivity. Further, when they examined the group as a whole at the end of the study, they found that even amongst the population that was not obese, there was a significantly greater risk of Type 2 Diabetes development for those with family history. Therefore, this supports that there is in fact a genetic influence that increases risk for Type 2 Diabetes and in some cases obesity is not a necessary factor for developing the disease. If obesity, the most often associated environmental factor, is not always the key factor in developing diabetes, then there must also be a strong enough genetic effect that can increase susceptibility for the disease.

Even with the large number of genetic biomarkers identified, there is more study needed to identify the mechanism that causes these small genetic changes to lead up to Type 2 Diabetes and how sets of SNPs can be sub-categorized and interact with other influential factors for the disease like the environment and the gut microbes.

One aspect that makes understanding the influence these SNPs play in diabetes development is that they often lie in non-coding regions like introns and upstream or downstream from a gene (Imamura & Maeda, 2011). This could be why these genetic mutations do not have as big an impact by themselves, but rather have a small effect that adds up to disease. One theory is that the SNPs are located in and around genes that play a role in either  $\beta$ -cell development and function or in insulin and other hormone signaling and sensitivity (Bonnetfond *et al.*, 2010). This mode of action would support the multifactorial nature of Type 2 Diabetes in that many of these genetic factors could either add up enough to cause the disease or influence the system enough to increase susceptibility for the disease along with the right environmental factors leading to diabetes. From this, it is feasible to imagine that with enough study of these biomarkers, disease susceptibility could be identified early and preventative steps taken to avoid development of diabetes. With the impact that Type 2 Diabetes is having on human health and the economic strain of over \$200 billion per year in the US alone for treatment (CDC Diabetes Reportcard, 2014), it is essential to continue identifying the genetic biomarkers and mechanisms to prevent the disease from continuing to increase and to develop new therapies that are specific for the sub-category of Type 2 Diabetes to reduce the risk of complications and improve quality of life.

### 1.2.2 Gut Microbiome

When thinking about environmental factors that increase risk for Type 2 Diabetes, a more recent discovery is that the gut microbiome influences human health through its symbiotic relationship with the human host. The microbiome includes a diverse collection of microbes in and on the human body with the number of bacterial cells outnumbering human cells (Kelvin *et al.*, 2012). Thus these microbes must be essential to human life and function. In fact, the impact of the discovery of the human microbiome lead to the development of the Human Microbiome Project (The NIH HMP Working Group, 2009), which was created to identify all the microbes living in symbiosis with humans.

Within the gut, there is an equally diverse collection of hundreds of microbes in different ratios with one another and each individual has a unique collection in their gut (Qin *et al.*, 2010). Thus far, it is understood that the gut microbes are essential to human function. They support the development and regulation of the immune system, produce essential vitamins and nutrients, influence the energy balance in digesting food, and also impact how drugs are metabolized (Kinross *et al.*, 2011). Further, an imbalance in the gut microbiome, called a dysbiosis, can influence towards obesity, metabolic syndrome, and systemic inflammation, all factors likely to play a role in diabetes development (Woting & Blaut, 2016). In fact, there has been enough information to suggest there is a link between Type 2 Diabetes and the Gut Microbiome that the American Diabetes Association held a special session conference on the topic to further the discussion and research in this field (Semenkovich, *et al.*, 2015). Overall, the potential interactions

between the gut microbiome and the host are beginning to be understood. The gut microbiome can positively or negatively influence human health and a dysbiosis of the microbiome can lead to a variety of health issues including diabetes and autoimmune disorders. That is, because the gut microbes regulate the immune system, they can increase or decrease inflammation, which could affect the gut environment as well as periphery tissues like the  $\beta$ -cells in the pancreas. Further, the energy balance is influenced by the gut microbes and therefore they can influence how nutrients and calories are taken in and used. There is even evidence to suggest that certain microbes can influence towards obesity even with a healthy diet (Woting & Blaut, 2016). Thus understanding the activity of specific microbes could further elucidate the mechanisms involved in disease development and how to regulate the gut microbe environment to prevent or treat diabetes and related disorders.

### **1.3 A Survey of Current Detection Methods**

In order to understand how genetics and the gut microbes plays a role in Type 2 Diabetes, the ideal detection methods for biomarkers and function needs to be determined. There are a variety of options for measuring these two factors. Therefore, understanding which methods will provide the most information while keeping cost down will be the most effective choice for detecting causes of diabetes susceptibility.

#### **1.3.1 High-throughput Detection of Genetic Biomarkers**

Technology has advanced to a point that the entire human genome can be sequenced in three days (Illumina Sequencing, 2016). Thus genetic data can be measured in large scale in a high-throughput manner. When it comes to measuring genetic

variation, there are two common options used: whole genome sequencing or genotyping. Each has their own benefits and issues. A survey of each follows.

### **1.3.1.a Whole Genome Sequencing**

Whole-genome sequencing technology has improved dramatically in the past two decades. It is now possible to sequence a human genome for less than \$1000 and in less than a week. This technology provides the ability to examine the entire genome for all potential genetic variations in a person. Further, it allows for detection of mutations that are out of the norm like new SNPs and non-SNP mutations like insertions and deletions (Illumina Sequencing, 2016). For example, with Illumina's sequencing technology, genomic DNA is isolated and then a sequencing library is created by randomly fragmenting the DNA and ligating adapters on the 3' and 5' ends. This allows for identical primer binding sites for DNA amplification using PCR. The amplified DNA fragments are loaded into flow cells that contain complementary oligos that anneal with the adapters. Sequencing is performed using a polymerase and fluorescently labeled nucleotides which are imaged each time a base is added and the different colors associated with specific bases can then be detected in sequence order (Illumina Sequencing, 2016).

The benefit to sequencing is examining all possible variations and getting the whole picture of genetic risk for Type 2 Diabetes and other diseases. The downside to using whole genome sequencing for SNP detection is also acquiring extra information that is not necessarily needed to get an answer to the question being researched. This leads to extra cost in time for data processing and analysis as well as extra monetary cost

of measuring the entire genome. Thus, unless new mutations are being studied, whole genome sequencing is not often used for detecting SNPs associated with disease.

### **1.3.1.b Genotyping**

The more common method for SNP detection is called genotyping. In genotyping, only the SNP location is measured rather than the entire sequence. This focuses the identification of genetic variation to only the target locations. Illumina's genotyping technology uses a type of microarray detection that utilizes microbeads to load the oligos on a microchip. Each bead has a specific oligo with a complementary sequence to a target SNP. To measure SNPs, the genomic DNA is amplified using Illumina's proprietary amplification technology and then enzymatically fragmented. The DNA is then loaded onto the chip and the DNA fragments are allowed to anneal to their complementary oligos. The oligo sequence stops one base short of the SNP so that a polymerase can complete a single base extension at the SNP location with a fluorescently labeled nucleotide. Then the color of the fluorophore is associated with a specific base using bioinformatics software (Illumina Genotyping, 2016).

By only measuring the SNPs there is a significant reduction in data processing and analysis as well as cost of materials. Although the amount of time needed to measure the data takes about the same amount of time as a whole genome. Genotyping has become the standard method used to detect SNPs as long as novel variations are not required. There are whole human genome SNPs that cover the majority of single nucleotide variations in the human genome as well as custom genotyping arrays for targeted and cheaper detection of certain regions of the human genome. The main issue

with using genotyping over sequencing is that novel variations cannot be detected and not all human SNPs can be placed on a genotyping chip. The issue of coverage is typically addressed with bioinformatics tools. In particular, imputation is used to estimate the variation present for SNPs not on the chip that have linkage disequilibrium (LD) with measured SNPs. LD essentially means that certain SNPs travel together when variation is passed down from generation to generation. Therefore, if one SNP is present, it can be assumed that those in LD with it will also be present (Wall & Pritchard, 2003). This allows for the detection of many SNPs without having to directly measure all of them. Although it is an estimation, imputation has become a standard practice to increase coverage without increasing cost. Thus, unless it is necessary to detect novel variation, genotyping is used for measuring SNPs for association with disease.

### **1.3.2 “Omics” for Microbiome Measurements**

In the past, microbial studies involved culturing each strain of microbe and studying it extensively. However, this method is no longer applicable when it comes to the human gut microbiome. Many of the microbes are very sensitive to different environments and therefore cannot be cultured in a lab. Further, the collection of microbes is so complex that it would be near impossible to study each individually. Therefore a different approach was taken to study the microbiome. It is called “omics” and involves identifying a trait about the collection of microbes as a whole. The benefit to this is taking real samples and studying their properties as they have been influenced by their natural environment. Further, it eliminates the need for culturing or isolating each strain of microbe. The most common forms of omics measurements are metagenomics,

metatranscriptomics, metaproteomics, and metabolomics. The first is used to identify the microbial composition and the latter three are used to detect microbial activity (van Baarlen *et al.*, 2013).

### **1.3.2.a Microbial Identity and Community Composition**

The initial way the human microbiome was measured was using metagenomics. This technique involves isolating the ribosomal RNA (rRNA) to identify the species of bacteria. This works because of the unique structure of rRNA in that it has regions of constant sequence with variable regions in between. This allows for a simple PCR primer design that will anneal to the constant regions while amplifying the variable regions. Each variable region is unique to a type of microbe and thus the identity of the microbes from the community can be determined as well as the relative abundance using next-generation sequencing (Morgan & Huttenhower, 2012). The benefit of this method is that it is relatively simple to build the sequencing library and it has become standardized with Illumina's sequencing technology, which further streamlines the process (Illumina microbes and metagenomics, 2015). Therefore this process has been done extensively on many microbiomes including those from many parts of the human body such as the gut and skin (Kelvin *et al.*, 2012, Yatsunenko *et al.*, 2012). Further, metagenomics has been expanded to also begin building reference genomes of the microbes in the human microbiome. However, this is a difficult task since the microbes cannot be easily separated from the community and therefore building the whole sequence is time consuming and costly. Further, reference genomes only show the potential for activity, not activity itself. To date, the majority of microbes in the gut microbiome have been

identified and many have reference genomes (Martin *et al.*, 2012). However, that only shows who's there, not what they are doing in the gut environment. Therefore, in order to determine how the gut microbiome plays a role in human health and disease, the activity or what the microbes are producing needs to be studied and identified.

### **1.3.2.b Microbial Activity**

If we look at the potential indicators for activity, the most logical three are the mRNA, proteins, and metabolites. Each represents a step in the activity of microbes from gene expression, to protein production, to what the proteins produce. Each detection target has their pros and cons when studying microbiomes. For example, metabolites are typically smaller molecules which can be easily detected using modern spectroscopy techniques like mass spectrometry and they provide an accurate image of the activity profile of the microbiome. At the same time, although extracting the metabolites does not need to maintain biological conditions, isolating all the diverse classes of molecules can be time consuming and difficult. Furthermore, the identity of the microbes producing the metabolites cannot be traced back and therefore only the activity of the microbiome, not the individual microbes, can be detected (Baker, 2011). Proteins, on the other hand, are the workhorse of the cell and would show a true and complete depiction of the activity in the microbiome. Moreover, proteins can be sequenced and identified by their structural components and therefore it is possible to trace the identity of the microbes to the activity. However, proteins are difficult to work with and since they are considerably larger, they are not as easy to measure as metabolites. Further, protein sequencing is complicated in a mixed culture and therefore proteomics is the most labor intensive of the

activity measurements (Sven-Bastiaan & Jehmlich, 2016). Finally, mRNA is the gene expression and thus represents the image of what proteins are being made. Using mRNA for activity detection is the easiest for tracing back to microbes since nucleic acid sequencing is well established and high-throughput. Thus metatranscriptomics could provide both activity information as well as identity of the microbes. However, RNA is highly unstable and easily degraded. Further, mRNA provides only a snapshot of activity at the time of collection and multiple proteins can be made from one mRNA so relative levels of activity are not as clear (Moran *et al.*, 2013). Since the goal is to determine activity in the gut microbiome as it relates to human health and disease development, identifying which omics method is most effective will be essential for moving forward with understanding the role the gut microbiome plays in Type 2 Diabetes.

## **1.4 Specific Aims**

### **1.4.1 Identifying the Role the Host Plays in Susceptibility**

In order to understand the complexity of Type 2 Diabetes susceptibility, all potential factors need to be examined. The first factor that likely plays a role in disease development is the genetic code. The genome influences the start of a person's development and phenotypes. If any environmental factor is to be understood, the way the genetic variation interacts with the environment needs to be understood first. Hence, for studying what the gut microbiome does for human health and disease, the host's genetics must be defined first. The first aim of this study is to complete a pilot study to examine the genetic variation that increases susceptibility for Type 2 Diabetes within a sub-population. By identifying the genetic patterns that influence disease development,

there can be a greater chance of early detection, new treatment, and understanding of the mechanisms involved in how these SNPs alter expression pathways in such a way as to increase susceptibility for Type 2 Diabetes.

#### **1.4.2 Building a Simple Model for the Gut Microbiome**

Once the host is defined, it will be possible to assess how the gut microbes interact with innate susceptibility towards disease. Due to the complexity of gut microbiome, developing a simple model is needed to study their activity and to develop new experimental approaches for determining how the gut microbes influence human health and disease. Thus the second aim is to build a simple model of the gut microbes and optimize an assay for detecting a specific activity in the model system.

## CHAPTER II

### PRINCIPLE COMPONENTS ANALYSIS OF A REFERENCE DATASET

#### 2.1 Introduction

##### 2.1.1 Sample Size for a Pilot Study

The majority of genetic studies involve hundreds to thousands of participants with large groups of researchers completing the task of processing the samples and performing bioinformatics based data analysis. However, when beginning a genetic study it is often the case that the lab resources are small and there is little preliminary data to support the hypothesis being tested. Therefore it is beneficial to complete a pilot study to assess if the hypothesis is worth investing in on a large scale. When completing a pilot study it is helpful to know the minimum participants necessary to get reliable preliminary results to test if there is genetic cause for the disease being studied. This was the first question we set out to answer. In order to determine how many participants we needed for our Type 2 Diabetes pilot study, we went to a reference dataset to mimic pilot study conditions. The dataset we chose was the Geneva study, which was a longitudinal study for Type 2 Diabetes. The study included nearly 6,000 participants with both case and control groups as well as nearly 1 million SNPs per sample using the Affymetrix Genome-Wide Human SNP Array 6.0. Through the large-scale data analysis, several SNPs were identified for Type 2 Diabetes susceptibility mostly in women of European decent (Zeggini, 2008). Since this is a validated genetic dataset, it can be used as reference data to isolate smaller

groups of the data to see if genetic variation can be detected using statistical tools and thus identify how many participants are needed to see some results.

### **2.1.2 What is Principle Components Analysis?**

The tool chosen to investigate sample size is called principle components analysis (PCA). It uses a statistical algorithm to look for patterns in variation in a dataset. Specifically, it projects coordinates along the greatest axis of variation and the projections are called principle components (pcs). There can be infinite pcs, but the majority of variation is explained using the first few pcs. These pcs can be plotted in a biplot and patterns in the variation can be detected by looking at these projections (Ringner, 2008). In the past, PCA has been used for genetic studies to examine patterns in population structure within a dataset. For example, Tian *et al.* used PCA to reduce false-positives in their genetic study. They did this by looking at ancestry specific SNPs to adjust their case and control analysis to account for variation between sub-populations. That is, by doing PCA on ancestral markers, they were able to adjust the case and control groups to reflect their sub-population identity and therefore account for natural variation between sub-populations that could normally be seen as relevant to the disease rather than difference in ancestry. Therefore, PCA is a useful tool for looking at population structure and has been proven to be compatible with genetic data. Thus PCA was chosen to analyze the reference dataset for detection of case and control groups based on genetic information alone. If the two groups form distinct patterns or populations in PCA using the genotypes alone without any other information, that would suggest that they are distinct enough to be identified as with or without the disease.

### 2.1.3 Bioinformatics Tools

Although PCA has been shown to be effective with genetic data, there are limits to the amount of data that PCA can handle. Many of the genetic studies that use PCA only analyze the ancestry SNPs for population structure and therefore a significantly reduced dataset is necessary for PCA analysis. In order to prune such a large dataset like that from the Geneva study, a bioinformatics tool is needed to handle the data. A common program used is called PLINK (Purcell *et al.*, 2007). It is an open-source program that runs on a Linux operating system and has a specific file format that is common in genotype studies. PLINK has the capability to modify the data, perform quality control measures, and complete analysis using various statistical tests. Thus it can be used to isolate smaller subsets of individuals to mimic a pilot study and complete measures to prune the data to a workable size for PCA. At the same time, even a pruned dataset will be significantly large and can thus create long processing times for a PCA program. Therefore a specific PCA program called flashPCA (Abraham & Inouye, 2014) was selected since it was designed to work with genetic data and can take PLINK files as input files without further processing. Since it was designed to work with larger genetic datasets, it also has a faster processing speed and can look at the genotypes for projecting the pcs. Thus these tools allowed for assessing the sample size and data pruning techniques for developing a genetic based pilot study.

## **2.2 Method**

### **2.2.1 Data Acquisition and Processing**

Data was obtained from the Geneva Study by submitting a dbGAP project request through the National Center for Biotechnology Information (NCBI). The request included the name of the principle investigator and affiliated institution, the title of the project, and a brief summary of the work to be completed on the data. After the request was approved, the data was accessible through a secure login on the NCBI website. Data was downloaded onto a computer with a 2 terabyte hard drive and Linux operating system. The data was encrypted to make downloading the large dataset easier and to protect the sensitive human information. Once downloaded, the data was decrypted using the provided protocols (dbGAP, NCBI).

### **2.2.2 Isolation of Subsets of Individuals**

A random subset of 100 individuals was removed from the larger data set. This set of 100 included 50 individuals with Type 2 Diabetes and 50 without diabetes. To ensure the reproducibility of the method, this was repeated two more times to create three different sets of 100 individuals each with the same case and control ratio. See Appendix A for all plink commands.

### **2.2.3 Data Pruning**

To filter the number of SNPs down to a size that is compatible with PCA analysis, two methods were tested. First, the data was pruned randomly using linkage disequilibrium (LD), which will select one SNP from a set of SNPs that have the same LD. That is, a single SNP was chosen at random from those sets of SNPs that travel

together when passed down generationally. Second, a chi-squared association analysis was used to select only the SNPs that were statistically significant below a 0.05 p-value. For equation 1,  $O_i$  is the observed frequency and  $E_i$  is the expected frequency, which is the average of all the observed frequencies. If  $X^2$  is above the critical value based on  $n$  observations, then the null hypothesis can be rejected and the difference in frequency is statistically significant and thus the p-value will be below the 0.05 threshold (Miller & Miller, 2010).

$$X^2 = \sum (O_i - E_i)^2 / E_i \quad (\text{Equation 1})$$

Following this, the phenotype information was removed from both pruned data files so the PCA would run blind using only the genotypes to determine the pattern of variation.

#### **2.2.4 PCA Analysis and Plotting**

Each no-phenotype file was run through flashPCA. The program produced principle components (pcs) for the data, which were used to create a PCA biplot using XLstat (Addinsoft). Specifically, the first two pcs were plotted in a scatter plot and then the points were colored based on case (green) or control (blue). The graphs were then assessed for separation of case and control. The process of completing the association analysis and PCA was done for the other two sets of 100 individuals.

#### **2.2.5 Reduction of Sample Size**

To determine if separation of case and control could be achieved with fewer samples, the three datasets were each reduced to 50 and 20 individuals. The 50% case and 50% control ratio was maintained with the smaller sample sets. Association analysis

was completed for each set and then the phenotype was removed from the files using the same PLINK commands. Finally, flashPCA was used to determine the pcs, which were again plotted using XLstat.

### **2.2.6 Further Reduction of SNPs**

Since PCA looks at a set of data for patterns in variation, the lowest limits were tested to determine the minimum amount of data required to see separation of the case and control groups. The SNPs were systematically reduced by taking sets of SNPs below lower thresholds of p-values and creating files for PCA analysis using the “extract” command as before. That is, SNPs were taken below a p-value of 0.005, 0.0005, and 0.00005 in order to take the number of SNPs down to the lowest p-values and therefore more statistically significant values. This process was completed for all three sets of 100 as well as the three sets of 50 and 20 individuals. These reduced sets were run on flashPCA and the pcs plotted.

## **2.3 Results/Discussion**

### **2.3.1 Data Processing**

The NIH data approval was completed, which allowed 1 terabyte of data to be downloaded onto a 2 terabyte hard drive with a Linux operating system. The data was decrypted into a usable format that included PLINK format files. This version of the data was used for all further analysis.

### **2.3.2 Isolating Small Sample Groups**

The idea behind this project was to mimic a pilot study using data that had already been acquired and published with genetic biomarkers for Type 2 Diabetes. This reduces

the amount of resources used when starting the project by first testing the methods and sample sizes on previously measured genetic data. Genetic studies can be costly due to working with human samples and using next-generation sequencing technology. Thus, by simulating our pilot study by taking smaller groups of samples from a larger study, the process can be tested before continuing on with our samples.

Three sets of 100 individuals with a 50/50 ratio of healthy and diseased in each were isolated from the nearly 6000 individuals in the Geneva study. The samples were randomly selected from the study pool and isolated from the larger data using command 1. The data was then tested and confirmed that all 100 samples with the proper phenotype was included in the new data files (data not shown).

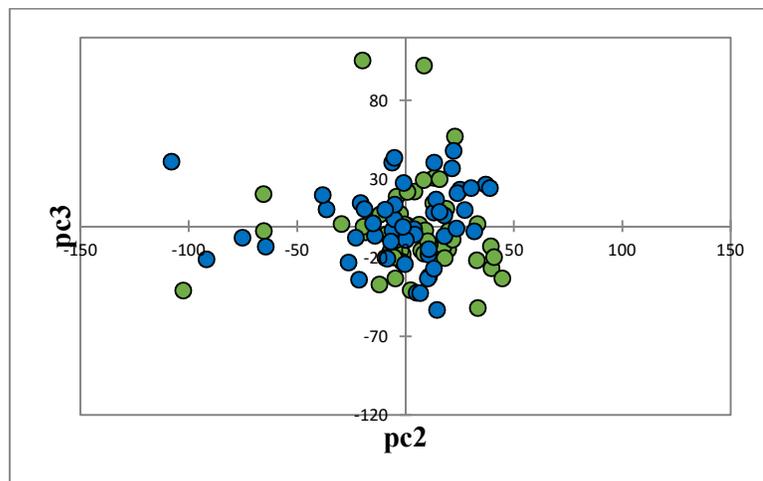
### **2.3.3 Pruning Methods**

Even with flashPCA, which has an extended capacity as a PCA program to accommodate larger genetic data sets, there is still a limit to what the program can handle. That is, the number of SNPs available from the chip was just under 900,000 and therefore a pruning method to reduce the number of SNPs for PCA analysis was necessary. The two methods chosen were meant to be algorithms that are common with genetic analysis studies and therefore not outside the norm of how genetic data is processed. The first method pruned the data based on LD using command 2 and reduced the SNPs to 11,391. The second method used association analysis with command 4 to look for statistically significant SNPs based on the comparison between the case group and the control group. This process reduced the SNPs to 37,319 below a p-value of 0.05. Without using the phenotype information to plot the samples, the three datasets were

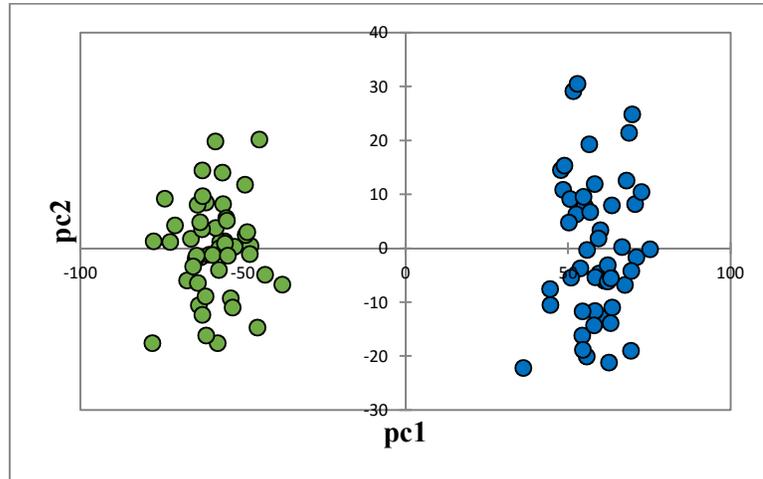
analyzed with flashPCA. The biplots are shown in Figure 1 with case samples colored green and the control samples colored blue. For the LD pruning (**Fig 1A**), the samples do not separate, but rather cluster together into the center. On the other hand, the association analysis separated the two groups (**Fig 1B**) into distinct clusters. Therefore using LD does not provide prune the data in such a way as to be able to detect patterns. Since association analysis is a statistical method, the data is essentially analyzed for genetic biomarkers and then the samples are pruned by their statistical significance. The biplot indicates that there is enough genetic variation to support a genetic difference between the case and control group and the groups can be separated with only 100 samples, which is much less than the 6000 samples in the Geneva study. Thus pilot studies can provide useful information with smaller sample sizes to establish initial genetic links to a phenotype or disease. Further, the association analysis pruning method was more effective for detecting patterns and therefore was used for all future work with PCA.

Figure 1. PCA Biplots of Two Pruning Approaches

A.



**B.**



A. LD Pruning. This biplot includes 100 samples and shows the set of SNPs selected using LD pruning of the data. There are 11,391 SNPs included in the PCA analysis. Green = healthy. Blue = diseased. B. Association Analysis Pruning. This biplot also has 100 samples with a p-value of below 0.05 as the cutoff value. There were 37,319 SNPs included in the PCA analysis. Two Individuals were removed as outliers due to lying significantly farther away from the rest of the samples. In both biplots, green is case and blue is control.

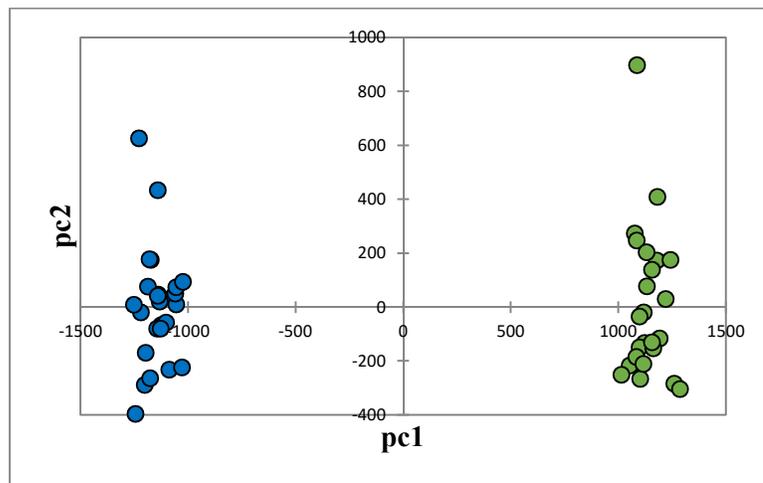
### 2.3.4 Testing the Reduction of the Data

After the 100 samples were confirmed to separate into two respective groups in the PCA biplot, there were two questions that needed to be addressed. One, could the number of samples be further reduced while maintaining the separation seen and two, could the number of SNPs put into the PCA analysis be reduced? For the first question, sample sizes of even 100 can still be quite large to handle for a small pilot study and therefore it would be ideal if fewer samples can be used while still detecting patterns in the genetic data. To test this, the three sets of 100 were reduced to 50 and 20 sample sizes with the same 50/50 case and control ratio. They were then pruned and analyzed as before with the association analysis and the PCA biplots created. The 50 samples had

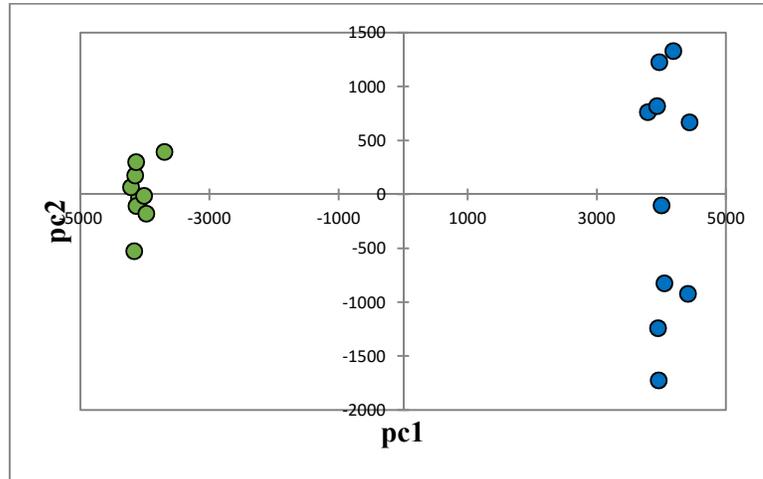
36,528 SNPs and the 20 samples had 32,085 SNPs. Thus the number of SNPs was relatively similar between the different sample sizes. Further, the PCA biplots show separation between case and control for the 50 samples (**Fig 2A**) and for the 20 samples (**Fig 2B**). However, the separation is not as distinct for the 20 samples and therefore a sample size between 20 and 50 would be ideal for the pilot study. The process was repeated on all three sets of 100 samples and the results were confirmed (**Fig 1 & 2, Appendix B**).

Figure 2. Reducing Sample Size

A.



**B.**



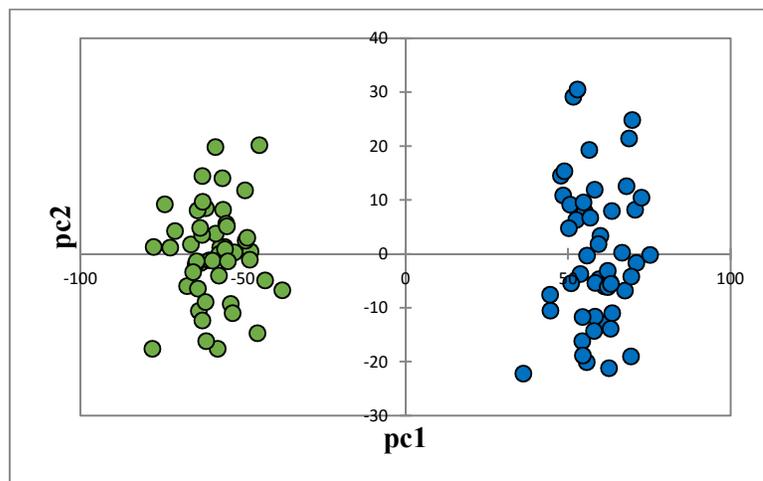
A. This biplot has 50 individuals with 36,528 SNPs using a p-value of 0.05 as the threshold. B. This biplot has 20 individuals with 32,085 SNPs using a p-value of 0.05 as the threshold. In both biplots, green is case and control is blue.

Although association analysis was an effective pruning method, the purpose of completing genotyping studies is to identify a small set of SNPs associated with a particular disease. Therefore we set out to further reduce the number of SNPs to determine if we could see separation of case and control and therefore how effective our analysis would be for genetic biomarkers in small sample sets. To systematically reduce the number of SNPs, the p-value was reduced by a factor of 10 each time to bring the SNPs to the lowest possible number while maintaining group separation. The p-values were reduced down to  $5 \times 10^{-5}$  which resulted in approximately a 10-fold reduction in the number of SNPs with each reduction in p-value. First, the set of 100 samples was tested (Fig 3) and separation was maintained down to the lowest p-value with 23 SNPs. The reduction of SNPs was then tested on the set of 50 samples (Fig 4) and the set of 20 samples (Fig 5). In each of these cases the two groups maintain their separation between

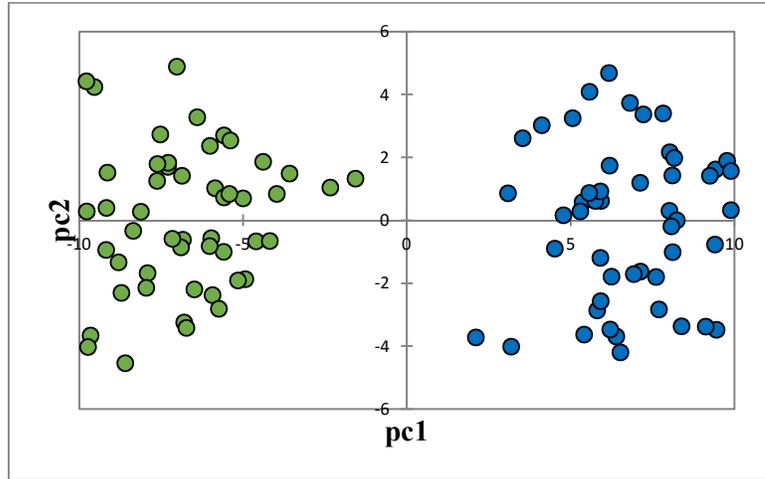
case and control. However, as the number of SNPs reduces to the lowest p-value, the samples start to spread out and the two groups look less distinct. This could suggest that there is a limit on how little data PCA can manage while creating viable pcs for biplots or it could support the idea that there are many SNPs that build up to Type 2 Diabetes and there is a limit to how few can be present while maintaining the difference between the groups. This process was repeated on the other two sets of samples and the results were the same (**Fig 3-8, Appendix B**). All in all, case and control maintained their separation even down to double digit number of SNPs and therefore the association analysis and PCA are feasible methods for data analysis in a genetic biomarker pilot study.

Figure 3. Set of 100 Samples with Number of SNPs Reduced

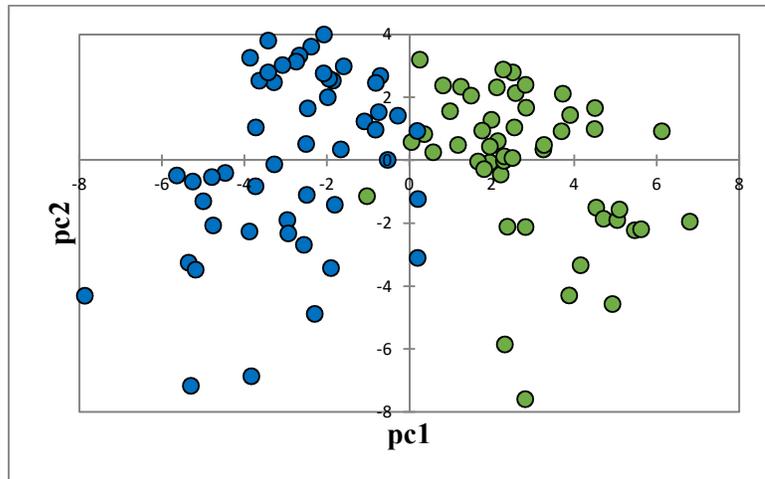
A.



B.



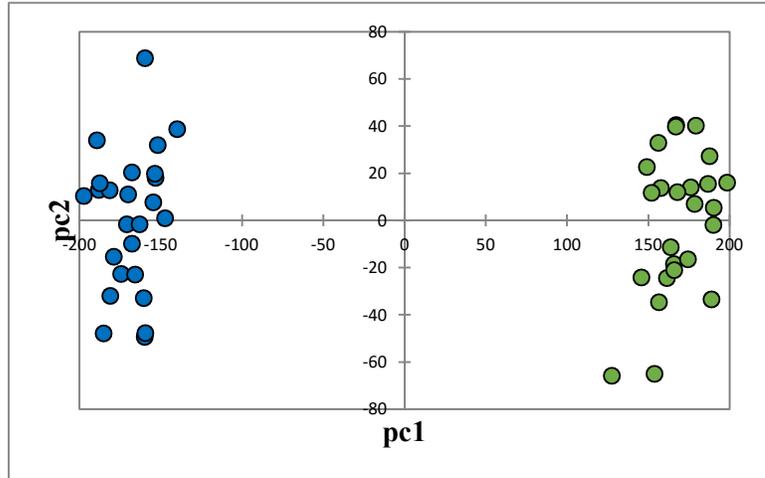
C.



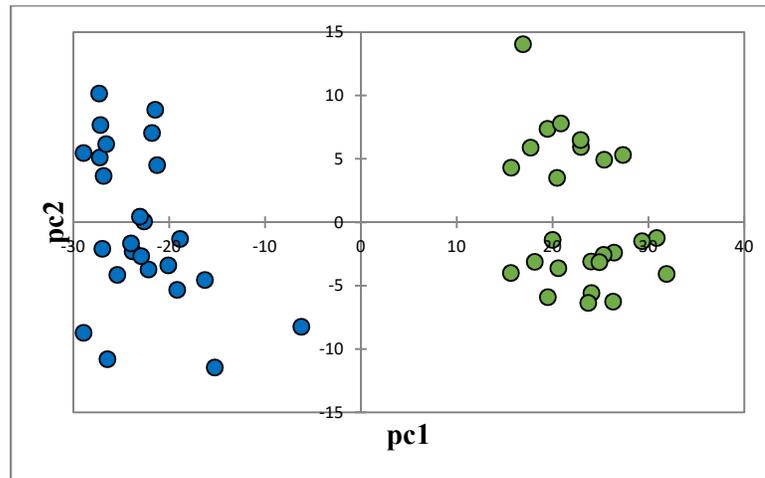
PCA biplots are of 100 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 3,032 SNPs for PCA analysis. Two samples were removed as outliers. B. P-value of  $5 \times 10^{-4}$  as the threshold with 234 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 23 SNPs. Separation of case and control is maintained for all reduction of SNPs.

Figure 4. Set of 50 Samples with Number of SNPs Reduced

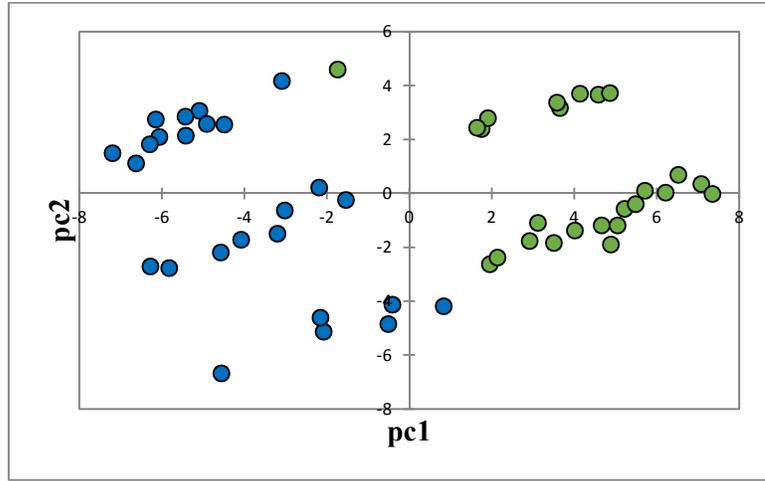
A.



B.



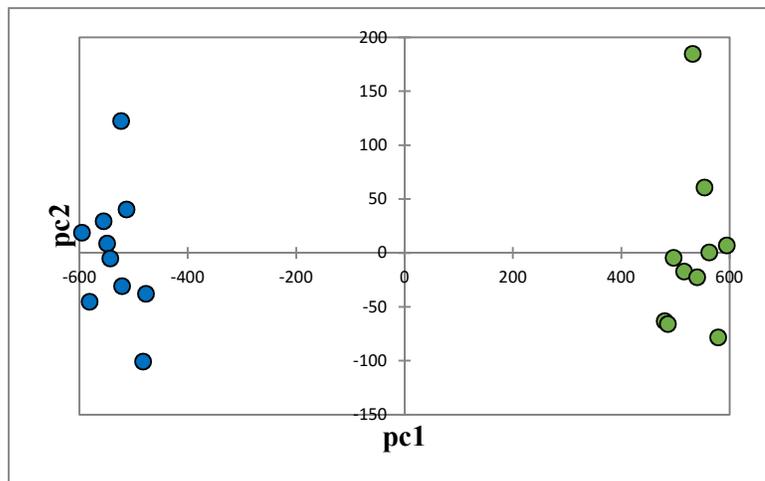
C.



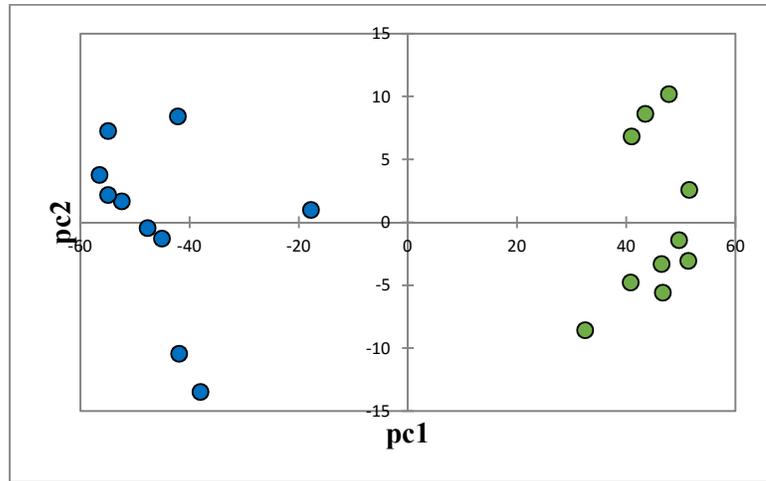
PCA biplots are of 50 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 3,050 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 282 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 23 SNPs. Separation of case and control is maintained for all reduction of SNPs.

Figure 5. Set of 20 Samples with Number of SNPs Reduced

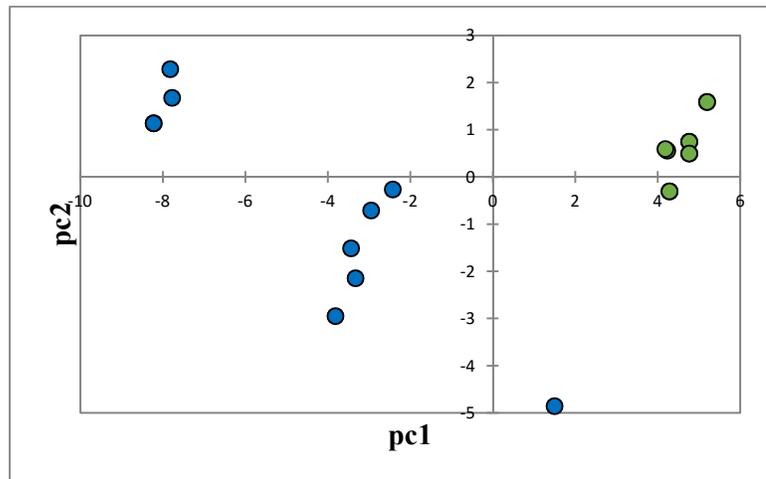
A.



B.



C.



PCA biplots are of 20 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 2,444 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 142 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 12 SNPs. Separation of case and control is maintained for all reduction of SNPs.

## 2.4 Conclusion

After simulating a pilot study with the Geneva data, we were able to use PCA to examine the sample size for a pilot study. The results indicated that between 50 and 20

individuals would be sufficient for detecting genetic patterns for disease. Therefore a small pilot study to investigate potential SNPs linked to Type 2 Diabetes is feasible. Furthermore, the results of this small study would be reliable preliminary data for beginning to understand how genetics plays a role in disease development. To confirm that, a small set of SNPs was isolated from the larger set by systematically reducing the p-values. The smaller the p-value, the more statistically significant and thus this technique was a logical approach to mimic the detection of the most relevant SNPs. The results indicated that even down to the lowest p-values given in the association analysis, the case and control separated. Therefore even small data sets are sufficient for results. Future work with this could include examining the use of PCA as a supplement to other forms of data analysis in genetic studies and using PCA to further sort or subcategorize SNPs for significance.

## CHAPTER III

### GENETIC BIOMARKERS FOR TYPE 2 DIABETES IN A HISPANIC POPULATION

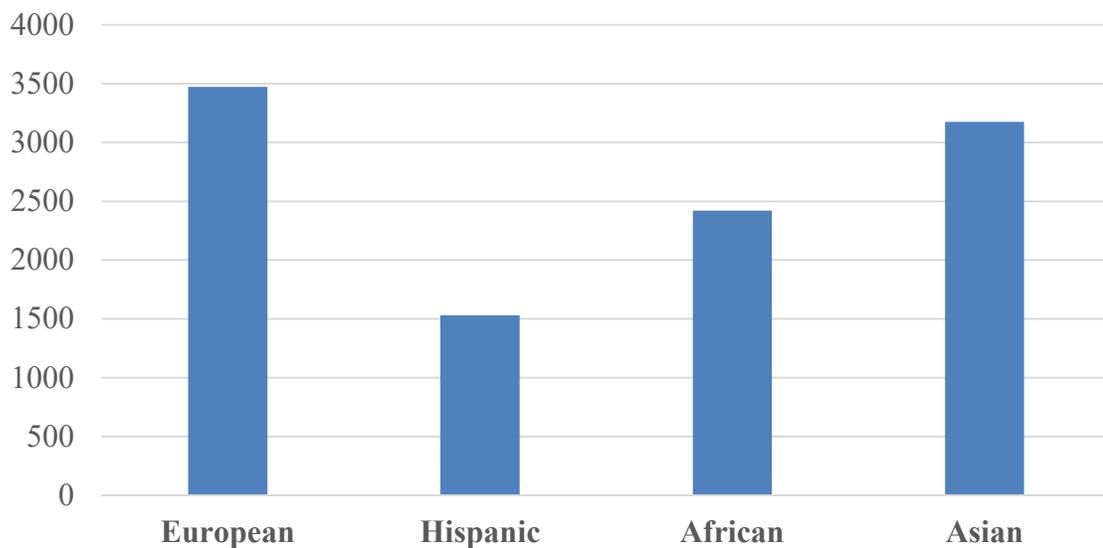
Parts of the work presented in Chapter III have been published in *Diabetes Research and Clinical Practice* and is referenced as: Watson AL, Hu J, and Chiu NHL (2015). Single Nucleotide Polymorphism in Type 2 Diabetes among Hispanic Adults. *Diabetes Research and Clinical Practice*, 108: e25-e27

#### 3.1 Introduction

It has been well established that Type 2 Diabetes is a worldwide issue affecting people of many cultures and environments. However, different populations are affected by Type 2 Diabetes more than others. In particular, Native Americans have the highest rate followed by Hispanic, African, and Asian heritage, with Caucasian heritage having the lowest rate of the disease (CDC Diabetes Statistics Report, 2014). However, if we look at the populations best studied for Type 2 Diabetes, they are mostly European Caucasian. Moreover, Hispanics have the second highest rate of Type 2 Diabetes, but they are the least studied according to the literature (**Fig 6**). Thus there is a need to study and understand why the Hispanic population suffers disproportionately from Type 2 Diabetes. Although it might be logical to assume that unique environmental factors are increasing Hispanics risk for developing diabetes, early studies showed that obesity, caused by the cultural diet, alone could not explain increased diabetes cases (Stern *et al.*, 1983) nor could socioeconomic status (Chakraborty *et al.*, 1986). In fact, Chakraborty *et*

*al.* found that with increased socioeconomic standing in a Mexican American population usually also coincided with a greater percentage of Caucasian gene markers compared to Native American gene markers. At the same time, it is known that Native Americans have genetic variations that increase their susceptibility to Type 2 Diabetes (Baier & Hanson, 2004) and Mexican American populations are a mixture of Native American and European Caucasian (Reed, 1974; Parra *et al.*, 2011). This suggests that environment alone cannot explain the increase in susceptibility to Type 2 Diabetes, but rather there is a genetic cause that plays a role in disease development for Hispanic populations. Since it has been demonstrated that there is a genetic factor that increases Type 2 Diabetes susceptibility for Hispanics and there is a need for further study in this population, the pilot study for identifying the genetics of the host was completed on Hispanic adults.

Figure 6. Number of Diabetes Studies by Population



Search results for studies done specifically for diabetes by population were done using the NCBI Pubmed site and was last updated September 2015. As can be seen, the European population is the most studied and the Hispanic population is the least studied.

## **3.2 Method**

### **3.2.1 Sample Selection**

A reasonable sample size of 36 participants was recruited for this pilot study at clinics and communities around central North Carolina. Eligibility included self-identification as of Hispanic descent, at least 18 years of age, and ability to speak Spanish or English. Of the 36, there were 18 with Type 2 Diabetes and 18 who did not have Type 2 Diabetes. IRB approval was completed before starting and all participants gave permission to use their de-identified information and blood sample for this study. Hispanic adults were interviewed for basic health background and family history of Type 2 Diabetes as well as tested for HbA1c levels with a cutoff of >6.5% as an indicator of diabetes (Tatsch *et al.*, 2012). Finally, participants provided a blood draw for DNA isolation using the PAXgene blood DNA tubes (Qiagen), which were stored at 4°C until further processed.

### **3.2.2 DNA Isolation and Characterization**

Genomic DNA was isolated from the blood samples using the PAXgene blood DNA isolation kit (Qiagen). To isolate the DNA, the blood from the collection tube was inverted to mix the blood and then poured into a tube with lysis buffer and mixed by inverting the tube 5 times. The mixture was then centrifuged for 5 min at 2500 x g at room temperature in a swing-out rotor centrifuge. The supernatant was removed and then 5 mL of washing buffer was added and the tube vortexed for a few seconds and then centrifuged for 3 min at 2500 x g. The supernatant was discarded and then 5 mL of digestion buffer with 50 µL PreAnalytiX Protease was added to the tube. The tube was

vortexed for 20 sec at high speed to fully dissolve the pellet. Then the mixture was incubated in a water bath at 65°C for 10 min. After 5 sec of vortexing, 5 mL of 100% isopropanol was added and the tube inverted 20 times until white clumps of DNA were visible. The tube was then centrifuged for 3 min at 2500 x g and the supernatant carefully discarded. The pellet in the tube was allowed to dry on a piece of kim-wipe for 1 min. Then 5 mL of 70% ethanol was added and the tube vortexed for 1 sec followed by centrifugation for 3 min at 2500 x g to wash the DNA pellet. The supernatant was removed and the pellet was again allowed to dry on a kim-wipe completely for about 5 min. Finally, 1 mL of resuspension buffer was added and the mixture incubated in a water bath at 65°C for 1 hour and then at room temp overnight. Once the DNA was thoroughly dissolved, it was stored at -20°C for future use.

Genomic DNA was characterized for concentration. This was determined using PicoGreen fluorescent dsDNA stain and a POLARstar OPTIMA plate reader (BMG Labtech). A standard curve was made with double stranded lambda DNA standard in TE buffer. The concentrations ranged from 2 µg/mL to 2 ng/mL with a TE buffer blank. Equal volumes of standard and PicoGreen reagent were mixed in wells in a 96 well plate. The solution was incubated for 2-5 min protected from light and then the fluorescence measured with excitation at 480 nm and emission at 520 nm. The blank was subtracted from each value and a standard curve was plotted of concentration versus fluorescence. For the DNA samples, each was diluted 1:5000 in TE buffer and then mixed with an equal volume of PicoGreen reagent in a 96 well plate. The mixture was incubated for 2-5 min protected from light and then the fluorescence measured as before. The concentration

in each diluted sample was determined using the standard curve and then the concentration of the original DNA samples were calculated based on the dilution factor. Each sample was diluted to 50 ng/ $\mu$ L for loading on the genotyping chip.

### 3.2.3 Genotyping

All 36 samples were genotyped on the Cardio-Metabochip (Illumina) using the protocols provided. The Cardio-Metabochip is a genotyping chip designed by a consortium of experts in Type 2 Diabetes and other metabolic and cardiovascular diseases (Voight *et al.*, 2012). The details of the chip are shown in Table 1. After genotyping, the raw fluorescent data from the chip was converted to genotypes for each SNP using GenomeStudio (Illumina).

Table 1. The Cardio-Metabochip

	Fine-mapping Targets		Replication SNPs	Total SNPs
	# Loci	# SNPs		
<b>Type 2 Diabetes</b>	34	16,717	5,057	21,774
<b>Cardio-Metabochip</b>	257	122,241	63,450	185,691

The SNPs included on the Cardio-Metabochip were selected by a consortium of experts on genetic biomarkers for diabetes and related cardiovascular and metabolic disorders. Fine-mapping targets are included to explore genes that are suspected to be involved in disease traits and replication SNPs are specific variations that have already been identified to be associated with a disease.

### **3.2.4 Data Processing**

There were several tests completed to check for the quality of the data and to remove any SNPs or samples that would cause a bias or error to the data analysis. First, SNPs were removed if they did not contain any chromosome information since it was essential for PLINK to complete the data analysis. At the same time, the data was converted to a binary format for faster processing speed. Then phenotypes and genders were added into the data to facilitate quality control and analysis. Each text file contained the appropriate information in the proper format required by the PLINK program to insert the new information into the data files. Finally the participant identification numbers were changed to further de-identify the data and to simplify the code for each individual to a single number.

### **3.2.5 Quality Control**

In order to reduce the possibility of false positives or bias in the data, several quality control measures were completed. First, a sex-check was done to ensure there was no crossover of samples on the genotyping chip by examining the included gender specific SNPs on the chip and cross checking with the sex listed in the data. Second, the data was checked for any missing SNPs due to an error during the measurements on the genotyping chip. Any missing SNPs were removed with the same command as with the missing chromosome information. Finally, the Hardy-Weinberg equilibrium test was completed to remove any SNPs that have unusual relatedness between samples due to possible relatedness of the participants.

### 3.2.6 Data Analysis

The data was analyzed using a chi-squared based association analysis in PLINK. This produced a list of SNPs and associated p-values. Then the data was converted to a format for Haplowview using gPLINK. In Haplowview, the data was uploaded and sorted for p-value from low to high. A Manhattan plot was created to view the entire set of data and visualize the variation by chromosome. Then the set of SNPs with the lowest p-values were compared to the Cardio-metabochip supplemental data to determine which were chosen for potential association to Type 2 Diabetes. The allelic frequency for each of the top associated SNPs was calculated using equation 2 and then the nearest gene for each was determined using the dbSNP database from NCBI.

$$\text{Allelic Frequency} = \sum \text{allele 1} / \text{total alleles} \quad (\text{Equation 2})$$

### 3.2.7 Comparison to Reference Populations

Reference populations were collected from the HapMap Program (International HapMap Consortium, 2010). Specifically, three populations were used: Hispanic, Central European, and Japanese/Chinese. Each of the top 26 SNPs were identified in the reference populations and the allelic frequency was recorded. The reference populations were considered healthy. Then the allelic frequency of the case group was compared to the reference populations by calculating a percent difference using equation 3.

$$\% \text{Diff} = \left| (S_{AF} - R_{AF}) \right| / R_{AF} \times 100 \quad (\text{Equation 3})$$

From equation 3,  $S_{AF}$  is the allelic frequency of the sample for the case group and  $R_{AF}$  is the allelic frequency of the reference group. After the %Diff was calculated for each SNP, it was compared to the reference populations. The average and standard deviation was calculated and the SNPs with a percent difference above the average were highlighted.

### **3.2.8 PCA**

The top 26 SNPs were placed in a Microsoft Excel spreadsheet with their allelic information. The letters were converted to numbers with 0 being homozygous for the ancestral allele, 1 being heterozygous, and 2 being homozygous for the SNP allele. The data was then measured using PCA with XLstat and the biplots and coordination plots were graphed.

## **3.3 Results/Discussion**

### **3.3.1 Participant Selection**

Based on the PCA data from the reference material, a sample size between twenty and fifty participants was acceptable for a pilot study of genetic variation. It was essential to have enough participants to be able to detect if there is a genetic factor that increases the susceptibility for Type 2 Diabetes in a Hispanic population and therefore we used the results from the initial PCA study to determine the number of participants to select. There were 36 participants, half of which had Type 2 Diabetes and half who were healthy to serve as a control group (**Table 2**). All were adults with an age range from 19 to 70 and about half of them had family history of diabetes. The HbA1c test is used to measure how glycosylated the hemoglobin has been over the past 3 months. It is an indicator of how

concentrated the blood sugar has been for a period of time and is easily measured with a small drop of blood. It is standard that an HbA1c of greater than 6.5% is an indicator for Type 2 Diabetes (Tatsch, 2012). All participants in the case group had an HbA1c above the threshold with an average of 7.61%. Those in the control group were below the threshold with an average of 5.47%. This could be deemed a bit high for a control group average, however a number of the healthy participants could be classified as pre-diabetic based on their HbA1c levels. Since they had only slightly elevated blood sugar levels and did not have diabetes at the time of measurement, they were kept in the control group.

Table 2. Participant Demographics

	Case	Controls
<b>n</b>	18	18
<b>Sex: Male</b>	5	4
<b>Female</b>	13	14
<b>Age (Mean ± SD)</b>	51 ± 13	43 ± 11
<b>Age Range</b>	19 - 70	25 - 62
<b>HbA1C</b>	7.61% ± 1.31	5.47% ± 0.49
<b>Family History</b>	11	8

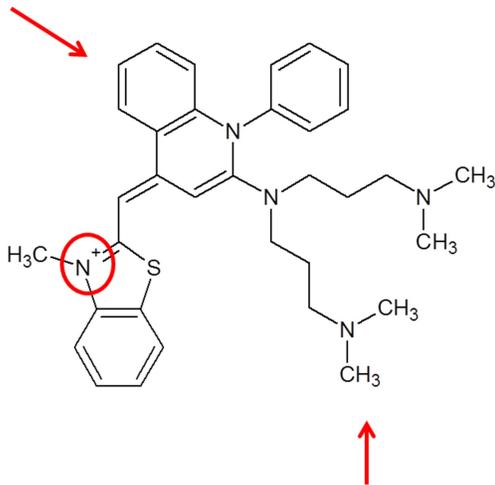
There were 36 participants in the pilot study. Specific information was recorded about each individual at the participation interview and is shown in the table above. All participants were adults and all those in the case group had an HbA1c above the 6.5% threshold. There were about 75% females and 25% males and approximately half of the entire group had family history of diabetes.

### 3.3.2 Genomic DNA Isolation

Each blood sample was extracted for genomic DNA. The kit used for extraction was designed specifically for isolating DNA from blood samples. This allows for an optimized and quick method for DNA isolation, which is beneficial when working with a larger number of samples and with complex biological samples. All blood samples were successfully extracted for genomic DNA. Further, determining an accurate DNA concentration in each sample was important. This is due to the fact that the genotyping chip works within a very specific concentration range. That is, in order for proper determination of the genotypes, the amount of DNA loaded on the chip has to be known exactly. Therefore, a highly sensitive fluorescent assay was used to determine the concentration. The DNA stain used was PicoGreen from ThermoFisher (**Fig 7**). It is a fluorescent stain that is highly selective for double stranded DNA (dsDNA) with nearly a thousand fold enhancement in fluorescence upon binding with dsDNA. Thus, even if single stranded DNA or RNA is present in the sample, only the concentration of dsDNA will be detected. PicoGreen also has a large linear dynamic range, allowing for detection of a variety of concentrations including very low concentrations, making it a very sensitive detection method. Using PicoGreen, the concentration of each DNA sample was successfully determined (**Table 3**). The concentrations between samples varied significantly. However, that can be explained by two factors. First, the entire blood sample was used to extract DNA. Therefore the volume was not measured and varied in each blood draw. Second, the DNA is specifically isolated from the white blood cells. Since blood is a complex mixture of many cells and compounds, the number of white

blood cells can vary drastically depending on the person. This variety between each sample is enough to cause a wide difference between DNA concentrations. Despite this variability in concentration, it is important to note that there was enough DNA isolated from each sample except number 8 to be able to get the ideal 50 ng/ $\mu$ L for the genotyping chip. Sample 8 was loaded undiluted onto the chip and was able to be measured despite the lower concentration.

Figure 7. PicoGreen Structure



PicoGreen has little fluorescence by itself, but has over 1000 fold increase in fluorescence upon binding with dsDNA. Further, the structure of the molecule has three main components that allows it to selectively bind with dsDNA. The quinolinium group intercalates into the DNA, the positive charge on the thiazol electrostatically interacts with the negatively charged backbone of the DNA, and the bimethyl aminopropyl groups act as arms to interact with the minor groove of the double helix structure to stabilize the PicoGreen-dsDNA complex (Dragan *et al.* 2010).

Table 3. Genomic DNA Concentration from Human Blood Samples

<b>Sample Number</b>	<b>Conc. DNA (ng/μL)</b>						
1	348.4	10	678.7	19	132.2	28	401.9
2	720.2	11	525.2	20	341.3	29	241.6
3	92.11	12	517.8	21	443.9	30	99.06
4	746.8	13	129.1	22	319.6	31	182.0
5	1114	14	449.1	23	119.6	32	514.3
6	856.3	15	292.6	24	146.7	33	273.5
7	253.6	16	191.7	25	564.5	34	151.0
8	32.50	17	1162	26	304.8	35	1162
9	1139	18	208.7	27	193.7	36	418.5

All blood samples were extracted for genomic DNA and then the DNA was characterized for concentration using the PicoGreen Assay. All samples except number 8 were diluted to 50 ng/μL for genotyping.

### 3.3.3 Data Preparation for Analysis

Data was obtained from GenomeStudio (Illumina) in a format specifically for use in PLINK. In order to complete the data analysis, certain processing had to be done to the raw data to include some phenotypic information and formatting for association analysis. First, any SNPs that did not have chromosome information in the MAP file were removed using command 10 (Appendix A). The file snplist.txt contained a single column list of the SNPs to be removed, which PLINK used as a reference to complete the

command. The “make-bed” option converts the data to a binary format, which combines the data into a new layout as well as converts the genotypes to a binary code. This format allows for a reduction in processing time for each new command and so it was maintained for the rest of data processing and analysis.

The next set of processes added all the necessary phenotype information since the raw data from the chip did not contain any. Command 11 and 12 had a related text file which contained the family ID, the individual ID, and then the phenotype or gender respectively. The phenotype was listed as either positive or negative for Type 2 Diabetes. The gender information was essential for completing one of the quality control checks and the phenotype was needed for the case-control association analysis. Finally, the individual ID was updated to further de-identify the samples and to simplify the code associated with each participant for easier note taking using command 13.

Table 4. Quality Control Results

<b>Quality Control Test</b>	<b># SNPs Removed</b>
Chip-reading error	50,261
Sex-Check	0
Hardy-Weinberg Equilibrium	2807
Total SNPs Available for Analysis: 132,623	

Three different measures were used to complete quality control assessments on the genotyping data in order to reduce potential bias or errors in the data analysis. Both SNPs

and samples were assessed for quality, but only SNPs were removed due to failing a quality control test.

### 3.3.4 Quality Control

When using a statistical tool to analyze a data set, it is important to perform quality control measures to remove potential sources of error or bias in the data that could result in a false-positive. The measures chosen checked for crossover contamination, errors due to the genotyping process, and bias that could be introduced due to relatedness of the participants to one another. This way both the measurements and the inherent genetics are controlled for error.

The first check was a “sex-check” using command 14. This produces a list of what gender each sample should be based on the gender specific SNPs included on the genotype chip. This can then be checked against their actual gender to make sure the check was accurate. All samples identified as their actual gender except sample 24. For that sample, it came back as no identified gender. Since this does not actually indicate a crossover of samples, merely an error in the chip read, and the sample size is fairly small, sample 24 was not removed from the sample set. Second, using command 15, a list of any SNPs with greater than 5% missing genotypes amongst the samples was created. This list was then transferred to a text file to be able to remove those SNPs using the same “exclude” command as in command 10. There were ~50,000 SNPs removed due to this error from the genotyping chip (**Table 4**). Finally, command 16 was used, which tests the data using the Hardy-Weinberg equilibrium (HW). This test looks at SNPs from a perspective of high frequency due to close relationship between participants. It is

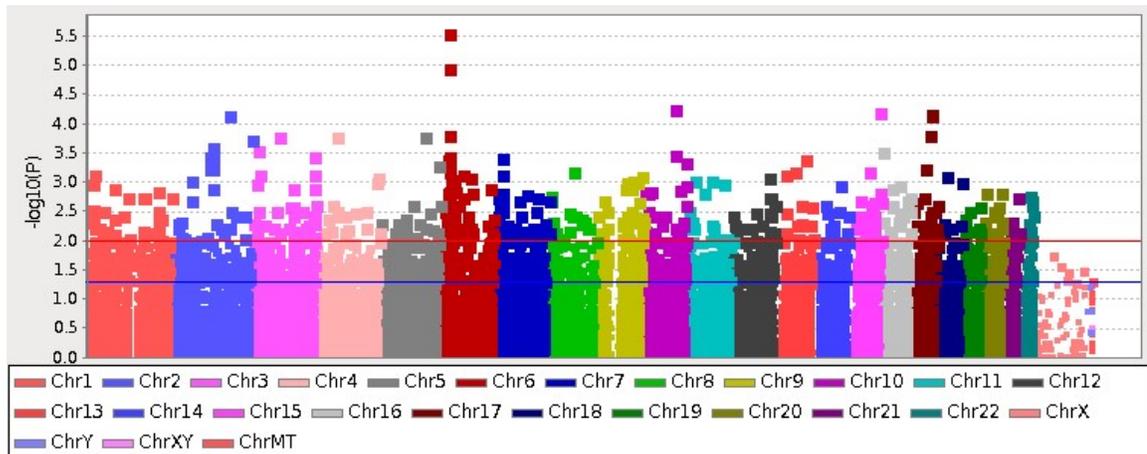
common for high HW p-values when there is inbreeding and lack of diversity in a population (Wigginton, 2005). Since we are examining a specific Hispanic population for genetic biomarkers, it was important to make sure there wasn't any bias introduced in the analysis if they all were closely related. There were about 2,800 SNPs removed due to failing the HW test (**Table 4**). This left a little over 132,000 SNPs per sample for data analysis.

### **3.3.5 Association Analysis**

Using command 17, the association analysis was completed on the data. This produced a list of every SNP and its associated p-value. The lower the p-value the more statistically significant the SNP is for the tested trait, which in this case is susceptibility to Type 2 Diabetes. In order to handle and analyze the data, it was converted to a form that can be used in Haplowview, which is a bioinformatics software that allows sorting and processing of genetic data. In Haplowview the association results were sorted by p-value from low to high and the data with the lowest 10% of p-values was selected for further processing. A Manhattan plot was created in Haplowview to visualize the significance in the variation (**Fig 8**) and this was used to assist in reducing the top SNPs to 10% of the lowest p-values due to the lessening of the SNP density at that point in the plot (red line). These SNPs were compared to the reference data associated with the Cardio-metabochip to determine which of these top 10% SNPs were associated with Type 2 Diabetes. During this process, the nearest gene for each of these top SNPs was also determined by using the dbSNP database from NCBI. This database contains all genetic variation for humans along with a variety of information known about each SNP including gene and type of

mutation present. The results of this analysis produced 26 SNPs associated with Type 2 Diabetes (**Table 5**). The rs# is a standard code used to identify SNPs. The table shows the name of each SNP, the chromosome where it is located, the two alleles with the SNP allele listed second, the nearest gene, and p-value. All SNPs have a significant p-value and are fine-mapping SNPs from the chip. Therefore, although these SNPs are located in regions that have been associated or suspected for Type 2 Diabetes susceptibility, none of the 26 SNPs have been directly linked to Type 2 Diabetes before. Thus we have identified new potential biomarkers for Type 2 Diabetes susceptibility in a Hispanic population.

Figure 8. Manhattan Plot of Genotyping Data



A Manhattan plot shows all the SNPs sorted by chromosome location after association analysis. The blue line is the minimum p-value for statistical significance and the red line is an arbitrary value chosen based on the lessening of SNP density. All SNPs above the red line were examined further for Type 2 Diabetes related SNPs.

Table 5. Top 26 SNPs

#	SNP ID	Chr	Alleles 1/2	Nearest Gene	p-value
1	rs17497477	17	A/C	TBC1D3P1-DHX40P1	1.39E-04
2	rs2185756	10	C/A	EIF2S2P3	2.94E-04
3	rs4383556	3	A/C	IGF2BP2	3.14E-04
4	rs17293846	3	C/T	IGF2BP2	3.14E-04
5	rs17826758	3	G/A	IGF2BP2	3.14E-04
6	rs73175565	3	C/T	IGF2BP2	3.14E-04
7	rs73175555	3	G/T	IGF2BP2	6.55E-04
8	rs191990500	3	C/A	IGF2BP2	9.50E-04
9	rs9834931	3	A/C	IGF2BP2	1.09E-03
10	rs7646419	3	A/G	IGF2BP2	1.09E-03
11	rs6780808	3	A/G	IGF2BP2	1.09E-03
12	rs6781019	3	T/C	IGF2BP2	1.09E-03
13	rs6457742	6	T/C	unknown	1.09E-03
14	rs2891226	14	T/C	NPAS3	2.14E-03
15	rs1320195	10	T/A	CDC123	2.70E-03
16	rs10875142	1	C/T	unknown	3.19E-03
17	rs12213132	6	G/A	CDKAL1	3.19E-03
18	rs7005401	8	A/G	CRISPLD1	3.19E-03
19	rs9848681	3	C/T	IGF2BP2	3.27E-03
20	rs2682919	15	C/T	unknown	3.27E-03
21	rs6791275	3	T/C	IGF2BP2	4.34E-03
22	rs289107	15	C/A	FLJ38723	4.43E-03
23	rs7300366	12	A/G	IFFO1	4.59E-03
24	rs7697417	4	A/G	BEND4	5.83E-03
25	rs72657613	6	G/A	CDKAL1	5.83E-03
26	rs188617336	6	T/C	CDKAL1	5.83E-03

After analysis there were 26 SNPs associated with Type 2 Diabetes in the Hispanic population studied. The rs# is the common code used for identifying SNPs. The two possible alleles are listed for each SNP as well as the nearest gene and p-value from the association analysis. All SNPs are located in introns or near the listed gene. Several SNPs are not close to a gene and therefore they are listed as unknown.

### 3.3.6 Comparison to Reference Populations

Since this study was a pilot study with a smaller sample size, secondary analysis was completed to attempt to validate the top 26 SNPs identified through the association analysis. This was done by comparing the allelic frequencies of the identified SNPs to the reference allelic frequencies of related populations. It is known that Mexican Hispanic populations have genetic similarity between Native American populations and European populations due to the colonization of the Americas by Europeans (Parra, 2011).

Therefore two populations were chosen to be similar in genetic background to our population: another Mexican population and a central European population. These sets of data were obtained from the HapMap project (International HapMap Consortium, 2010), which is a consortium developed to identify all genetic variation in all populations across the Earth. We also examined a population that should be genetically different from our Hispanic population, which was a Chinese and Japanese population group. All reference populations were assumed to be healthy. Using equation 3, the percent difference between these reference populations and the case group of our Hispanic population was calculated (**Table 6**). The average %Diff was calculated and used to identify those SNPs that were above the average and therefore significantly different between the two groups. There were five in each of the two similar populations that had a %Diff above the average with many more right around the average. This supports the fact that there is a difference in allelic frequency between other healthy populations and our Type 2 Diabetes group. Further, the %Diff for the Chinese and Japanese group was very different, many of which had values that were more than double the %Diff with related

populations. This is expected since this population is very different from our study population. Genetic variation frequency is expected to be different in disease, but since SNPs usually only occur in small percentages, a %Diff that is extremely large would indicate a difference in genetic background rather than a disease biomarker. This is why it is important to complete quality control measures. All in all, the results of this comparison supports that our Hispanic population is closely related to Mexican and European populations and that there is some significant variation between the two, which is likely an indicator that these SNPs can be used as biomarkers for Type 2 Diabetes in Hispanic populations.

Table 6. Comparison to Reference Populations

SNP ID	Case*	MXL	% Diff	CEU	% Diff	CHB-JPT	% Diff
rs17497477	0.444	0.2891	53.6	0.2235	98.7	0.125	255.2
rs2185756	0.667	0.6250	6.72	0.5588	19.4	0.342	95.0
rs4383556	0.694	0.8438	17.8	0.7941	12.6	N/A	--
rs17293846	0.694	0.8438	17.8	0.7824	11.3	0.975	28.8
rs191990500	0.333	N/A	--†	N/A	--	N/A	--
rs17826758	0.694	0.8438	17.8	0.7824	11.3	0.975	28.8
rs73175565	0.694	0.8438	17.8	0.7882	12.0	N/A	--
rs73175555	0.694	0.8359	17.0	0.7706	9.94	0.958	27.5
rs9834931	0.583	0.7109	18.0	0.5941	1.87	0.692	15.7
rs7646419	0.583	0.7109	18.0	0.5882	0.88	0.692	15.7
rs6780808	0.583	0.6953	16.2	0.5941	1.87	0.667	12.6
rs6781019	0.583	0.7109	18.0	0.5941	1.87	0.692	15.71
rs6457742	0.583	0.6016	3.09	0.5235	11.4	0.783	25.52
rs2891226	0.528	0.6563	19.5	0.5941	11.1	0.883	40.23
rs1320195	0.500	0.8125	38.5	0.6882	27.3	N/A	--
rs10875142	0.722	0.8203	12.0	0.6588	9.59	0.925	21.9
rs12213132	0.722	0.8438	14.4	0.6647	8.62	0.958	24.6
rs7005401	0.722	0.8984	19.6	0.8412	14.2	0.883	18.2
rs9848681	0.583	0.6875	15.2	0.5529	5.44	0.667	12.6
rs2682919	0.583	0.6797	14.2	0.5588	4.33	0.475	22.7
rs6791275	0.556	0.6797	18.2	0.5765	3.56	0.667	16.6
rs289107	0.389	0.5859	33.6	0.7188	45.9	0.658	40.9
rs7300366	0.639	0.7578	15.7	0.7412	13.8	0.733	12.8
rs7697417	0.694	0.7031	1.29	0.5647	22.9	0.633	9.6
rs72657613	0.694	N/A	--	N/A	--	N/A	--
rs188617336	0.694	N/A	--	N/A	--	N/A	--
<b>Ave ± SD</b>			18.4 ± 11.0		15.6 ± 20.7		37.1 ± 54.6

The %Diff is the calculated percent difference between the allelic frequency of the case group from the study population and the reference population. All reference populations came from HapMap Consortium data. MXL is Mexican from Los Angeles, CA, CEU is

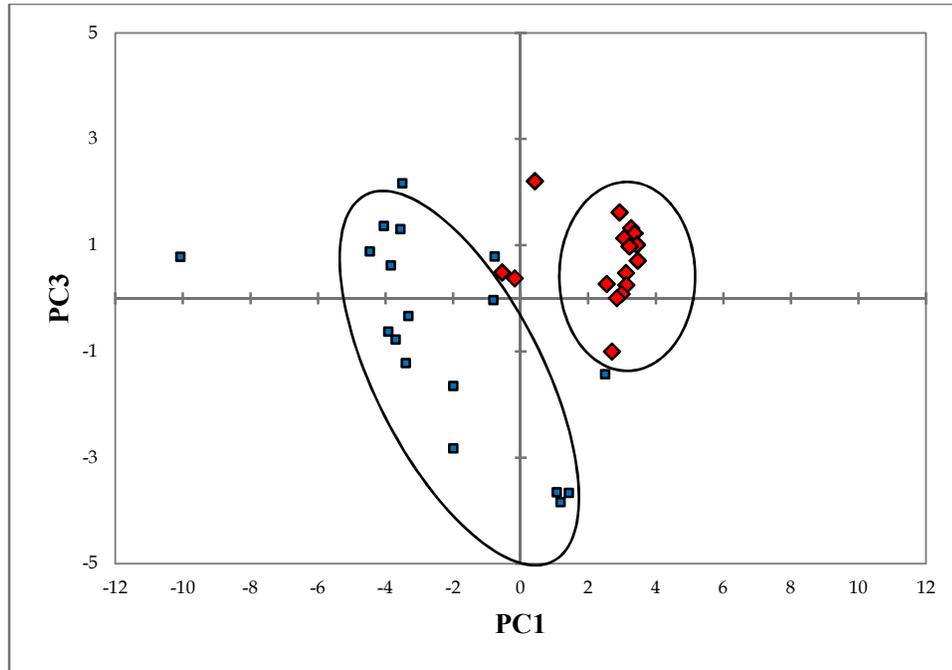
Central European from Utah, and CHB/JPT is Han Chinese from Beijing, China plus Japanese from Tokyo, Japan. The first two are for comparison with similar population groups and the third is for contrasting with a very different population group.

### 3.3.7 PCA

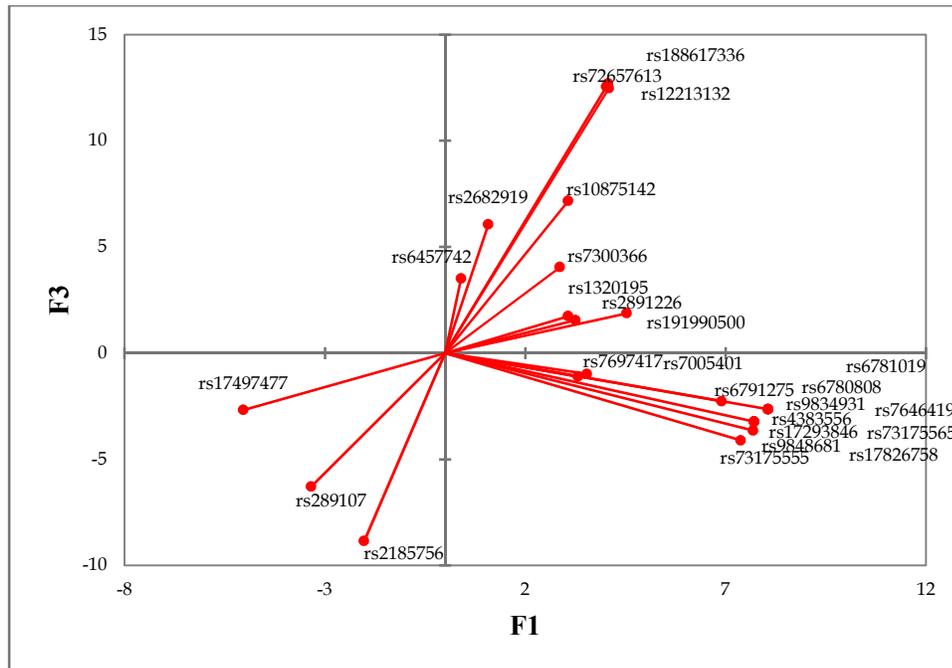
Finally, to further validate our top SNPS, the 26 SNPs were analyzed using PCA. Since the dataset was small, XLstat was able to be used for PCA analysis, which automatically graphs the results. Even with both a small number of samples and a small number of SNPs, the case and control group clearly separate (**Fig 9A**) between case (blue) and control (red). This supports that there is a genetic factor in the susceptibility for Type 2 Diabetes in a Hispanic population and that factor is strong enough to separate the sample groups even with a small set of variables. Further, a coordination plot of the SNPs (**Fig 9B**) shows that they separate into three groups. This could mean that there are subclasses of variation for Type 2 Diabetes as has been suggested in the literature (Murea *et al.*, 2012). For example, the three SNPs that separate out on the bottom left of the plot are rs17497477, rs289107, and rs2185756. The first two are the top two SNPs and the last is a SNP that had significantly different allelic frequency from all the reference populations. This further supports the idea that the top SNPs for Type 2 Diabetes could be put into subcategories. However, further work needs to be done to confirm this.

Figure 9. PCA of Hispanic Population Data

A.



B.



A. PCA Biplot of the top 26 SNPs. Case is blue and control is red. Even with a small sample size and 26 SNPs, the case and control separate into two groups. B. Coordination plot with top 26 SNPs. The SNPs separate into three groups. The bottom left group coincides well with the comparison to reference population data.

### **3.4 Conclusion**

After completing the pilot study for Type 2 Diabetes, there were 26 SNPs that were statistically significant for disease susceptibility in our Hispanic study population. This supports the hypothesis that there are genetic factors that increase susceptibility for Type 2 Diabetes in Hispanics. This was further confirmed in that 7 of the top SNPs also showed difference above the average in allelic frequency compared to similarly related reference populations. Further, the small number of samples and SNPs were able to separate case and control on a PCA biplot. In addition, the number of SNPs that were statistically significant would support the multifactorial nature of Type 2 Diabetes in that it is a combination of many SNPs that adds up to disease. The coordination plot might suggest that there are also subcategories to the set of SNPs and therefore in Type 2 Diabetes, but more would need to be done to conclude this. In future, using the pilot study data as preliminary results, the research can be expanded to include a larger number of participants to validate the 26 SNPs and possibly identify others that are too rare to be detected in the smaller sample set.

**CHAPTER IV**  
**OPTIMIZATION OF A WHOLE CELL BETA-GALACTOSIDASE ASSAY IN**  
**GRAM-POSITIVE BACTERIA**

**4.1 Introduction**

**4.1.1 Identifying the Role the Gut Microbes Play in Susceptibility**

In order to understand how the gut microbiome affects human health and disease, new methods need to be developed to identify the activity the microbes have in the gut environment including how they interact with food and medicine as well as how they live in symbiosis with the host. If patterns of activity can be detected, then perhaps how the gut microbiome influences the development of Type 2 Diabetes and other metabolic disorders can be unraveled. In order to study the patterns of activity, models of the gut microbes need to be built and tested to simplify the creation and optimization of new methods of detection.

**4.1.2 Modelling the Gut Microbiome**

Since most gut microbes cannot be cultured, the easiest option for model systems is probiotic bacteria. Probiotics are microbes that produce health benefits for humans. They can be cultured and are common in different food products like yoghurt (Tamang *et al.*, 2016) and many of them are also native to the gut, supporting their compatibility in food products. It is because of these properties that probiotics would make an ideal choice for developing a model for testing gut microbes. At the same time, a large

proportion of probiotic bacteria are gram-positive. This means they have a thick peptidoglycan layer around their cell wall. Because of this, there is a limit to what substrates can be used and gram-positive cells may be difficult to work with when performing assays (Delcour *et al.*, 1999). Therefore, starting with a simple model will help to understand how new methods will work on probiotic, and thus gut, microbes. To begin with, the simplest model is a single pure culture of probiotic bacteria. The strain chosen was *Lactobacillus helveticus* (ATCC 15009). It is a gram-positive bacteria that is commonly found in certain cheese cultures and has also been detected in the gut (Taverniti & Guglielmetti, 2012). It is from the *Lactobacillus* genus, which is made up of many other probiotics and it is one of the better studied genus' of bacteria (Kant *et al.*, 2010). Therefore studying *L. helveticus* will provide a good foundation for developing methods of microbial activity detection.

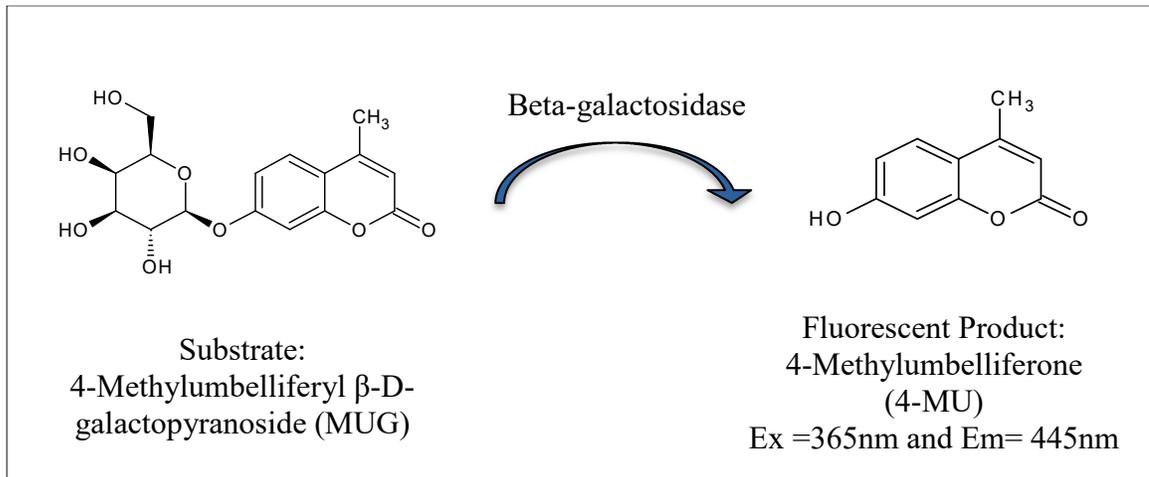
#### **4.1.3 Microbe Activity**

With three potential targets for activity detection, deciding which is the most appropriate for the goal is the first step to begin designing new methods. Since this is the first thing that needs to be answered, a known substance was chosen to measure and compare between the three targets. Beta-galactosidase is an enzyme that is part of the metabolism of many organisms. Its main reaction is to cleave lactose into glucose and galactose so it can be used as an energy source. The enzyme is well-studied and has been used extensively as a reporter molecule with a variety of colorimetric and fluorescent substrate options (Alam & Cook, 1990; Miranda et al. 2010). Further, the lac operon, the gene that codes for beta-galactosidase, is well understood and is naturally repressed in

bacterial systems. The repressor is turned off by the presence of lactose and thus using a lactose based media to grow the cells could activate the target of detection and simulate an activity scenario in the gut (Citti *et al.*, 1965; Fortina *et al.*, 2003). At the same time, using beta-galactosidase allows the detection of all three possible activity targets for gut microbes. That is, the enzyme, and thus the metabolites it produces, as well as the mRNA for the gene are known and therefore can be studied to help determine which target would be ideal for tracking the activity of gut microbes.

Since there are already designed substrates for detecting enzymatic activity of beta-galactosidase, the protein/metabolite target was chosen as the first part of the study. There are several options for substrates, but a fluorescent one was used due to the increased sensitivity possible with fluorescent measurements over colorimetric measurements. The substrate is 4-Methylumbelliferyl  $\beta$ -D-galactopyranoside (MUG) and the structure and reaction with beta-galactosidase is shown in Figure 10. MUG has a galactose moiety, which the enzyme recognizes and binds in its active site. The enzyme then cleaves the bond between the saccharide and the fluorophore to release them and enhance fluorescence to a detectable level (Grange & Clark, 1977). Since MUG is a small molecule with a similar structure to lactose, it should not be hindered from passing into the cell and being acted on by the enzyme. The goal then is to turn on the production of the enzyme and measure the activity using MUG. If an assay can be optimized to detect relative levels of enzyme activity, then the viability of the assay as a detector for activity can be determined and compared to the measurement of the mRNA for the beta-galactosidase gene for bacterial activity.

Figure 10. Reaction of MUG with Beta-galactosidase Enzyme



Structure of MUG substrate and the reaction with beta-galactosidase. MUG has a similar structure to lactose with a galactose moiety (left half of substrate) with the fluorescent moiety in place of glucose. The enzyme cleaves MUG to produce 4-MU, a fluorescent product, which can be detected, indicating enzyme activity.

## 4.2 Method

### 4.2.1 Bacteria Stock

*Lactobacillus helveticus* (ATCC 15009) was purchased from the American Type Culture Collection (ATCC) and the culture was started in 5 mL of deMan, Rogosa, and Sharpe (MRS) media using the provided protocol. Growth was at 37°C with 5% CO<sub>2</sub> for 24 hours. After this initial growth, 10 μL of *L. helveticus* was inoculated into a fresh 5 mL tube of MRS media and grown until the late log phase based on optical density (OD) at 600 nm. Then the culture was decanted into a 15 mL centrifuge tube and centrifuged at 3000 x g for 15 min at 4°C. The supernatant was removed and the cells were then resuspended in an equal volume of fresh MRS media containing 20% (w/v) sterile glycerol. The stock was aliquoted into cryogenic tubes and stored at -80°C.

#### **4.2.2 Growth Curve**

A growth curve was plotted to determine how long it would take for the bacteria to reach the log phase of growth. *L. helveticus* was grown in MRS media for 48 hours with measurements of OD at 600 nm approximately every 2 hours. Further, to see how the cells respond, a growth curve was also done for *L. helveticus* in MRS media that contained lactose as the only sugar source (LMRS) using the same parameters. Each growth curve used a 10  $\mu$ L inoculation in a 5 mL volume of media in duplicate and OD was tested directly in the culture tube in a Genesis10 UV/Vis Scanning Spectrometer (ThermoSpectronic). The average OD for each time point was calculated and plotted against the time in hours.

#### **4.2.3 Initial Assay Conditions**

The initial assay was based on a gram-negative method from Vidal-Aroca *et al.* *L. helveticus* was grown in 5 mL of MRS media at 37°C with 5% CO<sub>2</sub> to the late log phase to bring the cells out of a freezer state. Then 10  $\mu$ L was inoculated in 5 mL LMRS and grown overnight in the same conditions. The OD at 600 nm was measured and then cell aliquots of 20  $\mu$ L were dispensed into a 96-well plate and 80  $\mu$ L of z-buffer was added followed by 25  $\mu$ L of 1 mg/mL MUG in DMSO. The reaction mixture was incubated for 15 mins and then the reaction was stopped with 30  $\mu$ L of 1 M Na<sub>2</sub>CO<sub>3</sub>. The fluorescent product was measured on a POLARstar OPTIMA plate reader (BMG labtech) with excitation at 365 nm and emission at 445 nm. Z-buffer was made based on the procedure from Miller. After it was made, it was filtered through a 0.2 micron filter to remove

impurities and then stored at 4°C until use. A fresh aliquot of z-buffer was prepared for each assay by adding 14 µL of β-mercaptoethanol per 5 mL of z-buffer prior to use.

#### **4.2.4 Testing for Background Noise**

*L. helveticus* was grown and activated as before. To test for background caused by the cells, they were washed once in either 1X PBS buffer, z-buffer, or fresh LMRS media. Further, the two buffers and the LMRS media were added in place of cells in the MUG controls to look for any interaction. The assay was completed with the initial parameters and the results compared.

#### **4.2.5 MUG Concentration**

Different MUG concentrations were tested to optimize the signal-to-noise ratio. A serial dilution of MUG was created in DMSO from 5 mg/mL to 0.002 mg/mL and tested for beta-galactosidase activity. Controls were also made to include each MUG concentration without cells. The fluorescent signal was used to calculate the signal-to-noise ratio and based on that, an optimum MUG concentration was determined.

#### **4.2.6 Activation**

In order to test how activation works and how long is necessary, *L. helveticus* was grown in three tubes and tested at 1, 2, and 4 hours. For each time point, 4.5 mL of cells were centrifuged for 15 min at 3000 x g and 4°C and then resuspended in 500µL of LMRS and serially diluted until a minimum OD was reached. All samples were tested for beta-galactosidase activity. Using the serial dilutions, a standard curve was created of OD vs. fluorescent signal in order to normalize the fluorescent signal at each time point to be

comparable for beta-galactosidase activity rather than just cellular growth. The results were then compared to determine activation time.

#### **4.2.7 Linear Dynamic Range**

Using the newly optimized parameters, *L. helveticus* was activated for 4 hours, and then 15 mL of culture was collected and centrifuged at 3000 x g for 15 min at 4°C. The supernatant was decanted then the cells resuspended in 400 µL of z-buffer. The cell concentrate was serially diluted in 1:2 ratios with z-buffer and then each dilution was aliquoted into a 96 well plate. The cells were incubated with MUG for 15 min and then the reaction stopped as before. The dilution level was plotted against the fluorescent signal to determine the linear dynamic range of the assay for *L. helveticus*.

#### **4.2.8 Reproducibility of the Assay**

The reproducibility of the assay was tested on a different strain of gram-positive bacteria: *Lactobacillus reuteri* (ATCC 53609). Bacteria stock and a growth curve were created for *L. reuteri* using the same method as before. *L. reuteri* was then processed through the same assay parameters as for *L. helveticus* to create a linear dynamic range. The results were plotted and compared to determine if the assay was reproducible on this strain.

### **4.3 Results/Discussion**

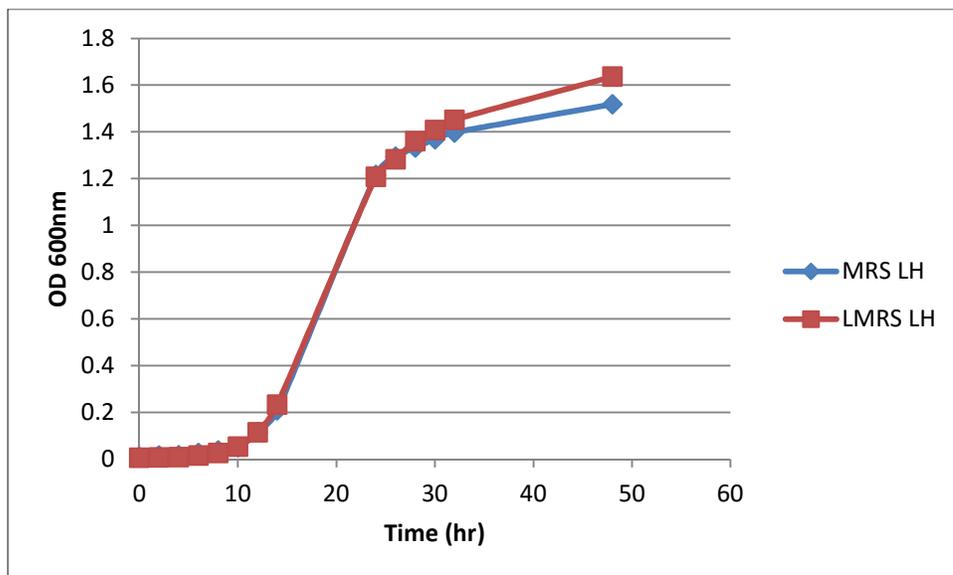
#### **4.3.1 Initial Growth of *L. helveticus***

The pure strain of *L. helveticus* was purchased from ATCC and successfully started using the provided instructions. The stock was created through a single transfer from the purchased strain culture. This was done to avoid contamination and to ensure

the stock was as close to the purchased culture as possible. Further, each assay was started from stock so that there was consistency for each experiment and thus comparability.

The growth curve was created for both MRS and LMRS media (**Fig 11**). *L. helveticus* reached the beginning of the log phase at around 10 hours and ended around 30 hours with the 10  $\mu$ L inoculation volume. Further, the bacteria responded nearly identically to both types of media with just a small variation at the end phase of growth. This indicates that this strain of bacteria can use glucose and lactose with equal efficiency and therefore the growth is not drastically altered by introducing the different sugar source. Further, this indicates that *L. helveticus* can grow on lactose and thus it has an active beta-galactosidase enzyme.

Figure 11. *L. helveticus* Growth Curve

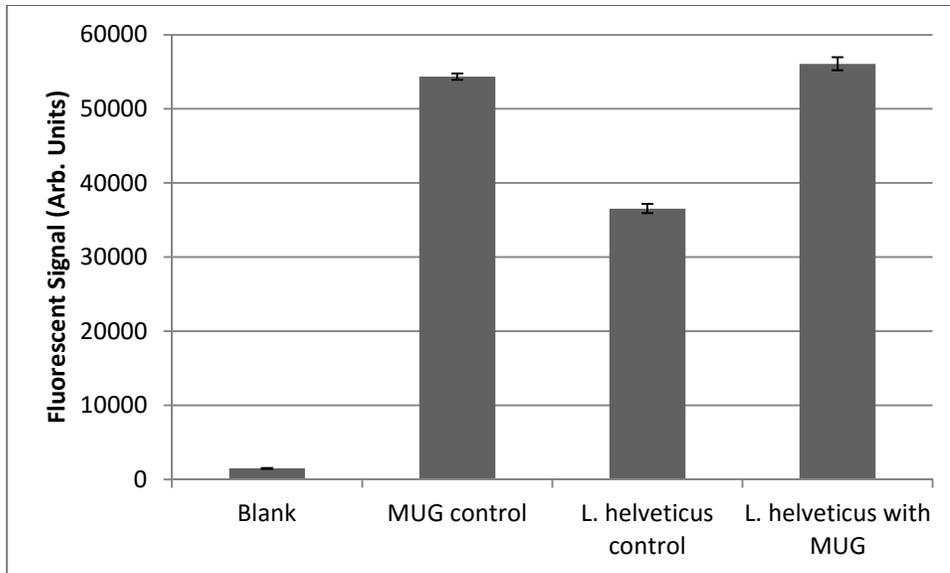


A growth curve was created to understand how *L. helveticus* (LH) responds to the two different sugar sources in the media. Blue is MRS media with glucose and Red is LMRS media with lactose. Cells were grown over a 48 hour period at 37°C with 5% CO<sub>2</sub>.

#### 4.3.2 Results of Initial Assay

The initial assay conditions were based on a gram-negative assay in whole cells. First, *L. helveticus* is started from stock in MRS media to bring the cells up to active growth from the freezer. Then they are inoculated into LMRS media to activate the beta-galactosidase enzyme. The assay uses a fluorescent substrate to increase sensitivity of the assay due to the possibility of low signal caused by working with whole cells. Before going through the process of optimization, the gram-negative assay was tested on the gram-positive bacteria to see if signal could be achieved. Using the parameters from Vidal-Aroca *et al.*, *L. helveticus* was tested for beta-galactosidase activity (**Fig 12**). The blank is z-buffer and the two controls are for MUG background noise, which included MUG with media in place of cells, and cell background noise, which included *L. helveticus* with DMSO in place of MUG. The buffer by itself has minimal noise, but both controls are at essentially the same fluorescent signal as the reaction sample. Thus signal above the noise was not achieved using the initial parameters and optimization was necessary. Specifically, optimization was needed to address the large background noise of the controls as well as to examine other aspects of the assay to improve the reaction signal and the assay time.

Figure 12. Initial Assay Results



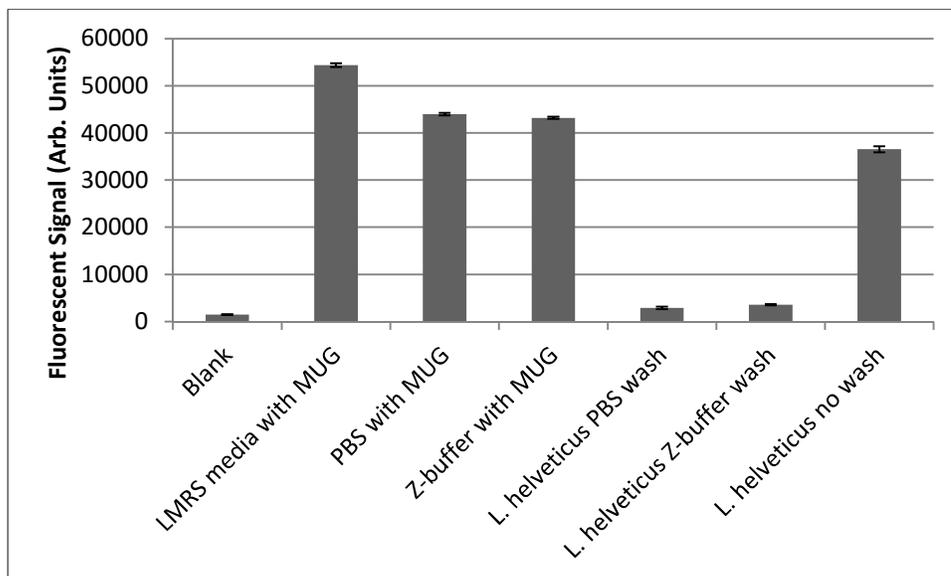
The initial assay parameters were tested and the fluorescent levels compared between the reaction sample and the controls. Blank is z-buffer, MUG control is everything but cells, and *L. helveticus* control is everything but MUG.

### 4.3.3 Background Noise

Due to the high background noise produced by the initial assay, several different aspects of the assay were tested. First, to examine if the cells or the media were producing noise, the cells were washed in once z-buffer, PBS buffer, or fresh LMRS media and then tested for fluorescence without MUG to see if the background noise was reduced. PBS buffer was chosen as a gentle buffer that works well with the majority of cell cultures without damaging them and z-buffer was used since it is the assay buffer so it is compatible with the rest of the assay. By also washing with fresh media, it could be determined if the cells were producing a compound that fluoresces at the same wavelengths as the substrate or if it was the media producing the noise. As can be seen in

Figure 13, washing just once with either buffer eliminates the noise in the cell controls. However, the control that was washed with media maintains a higher level of noise. Therefore the media produces background noise at the detection wavelength and washing with buffer removes the media and the noise. Both buffers worked equally well and thus the z-buffer was used in future to reduce the number of solutions needed for the assay. Second, the MUG was tested at 1 mg/mL with either of the buffers or fresh media added in place of the cells to confirm that the MUG gave high background noise and to see if there were any negative interactions between the buffers and the substrate. The results confirmed that the MUG is giving large background noise at the concentration used. The buffer did not have an effect on the substrate and the media noise is confirmed in the MUG controls as the signal from the MUG with media is higher than the MUG with either buffer. Thus further work needed to be done to address the MUG concentration.

Figure 13. Sources of Background Noise



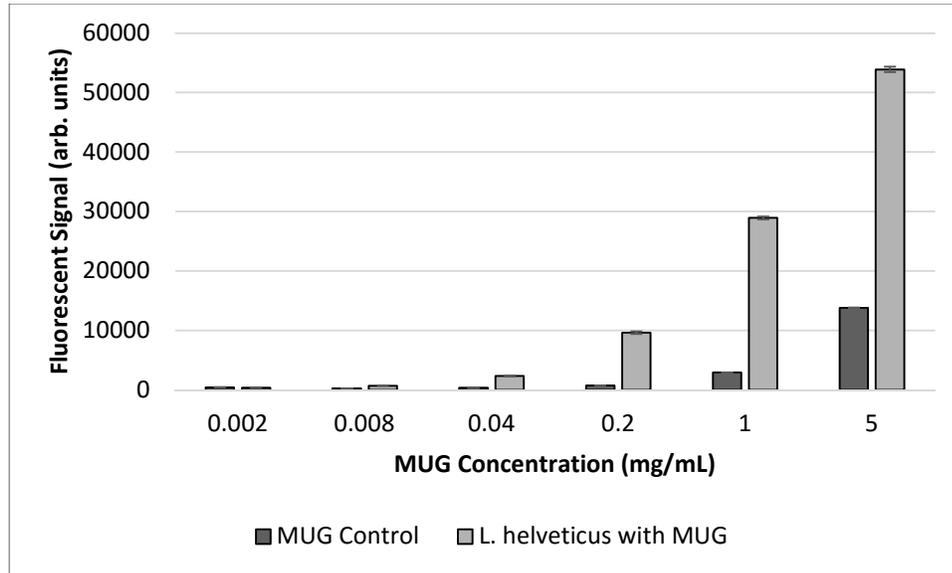
The MUG substrate and the media were tested for background noise in the assay. All samples listed are controls. Blank is z-buffer, the three MUG controls have LMRS media, PBS buffer, or z-buffer respectively in place of cells. The washed cells have DMSO in place of MUG and were washed once in either PBS buffer, z-buffer, or fresh LMRS media.

#### 4.3.4 MUG Concentration

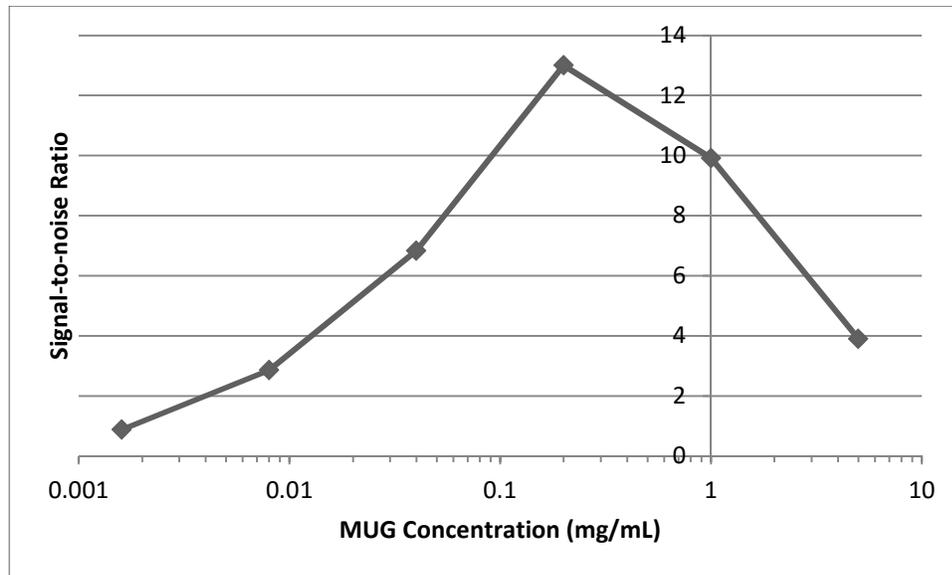
It was verified that the MUG at 1 mg/mL was producing too much background noise. Therefore, a serial dilution was created to assess what concentration would be best for the assay with gram-positive bacteria. The higher concentration was to assess if more MUG was needed to induce the cells to take it in and allow the beta-galactosidase to act on the substrate. The dilutions went down about three orders of magnitude to test a wide range of concentrations. The controls are MUG with z-buffer in place of cells and the reaction mixture has activated cells with the MUG (**Fig 14A**). The higher MUG concentrations produced more signal with the cells while the lowest concentrations produced neither signal nor noise. The lowest concentrations were either below the detectable level of fluorescence or not high enough to encourage diffusion into the cell and therefore could not react with the enzyme. To understand which of the higher concentrations was best, the signal-to-noise ratio was calculated and plotted against the concentration (**Fig 14B**). Based on the signal-to-noise ratio, 0.2 mg/mL was the best MUG concentration as the curve peaks at that concentration. The concentrations above this appear to have higher signal from the reaction, but the noise is also increased proportionally and therefore the ratio decreases. Thus the future assays used a MUG concentration of 0.2 mg/mL.

Figure 14. Changing MUG Concentration to Increase Signal-to-noise Ratio

A.



B.



A. A serial dilution of MUG was made from 5 mg/mL to 0.002 mg/mL in DMSO then each concentration tested on activated *L. helveticus* with their respective controls. B. The signal-to-noise ratio of the MUG concentrations was calculated and then plotted against concentration of MUG. The best signal-to-noise ratio is achieved at 0.2 mg/mL of MUG.

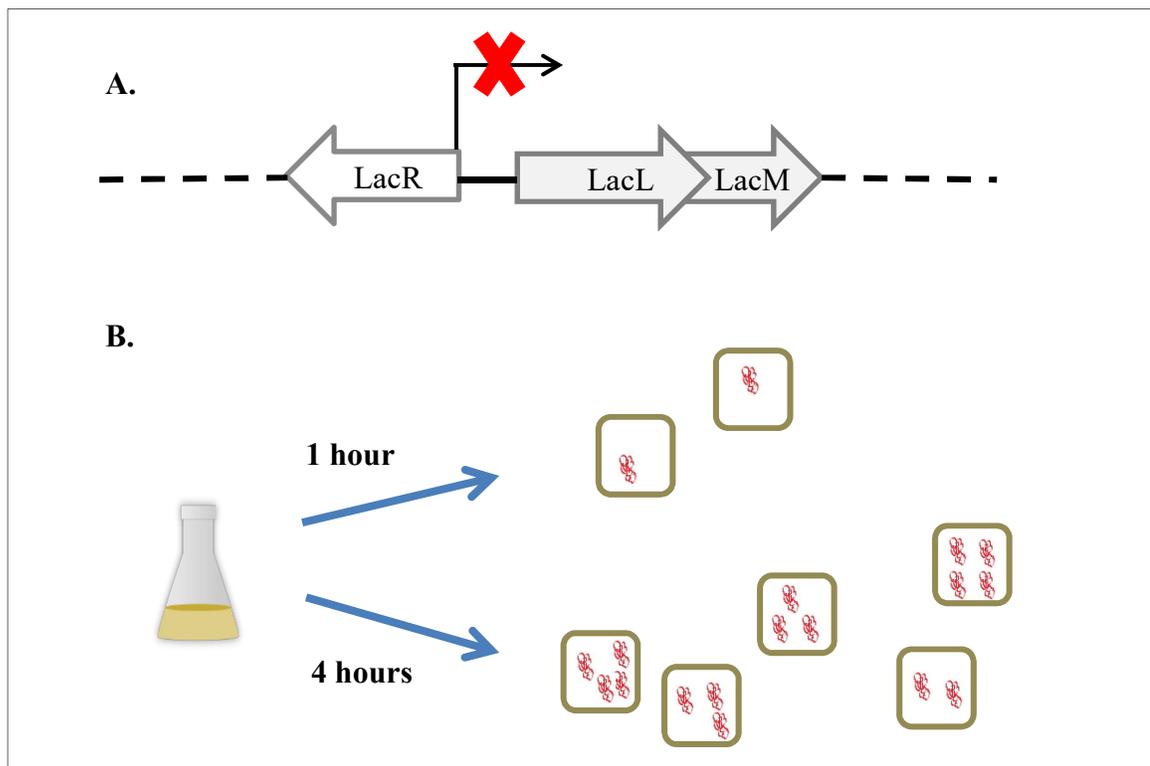
#### 4.3.5 Activation of Beta-Galactosidase

In addition to background noise, the time the assay took was a negative aspect to the original parameters. Since the cells are grown in MRS first and then activated overnight in LMRS, the assay takes several days to complete. Thus it would be ideal to have a shorter activation time to keep the assay to one day. We hypothesized that since the beta-galactosidase gene is off when the cells have their normal sugar source, the activation with lactose would be exponential. That is, cell growth will increase the amount of beta-galactosidase enzyme and over time the cells will also contain more enzyme. Specifically, the lac operon is repressed when lactose is not present (**Fig 15A**). When cells are inoculated into LMRS media, the lactose will turn on the gene and start producing beta-galactosidase enzyme. The form of the lac operon for *L. helveticus* is different than the normal lacZ type gene. It produces a heterodimeric enzyme that has the same activity, but different structure and stability. This is part of the reason for choosing a whole-cell assay and testing the activation of the enzyme. Figure 15B shows the hypothesized mode for exponential increase of beta-galactosidase enzyme. After only 1 hour of growth there will be cell replication, but since it is early in the growth period, only a few beta-galactosidase enzymes will have been made. After 4 hours, enough time has passed to produce a significant amount of enzyme per cell as well as more cell growth. Thus the increase in signal after the start of activation will be exponential.

To test this, the cells were activated and tested at 1, 2, and 4 hours. As would be expected, the amount of signal increases as the activation time increases (**Fig 16A**). In order to determine if this increase is due to just cell growth, and thus the amount of

enzyme remains constant, or if there is an exponential increase, the 1 and 2 hour time points were normalized to the same OD as the 4 hour time point (**Fig 16B**). With normalization the signal increases exponentially and thus our hypothesis was correct. Further, after only 4 hours there is a strong enough signal to be detectable and therefore activation overnight is not necessary. Further, since cells are started in MRS, more can be used to inoculate in the LMRS to achieve the log phase in a shorter time as was done here. By doing this, the activation time is reduced and as a consequence the whole assay time is reduced and can be accomplished in one day.

Figure 15. Beta-galactosidase Activation

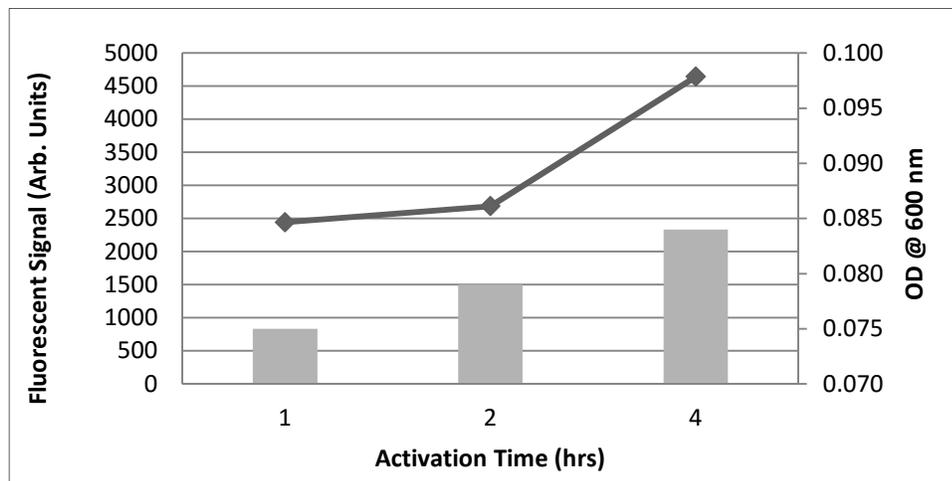


A. The LacL/LacM form of lac operon found in *L. helveticus*. The gene produces the beta-galactosidase enzyme, but gene expression is repressed unless in the presence of D-

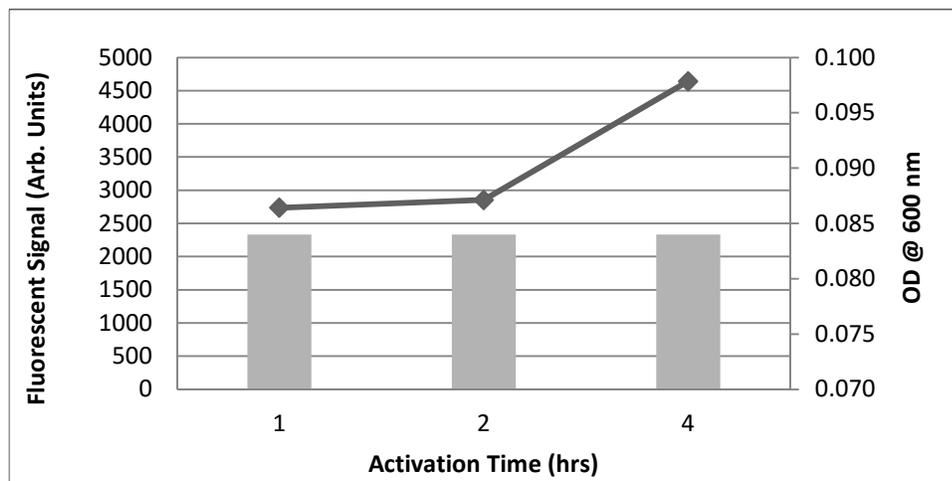
lactose because the repressor is located upstream (LacR). The enzyme is a heterodimer with the two genes overlapping each other (LacL & LacM). Other side components of the gene are located on either end of the dashed lines, but are irrelevant to the study. B. Hypothesized mechanism for activation over time. Red spots represent beta-galactosidase enzyme and the brown boxes represent cells. Figure depicts how exponential increase of beta-galactosidase production might occur.

Figure 16. Results of Activation Study

A.



B.



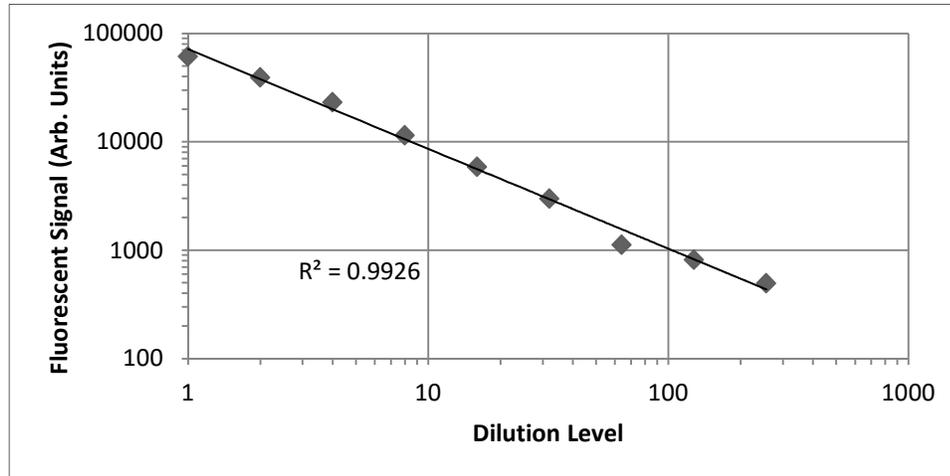
A. *L. helveticus* was tested at 1, 2, and 4 hours for activity. The bars represent the OD at 600 nm and the line represents the fluorescent signal. B. The three time points normalized to the 4 hour OD. The line continues to increase even when the OD is the same indicating beta-galactosidase production is exponential. Error bars are present, but not large enough to be visible.

#### 4.3.6 Linear Dynamic Range

After optimizing the assay, the linear dynamic range needed to be determined.

This shows the range of cell concentration that produces a detectable signal. The cells were concentrated to the maximum OD and serially diluted to determine this range. The plot is of fluorescent signal versus dilution level (**Fig 17**). It shows that the linear dynamic range is a little over two orders of magnitude. Thus the assay is sensitive over a large range of cellular concentrations. This validates the effectiveness of the optimization parameters and indicates that the assay can be used with a variety of cellular concentrations. This is useful since cellular growth is not always consistent and the assay can still be completed even if the cellular concentration is higher or lower than expected rather than lose a day of work to restart the cell growth and activation.

Figure 17. Linear Dynamic Range of Assay for *L. helveticus*



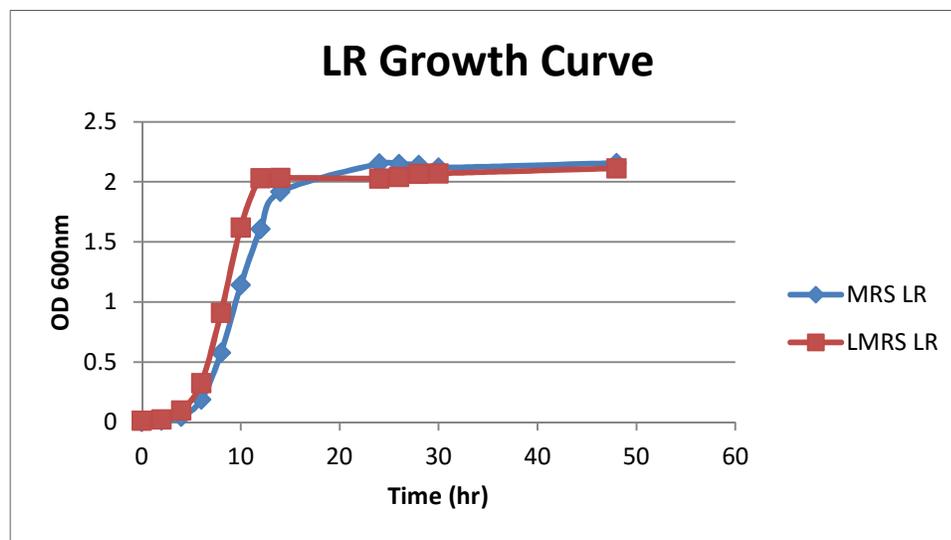
*L. helveticus* was concentrated to a maximum OD and then serially diluted to the lowest possible OD. After the assay, the resulting fluorescence was plotted against dilution level to determine linear dynamic range.  $R^2$  is shown to depict how well the trend line fits the data. Error bars are present, but too small to be seen.

#### 4.3.7 Reproducibility of the Assay in Another Gram-Positive Bacterial Strain

Since the goal of this assay was to be optimized for gram-positive bacteria, the new assay parameters were tested on a different strain to see if the results were reproducible. *L. reuteri* was purchased from ATCC and stock was made as before. *L. reuteri* is closely related to *L. helveticus* and is also a probiotic as well as a gut microbe. Thus it is expected to have an active beta-galactosidase enzyme. On the other hand, it is a different strain of bacteria and thus it likely produces different secondary metabolites has slightly different activity. To determine if *L. reuteri* is compatible with the LMRS media and thus has a robust beta-galactosidase enzyme, a growth curve was made using the same method as for *L. helveticus* to examine the growth rate (**Fig 18**). The growth curve shows that *L. reuteri* grows well in both types of media, with just a slight difference in

the log phase between MRS and LMRS. However the cells reach the levelling off of growth stage at the same time and reach the same level of OD and therefore *L. reuteri* can grow well in the LMRS media. Further, it grows faster than *L. helveticus* confirming that it has a slightly different metabolism despite being closely related. Knowing this, using *L. reuteri* will help determine if the assay can be used on a beta-galactosidase enzyme in gram-positive cells with different properties and therefore the assay is applicable to different strains of gram-positive bacteria.

Figure 18. Growth Curve for *L. reuteri*

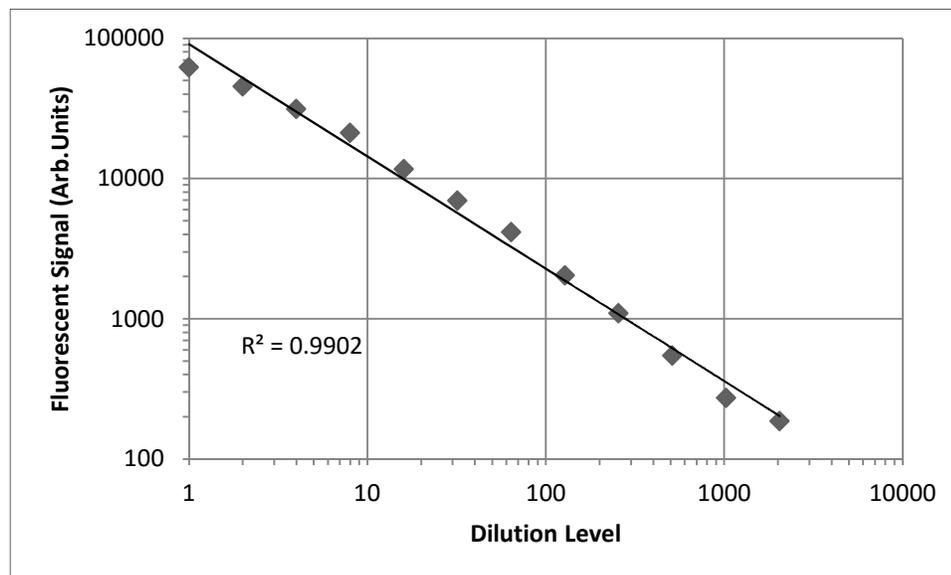


Growth curve completed using same method as with previous strain of bacteria. Blue is with MRS media and red is with LMRS media. The growth curve covered a period of 48 hours grown at 37°C with 5% CO<sub>2</sub>.

Since *L. reuteri* was confirmed to grow well in the presence of D-lactose, it was subject to the optimized beta-galactosidase assay parameters. The linear dynamic range was tested and plotted using the same conditions as for *L. helveticus* (Fig 19). For *L.*

*reuteri* the linear dynamic range is three orders of magnitude. Therefore, the assay is reproducible in a different strain of gram-positive bacteria. Further, the linear dynamic range is improved for this strain. Thus supports the idea that different strains will responds slightly differently to the assay due to the difference in their metabolism, but the assay is applicable nonetheless.

Figure 19. Linear Dynamic Range for *L. reuteri*



To confirm the repeatability of the assay on a different strain of bacteria, a linear dynamic range was determined for *L. reuteri* using the same assay parameters as before. Linear dynamic range increased with this strain of bacteria and thus the assay can be reproduced in other gram-positive bacteria. Error bars are present, but too small to be seen.

#### 4.3.8 Conclusion

An activity assay for beta-galactosidase enzyme was completed in gram-positive bacteria. The initial results proved that optimization was necessary to adapt a gram-negative method to work for gram-positive bacteria due to the difference in cellular

structure and possibly interaction between the substrate and the cells. Optimization was successful with a reasonable linear dynamic range and validation with a different strain of gram-positive bacteria. With the enzyme assay completed, the next steps of the microbiome model can be developed and optimized. This will include examining the mRNA of the beta-galactosidase gene and comparing it to the enzyme activity to determine which activity test is ideal. Further, a mixed microbe model will be tested for activity to mimic a simple gut environment and microbial interactions. Ultimately the goal will be to develop a method of detecting activity, that is which microbes are actively involved in human health and disease, and applying this method to gut microbe samples for the detection of patterns in activity that link to how the gut microbes increase susceptibility for Type 2 Diabetes and other diseases.

## **CHAPTER V**

### **CONCLUSIONS**

Overall, by using a newly developed microarray chip, the research in this dissertation has completed a pilot study on the genetic biomarkers for Type 2 Diabetes among Hispanic adults. The genetic biomarkers can be used to define the hosts in future studies on the relationships between the hosts and gut microbiome and other related clinical studies. Using statistics tools, the pilot study successfully identified 26 SNPs associated with Type 2 Diabetes in a Hispanic population. These results were supported by secondary analysis with reference populations and PCA. The PCA results indicated that there is a genetic basis for Type 2 Diabetes based on the clear separation between individuals with Type 2 Diabetes and individuals without. This was seen in both the reference Geneva data set as well as the Hispanic study population. In both data sets, there were indications that multiple SNPs were required to the development of diabetes as seen by the number of SNPs needed to maintain separation in the PCA biplots as well as the number of SNPs isolated in the Hispanic study. Furthermore, Type 2 Diabetes SNPs were detected in the Hispanic population and this suggests that there is a genetic factor that increases the susceptibility to diabetes. Future work will increase the sample size to validate the SNPs and possibly identify more SNPs as well as the mode of action these genetic variants have in the host.

The need for optimizing the beta-galactosidase assay indicated that gram-positive bacteria do behave differently from gram-negative bacteria, thus different assay conditions were needed to maintain the detectability of a specific activity. The assay time was also reduced to one day to improve the protocol and make it quicker and easier for future comparison of different activities. The beta-galactosidase assay is considered to be equivalent to measuring a specific protein or metabolite (4-MU). Now that the assay is developed and validated, the next steps will include measuring the mRNA of beta-galactosidase, and building a more complex model with mixed microbes and detect their corresponding activities. Ultimately, the goal is to develop a method for monitoring the activity of entire gut microbes and further our understanding on the relationships between the host and gut microbes on the development and/or treatment of Type 2 Diabetes.

## REFERENCES

- Abraham G and Inouye M (2014). Fast principle component analysis of large-scale genome-wide data. *PLoS One*, 9(4): 1-5, e93766
- Alam J and Cook JL (1990). Reporter Genes: Application to the Study of Mammalian Gene Transcription. *Analytical Biochemistry*, 188, 245-254
- Baker M (2011). Metabolomics: From Small Molecules to Big Ideas. *Nature Methods*, 8(2): 117-121
- Baier LJ and Hanson RL (2004). Genetic Studies of the Etiology of Type 2 Diabetes in Pima Indians. *Diabetes*, 53: 1181-1186
- Bonnefond A, Froguel P, and Vaxillaire M (2010). The Emerging Genetics of Type 2 Diabetes. *Trends in Molecular Medicine*, 16(9), 407-416
- Broad Institute (2013). SNP, from <http://www.broadinstitute.org/education/glossary/snp>
- Centers for Disease Control and Prevention. (2014). Diabetes Reportcard, from <http://www.cdc.gov/diabetes/pdfs/library/diabetesreportcard2014.pdf>
- Centers for Disease Control and Prevention. (2010). Health, United States, 2010: With Special Feature on Death and Dying, from [http://www.cdc.gov/nchs/data/10.pdf](http://www.cdc.gov/nchs/data/hus/10.pdf)
- Centers for Disease Control and Prevention. (2014). National Diabetes Statistics Report, from <http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf>
- Cerf ME (2013). Beta cell dysfunction and insulin resistance. *Frontiers in Endocrinology*, 4: 1-12, Article 37
- Chakraborty R, Ferrell RE, Stern MP, Haffner SM, Hazuda HP, and Rosenthal M (1986). Relationship of Prevalence of Non-insulin Dependent Diabetes Mellitus to Amerindian Admixture in the Mexican Americans of San Antonio, Texas. *Genetic Epidemiology*. 3: 435-454
- Citti JE, Sandine WE, and Elliker PR (1965). B-galactosidase of *Streptococcus lactis*. *Journal of Bacteriology*, 89(4): 937-942

Delcour J, Ferain T, Deghorain M, Palumbo E, and Hols P (1999). The biosynthesis and functionality of the cell-wall of lactic acid bacteria. *Antonie van Leeuwenhoek*, 76: 159-184

Fortina MG, Ricci G, Guglielmetti S, and Manachini PL (2003). Unusual organization for lactose and galactose gene clusters in *Lactobacillus helveticus*. *Applied Environmental Microbiology*, 69(6): 3238-3243

Grange JM and Clark K (1977). Use of umbelliferone derivatives in the study of enzyme activities of mycobacteria. *Journal of Clinical Pathology*, 30: 151-153

Goldfine AB, Bouche C, Parker RA, Kim C, Kerivan A, Soeldner JS, Martin BC, Warram JH, and Kahn CR (2003). Insulin Resistance is a Poor Predictor of Type 2 Diabetes in Individuals with No Family History of Disease. *Proceeds of the National Academy of Science*, 100(5), 2724-2729

Illumina (2016). An Introduction to Next Generation Sequencing, from: [http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina\\_sequencing\\_introduction.pdf](http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf)

Illumina (2015). Microbes and Metagenomics in Human Health: An Overview of Recent Publications featuring Illumina Technology, from [http://www.illumina.com/content/dam/illumina-marketing/documents/products/research\\_reviews/metagenomics\\_research\\_review.pdf](http://www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/metagenomics_research_review.pdf)

Illumina (2016). Technology Spotlight: SNP Genotyping “Infinium Assay Workflow,” from [http://www.illumina.com/content/dam/illumina-marketing/documents/products/workflows/workflow\\_infinium\\_ii.pdf](http://www.illumina.com/content/dam/illumina-marketing/documents/products/workflows/workflow_infinium_ii.pdf)

Imamura M and Maeda S (2011). Genetics of Type 2 Diabetes: the GWAS Era and Future Perspectives. *Endocrine Journal*, 58(9): 723-739

The International HapMap 3 Consortium (2010). Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature*, 467, 52-58

Kant R, Blom J, Palva A, Siezen RJ, and de Vos WM (2010). Comparative genomics of *Lactobacillus*. *Microbial Biotechnology*, 4(3): 323-332

Kinross JM, Darzi AW, Nicholson JK (2011). Gut Microbiome-Host Interactions in Health and Disease. *Genome Medicine*, 3(14): doi:10.1186/gm228

Kelvin L, Bihan M, Yooseph S, and Methe BA (2012). Analysis of the Microbial Diversity across the Human Microbiome. *PLoS One*, 7(6): 1-18

Lusis AJ, Attie AD, and Reue K (2008). Metabolic syndrome: from epidemiology to systems biology. *Nature Reviews Genetics*, 9(11): 819-830

Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, Orvis J, Sodergren E, Wang Z, Weinstock GM, and Mitreva M (2012). Optimizing Read Mapping to Reference Genomes to Determine Composition and Species Prevalence in Microbial Communities. *PLoS One*, 7(6): 1-15, e36427

Miller, J. (1972). *Experiments in Molecular Genetics*; Cold Spring Harbor Laboratory: NY; 352-355

Miller J and Miller J (2010). *Statistics and Chemometrics for Analytical Chemistry*, 6<sup>th</sup> ed.; Pearson: Harlow, England; 59-61

Miranda OR, Chen H, You C, Mortenson DE, Yang X, Bunz UHF, and Rotello VM (2010). Enzyme-Amplified Array Sensing of Proteins in Solution and in Biofluids. *Journal of the American Chemical Society*, 132: 5285-5289

Mohlke KL and Boehnke M, (2015). Recent Advances in Understanding the Genetic Architecture of Type 2 Diabetes. *Human Molecular Genetics*, 24(R1): R85-R92

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK, Meng J, Durham BP, Shen C, Varaljay VA, Smith CB, Yager PL, and Hopkinson BM (2013). Sizing Up Metatranscriptomics. *The ISME Journal*, 7: 237-243

Morgan XC and Huttenhower C (2012). Chapter 12: Human Microbiome Analysis. *PLOS Computational Biology*, 8(12): 1-14, e1002808

Murea M, Ma L, and Freedman BI (2012). Genetic and Environmental Factors Associated With Type 2 Diabetes and Diabetic Vascular Complications. *The Review of Diabetic Studies*, 9(1): 6-22

NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, and Guyer M (2009). The NIH Human Microbiome Project. *Genome Research*, 12, 2317-2323

Parra EJ, Below JE, Krithika S, Valladares A, Barta JL, Cox NJ, Hanis CL, Wacher N, Garcia-Mena J, Hu P, Shriver MD, DIAGRAM Consortium, Kumate J, McKeigue PM, Escobedo J, and Cruz M (2011) Genome-wide Association Study of Type 2 Diabetes in a

Sample from Mexico City and a Meta-analysis of a Mexican-American Sample from Starr County, Texas. *Diabetologia*, 54:2038-2046

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, and Sham PC (2007). PLINK: A Tool for Whole-Genome Association and Population-Based Linkage Analysis. *The American Journal of Human Genetics*, 81, 559-575

Qin J, Li R, Raes J, Arumugam M, Solvsten K, Burgdorf, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, and Wang J (2010). A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. *Nature*, 464, 59-65

Reed TE (1974). Ethnic Classification of Mexican-Americans. *Science*, 185(4147): 283

Ringner M (2008). What is Principle Component Analysis? *Nature Biotechnology*, 26(3): 303-304

Semenkovich CF, Danska J, Darso T, Dunne JL, Huttenhower C, McElvaine AT, Ratner RE, Shuldiner AR, and Blaser MJ (2015). American Diabetes Association and JDRF Research Symposium: Diabetes and the Microbiome. *Diabetes*, 64(12): 3967-3977

Stern MP, Gaskill SP, Hazuda HP, Gardner LI, and Haffner SM (1983). Does Obesity Explain Excess Prevalence of Diabetes Among Mexican Americans? Results of the San Antonio Heart Study. *Diabetologia*, 24: 272-277

Sven-Bastiaan H and Jehmlich N (2016). Proteomic interrogation of the gut microbiota: potential clinical impacts. *Expert Review of Proteomics*, 13(6): 535-537

Tamang JP, Shin DH, Jung SJ, and Chae SW (2016). Functional Properties of Microorganisms in Fermented Foods. *Frontiers in Microbiology*, 7: 1-13, Article 578

Tatsch E, Bochi GV, Piva SJ, Pereira RS, Kober H, De Carvalho JAM, Sangoi MB, Duarte MMMF, Moresco RN (2012). HbA1c as a Tool for the Diagnosis of Type 2 Diabetes: Comparison with Fasting Glucose. *Clinical Laboratory*, 58, 347-350

Tian C, Gregersen PK, and Seldin MF (2008). Accounting for Ancestry: Population Substructure and Genome-wide Association Studies. *Human Molecular Genetics*, 17(2): R143-R150

Valentina T and Guglielmetti S (2012). Health-promoting properties of *Lactobacillus helveticus*. *Frontiers in Microbiology*, 3: 1-13, Article 392

van Baarlen P, Kleerebezem M, and Wells JM (2013). Omics Approaches to Study Host-Microbiota Interactions. *Current Opinion in Microbiology*, 16(3), 270-277

Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burt NP, Fuchsberger C, Li Y, Erdmann J, Frayling TM, Heid IM, Jackson AU, Johnson T, Kipeläinen TO, Lindgren CM, Morris AP, Prokopenko I, Randall JC, Saxena R, Soranzo N, Speliotes EK, Teslovich TM, Wheeler E, Maguire J, Parkin M, Potter S, Rayner NW, Robertson N, Stirrups K, Winckler W, Sana S, Mulas A, Nagraraja R, Cucca F, Barroso I, Delouks P, Loos R, Kathiresan S, Munroe PB, Newton-Cheh C, Pfeufer A, Samani NJ, Schunkert H, Hirschhorn JN, Altshuler D, McCarthy ML, Abecasis GR, and Boehnke M (2012). The MetaboChip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLOS Genetics*, 8(8), e1002793. doi: 10.1371/journal.pgen.1002793

Wall JD and Pritchard JK (2003). Haplotype Blocks and Linkage Disequilibrium in the Human Genome. *Nature Reviews Genetics*, 4(8): 587-597

Watson AL, Hu J, and Chiu NHL (2015). Single Nucleotide Polymorphism in Type 2 Diabetes among Hispanic Adults. *Diabetes Research and Clinical Practice*, 108: e25-e27

Wigginton JE, Cutler DJ, and Abecasis GR (2005). A Note on Exact Tests of Hardy-Weinberg Equilibrium. *American Journal of Human Genetics*, 76: 887-893

Woting A and Blaut M (2016). The Intestinal Microbiota in Metabolic Disease. *Nutrients*, 8(4): 202. doi:10.3390/nu8040202

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Knight R, and Gordon JI (2012). Human gut microbiome viewed across age and geography. *Nature*, 486: 222-228

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin

L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Wellcome Trust Case Control Consortium, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for Type 2 Diabetes. *Nature Genetics*, 40(5): 638-645

## APPENDIX A

### PLINK COMMANDS

#### I. Chapter 2

(Command 1) `./plink --bfile Genevadata --keep Sample100.txt --make-bed --out first100data`

(Command 2) `./plink --bfile first100data --indep-pairwise 1000 50 0.05 --exclude range exclusion_regions.txt`

(Command 3) `./plink --bfile first100data --extract plink.prune.in --make-bed --out data_pruned`

(Command 4) `./plink --bfile first100data --assoc --out first100assoc`

(Command 5) `./plink --bfile first100data --extract pvalue05.txt --make-bed --out 05Sample100`

(Command 6) `./plink --bfile data_pruned --pheno First100NoPheno.txt --make-bed --out datapruned_nopheno`

(Command 7) `./plink --bfile 05Sample100 --pheno First100NoPheno.txt --make-bed --out First100nopheno`

(Command 8) `./flashpca_x86_64 --bfile datapruned_nopheno`

(Command 9) `./flashpca_x86-64 --bfile First100nopheno`

## **II. Chapter 3**

(Command 10) `./plink --file mydata --exclude snplist.txt --make-bed --out mydata2`

(Command 11) `./plink --bfile mydata2 --pheno Pheno.txt --make-bed --out mydata3`

(Command 12) `./plink --bfile mydata3 --update-sex Gender.txt --make-bed --out mydata4`

(Command 13) `./plink --bfile mydata4 --update-ids Recoded.txt --make-bed --out t2ddata`

(Command 14) `./plink --bfile t2ddata --sex-check`

(Command 15) `./plink --bfile t2ddata --missing --out missingstats`

(Command 16) `./plink --bfile t2ddata --hardy --out HWstats`

(Command 17) `./plink --bfile t2ddata --assoc --out assoc1`

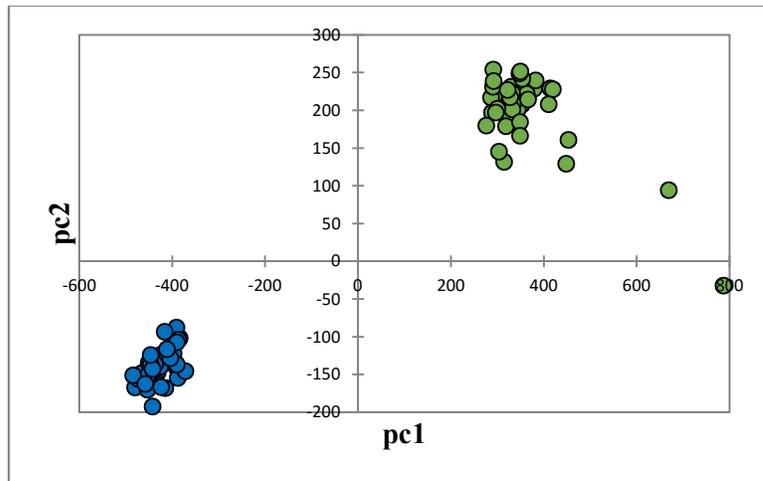
(Command 18) `./plink --bfile t2ddata --recodeHV --out haplowviewdata --gplink`

## APPENDIX B

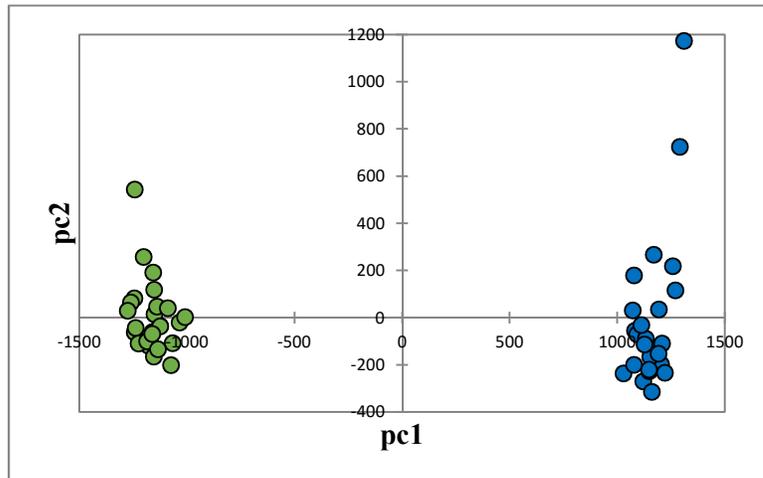
### PCA BILOTS FOR SAMPLE AND SNP REDUCTION

Figure 1. Reducing Sample Size for Second Set of 100 Individuals

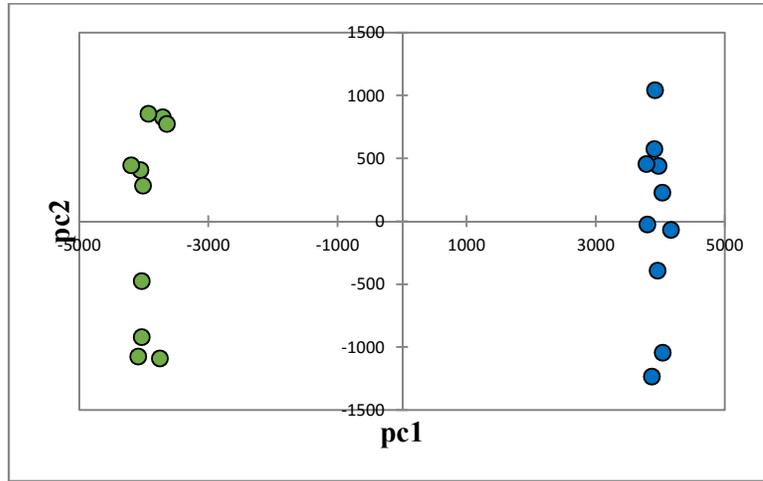
A.



B.



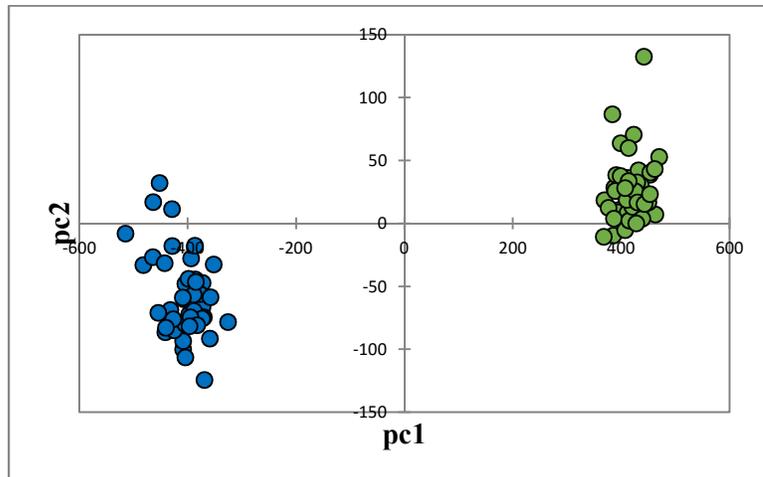
C.



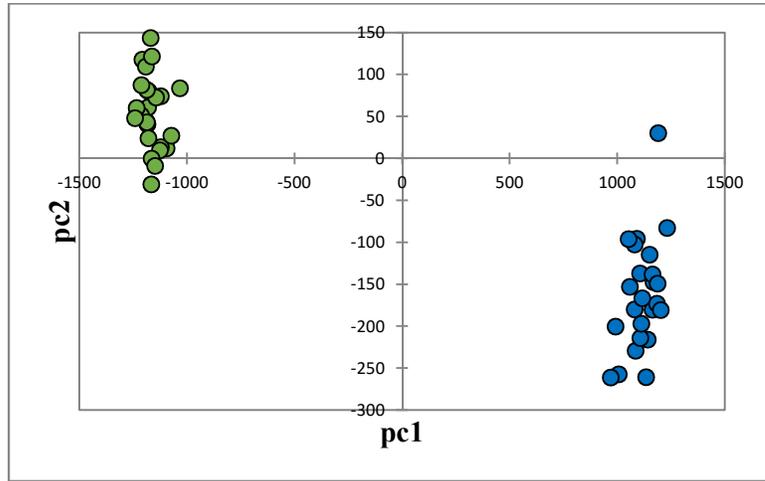
Repeat of the reduction in sample size with the second set of 100. A. This biplot has 100 individuals with 39,177 SNPs using a p-value of 0.05 as the threshold. There were 2 outliers removed. B. This biplot has 50 individuals with 37,205 SNPs using a p-value of 0.05 as the threshold. C. This biplot has 20 individuals with 30,945 SNPs using a p-value of 0.05 as the threshold. In all biplots, green is case and control is blue.

Figure 2. Reducing Sample Size for Third Set of 100 Individuals

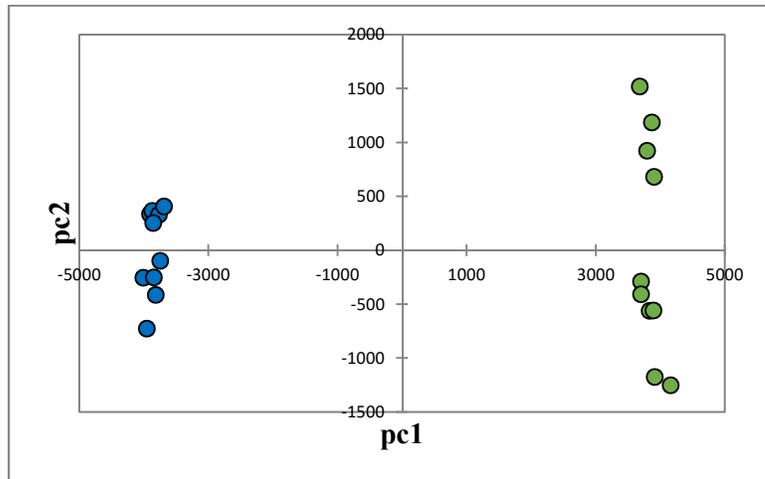
A.



**B.**



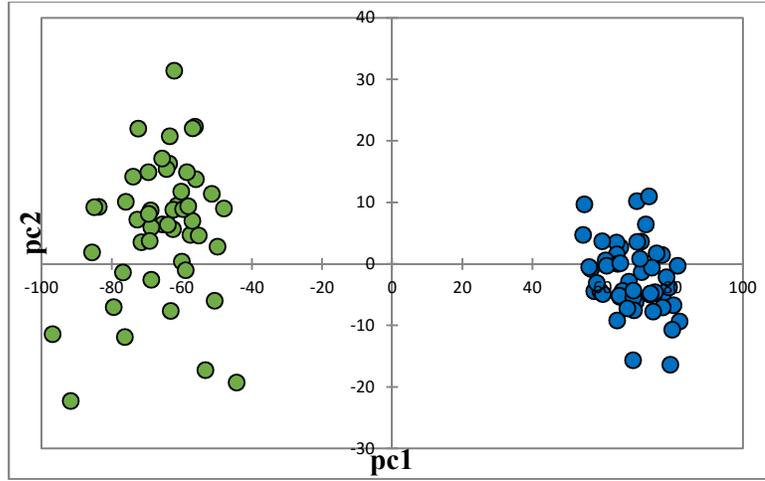
**C.**



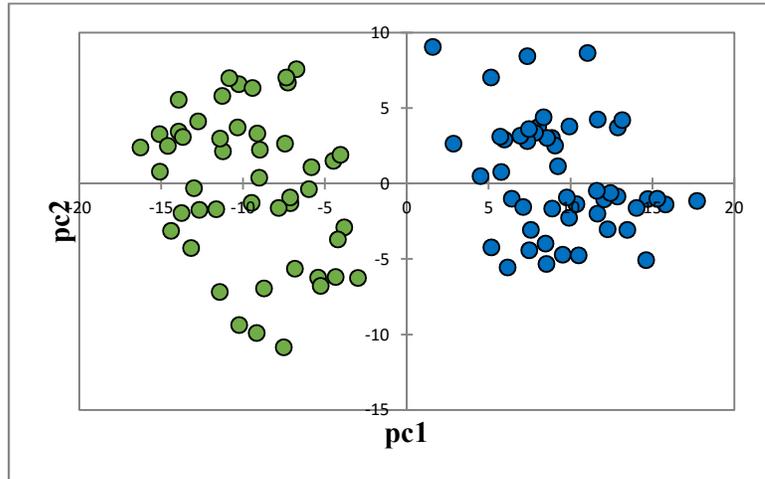
Repeat of the reduction of sample size with the third set of 100. A. This biplot has 100 individuals with 37,959 SNPs using a p-value of 0.05 as the threshold. There was 1 outlier removed. B. This biplot has 50 individuals with 37,189 SNPs using a p-value of 0.05 as the threshold. There was 1 outlier removed. C. This biplot has 20 individuals with 30,223 SNPs using a p-value of 0.05 as the threshold. In all biplots, green is case and control is blue.

Figure 3. Reducing Number of SNPs for Second Set of 100 Samples

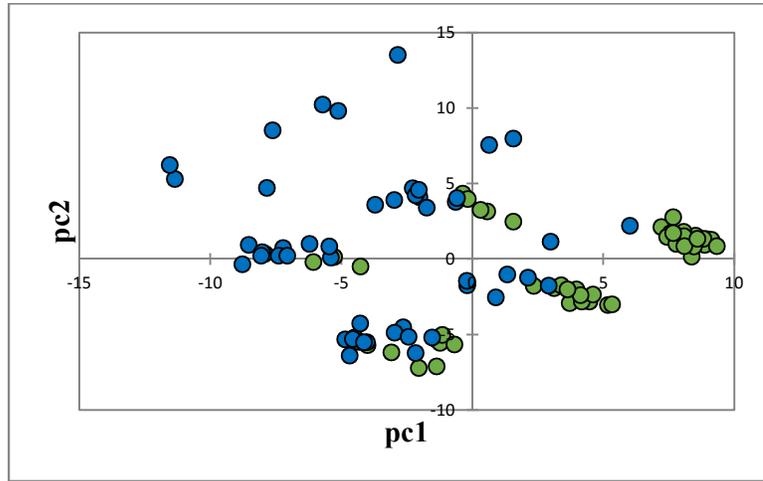
A.



B.



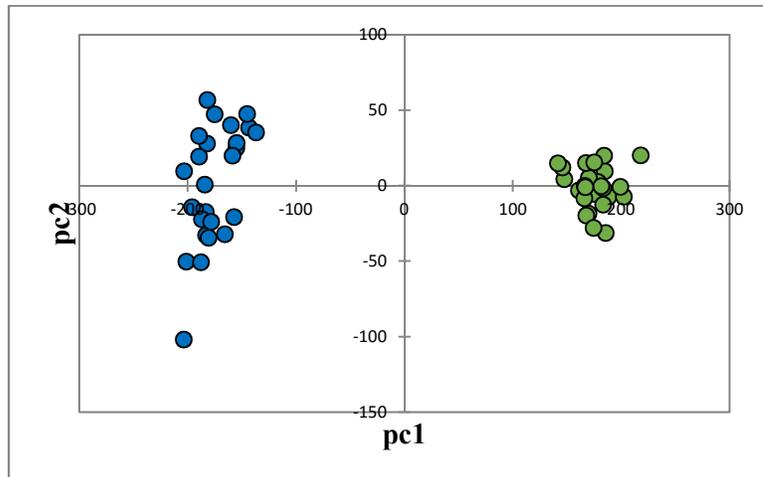
C.



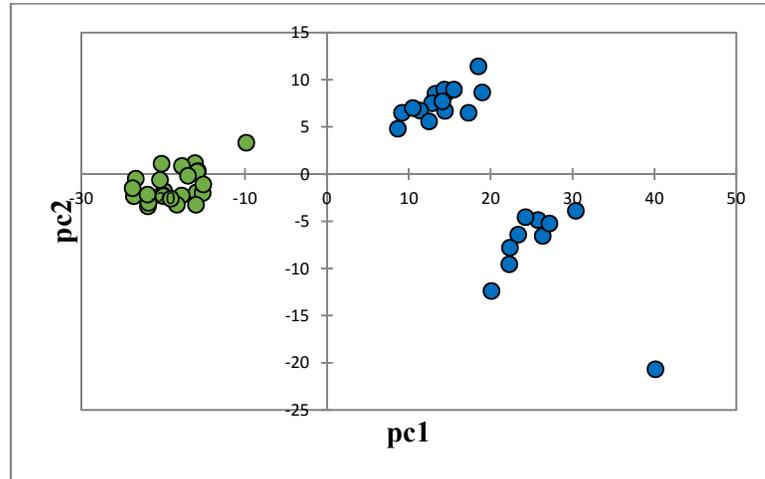
PCA biplots are of second set of 100 samples with case green and control blue. There were 2 outliers removed from all plots. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 3,483 SNPs for PCA analysis. Two samples were removed as outliers. B. P-value of  $5 \times 10^{-4}$  as the threshold with 316 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 45 SNPs. Separation of case and control is mostly maintained for all reduction of SNPs.

Figure 4. Reducing Number of SNPs for Second Set of 50 Samples

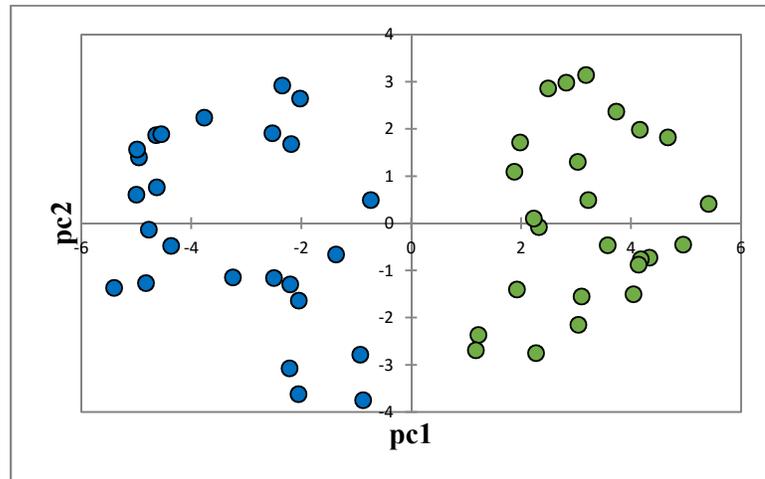
A.



**B.**



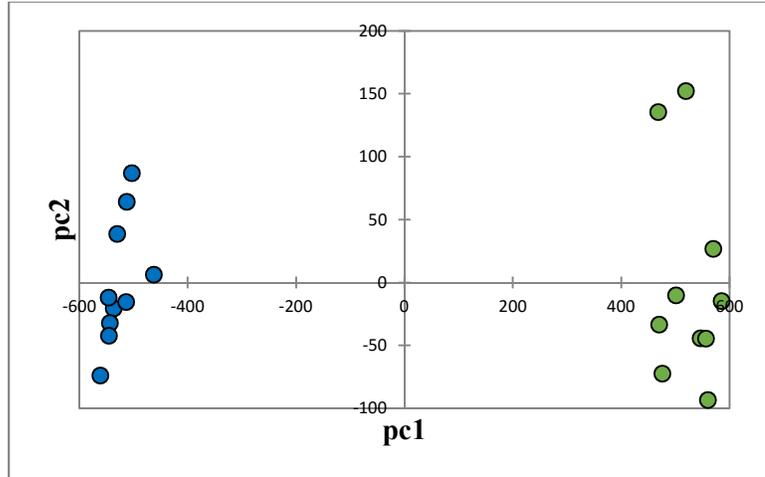
**C.**



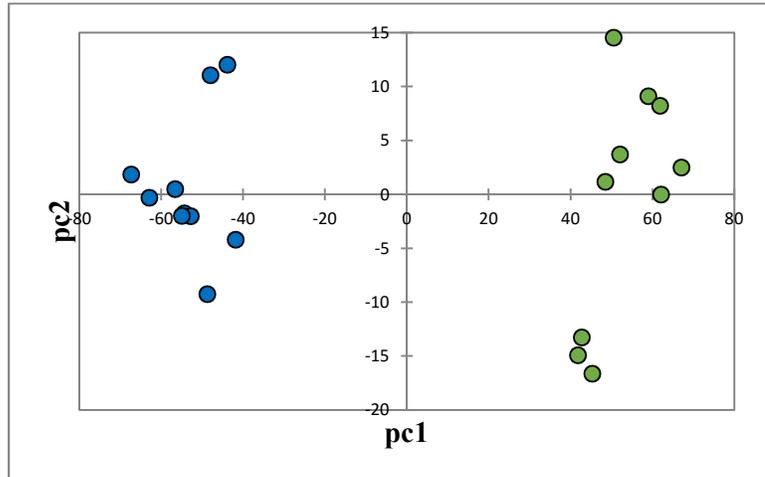
PCA biplots are of 50 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 3,220 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 235 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 14 SNPs. Separation of case and control is maintained for all reduction of SNPs.

Figure 5. Reducing Number of SNPs for Second Set of 20 Samples

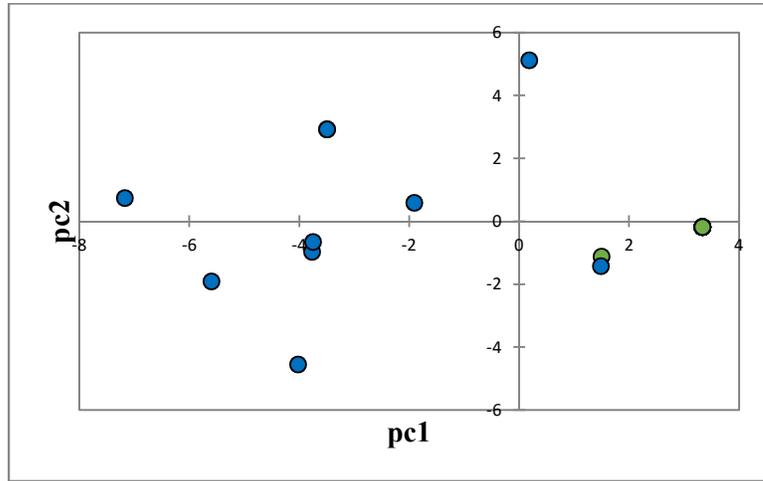
A.



B.



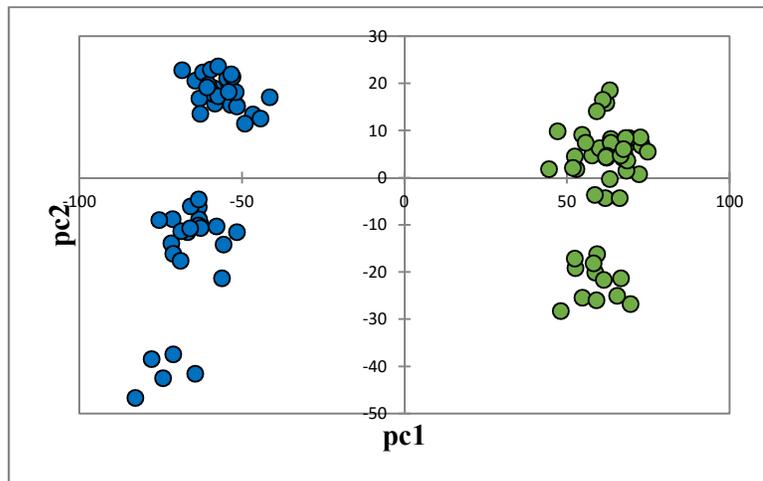
C.



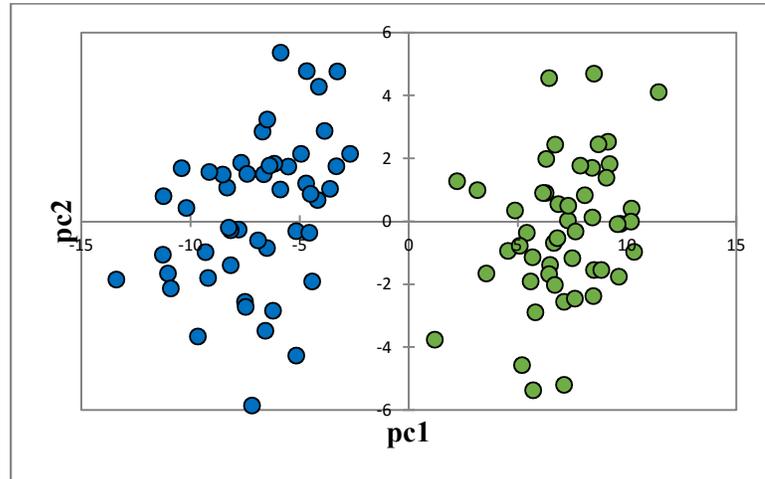
PCA biplots are of 20 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 2,396 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 169 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 6 SNPs. Separation of case and control is mostly maintained for all reduction of SNPs.

Figure 6. Reducing Number of SNPs for Third Set of 100 Samples

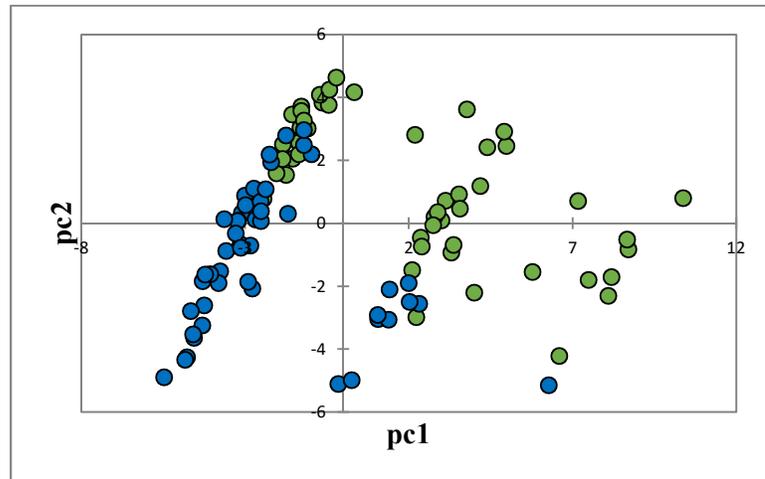
A.



**B.**



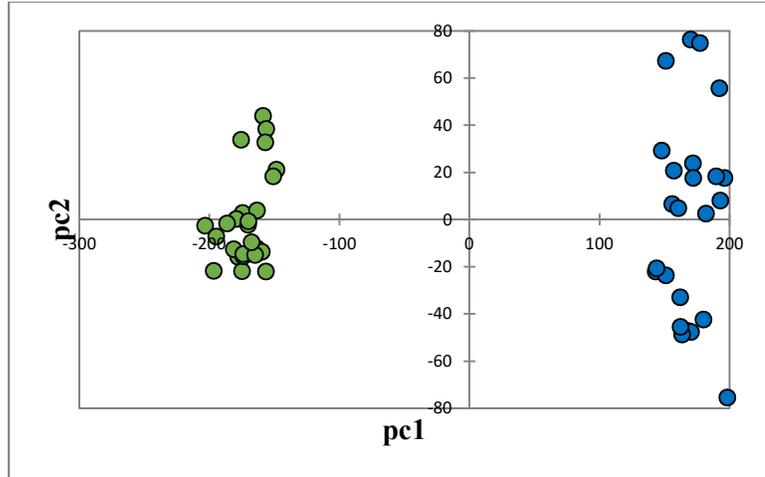
**C.**



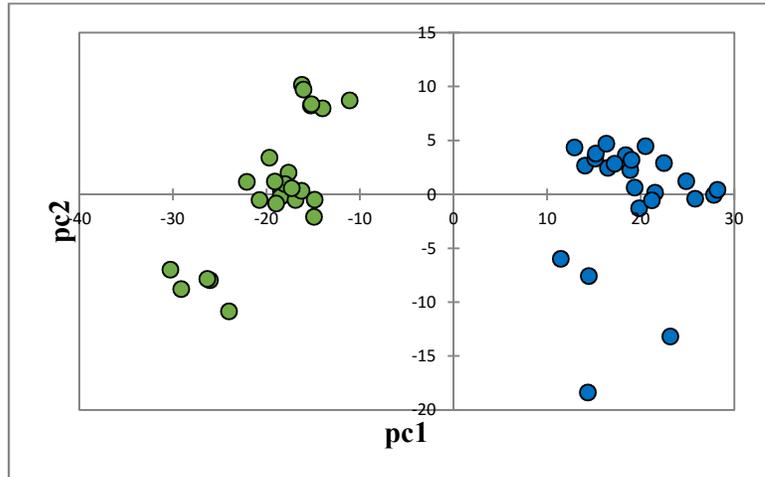
PCA biplots are of 100 samples with case green and control blue. There were 2 outliers removed from all plots. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 3,190 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 248 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 19 SNPs. Separation of case and control is mostly maintained for all reduction of SNPs.

Figure 7. Reducing Number of SNPs for Third Set of 50 Samples

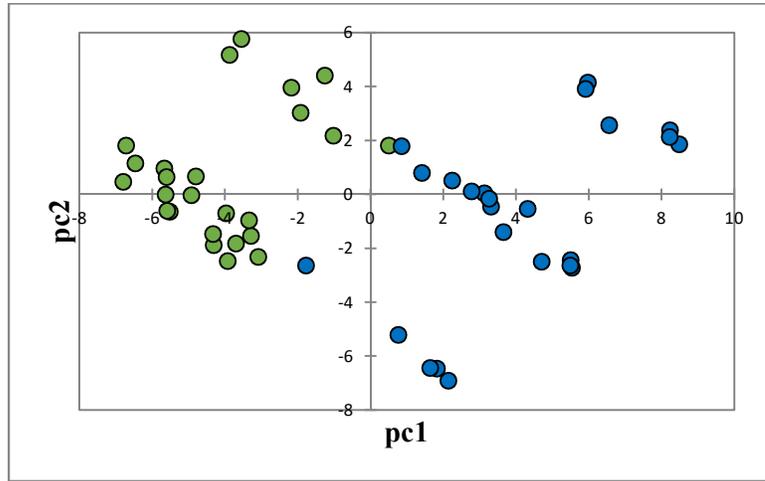
A.



B.



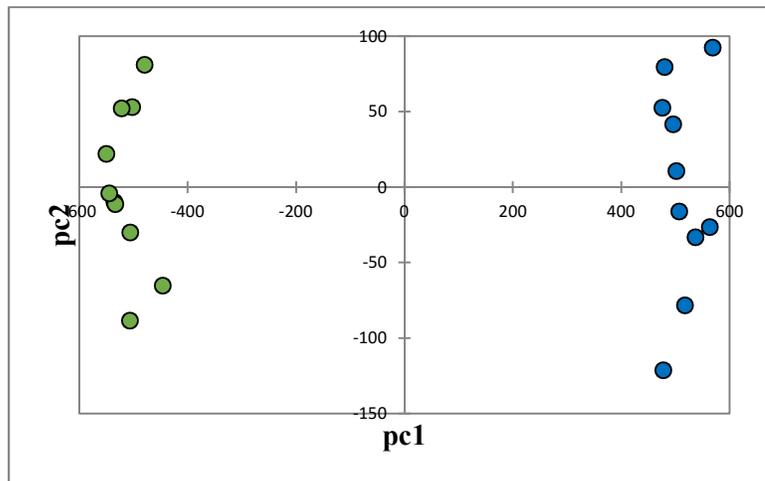
C.



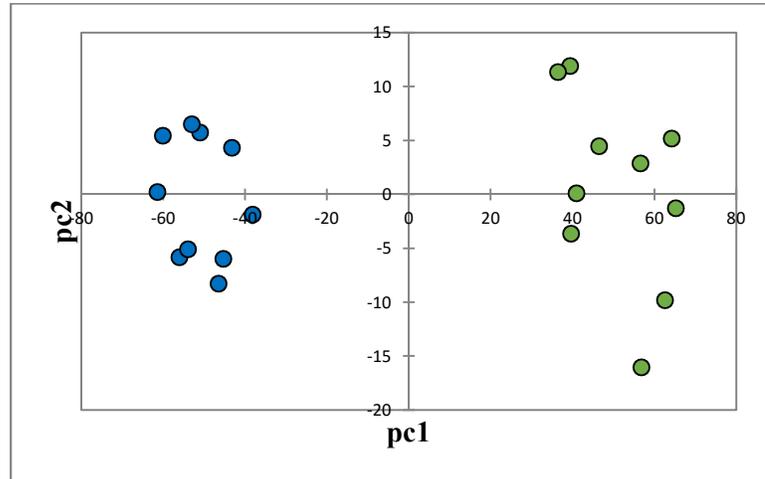
PCA biplots are of 50 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 3,136 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 234 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 24 SNPs. Separation of case and control is maintained for all reduction of SNPs.

Figure 8. Reducing Number of SNPs for Third Set of 20 Samples

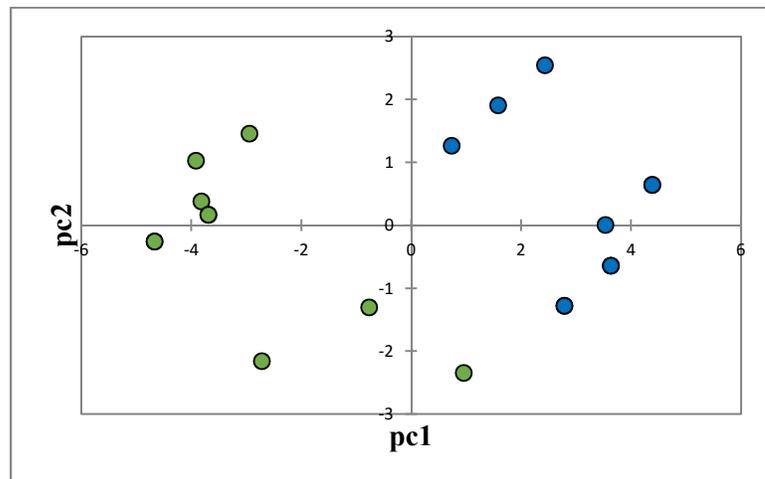
A.



**B.**



**C.**



PCA biplots are of 20 samples with case green and control blue. A. The p-value is  $5 \times 10^{-3}$  as the threshold giving 2,342 SNPs for PCA analysis. B. P-value of  $5 \times 10^{-4}$  as the threshold with 162 SNPs. C. P-value of  $5 \times 10^{-5}$  as the threshold with 5 SNPs. Separation of case and control is maintained for all reduction of SNPs.