8-2023

Exploring Experimental Design and Multivariate Analysis Techniques for Evaluating Community Structure of Bacteria in Microbiome Data

Kelsey Karnik
knkarnik@huskers.unl.edu

EXPLORING EXPERIMENTAL DESIGN AND MULTIVARIATE ANALYSIS TECHNIQUES FOR EVALUATING COMMUNITY STRUCTURE OF BACTERIA IN MICROBIOME DATA

by

Kelsey Karnik

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professor Kent Eskridge

Lincoln, Nebraska

August, 2023

EXPLORING EXPERIMENTAL DESIGN AND MULTIVARIATE ANALYSIS TECHNIQUES FOR EVALUATING COMMUNITY STRUCTURE OF BACTERIA IN MICROBIOME DATA

Kelsey Karnik, Ph.D.

University of Nebraska, 2023

Advisor: Kent Eskridge

The gut microbiome plays a crucial role in human health, and by working collaboratively with microbiologists, we aim to further our understanding of the human gut and its impact on human health. Promoting a diverse microbiome is emphasized throughout the microbiology literature, and involving a statistician in designing experiments to relate gut bacteria and some measured health outcome is crucial for ensuring valid and accurate results. By adopting new experimental design and analysis methods, researchers can begin to gain a deeper understanding of how the genetics of our food affects the composition of taxa within the gut microbiome. This dissertation is structured around three main objectives, demonstrating how applying new experimental design techniques and multivariate analysis methodologies could potentially benefit domain-specific researchers throughout the scientific process. This work developed a new experimental design structure for assigning treatments to well-plates. Multivariate analysis methods were used to analyze the data, creating new polymicrobial traits to introduce a community taxonomic effect into genome-wide association models. Finally, the effects of experimental parameters on statistical optimality criteria were explored. Our randomizations and experimental design structure exhibited increased efficiency over a design that included only replicate effects. After analyzing our taxonomic abundance data and decomposing the variability in multiple formats,

our new pseudo-multivariate phenotypes were included in our collaborators' GWAS models. We found that 57% of the calculated polymicrobial traits were included in the genome-wide association study (GWAS) models. Over half of the polymicrobial traits used as responses contained either a direct or related overlap with a univariate taxon on the same Major Effect Loci, where some of the unique and helpful relationships were explored more in-depth regarding taxonomic functions within the microbiome. Lastly, we developed a function that calculates the composite optimality criteria to compare design optimality for a multivariate linear mixed model with a covariance structure on the random genetic effects. In the future, similar models and optimal design functions could help researchers improve their experimental design layouts by leveraging their knowledge of genetic relationships in our diets and the relationships between taxa in the gut.

DEDICATION

For my uncle and Godfather, Jeffrey Allen. I hope you are looking down, proud of me now as you always were. See you again someday. Smooches.

ACKNOWLEDGMENTS

I want to thank many people for assisting me in my personal and professional growth from childhood through now. I would not be the person I am today without every one of these amazing individuals.

First, thank you to my advisor, Dr. Kent Eskridge. I appreciate all the thought and hard work you have put into assisting me with my research. You provided me with the best support system possible to thrive in becoming a researcher and statistician. Thank you to Dr. Yuhang Xu, Dr. Jane Meza, Dr. Reka Howard, and Dr. Kathy Hanford for serving as my committee members and supporting me throughout my graduate school career. Kathy, you have always been a champion for my work as a statistical consultant, and I have grown so much as a statistician, researcher, and scholar with your assistance. Your helpful network of other statistical collaborators helped me expand my knowledge of the field and greatly aided my professional growth.

Next, thank you to all my UNL instructors, classmates, and colleagues who have helped me along my journey. I owe so much of the foundation of my statistical knowledge to Dr. Walter Stroup. Without your classes and assistance at the SC3L, I don't think I would be the statistician I am today. Thank you as well to Brianna, Johnna, Tiffany, Ella, Ashley, Rachel, Miguel, Vamsi, Emily, and Alison. I cherish countless wonderful memories with all of you, moments that will forever hold a special place in my heart. Specifically, Brianna, Emily, Miguel, Vamsi, and Rachel, thank you for being by my side in growing the SC3L and being the best crew of colleagues. We helped so many people, and I grew immensely professionally along the way. Brianna, thank you for always being a mentor and open ear throughout my graduate school career. You showed me how to persevere through the hard times, and it has been a joy seeing you and your family grow over the last few years.

Jessica, you have been my roommate and closest friend in graduate school through all my ups and downs. Between all of the sporting events we attended, the many nights spent arduously worrying over one thing or another, and all of our inside jokes, I will never forget our time spent together. I can never thank you enough for all of your support. You will be a brilliant professor, researcher, and mentor at the Air Force Academy. I am so proud of you, and I know we are both going to do great things. I am so happy to have been a part of your story.

Thank you also to all my friends and family outside of graduate school. Taylor, you have been one of my biggest supporters and best friends since high school. Thanks for always checking in and talking through things when I need an outside perspective. Alex, meeting you in 2018 changed my life forever, and I cannot imagine going through these years of my life with anyone else. You make me feel like I can do anything and support me no matter what. I love you, and I like you.

Lastly, the biggest thank you goes to my family, who have helped me along the way. Especially my parents, Tammy and Ed Karnik, and siblings, Adam and Shaleigh. You are my people, my closest support system; without you, I would not be who I am today. Words cannot describe all you mean to me, and a million thank yous to you all.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Interest in gut-microbiome research has been growing and evolving over the last 15+ years with the expansion of scientific research in academia and industry. Statistical methodology and research have facilitated this growth by creating new methods to evaluate the relationships of the microbial responses to various aspects of diet, genotypic, and other host characteristics. Research into new analysis methods is ever-growing, but it seems no one has tried to evaluate the "best" way to design microbiome studies that collect data in a lab setting with a similar type of data. For various reasons, direct power analyses are difficult to conduct for microbiome data, so they are not often completed before conducting experiments [30]. In recent years, the creation of new technology for collecting omics data has made statisticians aware of the possibility of issues with confounding variables that could create unexpected differences in the final results[6]. Bailey (2019) also states that counter to microbiome research, "agricultural science has a long history of rigorous experimental design, power calculation and statistical evaluation", which will be helpful in this dissertation to aid in the creation of a new design for collecting data for microbiome experiments.

There has been an abundance of microbiome literature in recent years outlining

general, basic recommendations for best processes with how to analyze microbial data and some pinpointing basic design methods that could improve experiments [14, 80, 91, 132, 133, 134]. Different analysis methods are compared based on how models account for different response data types and what types of explanatory and random effects the models consider. The data's compositional and multivariate nature can cause modeling issues, and often "design and planning for congruence in diet and biological sample collection, along with a consideration of the analysis and statistical challenges ... will facilitate high-quality research and greater confidence in reported outcomes." [133, 83]

UNL Food Science and Technology researchers conduct genetic experiments investigating how different foods affect the bacterial makeup of the human microbiome composition. The idea for this dissertation was developed based on the need for new sample randomization for an experiment utilizing well plates in the lab of Dr. Andrew Benson. The study aimed to evaluate relationships between the gut-microbiome microbial makeup and the genetics of a diverse population of dry bean lines. Wellplates are used to create simulated gut microbiomes using provided stool samples from human subjects combined with samples from a variety of different bean lines. The work aims to develop new and novel statistical designs and analysis methodologies for researchers in microbiology who conduct experiments dealing with quantifying how food and diet can affect changes in the microbial diversity of the gut microbiome. The data within this dissertation focuses on quantifying how changes in the proportion of bacteria in a simulated human microbiome are correlated with unique attributes of bean line genomes. It is hypothesized that changes in the bacterial makeup in the microbiome could lead researchers to new information about how diet affects human health through the gut. It is a general goal that "by evaluating taxa presence, absence, and abundances, one can investigate the extent of the difference between the

compositions of communities between samples."[132]

Currently, few standards exist for incorporating statistical design principles into well-plate experiments since many take place on a large scale, making it more costly and time-consuming to use randomization, blocking, or other statistical design methodology. Occasionally there are different types of simple randomizations built into lab experiments. However, only randomizing treatments to wells within a plate can only account for so much extra variability. Often the randomizations included an attempt to account for well-to-well contamination, variability between lab equipment, or overall random noise. As studies become more intricate and comprehensive, relying solely on basic randomizations is insufficient to address the diverse range of variability, necessitating the implementation of blocking methods.

When discussing areas of possible growth in microbiology research with those in Dr. Benson's lab, ideas for research were developed based on the need for more intricate experimental designs in 96-well plate lab experiments. First, design protocols and standards should be created to minimize spurious variations from the data responses collected. For this dissertation, an initial sample design was created to satisfy the immediate need for a lab experiment randomization. Then, using knowledge learned from these initial experimental runs, the optimality of different design structures was evaluated for future experiments. Also, the analysis methodology used to relate the bean genetics to the output of bacterial abundance should aid in relating statistical results to the outputs from a genome-wide association study. In questioning how to integrate more randomization into microbiome research, an issue to consider is how best to implement the selected procedure, which relates to the optimality of experimental designs. Like a power analysis, researchers need guidance on how many samples will yield the most precise model estimates, the best way to replicate across different types of experimental units, what restrictions exist with the analysis of these

methods for multiple response variables, etc. Chapter 2 describes the methods used to create the design for our experiment conducted in 2019 and 2020. Descriptions of the analysis procedures and methods are outlined in Chapter 3. Chapter 4 evaluates optimal designs related to microbiome experiments across a community of multiple bacterial taxa.

1.2 Motivation and Contribution

From a statistical point of view, there is evidence from the microbiology literature that experimental design in microbiome research is lacking and is not often included in the experiment protocol. The motivation of this current experiment is to account for any extra variability that is not due to the relationship between the treatments and corresponding responses. Including experimental design in the data collection process will assist in answering the researchers' questions and act as a starting point for creating the most optimal designs to quantify how specific bean genetics are changing the makeup of the gut microbiota. Overall, researchers' goals are to investigate and quantify the relationships between the genetics of different bean lines and the microbial taxa created in a simulated microbiome. However, analyses often only consider univariate cases and do not include design effects or covariates accounting for the experimental units where the different treatments are applied.

Statistical design theory of the most optimal designs started with applications to agriculture and biological sciences in the 1920s, but new research has expanded the topic around how to create the best design for optimizing a specific function of the models' variance matrices [78, 79]. With this focus, optimal statistical designs can focus on specific data types and create a methodology for making the best use of the resources to find the most precise model parameter estimates. Hopefully, with

efficient estimates of the model variability, an "optimal" design could lead researchers to gain the most information from the data analysis. The theory of optimal design can give microbiological researchers specific directions as to what types of randomization schemas to introduce into their experiments by offering "a systematic way for finding an optimal or highly efficient design using all current information for the problem at hand" [13]. With the field of analysis for these types of experiments growing so fast, the need for "randomized, controlled experimental designs is crucial," and microbiome research specifically would benefit from statistical methods that utilize important genetic information in the experimental design [65].

Since the type of design described in Chapter 2 is newer in microbiome research, the initial data can act as pilot data for research into the "best" type of design with relative abundance microbial data measured across plates in this lab setting. In turn, a motivation of this dissertation is to outline methods for microbiologists to create statistically optimal designs that give the best estimation of the corresponding variances and covariances between bean lines and taxa while considering the limitations and restrictions created within the boundaries of this type of lab setting. Data collection for these microbiome studies comes at a price both with money and time. The current design took around 3-4 months per subject to complete, not considering the time for analysis and interpretation of the results. So, finding the most statistically optimal design that can reduce the number of replications within or across plates will be beneficial. Using fewer plates could reduce the time needed in the lab while also producing the most informative variance estimates.

Additionally, there is motivation for work in the analysis stage of these experiments. Microbiome bacterial abundance data is compositional and multivariate, meaning that the responses measured are all proportions of bacteria in the sample that sum to 100% of the total sample. Over the last several years, there has been a

plethora of research investigating how to account for these attributes in data analysis. [80, 53, 155] In 2021, Bharti and Grimm wrote that "analyzing microbial data is challenging due to its large and multivariate data structure," and it is "difficult to provide a best-practice pipeline for straightforward statistical analysis because it highly depends on the core objectives of the study and the underlying hypothesis." [14] This presents an interesting gap of knowledge in the subject matter that this dissertation will attempt to fill given information from our current experimental design, analysis, and optimal design simulations.

Together, background from microbiome literature and discussions with subject-matter researchers indicated gaps in knowledge in the field concerning experimental study design and multivariate analysis to quantify the effects of treatments on the community of bacterial taxa produced in gut samples. New technological advances with lab equipment have allowed for better-randomized designs and systematic allocation of samples to the typical 96-well plates used for the testing and analysis of the microbiome samples. With these new procedures, improvements can be made with how experimental design can be integrated into the study design and analysis phase of microbiological research with bacterial abundance responses. Since developments in the realm of experimental design are relatively new within this field, there is little to no research on the optimal way to design these types of experiments. This is a substantial reason why we believe this research is necessary to fill the knowledge gap in this area.

Overall, the contribution of this dissertation is to create methodology and recommendations assisting researchers with the design of microbiome studies that can identify the relationships between treatments and microbial relative abundances measured from well-plate NGS (Next Generation Sequencing) studies. This process started with creating a basic initial design, which led to evaluating the results from

that design and interpreting the corresponding outputs to answer the researchers' main questions. Finally, this data was used as pilot data to explore more optimal methods to design their studies.

1.3 Aims and Objectives

Given the background from microbiology literature and statistical methods typically used within the field, some challenges and questions of interest have been identified. We believe that addressing these questions and issues can assist researchers in creating better and more optimal experiments. Improved experiments will facilitate the process of identifying important relationships that exist between the genetics of specific treatments and the microbial taxa that are created in the microbiome. This knowledge could help identify areas of diet essential to a healthy gut.

The following research objectives will be addressed throughout the next three chapters of this dissertation:

1. To develop a new experimental design for next-generation sequencing 96 well-plate lab experiments, specifically dealing with the human microbiome.

2. To use multivariate analysis methods to analyze data outputted from the experiment designed from objective 1.

3. To identify and explore the effects of experimental parameters on statistical optimality criteria for analyzing data from microbiome well-plate experiments.

Chapter 2

Experimental Design

2.1 Background

Over the past 15 years, the proliferation of high-throughput sequencing techniques has opened up new avenues for researchers to explore the microbial realm, enabling them to investigate the intricate intricacies of the microbial world in scales and depths previously unimaginable just a decade ago. [149]. Since these types of experiments take place in a laboratory setting, statisticians working with researchers should try their best to incorporate design configurations that assist in explaining as much variability as possible. The more variability in outside factors the model can account for, the more likely it becomes to find insightful relationships between the explanatory and response variables. To make the most informative interpretations of the data, the selection of sample size is essential because "variability between similar samples makes it challenging to identify weak biological signals," thus, designs "based on statistical principles can certainly help to avoid biases and spurious interpretations." [14] Thus, reducing outside sources of variability with appropriate experimental design features will help researchers better interpret the relationships between the different bean lines and the bacterial abundance output.

As the science around microbiome research continues to grow alongside the

technology for DNA sequencing, researchers have begun to handle big data sets, often requiring new computational tools to collect and interpret the data. However, the "standards for data collection and analysis are still emerging in the field," and new methods are struggling to grow because the current methods can still yield plenty of beneficial results [80]. These authors point out that some of "the most fundamental issues that concern microbiome studies arise from statistical and experimental design issues," and researchers should "integrate new approaches that are unique to microbiome studies, while remembering standard practices that are broadly applicable to all scientific studies" [80]. Thus, this chapter's goal is to evaluate what types of experimental designs currently exist in the field and find what current literature identifies as important issues surrounding integrating experimental design into microbiological microbiome experiments.

First, in creating the most practical and insightful design, we had discussions with Dr. Benson and the graduate students working on the project to help set general guidelines and ensure we understood this experiment's constraints and what restrictions exist in the field. When talking to the researchers, it is essential to remember that the design should be reasonable for the lab to complete and allow for the most suitable statistical analysis. In the experiment, researchers should address extra variability that may be introduced into the bacterial response measurements from the samples because of location differences within and across plates. Many aspects of these experiments motivate a more intricate experimental design than is currently being used. This new design should explain the extra noise and improve the ability to find statistical and practical significance. Hopefully, reduced noise would provide a more precise understanding of how the different bean lines affect the bacterial makeup of all the human subjects' samples. Based on discussions with Mallory Van Haute, a Nebraska Food for Health Center fellow and Ph.D. graduate from the UNL Food

Science and Technology department, microbiologists do not generally utilize experimental designs with large-scale microbiome next-generation sequencing experiments. Similarly, there are instances in the literature that calls for better randomization techniques and sampling procedures to be used in this field of research. Many of these papers deal with how to account for well-to-well contamination and location effects, specifically within experiments that utilize 96-well plates to conduct genetic testing.

2.2 Literature Review

The review of the literature will be two-fold. First, information from biological resources will evaluate what experimental designs have been investigated. With this, it is essential to identify what potential issues can be solved by incorporating a more detailed experimental design methodology in creating the lab experiment, alongside taking note of what has been done and what has and has not worked well. Additionally, the design methodology employed to formulate our ultimate design for the bean line experiment will be elucidated, drawing upon relevant background information from the statistical literature. This methodology revolves around creating partially replicated (p-rep) designs augmented with an α - design for the extra replications.

2.2.1 Statistical Design in NGS Lab Experiments

The biological literature is particularly helpful in evaluating descriptions of lacking experimental design and describing methods that have been integrated into current research. This review will assist in pinpointing what researchers have been able to help with versus what still needs to be improved. There seems to be, first and foremost, a need for more structured experimental designs to reduce extra variability due to human error, sample contamination between wells, or sample differences due

to location, time, or other covariate effects. A pocket of literature also calls for more specific ways to randomize the treatments to the wells where the microbiome samples will be created and tested. These pieces show the need for a new and succinct method of design and randomization in sequencing experiments using well plates in a lab setting.

As first referenced in the introduction to this chapter, Knight et al. (2018) describes some of the best practices for experimental microbiome analyses and write that issues of confounding within human microbiome research need to be taken into consideration in order to be able to quantify meaningful results [80]. Confounding factors (such as sample position within a plate) can cause bias due to temperature gradients and specific properties of the plastic used in the plates and errors due to how the materials are transferred into the wells. [62]. Previous research has demonstrated that if location and position effects are not accounted for, there can be bias in the mean square error, which will cause invalid inferences from t or F tests (when simple analyses are used) [62]. This work from Hooton and Paetkau is based on another study that demonstrated that even with some random assignment of replicate wells to treatments within the plate, the replicates could help control extra position variability [62, 141]. These two papers also began to consider that researchers in a lab will perhaps be more willing to utilize design and randomization procedures if computer programs create the treatment layout.

Other authors have called for the expansion of randomization and structured design to account for well-to-well contamination that could also bias the responses measured from the human samples. For example, Minich et al. (2019) conducted an experiment specifically designed to "empirically characterize the frequency and nature of well-to-well contamination using different DNA extraction and sample handling protocols" [96]. Designed procedures looked for contamination due to automated vs.

manual extraction measures and whether there were differences between plates tested at two microbiome facilities. When looking for this type of contamination, plates contained "16 unique bacterial "source" isolates at high biomass in individual wells across plates of alternating low-biomass "sink" bacteria and no-template blank wells" so that researchers could evaluate well-to-well transfer events. Results demonstrated that "well-to-well contamination occurred in all six PCR replicate plates in both laboratories" and "that the highest rates of contamination occurred in the immediately proximate wells for both plate and tube extraction." The authors then concluded that since well-to-well contamination was largest in immediately adjacent wells, "sample location on plates should be explicitly considered in experimental design" and "it is important to block and/or randomize treatments across 96-well plates" to "better ensure that well-to-well contamination adds only noise and not bias to experimental designs" [96]. However, although the literature has advocated for design and randomization since the 1980s, currently standard design protocols have not been created for microbiome research to assist with well-plate location type effects.

Roselle et al. (2016) investigated the positioning/plate effects and found that their simple "block-randomized plate layout" as compared to a non-randomized plan demonstrated a "significant reduction in assay error including both reduced bias and reduced imprecision" [125]. The authors described that this design was chosen as a development while looking for techniques that would help "optimize plate layout and replication strategy," which previously had focused on complete randomization and variations of split-plot and strip-plot designs. These types of designs were discussed as strategies that have been "widely recommended as mitigation strategies to compromise between complexity and effectiveness" within the experimental design [125]. Results showed that the non-random plate assignments did return positional effects, and the randomized plates were more accurate with a negligible mean absolute rela-

tive bias as compared to the non-randomized plates. Thus, it appears that there are benefits to using plates as a block where treatments are randomly assigned. These authors also compared their design to a design proposed in a 2012 briefing chapter from the USP Pharmacopeial Convention, which described a type of split-plot design that assigns treatments to plate rows, secondary variable levels to wells within rows while leaving the outer wells (around the edge of the plate) completely blank [125]. In comparison with the standard block design, they discussed that even with the benefits of these randomizations, this type of split-plot comes "with a questionable trade-off: increased complexity and significant loss of throughput (three treatments per plate) balanced by only partial mitigation of positional effects." Thus, it seems pertinent to make sure that a final design uses an appropriate randomization and design scheme to account for location effects and that the design remains relatively straightforward so that biological scientists can easily implement it.

Above and beyond specific well-to-well contamination, batch effect issues exist across different plates (and sets of plates) tested over time. Suprun and Suarez-Farinas addressed this by attempting to control systematic batch variability with experimental design and randomization across different microplate batch runs [145]. They created an online application that designs a microplate experiment with randomization across the plates to control the batch effects. They allow researchers to select the specific experimental design by "specifying the randomization units (RU), the number and position of replicates, and the type of control samples" [145]. The PlateDesigner application currently allows treatments to be randomly assigned to wells or to be assigned to keep replicates together in rows or columns "as to ease the technician's labor." Users can set specific sampling weights to ensure that particular treatment groups appear a certain amount within each plate. The randomization schematic can be easily outputted to a PDF or CSV file which can then be read

into a “microplate reader software” to tell the machinery where to input each specific sample automatically.

To the best of the authors’ knowledge, their app was ”the first and only web-based application currently available to researchers that generates randomization schemes for microplate experiments” [145]. The benefit of a system like this is that the researchers who will be creating their experiments do not necessarily need a wide breadth of coding knowledge to use the application. It provides them with an easy way to utilize different (fairly simple) randomization schemes to create unbiased experiments that can potentially assist in balancing out any confounding effects due to location. Thus, research shows that there is currently a need for new and innovative experimental design standards and easier ways to implement these methods for research scientists to create better microplate experiments. These benefits of new designs and simple execution could hopefully sway the research scientists into utilizing more complex designs and randomizations more often by making them simpler to understand and easier to put into practice directly.

2.2.2 Augmented and Partially Replicated (P-Rep) Designs

In research with many levels of an explanatory variable, as in plant genetics, there are often hundreds of genotypes to evaluate, making it challenging to have enough replications of each treatment. Resource restrictions, such as lack of time, money, lab space, etc., can often be the main cause of the lack of ample replication. However, when replication is low, the variability goes up, and when variability is large, it is more difficult to find statistically significant differences between treatment levels. However, suppose there cannot be many replications across all treatments, but a subset of treatments can be replicated more often. In that case, the design can become a type of partially replicated or p-rep design. The literature reviewed below

describes designs that can be utilized to control the extra variability as a side effect coming from a lack of sufficient replication.

Implementation of partially replicated designs was first based on the literature surrounding augmented designs proposed by Federer in the 1950s and 1960s. This group of augmented designs was from Federer's work screening sugar cane strains used in growing pineapples [41]. This type of design was created because adapted any standard design (hence "augmented") to allow for more treatment replication "in the complete block, the incomplete block, the row, the column, etc." [42]. As Federer points out in his 1961 paper, up to this point, there were descriptions of designs that utilized additional treatments, but a general approach had not yet been formalized. Examples of this approach contain blocks and two different types of treatments, one being replicated more often. Then by randomizing both treatments and the checks within the adjusted design block, one can get an estimate of the error variance and can correct for block effects [42, 113]. An example layout of an incomplete block design from Federer can be found in Figure 2.1, where the standard treatments are represented with capital letters and the new treatments with the lower case. [42] This is also described as an "augmented triple lattice" with nine standard treatments and 15 of the not multiple replicated treatments.

	Replicate and incomplete block number								
	1			2			3		
	1	2	3	1	2	3	1	2	3
	A	D	G	A	B	C	A	C	B
	B	E	H	D	E	F	E	D	F
	C	F	I	G	H	I	I	H	G
	j	l	n	p	r	t	u	v	w
	k	m	o	q	s	-	-	-	x
$n_{.jh}$	5	5	5	5	5	4	4	4	5

Figure 2.1: Augmented Incomplete Block Design

This design is generally very similar to the one described later utilized for our

main bean line experiment. There are sets of incomplete blocks, and then within each block, some bean lines will have more replications than others. Talking with the subject-matter researchers, it is possible to obtain enough power to replicate the bean lines multiple times across each subject, including for bean lines that are selected as check lines. Therefore, instead of having some lines not replicated, we can set up a design with different total replications between the check and non-check lines, with all lines still having more than one observation. While this initial idea provides a simple starting design, the additional background information from the researchers clarified what attributes the design could contain.

Additionally, previous literature evaluates the benefits of augmented designs. Federer and Raghavara (1975) evaluated optimal augmented designs. These designs get closer to what is possible to institute in the UNL Food Science lab because it utilizes row-column designs, which are natural to employ for the bean line experiment because of how the 96-well plates are organized (8 rows by 12 columns of wells) [43]. However, Federer and Raghavara's design places treatments in a row-column design based on the ideas of a Youden design from Youden (1937), where the treatments are arranged in rectangular rows and columns where all treatments appear in each of the rows, at most once in a column (then in some total number of columns); Every pair of treatments also appear together the same number of times [160]. This type of design might be more reasonable when testing a smaller number of lines, but it is perhaps unreasonable in this experiment to have all lines appear in each column.

Another advantageous offset of augmented designs are augmented partially replicated (p-rep) designs. Smith et al. and Cullis e.t al. were some of the first papers to formally propose this style of design, which is used frequently in plant breeding, crop research, and genetic experiments. [138, 27]. These designs are "particularly useful when seed is limited, the trait of interest is expensive to measure and/or where

there are multiple phases of experimentation, as in a situation where lines are grown in the field and then further tested in a laboratory" [64]. Cullis and Smith each evaluated the benefits of different styles of p-rep design in comparison to fully replicated designs and designs with no replications [27, 138]. Smith found within multi-phase wheat breeding trials that a p/q-rep design, which replicated a proportion (p) of the genotypes in the field and a proportion (q) of the genotypes in the lab setting, was able to reduce error variance in comparison to a design that did not include any replication [138]. Benefits from the p/q rep design include that when spatial correlation is present with "with a relatively large variance and a genetic variance the same size or smaller than residual variance," the genetic gain was larger for the p/q design than a design where subsampling was used as pseudo-replication. Thus, if spatial correlation is present, there are benefits to p-rep designs (depending on the different variabilities in the experiment). One of the goals of randomizing the bean line experiment is to account for spatial correlation across plates and between rows and columns within plates, making this type of design more attractive as a path for the bean line experiment.

The experiment designed within this dissertation does not have a multi-phase aspect, as in Smith et al., so another paper (coauthored by Smith in the same year) provides more insight into the usefulness of p-rep designs as compared to typical field-grid plot style designs. Cullis et al.is a commonly cited paper as one of the first pieces of literature to formally propose and evaluate the p-rep designs with agriculture variety trials [27]. Based on the simulation, they demonstrated larger genetic gains for p-rep designs compared to standard grid plot designs, assuming that spatial/correlated data was present. The ability to account for spatial data is helpful because well-to-well variability across and between plates creates correlations between samples; thus, it is beneficial to allow for design methods that consider these extra

spatial relationships.

While Cullis' main topic of interest is early generation variety trials, they feature interesting design procedures that could benefit other research areas. Again, the basic principle is that a percentage of treatments are replicated (proportion indicated again with a p) and the remaining treatments are unreplicated. One benefit of this design style is flexibility because "other values of p and levels of replication may be used, and the basic design is easily modified to suit specific requirements" and "test lines to be replicated may be made completely at random or may be influenced by the availability of seed or interest to the breeder" [27]. Thus, p-rep designs allow for designs to be specialized based on individual resources or research goals.

The simulated example in this paper contained 120 possible lines with an available grid of 150 total plots laid out in a rectangle-sized 30 rows by five columns. A pre-selected proportion of replicated lines of $p = 0.25$ were used, which means that 30 of the 120 lines should be replicated, and there will only be one observation for the remaining 90 lines. Out of the 150 total plots, unreplicated lines will take 90 plots, leaving two plots per replicated test line. Another specification made is that these designs are resolvable by letting "all 30 replicated lines occurred once in the top half of each design (that is, rows 1 to 15) and then again in the bottom half (rows 16 to 30)." [27] Resolvable here means that some of the groupings of the treatments in the incomplete blocks can be combined to create a complete block of observations.

With the given optimality criteria (based on average pairwise prediction error variance of the test line effects), simulations with both fixed and random variety effects were conducted to look at the similarities and differences between three p-rep and three standard grid designs. Data were analyzed with the models used to generate the data. Comparisons found that the grid plot designs had a larger "percentage of analyses in which the genetic variance ratio was estimated as zero" than the p-

rep designs.[27] It is not good to have a genetic variance ratio at zero because this restricts the data from making appropriate genetic selections; thus, the p-rep designs were performing better. P-rep designs also performed better than the standard grid design for the values of the realized genetic gains. While these ideas expand on the benefits and potential of using p-rep designs to gather more useful genetic information from a study, the ease of creating the design from this layout was lacking. Cullis and the authors state that a subject of ongoing research was still how best to allocate the test lines to the experimental units [27]. In 2011, Williams, Piepho, and Whitaker focused on augmented partially replicated designs and utilized the benefits of α -designs to produce more efficient and easily created designs [152]. The idea is that information from Federer and Williams in the 1960s, 1970s, and early 2000s also from Smith and Cullis provide good background about why p-rep designs are beneficial, but they lack an easy way to pick the best augmentation and best way to construct the design.

Williams et al. indicate that the augmented designs as described above from Federer are good to provide the "local error control" from using some replicated control (check) lines within the analysis to augment the unreplicated entries [152]. Authors add that the extensions from Federer et al., 1975; Federer, 2002; Williams and John, 2003 allow for different types of row and column blocking, which again allow for minimal replication for particular entries, but "suffers from the fact that a considerable proportion of plots may need to be devoted to controls."; Thus, the new contribution from Williams is the combination of Federer's augmented designs, which replaced controls with other entries, and Cullis and Smith's designs that could be combined across different blocks (which make the designs adaptable when experiments take place over different locations) [152]. Williams, Piepho, and Whitaker then demonstrate how further adding in α-Designs from John and Williams (1995) helps

to "produce p-rep designs where at each location there is a two-replicate resolvable block design for a proportion of the entries and overall pairs of entries concur no more than once" [71, 152]. These new designs generate the design for the treatments across blocks and locations using α arrays. The replicated entries are thus spread out among the plots and not necessarily arranged in any particular manner within or across experimental units. Alpha-designs are flexible for these types of experiments because "there is no limitation on block size other than the unavoidable constraint that" the overall number of treatments divided by the block size must be an integer.

The α - designs from the John and Williams text are based on designs of the same name from Patterson and Williams (1976a) in a paper entitled "A new class of resolvable incomplete block designs."[71, 111] This new design is described as an extension of previously researched resolvable incomplete block designs and other lattice designs. Their α-designs were created in order to construct "a class of resolvable equiblock sized designs ... with no limitation on block size other than the unavoidable constraint that k must be a factor of v" (where v represents the number of treatments) and to provide a simple way to automate a computer algorithm to create plots for agricultural trials.

The creation of an α - design begins with a generating array that is comprised of a k x r array (called the generation array), where r is the number of replications desired and k is the number of plots per replication that contains "elements $a(p, q)$ in the set if residues mod s ($p = 1, ..., k; q = 1, ..., r$)." Then the columns are used to generate the following columns cyclically up until there are s total columns, where s is the total number of blocks. To make the process clearer, here is an example from the Williams and Patterson (1976) paper where a design is comprised of 20 treatments for four replications, each with four blocks of size five. [111] From the generating array (discussed a bit more below), the columns of α are used to create

$s-1$ columns in the intermediate array α^* by cyclic substitution. Thus, the resulting array has 12 columns, one for each column in the generating array and then three additional columns per column in the generating array. Then, the first four columns are created by cyclically repeating the zeroes from the first columns, using the values mod 3. Columns 5-8 in α^* come from the cyclical expansion of the second column of α and so on. Finally, "Plan" arrays are created by adding "s to all elements in row 2 of α^*, 2s to all elements in row 3, and so on."

Generating Array, α

0	0	0
0	1	2
0	2	3
0	3	1
0	3	2

*Intermediate Array, α**

0	1	2	3	0	1	2	3	0	1	2	3
0	1	2	3	1	2	3	0	2	3	0	1
0	1	2	3	2	3	0	1	3	0	1	2
0	1	2	3	3	0	1	2	1	2	3	0
0	1	2	3	3	0	1	2	2	3	0	1

Plan

Replication I				*Replication II*				*Replication III*			
1	2	3	4	5	6	7	8	9	10	11	12
0	1	2	3	0	1	2	3	0	1	2	3
4	5	6	7	5	6	7	4	6	7	4	5
8	9	10	11	10	11	8	9	11	8	9	10
12	13	14	15	15	12	13	14	13	14	15	12
16	17	18	19	19	16	17	18	18	19	16	17

Figure 2.2: α - Design Creation Example from Patterson and Williams [111]

The design now has values tied to each treatment, and columns are the blocks for each of the three sets of replications of the treatments.

The generating array α is a particular kind of "reduced" array that has all 0s in the first row and the first column and is isomorphic to the design. So, multiple generating arrays can be used to create different final designs. Williams and Patterson (1976) and John and Williams (1995) go into more depth about the creation of generating arrays.[111, 71] Different computer programs and procedures within SAS and R to create α-generating arrays and designs. With these pre-made algorithms, it is

very simple to create these designs. The compute program used for this dissertation's design is described later in depth.

Currently, several works in the literature proposed ways to combine the ideas of p-rep and α - designs together to get a design with the benefits of both. Williams et al. proposed using and adapting the above α - designs to form the new p-α designs that create α - designs, but then delete certain sets of rows from replicated designs to create the final p-rep design.[152] The authors choose which elements to drop from the final design based on which optimizes their efficiency factor based on the canonical efficiency factor from John and Williams (1995). This type of design has the benefit of giving a simple way to create a full design. However, it lacks a way to easily visualize where the more and less complicated treatments show up in the design and does not allow a simple way to choose the proportion of the treatments replicated more often (like a more typical p-rep design).

Authors also state that an alternative way to generate an augmented p-rep design across locations is to use an α - design for the replicated treatments at each location and then augment them by hand alongside the replicated entries. However, this "naïve" approach allows "concurrences greater than one".[152] Given all of the previous designs, we wanted to try and create an adaptation in which all replicated treatment pairwise combinations and unreplicated treatments occur an equal number of times with sets of plates as randomized blocks. With the experimental resources from the Nebraska Food for Health Center (NFHC) lab, it could be easy to pick which treatments will be replicated and place them in the design to ensure that treatment pairs are equally replicated. This background gives rise to our new combination of p-rep and α-design ideas where the randomization allows for replicated treatments, spread around a plate in an attempt to account for location effects. Researchers can then also superimpose unreplicated designs onto other more standard designs.

This process allows for "many different combinations of experimental designs and unreplicated experiments." [18]

Even though the final randomization for this dissertation is different from some of the current p-rep and α-design expansions, the combination of the two allows for better practical efficiency. Also, the designs together are more beneficial for agricultural and genetic designs instead of unreplicated or simple augmented designs. In the text, Applied Statistics in Agricultural, Biological, and Environmental Sciences, Juan Burgueño et al. comments on handling augmented designs when all the treatments cannot be replicated. Plant breeding is again mentioned as an associated area of research where these designs are useful because "systematic checks" should be utilized in order "to make better comparisons between genotypes." [18] Augmented designs from Federer were proposed to get better estimates of the experimental error because they allow for random allocation of treatments, "as opposed to the standard method in which checks were allocated systematically." [18] Issues can also arise from not knowing the total number of checks to select; however, there are also places within the literature showing that this is not a very worrisome issue overall. Burgueño references a paper in 2018 from himself in 2005 that tested data from "different experimental designs for a uniform experiment in which only one genotype was sown and found that the number of checks did not affect estimation of experimental error." [18]

2.2.3 Randomization Techniques

It would not be feasible to do all randomizations by hand when utilizing both p-rep and α - designs, especially given the number of treatments possible within microbiome experiments. The computer toolkit, GENDEX, has a wide variety of modules that use information based on the setup of the treatments and experimental

units and then output an optimal design based on the chosen randomization format. [101] There are 16 modules available, each of which was created in combination with published works by Dr. Nam-Ky Nguyen. Three different modules will be used to set up the randomization of treatments across plates and subjects for the current design. The ALPHA, IBD, and RRCD modules will be used to create different pieces of the final design. The ALPHA module creates optimal or near-optimal α - designs as introduced in Patterson & Williams (1976a, 1976), and it "uses a 2-stage optimization process" where "each stage of the optimization process uses an algorithm similar to the cyclic-coordinate exchange algorithm described in Nguyen (2002)." [111, 104] The IBD module creates an incomplete block design given a set of treatments, blocks, and the size of each block. Optimality criteria and the algorithm used to create the designs come out of Nguyen (1993, 1994). [103, 102] Lastly, the RRCD module creates resolvable row-column designs based on of the algorithm and optimality criteria proposed in Nguyen & Williams (1993). [105]

2.3 Proposed Experimental Design

This background explains why expanded experimental design is needed within microbiome lab experiments and how statistical methodology can assign 300 bean line treatments to the wells within plates. Mallory Van Haute, the former graduate student who ran the experiments, provided essential descriptions of design pieces that must considered for the design. Information directly from the those running the experiments is pertinent in creating a design to answer the subject-matter researcher's most critical research questions.

Three human subjects' stool samples are available to be added to segments of bean line powder to create simulated microbiomes. Overall, there are twelve 96-well

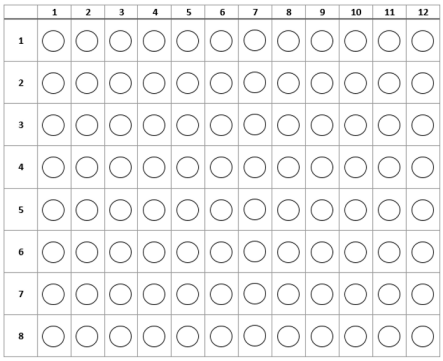


Figure 2.3: Example of 96-Well Plate

micro-plates available to be run through the experimental procedure. The plates are used as the medium where bacteria will grow in simulated microbiomes over time. Enough resources are available to conduct four separate "runs" of the experiment, which would allow for a total of 48 well-plates to be run over time. Figure 2.3 provides an example of an individual 96-well plate, with 12 columns and eight rows of wells where the beans and microbiome samples could possibly be assigned.

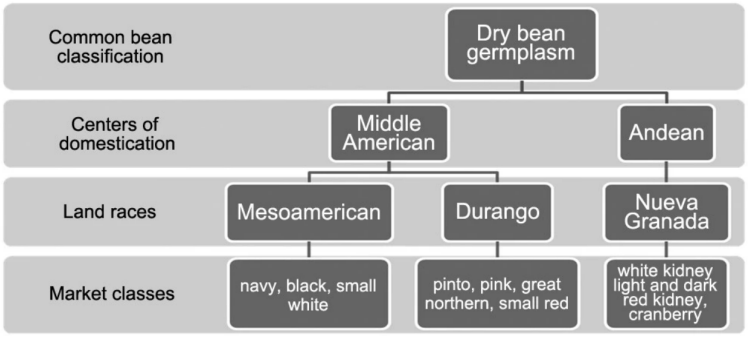


Figure 2.4: "Phylogenetic relationships between center of domestication and races" [95]

The treatments utilized in the experiment are 300 unique bean genotypes (or "bean lines") from the Middle America diversity panel. Within Figure 2.4, there is a breakdown of classifications and areas of domestication of different bean lines. The researchers are focusing on the Mesoamerican and Durango landrace bean lines, which are derived from the Middle American diversity panel. In the final data analysis

which will be discussed in Chapter 3, 297 different bean lines were included in the final data sets, 197 of the Durango bean lines and 100 of the MesoAmerican bean lines. These are also across multiple types of market classes, which are listed in Figure 2.4 as well. There were limited amounts of each bean genotype that could be created in powder for adding to the wells, so before design procedures were finalized, the proposed number of replications for each bean line was discussed with researchers. After finding out how many replicates per bean line were possible to test in total, the rest of the experimental steps were integrated into the design.

Researchers specifically wanted assistance reducing the effects of location between and within plates, so particular attention was paid to how the powders were assigned across plates. First, the 300 bean lines are ground into powder and placed into the well via a robotic lab machine. At this step, 12 plates can be organized as shown in Figure 2.5, with three rows by four columns of plates. With new software for the robot, it is possible to place any particular bean line into any specific well. Previously there was no ability to easily randomize the bean samples into the wells, as the placements were all done by hand.

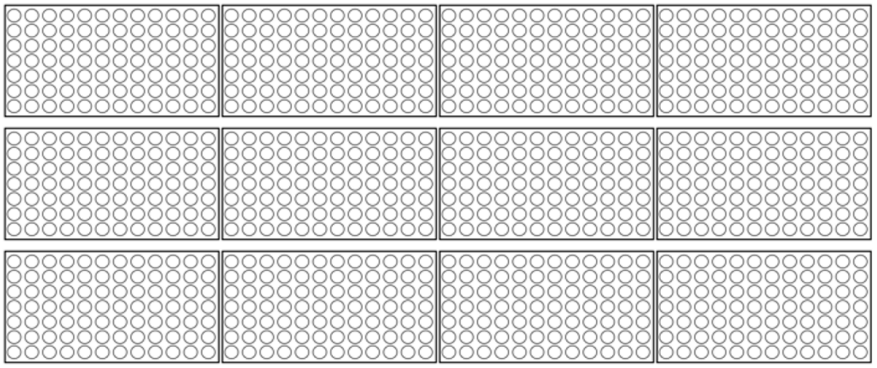


Figure 2.5: Plate Organization for Powder Dosing

Then plates go through digestion and dialysis steps, where four plates can be run together at a time. After digestion and dialysis, the samples proceed to the

fermentation step, when the microbiome samples from the test subjects are added to the wells. Twelve total plates can be run together at one time in the "swing" machine, which applies heat to the samples to simulate the human microbiome. The organization of the plates differed within this machine because it is tedious to get the robotic mechanism that distributed the samples to use that 3 x 4 plate configuration from the powder dosing step above. For the machine to run most effectively, the design for this step has the 12 plates in a new configuration, where numbered plates are placed next to each other in a clockwise circle pattern. After the samples have fermented, the genome sequencing takes place (four plates at a time), and the bacterial abundance data is outputted from each well. At this step, the experimental processes end, and the analysis can begin (which will be discussed in more detail in Chapter 3). After the responses are collected from this first run, another selection of 12 plates can begin through the dialysis steps. The experiment took place over several months in 2019 and 2020, as each experiment took about three to four months to run and be sequenced.

Initial consultation about the design procedure demonstrated concerns with the possible spatial/location effects overall, but specifically within the fermentation step. Across the machine, there could be differences in the bacterial outcome in the wells based on distance from the heating element or due to other well-to-well contamination. These issues are similar to those that were demonstrated in the literature. Now that there is a mechanism within the UNL lab to more easily place the bean samples into wells, more specific designs are possible that could assist in accounting for and measuring location effects. Previously, researchers would simply add the replications all next to each other in the wells with no randomization scheme of the treatments or the quality control wells across the plates. Realizing this could be improved, researchers and statisticians could work together to create better ways to organize

samples. In addition, there were other restrictions on how many wells were available through one run of 12 plates. It is necessary to have six wells for quality controls and a well for a "No Treatment" control within each plate. The quality control wells ensure that the biological processes are working correctly at each step. So, in the powder dosing stage, within a group of four plates, there are a possible $96 * 4 = 384$ wells to fill total, but at least seven must be taken out per plate for quality controls. This leaves a total of $(96 - 7) * 4 = 356$ wells to fill. It is important to replicate each bean line within the available wells, but it is not possible to have all bean lines replicated more than once. Partially replicated designs appear helpful at this point because selected proportions of treatments can be replicated different numbers of times. Researchers suggested it is ordinarily standard to ensure that bean lines are replicated in triplicate across a set of 12 plates in one run. This minimum is used as a baseline for the new design as well.

Moving forward, additional descriptions and terminology should be outlined to label the different experimental and observational units. As previously noted, twelve plates are utilized within one "run" of the experiment. Then, these plates are arranged within the dosing stage in three rows of four plates, as in Figure 2.5. Each four-plate "row" will be referred to as a "replication." The term "replication" is helpful because sets of four plates are the level at which some treatments are replicated. This terminology is beneficial in explaining the randomization process used across the experiment.

One main difficulty in selecting an augmented p-rep design is how to perform the augmentation, and nothing in the literature studied so far indicates any existing standards explaining the best way to conduct this type of augmentation.[152] So, this study was designed as a simple, direct, starting place for how to integrate more intricate experimental designs into these types of lab experiments. Based on the available

resources and the useful attributes of augmented p-rep and α - row-column designs, a complete design was created using a combination of both. The final design created is a combination of an incomplete block design with all 300 bean lines appearing across each replication within a run; Then, this design is augmented with a resolvable row-column α - design. The design contains 12 additionally replicated check treatments within each plate to account for location effects. The word "check" will indicate which lines are replicated in specific patterns across plates more often. Both the design and randomization schemes will be explained in more detail, discussing both the reasoning behind design choices and different options future studies can utilize depending on the researcher's overall goals.

First, for simplicity, the randomization begins with the incomplete block design (IBD) of 300 bean line treatments remains straightforward, as it is the most straightforward part of the design. Plates can only hold up to 89 total samples with the available resources, after accounting for quality control wells. So, to have all 300 lines replicated once, four plates will serve as a complete replicate of all 300 bean lines. However, this set of four plates with 300 lines leaves 56 empty wells across rep (89 open wells*4 plates)-300 treatments=56 empty wells). Another design can now be utilized to augment the typical IBD design to look for location effects and take advantage of these 56 remaining wells. The important part of the augmentation is to distribute the treatments appropriately across the available space to account for location effects. Different well locations are susceptible to different environmental factors throughout the experimental procedures, so these extra wells should be spread in such a way that we account for all these environments equally. For simplicity, the wells containing the extra replications of the same treatment in the augmented design will be referred to as check wells. These check wells act as replications of additional bean line treatments within the (at most) 56 open wells. If all 56 available wells are filled,

14 wells per plate would be included in the augmented design embedded within the IBD. The following design was chosen for the layout of the check wells across plates.

The numbers within the cells (Figure 2.6) are labels for the well numbers across all 96 wells in the plate, and the bold and highlighted wells are where the replicated check lines would appear.

1	**2**	3	4	**5**	6	7	**8**	9	10	**11**	12
13	14	15	16	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32	33	34	35	36
37	**38**	39	40	**41**	42	43	**44**	45	46	**47**	48
49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72
73	**74**	75	76	**77**	78	79	**80**	81	82	**83**	84
85	86	87	88	89	90	91	92	93	94	95	96

Figure 2.6: Design of Check Wells in a Plate

This setup will provide a spread of replicated wells across the whole plate and give an idea about edge and center effects while also being an easy way for replication of 12 additional lines. Two wells per plate are still empty (only filling up 48 of the 56 available wells across a replication), but researchers agreed that was not too much of an issue. The same 12 bean lines were replicated as checks across each four-plate replication. A different, independent set of 12 check lines are chosen for each of the four plate replications across the experiment. Thus, 144 bean lines (12 different check lines across 12 replications) are replicated more often. With the IBD, plates serve as incomplete blocks in a replication, and a line would be replicated in 12 different wells total: three times within a run of the experiment and across four different runs (where a run consists of 12 well plates). So, for example, Figure 2.7 displays the layout of plates during the powder dosing step, where crossed-out cells correspond to sets of check wells across one experimental run.

Crossed-out set of wells in Figure 2.7 correspond to the independent sets of 12 check lines. Since this is for one run, there are three sets of 12 check lines, 36 lines selected out of the total 300, with 12 being assigned to each “design replication” of

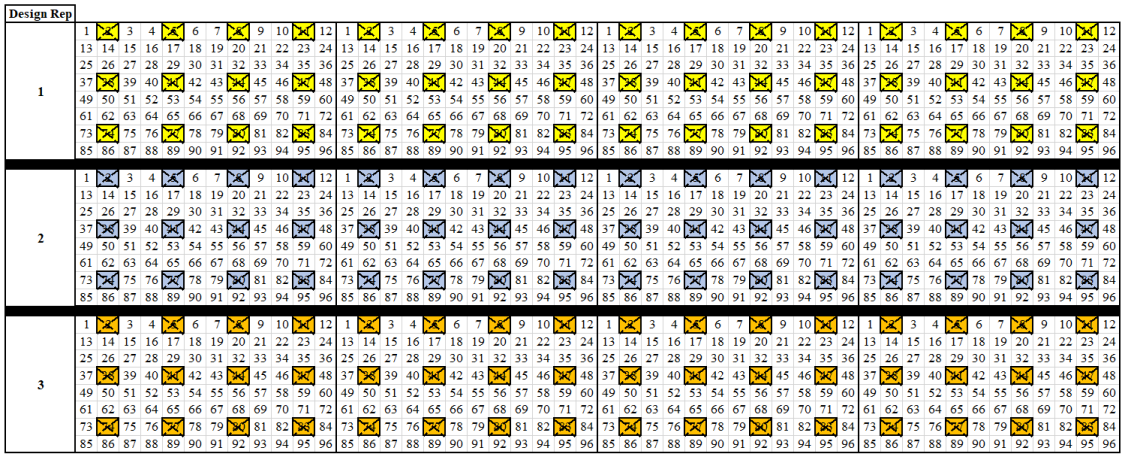


Figure 2.7: Check Line Configuration within the Powder Dosing Stage

four plates. So, across all runs of the experiment, there are 12 distinct sets of colored cells for all 144 additionally replicated lines. Within a replication in Figure 2.7, there are 1152 total wells 96*12=1152. Colored wells represent the extra replications of the 36 different check lines for this run. Out of the remaining non-marked cells, 900 wells (300 wells across each of the three replications) correspond to the IBD design of the 300 bean lines for each replication. Finally, the remaining 108 wells correspond to all the quality control, no treatment control wells, and two missing wells per plate due to design restrictions. In the next section, it is discussed how check lines were randomly selected, how checks were assigned to wells, and how the rest of the lines were assigned to the remaining wells. In 2006, Cullis et al. indicated that issues regarding distributing the test (check) lines across plots is still a topic of current research, so our choices remain fairly simple and individualized to this research. [27]

In addition to the above, each subject's microbiome was randomized to specific runs of the experiment. A combined IBD design was hypothesized, where all subjects samples would appear in each run. However, this was not possible because of the time between the experimental test steps and the timing effect on the microbiome samples. Dr. Benson explained that there is at least a month between each run of

the fermentation step, and while the fecal microbiome samples are kept at a low and consistent temperature, there can be changes in the samples over that time. For this reason, they wanted to ensure that each subjects' microbiome was run through the fermentation step at one time.

Lastly, he stated that the association mapping at the end of the experiment would be conducted by subject, assuming that any effects confounded with the subject would influence all wells equally and all bean lines comparisons would be valid. For this reason, the experiment only randomizes the order that the subjects are assigned to a run. So, there will still be confounding of time and the order in which the subjects were run. Ideally, all subjects would be tested simultaneously, thus eliminating bias. However, this could not be accomplished in this experiment due to time constraints. Processing all subjects in one run would require storage of subjects' microbiome, which is likely to increase uncontrolled variation. Other literature has shown this inconsistency in samples over time, no matter how good the storage methods may seem. Mallick studied how temperature directly affected variability of the subject's samples, stating that "microbiome samples of any type thus benefit from storage in a stabilization buffer, preferably with immediate homogenization" [91]. So, R was utilized to randomize the order in which a subject's microbiome samples would be used within the experiment.

2.4 IBD and α Design Randomizations

To randomize treatments and other miscellaneous well types (quality control, missing, and NTC), there needs to be a way to connect variables to wells. Three different randomizations were produced and combined to complete a full picture of the experimental design. Those three randomizations are 1) the IBD randomization

of all 300 lines to wells in four-plate replication sets within each run, 2) random selection of 12 independent sets of 12 check lines from the 300, along with randomly assigning each set to a resolvable row-column α - design within a four-plate replication and the 3) random assignment of the quality control, no treatment controls (NTC) and missing wells to wells within the experiment. All wells in each run are labeled with a number between 1 and 1152, as labelled within Figure A.1 in Appendix A. These labels are also used to identify wells for researchers and the machinery that distributes powder into wells.

We started the randomization by assigning the miscellaneous well types (quality control, missing, and NTC) to the same wells within every plate, keeping consistency across all runs. In R, nine random values between 1 and 96 were selected to represent which wells the miscellaneous controls would appear in, and then randomly assigned to the name of one of the types. The randomization produced the design of each plate in Figure 2.8, where: the yellow highlighted wells represent the check lines from the α – design, the blue wells are the six quality control wells for each plate, the orange wells represent the one NTC well, and the two red wells correspond to the "missing" or empty wells.

1		3	4		6	7		9	10		12
QC	14	15	NTC	17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	QC	33	34	35	36
37		39	40		42	43		45	46		48
49	50	51	52	53	54	55	56	EMPTY	58	59	QC
61	62	63	64	65	66	67	68	69	QC	71	72
73		75	76		78	79		QC	82		84
85	86	87	QC	89	90	EMPTY	92	93	94	95	96

Figure 2.8: Check, Missing and Quality Control Well Randomization

The next stage is randomizing the 300 bean lines to the uncolored wells (from Figure A.1 in Appendix A) across plates in each replication within a one run. In the GENDEX IBD module, there are options for number of treatments, number of replications, how many plots exist per block, and finally, how many overall blocks

are available. To get the entire experiment design randomization, the following information was put into the blanks. Across four runs, 300 treatments are assigned to 48 blocks (12 plates across each run). With each block, there are 75 "plots" (wells) to be filled, given the 96 total and taking away the control and empty wells. Lastly, there are 12 replications of each treatment throughout the entire experiment, as each treatment is replicated once over each set of four plates. With this information, GENDEX outputs 12 subsets of values (values 0-299, for each treatment) for each of the 12 four-plate reps. Since GENDEX numbers the treatments with values between 0 and 299, the specific bean line names were assigned to a corresponding number. This procedure is so that the researchers will know which bean line to place in which well associated with the final design randomization.

The final stage is creating the row-column design for the check lines within each plate. Given Researchers had no specifications about any particular bean lines that should be replicated more often, so bean lines were randomly selected to become checks. Twelve different distinct sets of checks were chosen (without replacement) from all 300 lines using the SAMPLE function within base R. [118] Check lines are also associated with bean line names, which assist in tying the R selections to the GENDEX randomizations.Then, the GENDEX randomizations for the check wells takes place in a couple different steps. First, the ALPHA module creates an α generating design. This module needs the following variables specified: plots per block, blocks per replication, number of replications, and number of treatments. The check line randomizations occur across four plates, so 12 treatments (checks) are replicated four times. Within each plate, we view the blocks and plots like rows and columns, where blocks are the three rows where check lines appear, and then the columns correspond to four plots per block (four columns per row). After inputting this information, GENDEX outputs an α - design randomizing the 12 checks in four sections,

with subgroups corresponding to the three rows and four columns within the plate.

Next, this α-design gets placed within a *.txt* file to be put back in the GENDEX RRCD module. The RRCD will take the α-design generating array and create a resolvable row-column design. The new design is in a similar form, with four groupings corresponding to each plate and segments corresponding to each row and column within the plates. Using the initial α - design generating array, this procedures was repeated 12 times for each of the 12 replications of four plates. For all output from the RRCD module, the lines are formatted with a value 0-11. These values are connected with the checks selected from R. For example, line number 136 was selected during the R sample step to be a check for the first run of the experiment. For the full randomization, this check line will be assigned to a value between 0 and 11 within the first output set from GENDEX for the α row-column design.

After completing all three randomizations, information was contained within three separate Excel files. All contained two main columns: one indicating the bean line name (or control group label) for the specific well where the bean line powder should be placed, with the corresponding well value between 1 and 1152 (shown in Figure A.1 from Appendix A). There are also indicator columns that label each observations run, replication, and plate value. The *rbind()* function from base R is used to combine the three separate randomizations. This final *.csv* file was provided to the researchers, and that machinery used this information to distributed bean line powder into all microplate wells.

2.5 Efficiency Comparisons

Chapter 3 provides an overview of the analysis methods and ANOVA tables utilized to initiate the data analysis of our experiment. With the obtained data, we

examined several efficiency characteristics by comparing the fitting of the analysis with terms for design rep, plate within the design reps, and the treatments of bean lines. All effects are considered fixed, and we compared two designs by utilizing the specified efficiency value. The first design incorporates both terms for design replication and the plate within the design replicate, while the second design includes only the effects of the replicates. This comparison allows us to determine the additional information accounted for by our incomplete block design structure, which spans the four plates in each replicate.

The efficiency value in Equation 2.1 will be used to determine if it was worth the effort to arrange the bean lines within the four plates in each design rep in an incomplete block format. The equation utilizes the mean square errors for the blocking effect, MS_b, and its associated degrees of freedom df_b. These values are pulled from our plate within design rep term, which is where we assigned bean lines in an incomplete block design structure. Terms are also included for the overall error, with the MSE and its associated degrees of freedom df_e. The ratio in Equation 2.1 will outline a percentage for

$$\text{Relative Efficiency} \simeq \frac{df_b * MS_b + df_e * MS_E}{df_e * MS_E} \tag{2.1}$$

In Table 2.1, there are calculated values of the efficiencies across 12 different cases. Analysis was conducted for each subject, because our randomizations change slightly for each subject, but overall in the same format. In addition, we compared difference cases within each subject. The first three cases within each subject select one of the overall taxa abundances used as responses, and focuses on the ANOVA tables and mean square values from three of the most abundant taxa. Information about how the data was treated and how the MANOVA analyses were conducted

can be found in subsection 3.5.3. In summary, the error values were pulled from the individual ANOVA tables within a MANOVA model fitting the overall 17, 15, and 9 taxa selected for each subject (respectively). The values in the PC1 rows of the table were obtained using an ANOVA model fitted to the first principal component analysis (PCA) score. To compute the PCA score, we applied PCA to the data of the overall most abundant taxa (i.e., the 17, 15, and 9 taxa for each subject) and calculated the first score as a weighted response value across the principal component weights of the selected taxa. This version of the ANOVA across the abundance data offers a comprehensive multivariate perspective on the impact of variability across all selected taxa, as well as the effectiveness of our experimental design in accounting for any additional variability across plates within replicates.

Subject	**Taxa**	**Efficiency**	**Percent**
1	**Prevotella9**	1.116	11.562
1	**Succininvibrio**	1.066	6.603
1	**Dialister**	1.049	4.907
1	**PC1**	1.15	14.976
2	**Bacteroides**	1.268	26.772
2	**Phascolarctobacterium**	1.292	29.202
2	**Faecalibacterium**	1.33	33.046
2	**PC1**	1.353	35.323
3	**Prevotella9**	1.068	6.832
3	**Megasphaera**	1.124	12.426
3	**Acidaminococcus**	1.344	34.4
3	**PC1**	1.034	3.408
		Average	**18.29**

Table 2.1: Relative Efficiency values across MANOVA models and Linear model on the first PC score across the taxa used within the MANOVA analysis.

Incorporating the term plate(designrep) to account for incomplete block variability increased efficiency compared to considering only the three replicates of the complete set of bean lines. On average, the relative efficiency comparing the designs was about 18.29%. Upon examining the values, it was evident that the efficiencies were significantly influenced by the specific subject's information analyzed and the

particular response within that subject. In some cases, the efficiency was as high as 35% or as low as 3.4%. This highlights the large variability in the taxonomic abundances across people's gut microbiomes and between different bacteria taxa within the same subject's gut. Introducing the additional incomplete block design within each replicate appears to be worth the effort overall. In some cases, the IBD structure was up to 35% more efficient than a design with just the replicates. Although the efficiency gain varies depending on the subjects and taxa, in this type of research, experiments are commonly conducted over multiple subjects and aim to identify trends across many bacterial taxa. Therefore, it's desirable to use a design that is most efficient across all cases.

2.6 Summary and Future Work

This Chapter aimed to develop a new experimental design for NGS 96 well-plate experiments dealing with the human microbiome. To account for the variability across plates and due to other constraints, this new method of design and random assignment of bean lines to wells was created and implemented in coordination with collaborators from the Food Science department. An incomplete block design, overlaid with a partially replicated (p-rep) α lattice design was used to randomly assign bean line treatments to wells across sets of 96-well micro-plates. Overall, we were able to successfully implement the design at this point in time because new technological advancements with lab equipment made it easier to distribute the bean line powders into specific wells. The final experiment was completed in 2019 and 2020, where data was analyzed as the individual subject data was collected. Chapters 3 will describe methods used to analyze and interpret the data from all three subjects and investigate how the design choices affected the final data set. The final chapter will

directly extend this chapter's ideas by outlining parameters to consider when creating the "most optimal" experimental design for evaluating the relationships between the bean line genetics and the community taxa structure.

To assess the efficiencies of the designs from the analysis we conducted that we describe in Chapter 3, and we compared how efficient the incomplete block design structure across plates was in comparison to only accounting for replications across the three design replicates (which are the batches of four plates). Relative efficiencies were calculated across each of the three subjects, and four differing response variables, including three of the most abundant taxa and one principal component score across all of the most abundant taxa included in the original analyses. We found that there was a wide variety of relative efficiency depending on subject and response, were at most the design with accounting for plates was 35% more efficient, or as little as only 3.4 % more efficient. Due to the significant variability of taxa between and within subjects, incorporating replicate and incomplete block designs can enhance efficiency in some cases. Because researchers can never know which taxa will be found in a subject's gut or how the variability changes between subjects, it would be best to make sure we always are able to account for this extra variability in the design. In the future, more efficiencies could be calculated to further evaluate the relationship between the row and column design structure of the check lines. However, this was not completed here because the work focusing on the design in Chapters 3 and 4 only focus on including the design replicate and plate effects. Additionally, we knew that Dr. Van Haute would account for a row and column structure within her analyses run on the data for the genome-wide association tests, so it was less interesting than the particular structure between the replication and incomplete block design randomizations.

Structured experimental design is relatively new to this particular field of mi-

crobiology and especially new to those that work with this data in the Food Science lab at UNL. This work directly contributes to researchers who conduct experiments measuring bacterial abundances from simulated microbiome samples using 96-well plates. Currently, design parameters were dictated given this specific project but are easily adjustable to fit many other projects. All three primary objectives of this dissertation aim to assist researchers in creating the best and simplest design necessary to both assigning treatments to wells while getting the most informative statistical information from all samples. Alongside figuring out the best design strategy, it is important to ensure that the design is easy to utilize and put into practice by researchers. The literature provided background from the microbiological sciences as to why it is necessary to have planned and thoughtful experimental designs in order to have "accurate and meaningful results" [80]. While not the most intricate or complex structure, this current design pleased Dr. Benson and Dr. Van Haute and now acts as a basis for the implementation of further research looking into other ways to improve microbiome lab experiments.

Creating the current experiment was quite tedious, requiring multiple steps across multiple computer programs. In the future, the creation process could be reworked and combined so that all three separate randomizations occur at once and output an easily readable file with the well placements for a specific number of treatments and plates. The randomization could occur in an app like the PlateDesigner app, which puts specific parameters into a web-based platform. This method could provide researchers with a quick and easy way to create randomizations associated with specific types of designs. [145] Integrating the actual experimental design structure with a randomization procedure together in one program would vastly improve the ability to implement more intricate experimental design structures. A goal of future research could use specific programming techniques to create a web page or

ShinyApp through R to output a final randomization for microbiologists. This app could return a complete randomization with graphics demonstrating the placements of treatments and well numbers and could be flexible enough to assist researchers in selecting the optimal number of plates or treatments, given the amount of variability in the responses.

Other future work could be tied into this last concept with the integration of optimality into the design procedure to explore more about the plate-to-plate variability, and if it would be more worthwhile for researchers to consider different blocking techniques across plates or across rows and columns within a plate. Other types of designs differing from α p-rep designs, with an incomplete block component, could also be investigated within the search. The optimal design will be more directly addressed in Chapter 4, and how the work may be able to tie into the procedures to help microbiologists have the ability to create their own designs. Overall, this design serves as a starting point to the broader discussion about how it will be best to integrate design into microbiome experiments in 96-well plates.

Chapter 3

Data Analysis

3.1 Introduction

Microbiology employs various approaches to analyze the relationship between the genetic material of food consumed and the microbial community in different areas of the human body. Many methods focus on the associations between phenotypic outcomes and the genetic information from food. There is a wealth of literature that summarizes current and novel techniques to analyze large genomic data sets. This research aims to identify the genetic basis for the changes in microbial phenotypes observed in the gut microbiome resulting from the consumption of different bean lines. Specifically, this chapter will highlight analysis methods for the inclusion of multivariate taxonomic information before genome-wide analyses (GWAS).

Much of the literature focuses on improving the GWAS model's ability to handle all facets of microbiome data. To address current challenges, improvements suggested include adjusting models to handle issues such as large data sets, non-normal data, multivariate responses, unbalanced data, and adding in effects for additional aspects of experimental design [144, 154]. This list could go on, but many statistical tricks and standards can be restructured to help deal with these issues. Properly interpreting microbiome data sets involves not only understanding the statistical results but also

considering their practical implications. In this chapter, we describe an analysis methodology that addresses data structure and generates practical outcomes that are easily interpretable by microbiologists in the field.

Researchers frequently use Genome-Wide association analyses (GWAS) to analyze sequencing data from microbiome experiments to explain the relationship between genomic and phenotypic information. Analyses directly evaluate the associations between the bacterial taxa found in each sample and various fixed and random effects in a mixed model. Effects may include variables such as batch, plate, rows/-columns within a plate, and, most importantly, information variables containing the SNP data of interest. Models of interest can feature many different structures, consisting of frequentist or Bayesian frameworks, linear or generalized mixed models, and varying methods for including the variables of interest in the model (univariate vs. multivariate responses and sets of explanatory variables). Univariate methods have previously been more common with GWAS and microbiome analyses, but multivariate methods have been more of interest in recent years and will be particularly better overall for this work [164, 70, 108].

This chapter reviews analysis methods for similarly designed experiments and highlights the importance of using multivariate analyses to evaluate community relationships. Previous research has relied on standard univariate analysis methods, but recent articles have demonstrated the benefits of integrating multivariate approaches into genome-wide association analyses for microbiome research. In this chapter, we describe the background and current methods used for genetic data analysis in this field and explain how the taxonomic abundance from our experiment was analyzed. We provide a detailed account of the proposed and actual analysis methodology and demonstrate how the multivariate response variables created for the GWAS show similar and additional information from the univariate methods. Our discussion and

summary highlight the importance of these methods in spotlighting how the community structure of the gut microbiome is affected by the genetics of P. vulgaris. Our work aims to demonstrate how this methodology can contribute to the field by employing multivariate analysis techniques with the initial data before running corresponding GWAS models, which can help emphasize the importance of the relationships between taxonomic abundances.

3.2 Outline of Analysis Goals

In studying the gut, researchers have shown that the composition of the bacterial taxa from person to person stays relatively steady over time, but multiple health conditions are related to changes in a stable community of bacteria within a person [37]. Eetemadi and coauthors also note that work must still be done when describing these conditions affecting the gut to "establish direct links between these conditions and the composition of microbial communities in the gut [37]. In the larger body of work, we hope to contribute to evaluating the relationship between a diet containing *P. vulgaris* varieties and the gut microbiota's resulting complete bacterial makeup. Collectively, this new methodology for quantifying and genetic analysis of human gut Microbiome-Active Traits (MATs) provides a new approach for mapping traits in crop plants associated with human wellness and predisposition to particular diseases.

Our study aimed to investigate how the genetic features of bean lines impact the human gut microbiome. Our approach involved identifying appropriate statistical techniques to analyze the association between microbiome phenotypic information and specific bean races' genetic diversity and population structure. To select the most suitable technique, we considered the data collection process (as outlined in chapters one and two) from a Food Science perspective. This process began by

collecting specific bean line cultivars of interest from the Mesoamerican, and Durango races, representing six major market classes [59]. Researchers created a new method called the "Automated in vitro Microbiome Screen (AiMS)" to phenotype microbiome responses across many different genotypes/cultivars of the P. vulgaris lines. This method was necessary due to limited resources for conducting fecal microbiome fermentations [59]. The AiMS method generated quantitative phenotypic measurements of the abundances of different taxonomic and metabolic features of the microbiomes. These measurements were critical for use in statistical analyses to quantify the interactions of the *P. vulgaris* components with human gut microbiomes [59]. At the core, the AiMS method creates microbial abundance values for analysis to relate beans consumption to output in the gut microbiome, which is useful for highlighting relationships between diet and the community structure of bacteria living in the gut.

Researchers run genome-wide association studies for individual taxa as phenotypes for data analysis to identify Major Effect Loci (MEL). MELs are areas of genetic variation in the P. vulgaris genome within different chromosomes with significant associations with phenotypic information (quantified as relative bacterial abundances) using the AiMS method. This new collaborative project aimed to develop a technique that could be incorporated as a novel approach in the GWAS models to account for the abundance information of phenotypes across a multivariate range of abundance values. Thus, the results would represent a "community" effect of bean lines across multiple taxa. Looking at groupings of taxa would assist researchers in identifying MELs that contain a gene or genes strongly affecting the bacterial communities in the gut, potentially highlighting how the food we digest changes the makeup of the gut microbiome.

The second research objective from section 1.3 states that in this chapter, we

will use multivariate analysis methods to analyze data outputted from the experiment designed from objective one. This first objective was to develop a new experimental design for next-generation sequencing 96 well-plate lab experiments, specifically dealing with the human microbiome. Following the data collection process outlined in chapter two, we aimed to investigate whether the natural decomposition of various covariance and correlation matrices could facilitate our understanding of how genetic variation across a diverse set of common bean lines is associated with the composition of microbial communities in the gut of three subjects. To achieve this goal, we collaborated with microbiologists to develop a comprehensive analysis methodology to enhance their understanding of the data.

To do this, we use different variability structures from MANOVA models analyzing the relationship between the taxonomic abundances and the bean lines of interest, accounting for the experimental design effects. We then use the decompositions of the variations to create weights for the final calculation of a variable representing a multivariate community effect to use in Dr. Van Haute's GWAS models. Then try and describe any patterns where the community taxa effects are significantly associated with areas on the *P. Vulgaris* bean genome. Through discussing results with the domain experts, we can highlight commonalities or unique attributes where our multivariate traits are significant in relation to all the univariate taxonomic traits also used in the GWAS models. Our primary objective was not to create a standalone method for the genetic-wide analysis that links the bean genetic structure to the taxonomic abundances. Instead, we aimed to work alongside the researchers to develop a new multivariate value that could be used with their univariate abundances for standard GWAS models.

3.3 Literature Review

3.3.1 Background

This review is divided into sections discussing techniques for analyzing microbiological data and their relevance to creating statistically sound and innovative methods for interpreting the data collected in Chapter 2. Statistical context is provided for geneticists' commonly used approaches, and we demonstrate how our analysis is similar to current research while utilizing other factors from different domains (such as animal and plant breeding). One of these common approaches for analyzing microbiome relative abundance data is based on linear mixed models, which incorporate random effects based on the experimental design. Mixed models are a straightforward way to integrate experimental design factors such as plate, batch, row/column, different runs, etc., into the model. For the response variables, in practice, the focus is often placed on univariate genome-wide association tests, using phenotypes within models one at a time [40, 106, 165]. While univariate methods are more common, new research focuses on multivariate analyses because of the strong relationships between phenotypic variables of interest. Multivariate mixed models and statistical dimension reduction methods can assist in analyzing complex microbial datasets and interpreting the gut community's genetic composition. While the literature does contain previously utilized multivariate techniques, there remains a "multitude of possible statistical choices makes it a daunting task for an investigator not experienced with these tools to pick a good technique to use". [109]

3.3.2 Analysis of Taxonomic Abundance in Microbiome Experiments

In light of the collaborative goals specified in section 3.2, it is crucial to explore prior efforts in addressing comparable research questions and identify potential adap-

tations or alternative approaches. Having collected relative abundance data from various subjects through our designed experiment, it is important to determine the key considerations when dealing with similar datasets. These outputs from microbiome experiments are most often multivariate, whether or not it is analyzed as such. Thus multivariate tools available to scientists need to grow as the data collection methods in the field are also increasing [124]. In combination, issues can arise from the data non-normality (sometimes measured in counts or as relative bacterial abundance), sparsity of the responses from too many zeroes and undersampling, and possible false positive outcomes, which could lead to bias [91]. Mallick et al. provide information about methods used to deal with differing types of microbiome studies, including clustering analyses (and unsupervised learning techniques), analyses that could include covariates such as nonparametric MANOVAs (PERMANOVA), and newer methods such "edgeR, DESeq2, metagenomeSeq, limma-voom, and MaAsLin" which can include multivariate associations. These methods are all varieties of linear models or nonparametric methodologies that can analyze microbial abundances. For those looking to analyze a similar style data set, a table in the paper provides a summary of the types of responses addressed by each model, the statistical distribution employed, and whether the model considers multivariate responses, or random effects, among other factors.

In the Mallick paper, there is also a discussion about how even though many researchers conclude that published methods in the domain literature are multivariate, many of the "analytical tools in microbiome epidemiology are essentially univariate" [91]. This is worrisome because there are statistical issues with treating multivariate data in a univariate form. One of the most significant concerns is that information about the relationship and correlation between the responses gets lost in this process. Other authors notice similar issues in finding multivariate models to fit the

data. La Rosa et al. discuss that developments in analyzing metagenomic data are developing slower than processes for collecting data [124]. Specifically, methods based on "exploratory cluster analysis, bootstrap or resampling methods, and application of univariate and nonparametric statistics" are being proposed, but many of them require reduction of the data, which in turn ignore the correlations of the responses and the multivariate structure of the data [124]. So even though many researchers use these common methods, there is room for improvement with how the approaches deal with the multivariate data and a need for models accounting covariates (such as effects introduced from more intricate experimental designs). The authors propose a parametric multivariate approach to the analysis based on the Dirichlet-Multinomial distribution. Although this approach improves the development of advanced multivariate models that rely on the biological characteristics of the data, it fails to consider any supplementary covariates, which would allow for the consideration of potential effects resulting from experimental design.

Separate from the more generic linear models, machine learning techniques are used for visualization and data clustering to quantify the relationships between the genotypes and their bacterial responses. Standard methods utilized are clustering techniques as well as principal coordinates analysis or principal component analysis [80]. Based on the nature of these techniques, it is possible to utilize them for variable reduction to reduce the size of either set of independent or dependent variables. Principal component analysis, for example, can be used to reduce the size of a data set into a few components made up of linear combinations of the larger data set, "called principal components, along which the variation in the data is maximal: in order to see in what ways the samples can be grouped together" [120]. Similar methods, such as different types of canonical correlation analyses and partial least squares regression methods, can also be used as dimension reduction methods sets of large data [94].

Statistical hypotheses of interest can also guide the particular route for selecting the type of analysis to look for relationships between SNPs and phenotypes. Many summary and review papers exist, highlighting common methods and issues that need to be resolved. Summary statistics can be calculated to summarize microbial taxa, including alpha, beta, Shannon, Inverse Simpson, Faith's diversity indices, and along with the raw abundance values, can be analyzed with t-tests or non-parametric type Kruskal Wallis or Wilcoxon rank sum tests when comparing microbiomes across groups [154, 51]. Galloway-Pena and Hanson highlight different ways to characterize (as in the previous diversity indices) and analyze types of genomic data from microbiome studies, including discriminant analysis, PERMANOVA, Regression methods (sparse regression, Dirichlet-multinomial regression, sparse regression, generalized boosted linear models), correlation network models, and a variety of different machine learning models (such as random forest or classification and regression trees) [51].

Although the existing resources are useful in the domain, there is a need for novel statistical techniques that concentrate on hierarchical, spatial, and temporal types of data and designs, and there remains a lack "of systematic studies aimed at the evaluation of multiple-covariates, [and repeated-measures] methods for microbiome epidemiology, with no clear consensus to date" [91]. New methodologies could be useful for researchers seeking to develop methods that more accurately address spatial or statistical random design variability. Secondarily, much focus falls directly on methods for genome-wide association analyses and recent research has been done investigating more multivariate methods, including "new robust and powerful statistical methods for testing association between a microbial community/clade of multiple taxa and an outcome of interest" [8]. In general, we acknowledge the approaches used by statisticians and microbiologists when dealing with abundance data. However, our analytical objectives center on an intermediate phase that follows data collection but

precedes GWAS modeling, with the aim of producing a novel pseudo-multivariate variable that can be employed in the traditional GWAS model.

3.3.3 GWAS and Genetic Association Methodologies

Our analysis objectives revolve around how the overall bacterial makeup in the gut, collected with the AiMS methodology, is affected by the genetics and SNP information from the bean lines. Specifically, we have focused on a multivariate extension completed before the GWAS. In looking for an analysis that could be utilized to emphasize community structure in microbial outputs, other fields of literature shed light on helpful modeling techniques and what information from the data will be essential to consider. Given that this research relies on input and output data generated by GWAS modeling, it is advantageous to provide an overview of the models' capabilities and their usefulness in this and other fields.

GWAS research is not only valuable for human health studies but also has applications in a variety of fields. With these models, the aim is to find relationships between an output of an observable characteristic of an individual (described most often as a "phenotype") to some genetic information from the same individual. While these can be stand-alone analyses, they usually fit under the description of GWAS or Genome-Wide Association analyses, which "test hundreds of thousands of genetic variants across many genomes to find those statistically associated with a specific trait or disease" [148]. This method has become increasingly common for identifying noteworthy genetic variants in plants and animals. Examining these relationships between genotypic and phenotypic traits can enhance our understanding of various crops and types of livestock, potentially improving their yields and characteristics [135, 84]. Tying into our work, the collaborating microbiologists are interested in relating characteristics of the Mesoamerican and Durango bean lines to the changes

in the taxa created by the gut microbiome. Comprehensively, this work spans health and agriculture research fields, making it of interest to various stakeholders.

As previously mentioned, our work will be combined with a GWAS to investigate potential correlations between bean genetics and abundance phenotypes measured by AiMS. We chose to work on an extension of GWAS as it is a widely accepted methodology for exploring the relationship between dietary genetic information and changes in human health, specifically in terms of the impact on the microbiota of the human body. Advances in DNA sequencing technology have also helped to allow this field of research to grow as it continues to become more manageable, cheaper, and faster over time to collect genetic information associated with humans and the human body (along with information about what we consume) [147]. Overall, we can focus on a description of what these methods are used for with regard to this research and why we are not adjusting the particular processes.

Data for microbiome studies consist of genetic (SNP) information with a large amount of chromosomal information collected from plants, animals, people, etc., of interest and compared to a phenotypic outcome of interest, searching for significant associations between the two [114]. Statistical methods can be broken down into three groups to search for the association of interest. A GWAS can be conducted individually on phenotypes (results can be interpreted over all tests), more than one trait can be converted into a "composite score" for the GWAS, or multiple phenotypes can be combined for the GWAS [156]. Researchers also want to "characterize the relationship between microbiome features and biological, genetic, clinical or experimental conditions" or "identify potential biological and environmental factors associated with microbiome composition" [154]. Extensive literature exists about the specific GWAS models because it is so common, but also since researchers are commonly dealing with the issues of complex data, including high-dimensional, zero-inflated, composi-

tional responses, where methods can often lack statistical power and the ability to control for confounding variables [51]. More information about multivariate methods commonly used for these models is presented in the next section.

Looking at GWAS models separately from models in subsection 3.3.2 is helpful because the goal of models for the abundances is to compare the attributes of different microbiomes depending on what types of groups or categories the person (or subject) was exposed to, like a treatment vs. control, etc. [51]. In this current research, the objective is to simply focus on the associations between chromosomal information from the bean lines compared to the outcomes from the microbiome. Microbiome literature provided a good background for knowledge to remember when utilizing taxonomic abundances as responses. Common genome-wide association analyses to study diseases with multiple phenotypes include "multivariate analysis of variance (MANOVA), the principal component analysis (PCA), the generalizing estimating equations (GEE), the trait-based association test involving the extended Simes procedure (TATES)" [156].

Because the abundances have become our traits of interest, GWAS methods of interest instead boil down to what would be used in plant breeding and genetics. Essentially, we could replace the AiMS-generated "Microbiome Active Traits" with any other trait/phenotype of the bean lines, and GWAS would be similar to those completed within agronomy and horticulture. This work is a small extension of Dr. Van Haute's overall work, where she collaborated with UNL agronomy and horticulture department members to provide the necessary background for the GWAS models, bridging the gap between plant and health studies [59]. When discussing the benefits of GWAS, she highlighted that the studies "have developed into a powerful tool for the investigation and identification of candidate genomic regions associated with traits of interest, " and while there have been over 50 publications focusing

on GWAS with the *P. vulgaris* line, none included microbiome "nutrition-oriented" traits in their study [59]. More information about Dr. Van Haute's specific GWAS models can be found in her dissertation and in subsection 3.5.5. Regardless of our possible extension, univariate models would be used as the primary way for Dr. Van Haute to look for general associations between taxa and the bean genome. However, to expand our research on community effects, we can explore multivariate models and consider how to incorporate community effects into phenotypic information for GWAS across various disciplines.

3.3.4 Multivariate Methods and Analysis of Community Relationships

The literature on microbiome research emphasizes the importance of utilizing multivariate methods to analyze taxa communities. Our collaborators aim to analyze more than one taxa at a time, making multivariate methods in linear or generalized linear mixed-method models relevant. Furthermore, summarizing multiple abundance values into a single variable becomes useful when creating a new variable for GWAS to consider community effects. Thus, I extensively reviewed the uses of multivariate and dimension reduction methods in statistical and domain contexts, which informed the proposed analysis methodology outlined in section 3.5.2. Using multivariate methods is necessary to gain a deeper understanding of how different bacteria in the gut interact with each other. In the field of medical research and healthcare, being able to measure the diversity of gut bacteria, along with their distinct mechanisms within the human body and the way they vary from one individual to another, could provide valuable insights for assessing disease risk and customizing treatments [32].

Numerous methods for analyzing bacterial abundances are designed to directly model the relationships between phenotypic output and genetic SNP information. Often, researchers use multivariate linear mixed models to evaluate community structure

in the phenotypes while also having the ability to "control for population stratification in genome-wide association studies"[164]. Linear mixed models, which can be extended to include multivariate LMMs, have become increasingly popular in GWAS due to their unique ability to account for the relatedness among individuals while simultaneously testing for associations between genetic markers and traits [57]. Stephens, in 2013, pointed out that multivariate linear mixed models are frequently used in genetics, but despite their increased power, interpreting these association analyses can be challenging, and not finding significance can keep researchers from being able to answer the question of which phenotypes are associated [142].

In addition, while a linear model like a MANOVA can provide specific outputs regarding fixed effects, these results may not be particularly interesting. Instead, the focus is placed on the extent of variability across response taxa while considering design effects and bean lines of interest. GWAS models employ this thinking by utilizing information from the random effects. Information is extracted from BLUPs (Best Linear Unbiased Predictions) with a variance structure using a kinship matrix between the effects of interest (the bean line genetics in this case), introduced by Henderson[61]. Depending on the method selected (many methods exist both within frequentist and Bayesian frameworks), BLUPS can be calculated for each genotype from the model and are used as a response when testing each genetic SNP marker individually as the fixed effect. Each BLUP test evaluates genetic associations and identifies regions of the genome that display statistically significant associations with specific traits or diseases. The summary plot (Manhattan plot) of all model p-values across the genome reveals peaks of the most significant hits[128]. These areas indicate to researchers which bacteria, or communities of bacteria, are highly related to the genetic makeup of the material of interest.

With the use of SNPs in the GWAS models, the "genotypes are well-defined

biological entities," whereas "phenotypes are defined more subjectively and can relate to numerous biological processes" [106]. Efforts have been made to determine well-defined phenotypes that are associated with specific biological functions for use in GWAS, however, there still lacks a clear and consistent process for defining the phenotypes[106]. However, O'Reilly describes a process for utilizing regression methods to highlight multivariate relationships across phenotypes. They observe that "the association between linear combinations of phenotypes and the genotypes at each SNP" could allow researchers to better identify associations that are not easily interpreted by GWAS models on a single phenotypes [106]. The issue at hand becomes how to statistically model related multiple phenotypes to the specific genotypic information of interest. O'Reilly et al. introduce MultiPhen, which uses ordinal regression to jointly model multiple phenotypes as predictors of SNP genotypes and makes no assumptions about the phenotype distribution, accommodating both binary and continuous measurements. The method employs a likelihood ratio test for model fit to determine evidence of an association between the SNP and the phenotypes, with the usual genome-wide significance level applied. Utilizing linear combinations of phenotypic information is a point of interest moving forward, but this could also be achieved with a variety of different dimension reduction techniques like principal component analysis or canonical discriminant analysis.

Another benefit is that analyzing the data across multiple phenotypic responses "can increase power not only to detect pleiotropic genetic variants but also genetic variants that affect only one of multiple correlated phenotypes" [164]. Pleiotropy can be defined as an outcome where a single locus affects more than one phenotypic trait, and associations detected may not all be biologically meaningful[139, 57]. In 2014, Zhou and Stephens created more efficient and flexible algorithm extensions for GEMMA software (Genome-Wide Efficient mixed Model Association) that are

extensions of multivariate linear mixed models with corresponding likelihood ratio tests for finding significant genotypic effects [164]. In addition, methods such as GEMMAemma, an extension of GEMMA, offer improved computation speed, power, and P-value calibration over existing methods and focus on including more than two phenotypes [164].

In 2013, Stephens wrote about a framework for analyzing more than two related phenotypes. The goal was to "consider the problem of assessing associations between multiple related outcome variables" and create models "based on Bayesian model comparison and averaging for multivariate regressions" [142]. The analyses were created to handle simple multivariate analyses with one genetic variant and "a modest number of phenotypes (e.g., up to 10)"[142]. The analysis utilizes Bayesian Multivariate Regression techniques to evaluate which genotype has significant effects on the groups of phenotypes. Others have investigated utilizing dimension reduction techniques in order to summarize multiple phenotypic data before running genome-wide association studies. Wang et al. discuss dimension reduction as a method to reduce the number of tests conducted, as the number can grow exponentially when interested in more individual traits[150]. The authors refer to several studies and note that for multivariate tests, while the number of tests decreases, there could still be challenges in identifying the precise effect of host genetics on particular microbes [150]. The methods used can include PCA, PCoA (principal coordinate analysis, and MDS (multidimensional scaling), where PCoA is often used more for microbiome data [150, 94]. However, many dimension reduction methods do not consider outside modeling effects. Methods such as PCA could assist in reducing the number of taxa into one value "while preserving as much 'variability' (i.e., statistical information) as possible" such that we are finding a new summary value that is a linear function across information about the original taxonomic abundances, "that successively maximize

variance and that are uncorrelated with each other" [72].

Other research areas focus on using statistical dimension reduction methods to reduce the large amount of data produced in microbiome studies. In research dealing with genetics from microbiome samples and a food source, such as the common bean (P. vulgaris), models evaluate the relationship between multiple phenotypic responses and large sets of related explanatory variables (SNPs). Liu, Barnett, and Lin examine how different PCA models react when using large sets of phenotypic data and when a model uses PCA to reduce the number of SNPs as explanatory variables [87]. Conversely, reduction methods can also be utilized for the phenotypic responses highlighted in the aforementioned papers. In genetic analyses with SNPs and phenotypes such as this, there are differences between what can be gained from summarizing either the explanatory or response variables based on a PCA or canonical discriminant analysis. But, in the literature, not much has been discussed "how the performance of PCA differs in multiple phenotype regression as compared to multiple SNP regression," as well as when more specific details about fine-tuning the method and trying to compare how different unsupervised dimension reduction methods perform [87].

The methods above give a broad overall picture of how multivariate methods offer helpful strategies to answer researchers' questions about the community structure of taxonomic phenotypes. These analyses can also identify relationships between multiple taxa and the SNP genetic information from the bean genome. After acknowledging these benefits with the additional information from the microbiology literature, we can better outline the methods for an initial analysis of taxonomic abundance data from the experiment from Chapter Two. To guide the focus of this specific project, we can identify the most effective analysis approaches and determine what works best at matching (or improving) the detection of significant genetic relationships from the

univariate methods. Working alongside Dr. Benson and Dr. Van Haute, this research aims to assist with their primary analyses, combining taxa abundance data into multivariate traits before the GWAS analyses. The details and knowledge presented from the literature provide beneficial descriptions of how multiple phenotypic traits can be summarized and how we can work to summarize community taxonomic relationships.

3.3.5 Additional Methods Incorporating Experimental Design

A simple first step in methodology would be to consider a multivariate linear model technique to best integrate and test for possible effects of plates, rows, and columns within plates and differences between subjects. There was little discussion about integrating experimental design variability into the analysis across the literature, even though many resources called for further statistical design aspects to be implemented. However, Mallick et al. made several different tools for analyzing abundance data, where some can account for both multi-variable associations and random effects, which could help account for the design effects [91]. Normalization and data transformations are commonly employed in analyzing microbiome abundance data and can be particularly useful in this context. For this reason, it will be worthwhile to investigate techniques created or proposed to combine experimental design methods and analysis techniques within statistical modeling.

In previous studies, researchers have employed zero-inflated non-normal models to analyze relative abundance proportions, incorporating multi-variable associations and random effects. A paper from Chen and Li proposes a model that accounts for repeated measures data for microbiome changes over time with subject-level covariates[22]. While initially designed for fitting a correlation structure to repeated measurements over time, the approach discussed here can be adapted to address other design effects. Furthermore, the structure of repeated measures data is akin to that

of multivariate data, in which multiple responses are recorded across different observational units. As a result, there may be some overlap between models used for longitudinal microbiome sampling and those used for multivariate analyses.

Grantham et al. expand on the mixed effect model methodology for designed experiments. Authors point out that "permutational multivariate analysis of variance (PERMANOVA) with pairwise differences between samples" from McArdle and Anderson in 2001 "is a popular tool to test whether environmental covariates are associated with" differences in the taxonomic responses [56, 93]. However, "PERMANOVA does not yield inferences about how the environment affects individual microbes" [56]. After reviewing the methodology used for multivariate models with covariates, the author proposed their new method, "microbiome mixed model (MIMIX)," a Bayesian-based model using Bayesian factor analysis to "capture complex dependence patterns among microbial taxa" [56]. According to Grantham, the impetus for the design arose when attempting to fit data from a randomized complete block design across multiple locations. Although not identical to the design presented in chapter 2, the principles underlying it could be extended to accommodate the distinctive features of this design. When using a Bayesian model structure, it is important to consider the amount of statistical expertise and background knowledge about expected responses and variability that the researcher should have to interpret the results. Although more complex models may better fit the data, they may not always be the most practical methods for subject matter researchers to use.

3.4 Data Collection, Descriptive Statistics, and Sources of Variation

The collected data were organized, cleaned, and structured for statistical analysis in Excel by Dr. Van Haute. Data sets for each subject contain variables from the design and background descriptive variables with details about the specific bean lines and some subject information. Variables included are listed in Table 3.1, separated between the variables that specifically pertain to the design and extra variables that are more descriptive and do not necessarily need to be used within the analysis. Once the data was received for each subject, the replication, row, and column variables (italicized in Table 3.1 were added to indicate each observation's replicate batch, row, and column value within a plate. The row and column variables were added to evaluate the levels of extra "noise" and variability in bacterial abundances based on well distribution. The replication label corresponds to each set of four plates run through various stages of the experimental protocol together. As discussed in Chapter 2, each replicate was designed as an incomplete block design, with one complete replicate of all 300 bean lines and four additional replications for 12 randomly selected "check" lines. The response variables included were the initial taxonomic abundances at the genus level.

The design was organized so that each subject's microbiome samples were put through the experimental protocol one at a time. Data for all subjects were received over two and a half years. Additionally, due to a lack of information to get the appropriate genetic output, three of the 300 total bean lines were left out of the randomization (BelMiNeb1, BelMiNeb2, and Indeterminate Jamaican Red lines). Lab researchers mentioned that some information could be recovered for these lines in the future, but there will now only be 297 bean lines for all analyses. Statistically, more

Explanatory Variables	
Design Related	**Descriptive**
Subject	Bean Race
Bean Line	Market Class
Check Line (indicator)	Seed Color
Plate	Bean Release Year
Well	Subj. Gender
Replication	Subj. Country Origin
*Row**	Bean Weight
*Column**	
**Within Plate*	

Table 3.1: Variables Within the Subject One Data Set

attention could be paid to the information about these lines since they were not missing at random, but since the lines are missing entirely, it could be viewed like these lines were not in the experiment at all. Also, some individual wells from other lines were "missing at random" due to more typical errors in the lab data collection. These wells could have been missing due to a lack of material to test at the end of the experiment or if the sample did not have any measurable bacterial taxa. One last note is that two of the completely missing lines were designated as check lines in some of the randomizations, each from a separate set of four plates. However, the statistical modeling will make corrections in the analysis to balance all the missing data in the associated variability estimates.

For each subject's data, the explanatory variables remained the same, and the output columns representing abundances for bacterial taxa measured on the samples from the simulated microbiome changed from subject to subject. There are over 50 columns per subject containing the relative abundance proportion of different strains of bacteria in each sample. Given the list of all bacteria compiled, researchers evaluated the sequenced data and selected a simplified number of top and most essential taxa to assess for potential use within individual univariate GWAS tests. For each

subject (1, 2, and 3), 36, 34, and 22 (respectively) of the top taxa were selected to be evaluated in further analyses.

Starting with a simple multivariate model, using more than 20 responses seemed unnecessary when analyzing the relationships between the experimental design variables and the relative abundance responses. So, it was important to see what taxa accounted for most of the abundance across each of the individual well samples and which would provide the most informative gut microbiota representation. This statement is related to the previous discussion, which highlighted the need for a variable reduction technique to more effectively comprehend and consider the relationships between the taxa. From the columns with less than 50% zeroes, I have included a table including the mean, max, and median relative abundances, ordered from largest to smallest median value and including only the top 22 responses (since only 22 were included in the subject 3 information). The summary values fit what we would expect to see with relative abundances and indicate that the responses will be small proportions with the potential for right-skewness. This outcome aligns with the literature's focus on non-parametric and non-normal methods for data analysis.

Subject 1	Mean	Median	Minimum	Maximum	Subject 2	Mean	Median	Minimum	Maximum	Subject 3	Mean	Median	Minimum	Maximum
Prevotella9	0.4459	0.4511	0.0084	0.6412	**Bacteroides**	0.2398	0.2236	0.1049	0.4976	**Prevotella9**	0.4755	0.4714	0.1428	0.7039
Succinivibrio	0.0639	0.0581	0.0207	0.4988	**Phascolarctobacterium**	0.2219	0.2046	0.0093	0.5258	**Megasphaera**	0.1113	0.1117	0.0092	0.2215
Dialister	0.0627	0.0623	0.0079	0.1358	**Faecalibacterium**	0.1134	0.1166	0.0039	0.2698	**Acidaminococcus**	0.0917	0.0929	0.0022	0.1851
Enterobacteriaceae	0.0618	0.0611	0.0096	0.1579	**Clostridiumsensustricto1**	0.0862	0.0582	0	0.4988	**EscherichiaShigella**	0.0837	0.0818	0.0285	0.2513
Bacteroides	0.0367	0.0361	0.002	0.0673	**Sutterella**	0.053	0.0471	0.0182	0.2761	**Bacteroides**	0.0826	0.0797	0.0286	0.1884
Coprococcus3	0.0356	0.0367	0.001	0.0759	**Bifidobacterium**	0.0375	0.0413	0	0.2533	**Clostridium sensu stricto 1**	0.0399	0.037	0.0011	0.1993
Anaerostipes	0.0328	0.0322	0.0025	0.1106	**Coprococcus3**	0.035	0.0401	0.0003	0.0898	**Phascolarctobacterium**	0.0211	0.0188	0	0.273
Sutterella	0.0314	0.0311	0.0024	0.0571	**Dorea**	0.0296	0.0359	0	0.0751	**Sutterella**	0.0179	0.0158	0.0065	0.0503
Prevotellaceae	0.0219	0.0218	0.0007	0.0459	**Alistipes**	0.0251	0.0242	0.0044	0.0496	**Megamonas**	0.0152	0.0158	0.0018	0.0385
Coprococcus1	0.0206	0.0203	0.0004	0.0465	**Roseburia**	0.0193	0.0188	0	0.0748	**Prevotella7**	0.0133	0.0129	0	0.0378
Bifidobacterium	0.0191	0.016	0.0006	0.2913	**LachnospiraceaeUCG004**	0.017	0.0176	0	0.0513	**Bifidobacterium**	0.0072	0.006	0.0003	0.1801
Lachnospiraceae_NA	0.0176	0.0168	0.0022	0.1754	**Blautia**	0.0167	0.0198	0	0.0452	**Parabacteroides**	0.0051	0.0049	0.0008	0.0129
Dorea	0.0168	0.0168	0.0004	0.0373	**Parabacteroides**	0.0098	0.009	0.0038	0.0213	**Dorea**	0.0031	0.0028	0	0.0153
Blautia	0.016	0.0158	0	0.0418	**Eubacteriumhallii**	0.0097	0.0079	0	0.038	**Paraprevotella**	0.0028	0.0028	0	0.0096
Roseburia	0.0159	0.016	0	0.0279	**Coprococcus1**	0.0077	0.0072	0	0.0321	**Blautia**	0.0028	0.0025	0	0.0205
Fournierella	0.0091	0.0083	0	0.0262	**Parasutterella**	0.007	0.0068	0.0015	0.015	**Coprococcus3**	0.0027	0.0014	0	0.0438
Faecalibacterium	0.0082	0.0057	0	0.0767	**Agathobacter**	0.0069	0.0067	0	0.0187	**Enterococcus**	0.0022	0.0022	0	0.009
LachnospiraceaeUCG004	0.0078	0.0078	0	0.017	**Anaerostipes**	0.0052	0.0043	0	0.0306	**Butyricicoccus**	0.0022	0.0024	0	0.0071
Clostridiumsensustricto1	0.0077	0.0046	0	0.2699	**ChristensenellaceaeR7group**	0.0051	0.0049	0.0017	0.0147	**Allisonella**	0.0018	0.0017	0	0.0063
Agathobacter	0.0054	0.0053	0	0.0137	**Odoribacter**	0.004	0.0039	0.0008	0.0106	**Coprococcus1**	0.0017	0.0013	0	0.0586
Lachnospira	0.0046	0.0043	0	0.0143	**RuminococcaceaeUCG002**	0.004	0.0036	0.0006	0.0267	**Lachnospiraceae UCG-004**	0.0017	0.0011	0	0.0341
Prevotella2	0.0046	0.0043	0.0007	0.0153	**PrevotellaceaeUCG001**	0.004	0.0035	0	0.016	**Faecalibacterium**	0.0016	0.0013	0	0.1326

Table 3.2: Top Twenty-Two Taxa by Mean and Median Abundance

When a specific design is utilized in an experiment, it is important to ensure that all sources of variation are accounted for as incorporated in the design. Given the specific assignment of treatments, it is crucial to incorporate relevant features in

the initial analysis to accurately account for their potential effect. Since plates are incomplete blocks for the treatments (bean lines), we can account for the plate-to-plate variability. The check replicates are also aligned within the rows and columns within a plate; thus, it is possible to account for within-plate variation. Writing out tables with the sources of variation is a good way to note all the aspects of the experimental design that can be included in the statistical model. In a perfect world, subjects could be all run through the experiment simultaneously, and differences between subjects could be evaluated within the analysis. We could include a subject term and any appropriate interactions across subjects in the sources of variation. This design layout would be the same as our final design, except a different subject would be randomly assigned to each replication of four plates within an experimental run. However, since running the samples from multiple subjects at one time is impossible (due to the time each subject's samples are viable), the sources of variation can be evaluated for an individual subject.

Without the subject effect, we can easily see in Table 3.3 what occurs for one experiment run across three replications of four plates. Included in the design are three replicates, with four plates each, used to calculate the degrees of freedom for the first two rows in Table 3.3. A total of 300 bean lines are included, and the line error is calculated by taking the degrees of freedom available for bean lines across the two replicates and accounting for the effect of the plate. There are 12 different check lines within each plate, which differ for every replicate, leaving 144 degrees of freedom to account for any check line effect. Variation among wells within a plate is described using 12 checks in a row/column α - design (with three rows and four columns), producing the lines for rows and columns under wells. Plate error is the leftover degrees of freedom remaining for wells after accounting for the rows and columns for the check design.

Source	Df	Calculation	Description
Rep	2	$3-1$	
Plate(rep)	9	$3*(4-1)$	
Lines	299	$300-1$	
(Lines X Rep) - Plates(Rep)	589	$(299*2)-9$	Line Error
Wells(Checks)	144	$12*4*3$	12 checks, across 12 plates
Row(plate)	24	$(3-1)*4*3$	Row trends in plates
Col(plate)	36	$(4-1)*4*3$	Col trends in plates
Plate error	84	$144-(24+36)$	
TOTAL	1043	$1044-1$	

Table 3.3: Sources of Variability for Experimental Design of One Subject

However, moving forward, check effects have been removed from the final MANOVA model for analysis to create the multivariate traits. Initial data exploration revealed non-practically significant differences between rows and columns and the check lines. While researchers were curious about any within-plate differences that could be characterized in the analysis, of particular interest to them with this extension to their primary analyses. In Dr. Van Haute's GWAS models, they account for overall row and column effects within a plate, so for the adjusted goals of our analysis, the MANOVA model was simplified to the sources of variation in Table 3.4.

Source	Df	Calculation	Description
Rep	2	$3-1$	
Plate(rep)	9	$3*(4-1)$	
Lines	L	$L=Lines-1$	Will change depending on how many lines are included
Error	E	$E=1043-(2+9+L)$	Combined Lines and Well Check Effects
TOTAL	1043	$1044-1$	

Table 3.4: Reduced Sources of Variability for Experimental Design of One Subject

3.5 Analysis Methodology

3.5.1 Overview

Given our proposed data structure and goals, it is necessary that the analysis highlight community taxa effects on the response abundances. The final model summarized in this chapter accounts for design effects and bean line variation across multiple bacterial taxa to parse out variability between them. This methodology outlines the process of taking covariance outputs from a MANOVA to create pseudo-multivariate phenotypic values accounting for taxonomic relationships. These new responses are called *Polymicrobial Traits*, a term coined by Dr. Van Haute for her specific research vocabulary. Polymicrobial traits are new responses for GWAS models now accounting for multiple of the most abundant taxa simultaneously, to be added to the standard added to univariate taxa phenotypes. Community structure within the gut can then be evaluated via taxa associations and their relationships across the bean lines by utilizing these new traits.

New values for every observation are created as weighted combinations of the abundances across a reduced list of selected taxa. Weights will be calculated using loading vectors from a principle component analysis on variability matrices outputted from a MANOVA model. Ultimately, results from univariate GWAS models analyzing microbiome phenotypes from individual relative abundances for specific taxa of interest will be compared to output from the polymicrobial traits. Similarities and differences between the results will inform researchers about if this new method can assist in informing scientists about which bean lines affect changes in the gut microbiome. Also, we will be able to verify if the method can identify the same important chromosomal areas as the individual phenotypic models.

3.5.2 Proposed Analysis Methodology

As observed from a general review of the literature, a lot of research on GWAS methodology focuses on exploring the connections between genetic information and observable characteristics. However, this study concentrates on a pre-processing analysis approach that generates valuable and easy-to-interpret "pseudo" phenotypic variables for GWAS models. For this project, researchers focus on the association between the bacteria output in the gut and the SNP region on the chromosome, where the outcomes are related to the genetic material associated with the P. vulgaris genome. The taxa traits from the AiMS phenotyping will be multivariate as they contain information across a community of bacteria from each of the subjects' microbiome samples. When designing the experiment and proposed analysis with Drs. VanHaute and Benson, we knew there were attributes that we could address with statistical methodology.

While the responses are expressed as proportions, we investigate treating a transformed version of the abundances as normally distributed for simplicity. To better capture the community effects, multiple taxa should be considered simultaneously. Using a MANOVA model will enable the evaluation of the variability between taxa in relation to the bean lines. We would include terms to account for plate and batch effects in the analyses, separated into analyses for each of the three subjects to match how UNL microbiologists perform their analyses. After accounting for design effects, each model's covariance output could be used to highlight relationships between taxa. Finally, principal component analysis (PCA) and canonical discriminant analysis (CDA) can be used with the covariance outputs to summarize the community relationship into component vectors of "pseudo" phenotypic summaries weighted across multiple taxa. PCA is an unsupervised technique used to identify the under-

lying structure in a data set by transforming the data into a new coordinate system that explain the maximum amount of variance in the data. PCA aims to simplify the data while retaining as much information as possible. CDA, on the other hand, is a supervised technique used to identify the directions in a data set that best discriminates between two or more classes. CDA is used to find a linear combination of features that maximizes the separation between classes. The result is a new set of features that can be used for classification. In summary, PCA is used for exploring the data structure, while CDA is used for classifying the data based on the class labels. The final "pseudo" phenotypic summaries will be used responses alongside univariate taxa to be run through Dr. Van Haute's GWAS procedures.

This work will compare the results of the GWAS outcomes from our multivariate responses to those from individual univariate taxa. The comparisons will help us determine whether community variables identify the same or different significant SNP areas as the univariate models. Confirming similar results in SNP areas would demonstrate that our new models can effectively capture relevant information compared to standard statistical techniques. The results will be discussed to identify the valuable knowledge that statisticians and microbiologists can gain from these new models and how they can be used in practice moving forward. Overall, this study aims to provide a novel approach to analyzing complex data sets, and its findings are expected to advance our understanding of the genetic basis of complex microbial communities.

3.5.3 Data Cleaning and Multivariate Analyses

The process begins by selecting the most prevalent taxa accounting for a majority of the abundance across each subject's microbiome samples. These values will be utilized to evaluate the variability between the subset of taxa and how it is affected

by the experimental design and bean line effects in a multivariate analysis of variance (MANOVA) model. To select the most prevalent taxa, the raw abundances of all taxa were summed for each observation across varying groupings of taxa columns. Taxa were ranked from most to least average relative abundance, and groups of columns were selected by adding in the most abundant taxa one at a time. Final subsets were selected based on which groups of taxa accounted (across all observations) for at least 90% of the overall abundance. On average, around 92% of the overall abundance was accounted for with the top 17 most abundant genera for subject one, 15 for subject two, and nine for subject three.

Faecalibacterium was required to be chosen for every subject, as domain experts were particularly interested in this taxon. Subject three is the only exception to the selection process, as Faecalibacterium did not fit the initial inclusion criteria. This means that it was included in the final taxa set even though it had a smaller overall average abundance versus other taxa that were not included. The final sets of all selected taxa are listed in Table 3.5. These sets of taxa are used as response matrices in a MANOVA to evaluate the variability between the taxa and how the taxonomic outcomes are affected by the plate, digestion batch (same as replicate), and bean line effects.

After selecting the response taxa, but before fitting the model, raw abundances were transformed using a centered log-ratio (CLR) transformation, a common transformation with microbiome abundance data and compositional data [53]. This data is compositional because, across each sample, the abundances for all combined taxa should sum would to one. All taxa found when sequencing represents the entire sample microbiome output from each well in the plate. The univariate GWAS analyses completed for Dr. Van Haute's dissertation utilized a simpler $log2$ transformation for the abundances, with a cutoff of 0.0015. However, the CLR transformation was

Taxa Selected for Modeling by Subject

Subject 1 (S770)	Zeroes	Subject 2 (S776)	Zeroes	Subject 3 (S768)	Zeroes
Bifidobacterium	0	Bacteroides	0	Prevotella9	0
Bacteroides	0	Phascolarctobacterium	0	Megasphaera	0
Prevotella9	0	Faecalibacterium	0	Acidaminococcus	0
Prevotellaceae	0	Clostridium sensu stricto 1	5	EscherichiaShigella	0
Anaerostipes	0	Sutterella	0	Bacteroides	0
Blautia	2	Bifidobacterium	6	Clostridium sensu stricto 1	0
Coprococcus1	0	Coprococcus3	0	Phascolarctobacterium	1
Coprococcus3	0	Dorea	1	Sutterella	0
Dorea	0	Alistipes	0	Faecalibacterium	73
Roseburia	3	Roseburia	2		
LachnospiraceaeNA	0	LachnospiraceaeUCG004	1		
Faecalibacterium	4	Blautia	19		
Fournierella	5	Parabacteroides	0		
Dialister	0	Eubacteriumhallii	66		
Succinivibrio	0	Coprococcus1	4		
Sutterella	0				
Enterobacteriaceae	0				
Total N	1003		1016		990

Table 3.5: All Taxa used in MANOVA with Number of Transformed Zeroes

used for this analysis to get close to normality since it was highlighted as an appropriate transformation across the literature. Aitchison introduced the method in 1982 [1]. For the transformation, instead of taking the ln of only the abundance value, the CLR uses the formula in Equation 3.1 where a is a given abundance value for a specific taxon, and $g(a)$ represents the geometric mean of a, which are all of the taxa abundance values for a given observation in the data set.

$$a_{CLR} = ln[\frac{a_i}{g(a)}] = ln(a_i) - ln(g(a)), i = 1, 2, ...m \tag{3.1}$$

To account for zero values in the data, which cannot be used with the CLR transform, we replaced all responses containing zero abundance values with a value of $10e^{-8}$. These replacements allowed us to include all observations in the analysis without losing valuable information. Table 3.5 contains columns containing the total number of zero observations adjusted before the CLR transformation. Next, we used the transformed abundances as the response matrix in a MANOVA to evaluate their

relationship with fixed effects of the bean line, digestion batch, and the plate within each of the three digestion batches. As described in Chapter 2, digestion batches consist of four plates that undergo the same digestion, dialysis, and fermentation steps to mimic the passage of food components through the gastrointestinal tract [59].

Analyses for all subjects were fit using the PROC GLM procedure within SAS 9.4 [146]. A general sketch outline of the model effects is outlined in Equation 3.2 with the skeleton ANOVA in Table 3.6.

$$
\begin{aligned}
CLR\ Abundance &= Design\ Rep + Plate(Design\ Rep) + Bean\ Line + Error \\
Y_{n\ x\ p} &= X_{n\ x\ 307}\beta_{307\ x\ p} + E_{n\ x\ p} \\
Y \sim\ & MVN(X\beta, \Sigma), E \sim\ N(0, \Sigma)
\end{aligned}
\tag{3.2}
$$

This model evaluates the relationship between the taxa resulting from changes across the bean lines while accounting for variability due to the major experimental design factors. To examine this relationship, we output the hypothesis and error sums of squares and cross-products (SSCP) matrices from the MANOVA, H_3, and E in Table 3.6. Specifically, these are the hypothesis SSCP matrices for the overall bean line effect and the error SSCP describing the variance in the abundances after accounting for the fixed effects. In addition, a Canonical Discriminant Analysis (CDA) is outputted from the analysis [159, 67]. The process for using the raw and standardized coefficients from the CDA is described in the following section. In summary, the CDA results will be used alongside the principal component vectors to create scores for additional polymicrobial traits calculated using the appropriate transformations of the H_3 and E hypothesis and error matrices as described in the

following paragraph.

MANOVA Skeleton Sources of Variation

Source	**df**	**SSCP**
Design Rep	2	$H_1 = Y'A_1Y$
Plate(Design Rep)	9	$H_2 = Y'A_2Y$
Bean Line	296	$H_3 = Y'A_3Y$
Error	$(n-1)-307$	$E = Y'[I - X(X'X)^{-}X]Y$
Total	$n-1$	$T = Y'Y$

Table 3.6: Skeleton Sources of Variation across all Subject MANOVAs

After analysis, hypothesis and error covariance matrices were calculated by dividing the SSCP matrices by their corresponding degrees of freedom. Correlation matrices were calculated directly from the corresponding covariances. The covariance from the hypothesis SSCP is interpreted as an estimate of both genotypic variabilities across the bean lines and additional error. An additional "genetic" covariance was derived as a multivariate analogy to the univariate broad-sense genetic variance to obtain a more targeted estimate of the genotypic variability. This new matrix was computed by substituting the observed genotype hypothesis covariance for the bean line expected mean squares and solving for the genetic covariance matrix. This calculation can be seen in Equation 3.5, where the error covariance is subtracted from the bean line hypothesis covariance and divided by a constant value ($Constant^{+}$) calculated from the Expected Mean Square for Bean Lines. Observing the variability in this way is similar to assessing the partitioning of variability for univariate sums of squares. A general calculation overview for the covariances and correlations can be seen in Equations 3.3 to 3.7. Within the equations, n represents the sample size for each subject's data set, and p is the number of taxa included in the response matrix. The number of taxa in each subject can be found above, and the sample sizes for subjects 1, 2, and 3 are 1003, 1016, and 990, respectively.

$$Hypothesis\ Covariance = S_{lines_{pxp}} = \frac{1}{296}\ \mathbf{H_3} \tag{3.3}$$

$$Error\ Covariance = S_{Error_{pxp}} = \frac{1}{(n-1)-307}\ \mathbf{E} \tag{3.4}$$

$$S_{G_{pxp}} = \frac{S_{Lines} - S_{Error}}{Constant^{+}} \rightarrow GenCov = \frac{HypCov - ErrorCov}{Constant^{*}} \tag{3.5}$$

$$Hypothesis\ Correlation = \sqrt{diag(H_3)}^{-1} * H_3 * \sqrt{diag(H_3)}^{-1} \tag{3.6}$$

$$Genetic\ Correlation = \sqrt{diag(GenCov)}^{-1} * GenCov * \sqrt{diag(GenCov)}^{-1} \tag{3.7}$$

The following section outlines the utilization of covariance and correlation matrices to generate score values, which in turn determine polymicrobial trait responses. These responses are employed to evaluate community structure based on taxa relationships, taking into account variation among the bean lines.

3.5.4 Loadings, Scores, and Polymicrobial Traits

Next, we utilized the MANOVA model outputs as the inputs for both CDA and PCA analyses. The loadings calculated from the models will become weights on the transformed taxa abundances to produce the polymicrobial traits. The weights are calculated using statistical information from both PCA and CDA analyses, which identify the key taxa responsible for the most separation in the bean lines. PCA differs from CDA in that PCA "unless specifically manipulated to do so, does not partition the data matrix to take account of any experimental design structure," whereas it "maximizes the proportion of the total variance of the data set expressed by successive principal components" [92]. Canonical coefficients within the discriminant vectors work similarly to the PCA loadings. Canonical Discriminant Analysis derives the linear combinations (i.e., canonical functions) of the variables that discriminate the

best (i.e., maximize the variation) among the groups [159]. In this case, our variables are the taxa, where we are maximizing the variation among the include bean line groups. PROC GLM in SAS directly calculates the different weights using the error and hypothesis SSCP matrices from the model to maximize the vector $\underline{a}$ in Equation 3.8. Each value in $\underline{a}$ represents the vector of CDA coefficients to be used further as weights for the "scores", which become our polymicrobial traits. Two types of discriminant vectors are outputted from the MANOVAs, with raw and standardized coefficients.

$$\max_{\underline{a}} \ \frac{\underline{a}'\mathbf{H_3}\underline{a}}{\underline{a}'\mathbf{E}\underline{a}} \ , \ \underline{a} \ \rightarrow \ \textit{Loadings} \tag{3.8}$$

There are multiple analyses for the PCA models with various input matrices. The *princomp* and *eigen* functions in R were used to conduct principal component analyses (PCA) with the hypothesis and genetic covariance/correlation matrices as inputs [118]. Analyses are computed directly with each hypothesis covariance and correlation matrix using the *covmat* command in the *princomp* function. Direct eigendecomposition using the *eigen* function found the principle component loadings for the genetic covariance and correlation matrices. The loadings from all models (PCA and CDA) were used as weights in linear combinations of taxa abundances, creating the corresponding scores for each observation. Larger weights demonstrate that the corresponding taxon accounts for the majority of the variability of the input matrix. Therefore, the final scores give a value heavily weighted by taxa that account for the most variability in the data while still including smaller weights across all taxa.

Scores were calculated by multiplying the loadings vector by the centered (or centered and scaled) CLR relative abundance values, as in Equation 3.9 below, where

W is a matrix containing all PC loadings. This calculation is repeated for each of the separate PCA models for the differing covariance and correlation matrices. Centered CLR abundances ($Y^* = Y_i - \overline{Y_i}$) were used to calculate scores for the covariances, and the centered and scaled CLR abundances ($Y^* = Y^* * \sqrt{diag(\Sigma)}^{-1}$) were used for the correlations (where Σ is the input matrix from the corresponding PCA). Since canonical discriminant loadings come directly out of the MANOVA, only the scores need to be calculated, and centered data was used for these values.

$$\text{PCA Score}_{n\ x\ p} = {Y^{*\prime}}_{n\ x\ p}\ *\ W_{p\ x\ p} \rightarrow Polymicrobial\ Traits \tag{3.9}$$

Both PCA and CDA result in a total of p loading and score vectors. The first vector in each is the linear combination of taxa that accounts for the most variability overall in the input data. Individually, this variability gets smaller from the 1st to the p^{th} loading, cumulatively increasing until the loadings account for 100% of the variability. Therefore, including all loadings (which would be the same as using all p taxa) is unnecessary. Instead, only the top four components from each method were used to calculate the individual polymicrobial traits.

Twenty-Four new columns of traits with values for each observation were added to each subject's data set. Sixteen columns are contributed from the PCA models and eight from the CDAs. For the PCAs, four columns come from each of the four types of input matrices: hypothesis covariance, hypothesis correlation, genetic covariance, and genetic correlation, creating the 16 columns. The eight CDA components make up four PM traits calculated with raw and standardized coefficients from the top four canonical discriminants. All of the types of traits were provided to the researchers for further testing, but further work could be done to assess if each method is necessary for inclusion in the GWAS. Similar GWAS peaks attributed to several trait varieties may

indicate redundant information about the bean line's variability. In such instances, it may be unnecessary to include all the values in the analysis. Further descriptions of the usefulness of all methods will be provided in the following results and discussion sections.

Overall, we described the final polymicrobial traits to the domain experts as values calculated as weighted sums of the abundances of selected taxa, with weights that optimize the trait's capacity to account for variation across the taxa. Dr. Van Haute emphasized that the polymicrobial traits are values that can "account for several of the most abundant genera after in vitro fermentation" at once, and there exist different styles of trait depending on how "variation" was defined" [59]. For clarity in the results, we labeled the six types of polymicrobial traits from each method in the following way: hypothesis covariance (HypCov), hypothesis correlation (HypCorr), genetic covariance (GenCov), genetic correlation (GenCorr), raw CDA, and standardized CDA. Additionally, the results indicate which component or discriminant vector (1 - 4) was used to calculate the trait values. For readers from a domain perspective, we provided the following explanation for the process of why the methods were separated as they were:

To provide domain-specific readers with a better understanding of the methodology, we assisted Dr. Van Haute in creating the following explanation for why the methods were separated in the way that they were:

> *For the "hypothesis methods," variation was covariation across the genera due to both genetic and unexplained causes, also described as residual error. The variation for the "genetic methods" was covariation across genera due only to the genetic causes. For both methods, "covariation" was based on a covariance matrix (i.e., HypCov, GenCov) of transformed taxa*

abundances or a correlation matrix (i.e., HypCorr, GenCorr) of standardized transformed taxa abundances to account for differences in variances of the genera. For the canonical discriminant analysis (CDA) method, the weights were chosen to maximize the genetic and error covariation relative to error covariation [59].

3.5.5 Genome Wide Association Studies (GWAS)

Genetic Wide Association analyses were used to find areas on the bean genome where variation in the P. vulgaris lines have statistically significant effects on changes in the different categories of selected traits. After each subject's polymicrobial traits (represented with "PM traits" for simplicity) were created, the values became responses alongside univariate taxa using the same GWAS processes. The GWAS models were run by Dr. Andrew Benson's lab, and the modeling process was fully described within Dr. Van Haute's Ph.D. Dissertation [59]. However, this section will summarize what was conducted related to the PM traits.

The PM traits are included as one of six categories of responses for the GWAS models. Other groups utilized as responses included α-diversity, ASV (Amplicon Sequence Variance values or 100% OTUs), genus and family level abundance values, and short-chain fatty acid information[59]. The number of traits selected for GWAS modeling was reduced based on a broad-sense heritability minimum value cutoff of 0.1. Broad sense heritability (H^2) was calculated by dividing the variance explained by the P. vulgaris genotypes by the variance explained by all other random effects, which can assist in evaluating "the extent to which the phenotype can be predicted by identified variants"[59, 115]. A summary of the number of trait values utilized in GWAS can be found in Table 3.7.

For each subject, 36, 34, and 22 of the top individual taxa genera were chosen

Number of Traits Per Subject

Subject	**1: S770**		**2: S776**		**3: S768**	
Category of Trait	**Top #**	$H^2 > 0.1$	**Top #**	$H^2 > 0.1$	**Top #**	$H^2 > 0.1$
Univariate Genera	36	23	34	17	22	13
Polymicrobial Traits	24	18	24	11	24	12

Table 3.7: **Univariate and Multivariate Microbiome Traits.** Total number of microbiome traits resulting from the subject's microbiome. Heritability columns ($H^2 > 0.1$) indicate which final traits were used in the genetic association analysis.

as possible univariate traits, and 24 total polymicrobial traits exist for each subject. Table 3.8 lists heritability values for each PM trait. Out of a possible 72 PM trait variables across subjects, 41 were selected for final modeling. Values marked with an asterisk (*) in Table 3.8 indicate traits that had heritabilities larger than the cutoff for all subjects. All three subjects had traits that were included in the final models.

Researchers ran the selected PM traits through the procedures for genome-wide association analyses. Full descriptions of the GWAS procedures can be found in Dr. Van Haute's dissertation in section 3.11. However, the following information provides a general summary of the methodology, which is necessary for the remaining results and for the background of how useful researchers found the use of the PM traits. First, BLUPs (Best Linear Unbiased Predictors) are calculated for each P. vulgaris genotype using "using the first three PCs of the SNP data as fixed effects" as well as random effects from the experimental design (bean genotype, digestion batch, plate, and row and column within a plate)[59]. The kinship matrix for the genotypes was utilized in their mixed models in the covariance effects for the random bean genotypes, similar to what will be expanded on in our chapter four. Separate GWAS models were run separately for the BLUPs calculated from each phenotypic trait using a FarmCPU algorithm ("iterated Fixed and Random Model Circulating Probability Unification algorithm using") using the first three principal components of 132,314 SNPs and the centered kinship matrix [59].

Heritability Values for Polymicrobial Scores				
Trait Group	**Vector**	**Sub 3 (S768)**	**Sub 1 (S770)**	**Sub 2 (S776)**
Genetic Correlation	1*	0.189	0.311	0.175
	2	0.05	0.112	0.035
	3	0.063	0.136	0.13
	4	0.08	0.037	0.058
Genetic Covariance	1*	0.133	0.38	0.319
	2	0.069	0.14	0.03
	3	0.024	0.082	0.103
	4	0.094	0.079	0.054
Hypothesis Correlation	1*	0.103	0.313	0.073
	2	0.147	0.09	0.19
	3	0.008	0.137	0.101
	4	0.078	0.113	0.084
Hypothesis Covariance	1	0.125	0.354	0.214
	2	0.101	0.089	0.097
	3	0.022	0.071	0.096
	4	0.05	0.164	0.036
Raw CDA	1*	0.432	0.504	0.679
	2*	0.31	0.48	0.296
	3*	0.222	0.403	0.243
	4	0.076	0.176	0.139
Standardized CDA	1	0.144	0.492	0.008
	2	0.112	0.347	0.084
	3	0.083	0.317	0.092
	4	0.115	0.23	0.07

Table 3.8: Heritability of Microbiome Polymicrobial Phenotypic Traits resulting from the AiMS platform and broad sense heritability values of each trait from the three donor microbiomes. * Trait is identified in all three microbiomes with $H^2 > 0.1$

There are several outputs from the GWAS models, but the specific results of interest are in sections of significant major effect loci (MEL) regions on the chromosome. This strategy for analyzing the traits by subject was developed to deal with the many phenotypes being analyzed. With the MEL areas, researchers can visualize "significant associations arising from multiple traits from multiple subjects across the *P. vulgaris* genome" and highlight narrow regions on the genome affected by particular sets of traits (including those from univariate or multivariate methods

[59]. Linkage Disequilibrium (LD), calculated (as R^2), helps determine the relatedness of nearby SNPs and is used to determine the strength of LD surrounding bins with high pleiotropy (the ability of a gene to affect multiple traits) to define major effect loci (MEL). MEL regions were defined as selected regions where $R^2 \geq 0.75$. Specifically, researchers looked for traits associated with significant SNPs within the MEL, where the focus was on finding MEL regions where there were "the most significant marker-trait associations" [59]. Because of the large number of included SNPs, to find MEL regions of interest, researchers binned together sections of the genome to observe "significant associations of the traits with SNPs that are immediately adjacent to one another" [59]. In the end, seven major effect loci (MEL) areas were selected, including significant associations of 10 or more traits on each. Results of interest will highlight areas where the polymicrobial traits were significant and how their significance overlaps with similar or different individual traits from the GWAS models.

3.6 Results and Discussion

3.6.1 MANOVA, PCA, and CDA

We used individual MANOVA models to analyze the selected subsets of taxa for each subject and obtain polymicrobial traits. The final sets of taxa utilized for each subject can be found in Table 3.5. The main outputs of interest were the SSCP matrices of hypothesis and genetic variance, which were utilized to generate the corresponding covariance and correlation matrices. These matrices, in turn, served as inputs for principal component analyses. Discriminant vectors come out of the CDA analysis from PROC GLM with the canonical option in the MANOVA statement. The cumulative proportion of variance in the data explained by the first four components

or vectors from all PCA and CDA models was outputted and compared. For subjects 1, 2, and 3 (labeled S770, S776, and S768, respectively), the first four canonical discriminant vectors accounted for cumulative proportions of 54.59, 55.36, and 66.70 percent of the variability in the responses (respectively). The canonical discriminants explain roughly the same amount of variation across subjects. Table 3.9 contains the individual cumulative proportions values in each of the first four components for the PCA models. A fair amount of variability is accounted for across the first four components. For each subject, there were a total of 17, 15, and 9 components overall. Given that the analysis included almost half of the possible components, it is reasonable that subject three exhibits a large proportion of variance accounted for in the fourth component.

Cumulative Proportion of Variance

Method	Subject	PC1	PC2	PC3	PC4
HCOV	1	0.41	0.64	0.76	0.83
	2	0.56	0.72	0.83	0.88
	3	0.86	0.94	0.96	0.98
HCORR	1	0.41	0.6	0.68	0.75
	2	0.44	0.65	0.74	0.81
	3	0.49	0.74	0.84	0.9
GENCOV	1	0.53	0.67	0.74	0.81
	2	0.58	0.71	0.81	0.9
	3	0.57	0.79	0.82	0.88
GENCORR	1	0.56	0.71	0.8	0.85
	2	0.61	0.76	0.87	0.93
	3	0.59	0.83	0.9	0.95

Table 3.9: Cumulative Proportion of Variability accounted for by the PCA models across different covariance and correlation matrices

After obtaining the loadings and discriminant vectors, we proceeded to assess the relationships between specific components across multiple methods. Observing the magnitudes and directions of the vectors is important as the values serve as weights in

creating final polymicrobial traits for the genome-wide association analyses. Simplified biplots were used as graphical representations of the loading/discriminant vectors from the PCA and CDA models. Each vector is represented by an arrow pointing toward the maximum variation for that specific component or discriminant, and the arrow's length represents the variation's magnitude. The positively correlated variables are placed close to each other on the plot, while those negatively correlated are placed far apart. Visualizing the relationships between different component vectors generated from the analyses through such plots is a useful tool for exploring and interpreting multivariate data.

To visually compare the weight and effect of each taxon on the corresponding polymicrobial trait, variations of biplots were evaluated for each model. Each plot's x and y axes represent combinations of varying component vectors from one of the PCA or CDA analyses. For example, when comparing PC 1 vs. 2 for the Hypothesis Covariance matrix in Figure 3.1, there are strong loadings in the direction of Eubacterium halli (shortened to Eubact) for the PC1 loading, drawing the vector out to the left. However, there are smaller loadings for this taxon for PC2, as evidenced by the arrow staying flat around zero in the y-axis direction. For PC2, the loading value for Eubacterium halli is -0.04, relatively close to zero compared to the directions of the other taxa along the y-axis. In comparison, PC1 has a larger loading value of -0.79. In the comparison of PC1 versus PC2 for subject 2 using the hypothesis covariance matrix, we have identified several taxa that are heavily weighted, as well as groups of taxa that share similar relationships. We can further evaluate other associations by plotting all the different combinations of PCs across various methods. Based on these plots, we found that there may be particularly interesting regions where one or two specific taxa heavily influenced the polymicrobial traits, as opposed to communities of taxa with more evenly distributed weight across their responses. Domain experts

expressed interest in having the ability to highlight what types of relationships were occurring across subjects and then how these relationships tied into which methods created significant associations with the bean genome.

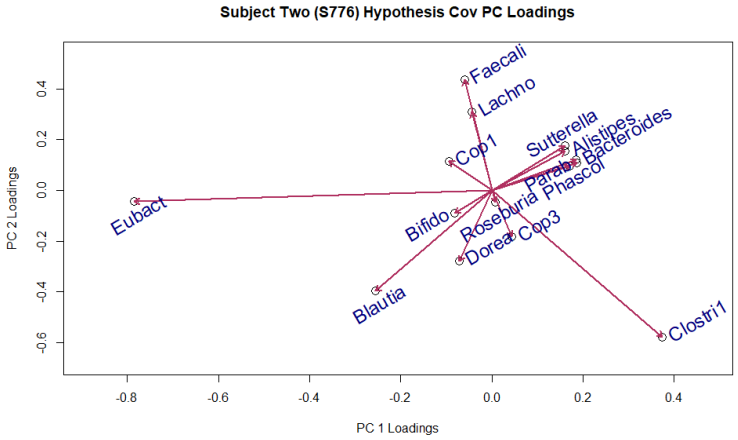


Figure 3.1: Subject 2 Biplot for PC1 vs. PC2 from PCA on the Hypothesis Covariance, where PC1 and PC2 account for 72% of the cumulative variability, from Table 3.9

Dr. Van Haute noted the following plots, which demonstrate other interesting relationships in the PC loadings. For Subject 3, when comparing loadings for PC1 vs. PC4, there is a strong influence of the Faecalibacterium, as seen in Figure 3.2. For the remaining taxa, there are smaller weights, but compared to PC 4, we can see some groupings together with positive and negative components along the y-axis. In the positive direction, there is a combination of Bacteroides, Sutterella, and Acidaminococcus. The negative weights consist of Clostridium, Phascolarctobacterium, and a more spread-out group of Prevotella9, Escherichia, and Megasphaera. Gaike et al. found that compared to healthy non-diabetics, there was an increased abundance of Megasphaera, Escherichia, and Acidaminococcus and decreased abundance of Sutterella in people with diabetes on antidiabetic treatments[50]. Separately, genera like Prevotella9, Escherichia, and Megasphaera could be of interest because elevated levels of proteobacteria (like Escherichia/Shigella) is regarded as a potential feature

of imbalance of organisms in the gut, related to inflammation and cancer, whereas taxa such as Prevotella9 and Megasphaera are generally regarded as beneficial genera [162, 137].

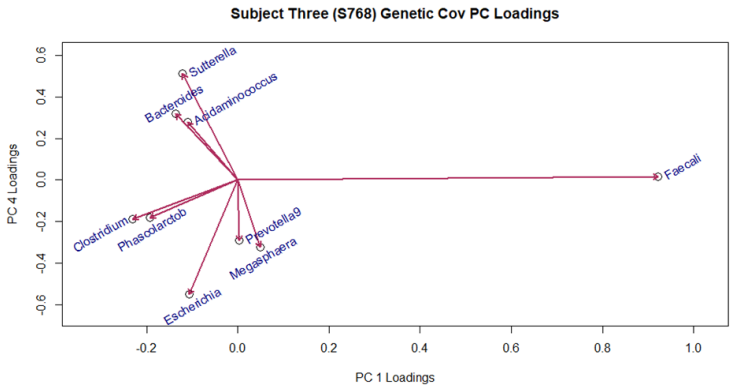


Figure 3.2: Subject 3 Biplot for PC1 vs. PC4 from PCA on the Genetic Covariance, where PC1 accounted for 57% of the variability and PC4 accounts for an additional 6% of the cumulative variability, from Table 3.9

There can also be component vectors of interest with substantial or closer to equally weighted values for taxa "communities", as in Figure 3.3. Component 2 seems primarily driven by values from Prevotella9, Megasphaera, and Escherichia, while component 4 has larger weights for Escherichia and Sutterella. There could also be cases where no individual taxa seem to drive the vector, and all values across taxa are fairly small.

Since it was not feasible to plot and visualize all comparisons in the two-dimensional space, Figure B.1, Figure B.2, and Figure B.3 list out the loadings for all subject and method combinations. The values in the tables demonstrate the same relationships as in the selected biplots, where we can observe whether components are highly affected by one or two major taxa (as in Figure 3.1 and Figure 3.2). To summarize how the taxa are associated, we reviewed loadings and selected values with Dr. Van Haute that represented the more heavily weighted taxa for the polymicrobial

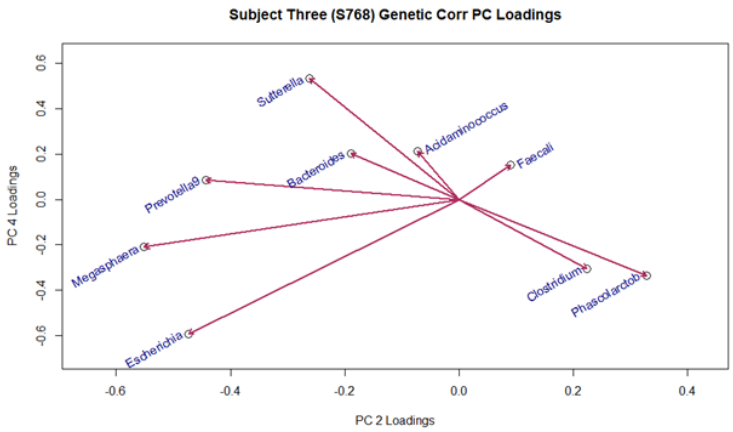


Figure 3.3: Subject 3 Biplot for PC2 vs. PC4 from PCA on the Genetic Correlation, where PC2 accounted for 24% of the variability and PC4 accounts for an additional 5% of the cumulative variability, from Table 3.9

trait values. Across the three figures, different values are highlighted in red and blue. PCA loadings <-0.4 were shaded blue, and >0.4 were shaded red. CDA loadings <-2.5 were shaded in blue, and >2.5 were shaded in red. The cutoff values were selected by observation based on the range of values observed in the vectors within the PCA or CDA output. After evaluating the weights for the corresponding cutoffs, the taxa names noted as the heavily weighted values were placed into Table B.1 and Table B.2 within Appendix B. Dr. Van Haute utilized these tables after GWAS to evaluate which taxa possibly overlapped with any significant univariate taxa.

For each area along the genome identified as important by Dr. Van Haute (described as major effect loci), these taxa may or may not be related to other univariate taxa identified as significant across the chromosome positions. Thus, it is not helpful to identify the effects of the terms within Table B.1 and Table B.2 because from this information alone, we cannot see where the traits overlap with the univariate information. The following section will describe this information below by describing tables B.3 through B.9. For each of the polymicrobial traits identified as significant from Table B.3 through Table B.9, we will identify if the trait of interest contains

univariate taxa that were also identified as significant.

3.6.2 Polymicrobial Traits Heritability and Genome Wide Association Studies (GWAS)

After the PM traits were calculated using information from subsection 3.5.4 provided to researchers, Dr. Van Haute calculated the trait's heritability values and plotted a heat map that can be used to demonstrate some initial relationships between the traits across subjects and methods.

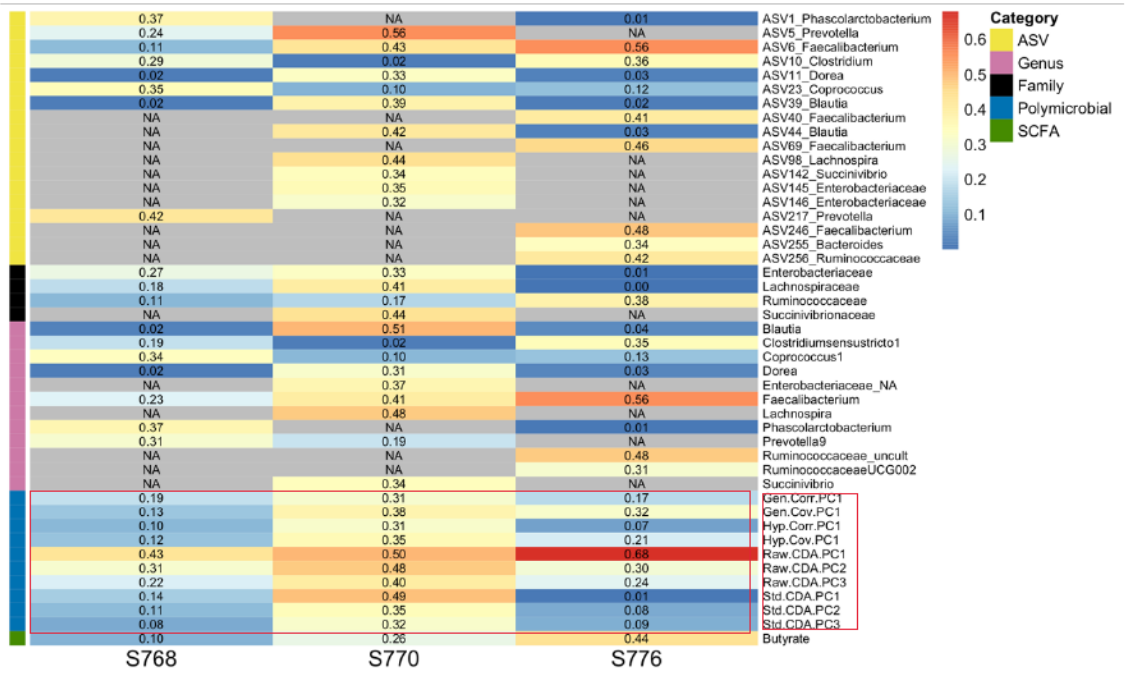


Figure 3.4: Heatmap of heritability values of microbiome features with values greater than 0.3 in at least one microbiome. Subjects 1, 2, and 3 are labelled as S770, S776, and S768 respectively.

Figure 3.4 includes the heritability values larger than 0.3 from Table 3.8, along with those from univariate traits. Coloring across the heat map indicates the magnitude of the heritability values according to the scale on the right. For Subject 1 (S770), the polymicrobial traits towards the bottom (in the red box on the right) match others with larger values in the yellow and orange colors. We can also note

that the first raw coefficient from the CDA for Subject 2 (S776) had a large heritability compared to some of the other groups, with a value of 0.68 highlighted in red. One of the Faecalibacterium groupings towards the top also had a fairly large heritability. While it was not one of the larger weighted taxa, it is interesting to note that for the Raw CDA coefficients, the first component had a negative loading for Faecalibacterium, different from all the other taxa. This could perhaps be a potential relationship to the individual Faecalibacterium univariate outcome.

As seen in Table 3.8, previously in subsection 3.5.5, only 13 total traits out of 226 were shared across all three microbiomes (minimum H^2 of 0.1). The microbiome features included the univariate terms of Shannon Diversity, ASV6 Faecalibacterium, ASV23 Coprococcus, Coprococcus1, Faecalibacterium, Ruminococcaceae family and butyrate, and the following PM traits: Genetic Correlation PC1, Genetic Covariance PC1, Hypothesis Covariance PC1, Raw CDA PC1/PC2/PC3. For the univariate traits, ASV stands for Amplicon sequencing variants (ASVs) which have been proposed as an alternative to operational taxonomic units (OTUs) [130]. The much smaller number of traits with larger heritabilities across all subjects demonstrate that "the individuality of the microbiomes have a major influence on AiMS traits defined from each microbiome," and there is a large amount of variability in the magnitude of H^2 for the same trait across microbiomes [59]. Only a few of the 13 traits significant traits across the microbiomes had associations with the same MEL. Still, there were important instances of other overlapping trait GWAS peaks within the same MEL, specifically including overlap from individual genera corresponding to those taxa contributing heavily to the variation of the PM traits [59].

All detailed GWAS results can be found in Dr. Van Haute's dissertation, with in-depth details about univariate traits with significant associations across the seven identified Major Effect Loci (MEL). This work focuses on outcomes for evaluating

how beneficial the polynomial traits are to the researchers. We want to address how researchers can explain the impact of community taxonomic effects and the chromosomal regions on the bean genome where they produce significant relationships. In each MEL (major effect loci), there are areas where significant associations exist for both the individual taxa and PM traits. Combined with the microbiological focus on the relationships between the individual traits and the bean genetic information, we aimed to show whether the PM traits would demonstrate similar relationships as the univariate. The overlap, or lack of overlap, between the traits can provide information about whether the PM traits are picking up signals not accounted for in the univariate GWAS analyses. In the appendices, for each MEL, Tables B.3 through B.9 summarize the significant associations for univariate and polynomial traits for each of the three subjects. For the PM traits, next to the significant method's name (such as hyp.cov.comp1), there is a list of the taxa determined to be heavily weighted from the loadings in calculating the traits.

According to Table 3.7, 41 of the 72 polymicrobial traits (57%) met the heritability cutoff of 0.1 and were included in the GWAS modeling. Of these 41 traits, 18, 11, and 12 were significant within each subject, respectively. This means that 75%, 45.8%, and 50% of the total possible polymicrobial traits (24 per subject) were included in the GWAS models for the three subjects, respectively. In Tables B.3 through B.9, symbols have been added to the names of the polymicrobial traits to identify several different category labels. If the trait name has a "+", one of the largely weighted taxa used to create the polymicrobial trait overlaps directly with a matching univariate trait within the same position or overall on the same MEL. Traits with a "$\sim$" contain largely weighted taxa related to one of the univariate traits on the same MEL. For example, the trait *hypcovComp*1 on MEL A was identified as significant on position 41,155,494. It has a large weight for *Eubacteriumhallii*. There

are also significant univariate traits of $ASV95_Eubacterium$ in the same position and $ASV58_Eubacterium$on the same MEL. While not the same taxa, they could be identified as related, so it was marked. Lastly, if the trait had no overlapping taxa with univariate effects, it was marked with a "$-$". Traits that were labeled with N/As were not given a symbol. These traits did not have any taxa weights from the tables in Figure B.1, Figure B.2, and Figure B.3 that were outside the cutoffs identified as "large" weights. Among these polymicrobial traits, some display a more balanced distribution of weights across all included taxa, while others simply did not have any taxa that were more largely weighted compared to other related taxa.

We examined all significant traits within each MEL and noted any matches between those traits and a univariate taxon in the same subject's information. Of the 41 polymicrobial traits used as responses in the GWAS models, 29.27% contained at least one taxa overlapping with a univariate taxon on the same position. 19.51% of the traits had at least one of their taxa directly overlapping with a univariate taxon not in the same position but within the same MEL. Next, 4.88% of the points were marked with the "$\sim$" and had a related taxon identified as significant in another position within the same MEL. Combining these proportions, over half of the traits (about 54%) contained either a direct or related overlap with a univariate taxa on the same MEL. Only 12.2% (5 out of 41) of the significant polymicrobial traits did not have any of their largely weighted taxa overlapping with the univariate taxa. Lastly, 34.15% of the significant traits had the "N/A" specification. Due to a large number of weighted taxa, checking for overlapping information became tedious for these groups. Therefore, the values pertaining to overlapping information were not included in the specific counts mentioned above. Instead, our breakdown focuses on identifying the cases in which the largely weighted taxa overlapped. However, we could say that about 34% of the significant polymicrobial traits were found to indicate areas in

which the genome exhibited a statistically significant association with a more diverse community of taxa, which were identified as highly abundant in the microbiome of each subject.

Significant associations were found between the SNPs and PM traits "that did not have specific taxa driving the variation, indicating that polymicrobial traits can capture variation in microbial interactions that cannot be identified solely by the abundances of individual taxa" [59]. Specifically, the associations where the final PM trait is driven by multiple microbes together (as seen in the loadings and some of the biplots in previous sections) illustrate "how variation in *P. Vulgaris* lines can affect the overall topology of microbial communities" [59]. Additionally, across all microbiomes, an interesting relationship was identified between traits and a specific SNP location within MEL C. The significant traits included 30 taxa (including Bacteroides, Prevotella, and several butyrate-producing Lachnospiraceae and Ruminococcaceae, including Faecalibacterium), two short-chain fatty acids, and two of our polymicrobial traits. It was noted that "the apparent high pleiotropy of this locus (controlling multiple traits from multiple microbiomes) provided strong evidence that genuine variation at this locus indeed has a major effect on the microbiome"[59]. The two significant PM traits were created from the second loading of the hypothesis correlation matrix and the first loading from the genetic covariance matrix. A biplot comparing these two sets of loadings can be found in Figure 3.5, and their individual loading values can be found in Figure B.2 in the Appendix.

Subject 2 (S776) on MEL C shows overlapping associations between individual and polymicrobial traits, as indicated by hits from HypCorrPC2 (a trait heavily influenced by Faecalibacterium, Coprococcus3, and LachnospiraceaeUCG004) as well as from the individual taxonomic traits of Faecalibacterium and Lachnospiraceae. The association of various microbiome features from different subjects with the same

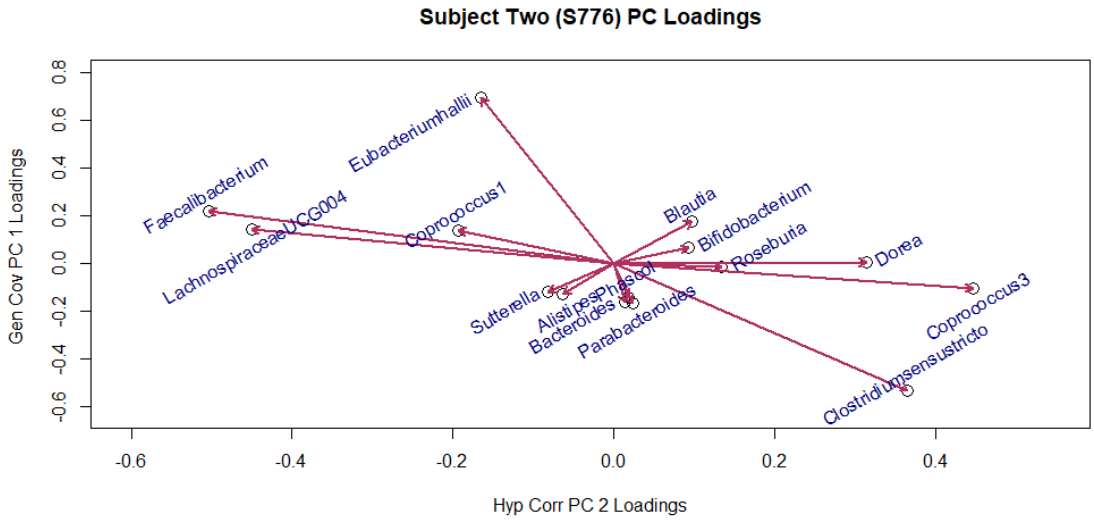


Figure 3.5: Biplot Comparing Significant polymicrobial trait loadings on MEL C

SNPs or MEL in the *P. vulgaris* genome strengthens the evidence for causality in the regions affecting the microbiome, which supports the utility of the PM traits [59]. These significant associations helped to identify genes related to glycyrrhizinate biosynthesis, where glycyrrhizin has been shown to have "anti-cancer, anti-inflammatory, anti-oxidative and antimicrobial" in some studies [59]. The PM traits on this MEL were driven mainly by some of the taxa directly related to these features, such as Lachnospiraceae, Faecalibacterium, and Clostridium. Lachnospiraceae and Faecalibacterium are bacteria that can produce butyrate in the gut, "a microbial metabolite with known health benefits in reducing susceptibility to colon cancer, gut immune homeostasis and gut barrier integrity" [59]. Overall, these methods can help identify important candidate genomic regions. Significant associations on each MEL with multiple microbiome features from each of the three subjects further support the conclusion that these MELs define areas with the most significant overall effect on the human gut microbiome.

3.7 Conclusions, Contributions, and Future Work

3.7.1 Summary

In this work, we assisted microbiologists in investigating the relationship between the consumption of different varieties of *P. Vulgaris* beans and the composition of bacterial communities in the human gut microbiome. The study aims to develop a methodology for quantifying and genetically analyzing human gut Microbiome-Active Traits (MATs) to map traits in crop plants associated with human wellness and disease predisposition. Our final results have outlined how we answered our goals from section 3.2, of investigating whether the natural decomposition of various covariance and correlation matrices could facilitate understanding how genetic variation across a diverse set of common bean lines is associated with the composition of microbial communities in the gut. Multivariate analysis methods analyzed the data outputted from the experiment designed from objective one. We then created sets of polymicrobial traits as multivariate community traits used alongside the univariate abundances in standard GWAS models. With the microbiologists, we identified areas on the *P. Vulgaris* bean genome, where our polymicrobial traits had significant associations, and discussed potential areas where the overlap was or was not associated with similar univariate traits.

Chapter 1 outlined the experimental design, and our goal was to identify appropriate statistical techniques to analyze the association between microbiome multivariate phenotypic information and specific bean races' genetic diversity and population structure. The study used multivariate analysis methods to investigate how genetic variation across a diverse set of common bean lines is associated with the composition of microbial communities in the gut of three subjects by analyzing data outputted from the simulated microbiomes. Our primary objective was to determine if utilizing

PCA and CDA methods on variations of covariance and correlation matrices would aid in determining the relatedness of the taxonomic abundances measured from simulated microbiome samples. Ultimately, we aim to show our results from this new technique could be incorporated into genome-wide association studies to account for the abundance information of phenotypes across a multivariate range of abundance values and highlight how the food we consume affects the gut microbiome's composition.

We developed polymicrobial traits that were incorporated as a novel approach in the genome-wide association studies accounting for the abundance information of phenotypes across a multivariate range of abundance values, potentially highlighting how the food we digest changes the makeup of the gut microbiome. To highlight community relationships, researchers wanted to establish how these new "traits" could "quantify interactions of P. vulgaris components with human gut microbiomes," and "provide an entirely new approach for mapping traits in crop plants that are associated with human wellness and predisposition to disease" [59]. The development of what we now refer to as "polymicrobial traits" acted as one of the several approaches to finding microbiome features for phenotypic traits in the GWAS models. Calculating these traits begins with modeling a multivariate response comprising a combination of the most abundant taxa across the well-plate samples. These traits were created by estimating variance components from a MANOVA model evaluating the relationship between the taxonomic abundance for the most abundant genera in each subject's data set. Because the types of taxa present in each subject were quite different, and future analyses would be done by subject, the multivariate set of taxa for the models was chosen separately for each subject. A variety of Sums of Square and Cross Products matrices were utilized from the MANOVA since the community structure may be described via taxa associations due to variation among the bean

lines. Once the response set was chosen, the data were analyzed to evaluate the variance across taxa explained by the difference between bean lines and an estimate of the genetic variance after accounting for the bean line covariance.

Dimension reduction analyses (PCA/CDA) from the MANOVA were used to create weighted linear combinations of the community of taxa selected within each microbiome. From the PCA and CDA results, biplots were produced, demonstrating a variety of community relationships within each subject. Discussions with Dr. Andy Benson and Dr. Mallory Van Haute provided background about the relationships among the taxa loadings. The biplots gave an initial picture of which taxa would provide more weight to the final calculations of each PM trait. Dr. Van Haute then used the biplot discussions and the loadings directly to highlight specific taxa driving the values of the PM traits. Lastly, in the GWAS results, we looked for areas on the genome with overlapping significant univariate and PM traits.

3.7.2 Polymicrobial Traits and overlap with Univariate GWAS

This new method sheds light on the significant associations between polymicrobial traits and variations in the *P. Vulgaris* genome, providing valuable insights for researchers. PM traits included in the GWAS models (with $H^2 > 0.1$) have associations between four and nine chromosomes and significant associations across the selected seven Major Effect Loci (MEL) areas on the chromosome. Out of 72 traits, 41 met the heritability cutoff and were used in the GWAS modeling. Of these, 18, 11, and 12 were significant in each subject, respectively. Symbols were added to the trait names to identify different categories, such as overlapping with univariate traits. Of the significant polymicrobial traits, 54% had an overlap with univariate taxa on the same MEL, while 12.2% did not overlap at all. About 34% of the significant polymicrobial traits were found to indicate areas in which the genome exhibited a statistically

significant association with a more diverse community of taxa. In general, we found that the vast majority of cases (36 out of 41) showed either a community-type effect or an overlap with univariate taxa that were also significant within the same MEL. This suggests that the polymicrobial traits we analyzed were able to detect the same associations as the univariate approach. However, we did observe five cases in which the polymicrobial traits did not overlap with any significant univariate taxa. It is possible that these cases picked up on a significant association that the univariate traits were not able to identify. For these polymicrobial traits, we recommend further investigation of the taxonomic relationships involved and a review of previous literature to determine if there is evidence for non-overlapping significant effects.

Results demonstrate that PM trait responses have several significant associations with different SNP markers across the chromosome. These associations reveal a correlation between the genetic makeup of beans and the bacterial taxa found in the microbiome samples of each subject, suggesting that bean consumption impacts the community composition of the gut microbiome. The analysis methods outlined in this chapter aided in describing how the combinations of bacterial taxa into one "phenotypic" summary can provide new information about how a taxonomic community structure relates to the *P. Vulgaris* genetic information. Then, together with both individual and combination phenotypic information, the results provide "tremendous implications for future application of this approach in genetic analysis in crop plants", including other types of species and genetic panels of bean lines, to more "comprehensively understand how genetic variation in the species affects the human gut microbiome" [59]. Our collaboration enabled researchers to demonstrate how data science can be utilized to measure intricate combinations of individual and group microbial organisms, which helps reflect the metabolic and ecological characteristics of microbiomes in response to changes in microbiome-active components found in crop

plants [59].

Adding the polymicrobial traits can help researchers in microbiology summarize how changes in diet play a role in changing the collective features of bacteria within the gut microbiome. In Dr. Van Haute's research, many of the traits selected for GWAS had one or more significant associations with locations across the bean genome. They found evidence that the "general construction of the bean genetics works together to affect variation in the feature results (including the PM traits), which in turn affects different health outcomes in individuals" [59]. Dr. Van Haute states this reflects many other studies that use microbiome features as phenotypic traits in a Genome-Wide Association study. In the gut, many taxa act together to affect different outcomes in the human body. Some are more beneficial microbes to human health performance, and others can produce undesirable health outcomes. The PCA and CDA loadings can demonstrate how the variability across the taxa shows when single taxa are more prevalent in the microbiome versus where the loadings are more evenly weighted across a smaller subset of taxa. Additionally, across subjects, we observed only a very small proportion of traits that had a large enough heritability to be included in GWAS for all three microbiomes, underscoring that the distinctiveness of microbiomes has an impact on the bacterial outcomes in the gut.

This study employs univariate and multivariate analyses of bacterial taxa to offer a more nuanced understanding of community diversity in the gut microbiome. The findings also underscore the importance of examining the interactions among different microbes, as these relationships can significantly influence community structure. Thus, the study highlights the need to shift from an exclusive focus on individual microbes to a more holistic approach considering polymicrobial traits. Significant findings containing the PM traits weighted by subjects of the overall taxa highlight the potential impact of variation in P. vulgaris beans on the overall microbe

community. The result of this current work helps to illustrate why univariate and multivariate analyses of the microbiome taxa can work together to create a more comprehensive description of the taxonomic variability in the subjects' gut measurements. Moreover, the significant associations of polymicrobial traits not driven by a single microbe illustrate how variation in diet can affect the overall topology of microbial communities.

3.7.3 Contributions

This work is a collaboration between statisticians and microbiologists, focusing on the design and analysis work presented in Chapters 2 and 3. Within each subject, the gut microbiome comprises an entire community of bacterial taxa, consisting of many individual taxa, and plays a pivotal role in overall human health. By considering these communities' statistical interdependence and correlation, researchers aim to measure the relatedness of bean genetics with the bacteria. Identifying ways to expand and diversify the gut bacterial community is crucial since more diverse microbiomes are associated with healthier individuals, while less diverse communities can be linked to health issues such as diabetes and obesity [90]. In addition to health outcomes, plant breeders can benefit from this research to develop bean lines that may potentially diversify the positively influential taxa in the gut.

The statistical methods presented in this chapter have been utilized across disciplines for many years. However, the benefits of this specific utilization are unique to this type of data and subject matter. The practical significance of this work stems from the collaborative efforts of a multidisciplinary team comprising microbiologists, food scientists, agronomists, plant breeders, and statisticians. Alone, this analysis of the taxonomic abundance data from the simulated microbiomes would yield little intriguing information. However, with background knowledge about the bean lines

of interest and results from combining these more basic statistical measures with the genome-wide association analyses, we can make broader interpretations of the relationships between bean genetics and the gut microbiome. Overall, this work can contribute to the establishment of "new methodology and approaches that enable studying MATs [microbiome active traits] as complex traits of crop plants", which can also lay the groundwork to allow food scientists to incorporate "human health-associated traits (MATs) into crop improvement programs" [59].

We highlighted two types of significant relationships for the polymicrobial traits utilized in the GWAS models. First, there were places where the polymicrobial traits appeared as a significant term on a section of the chromosome overlapping with a univariate trait that was a highly weighted value on the PM trait. Secondly, significant PM traits are not featured alongside any of the heavily weighted univariate taxa. Appearances of significant PM traits highlight two practically significant pieces of information for domain researchers. First, it emphasizes "the importance of using both individual taxonomic abundances as well as group measures of microbiome features to assess phenotypic variation in the microbiome"[59]. Secondly, associations not heavily weighted by one microbe "illustrate how variation in beans can affect the overall topology of microbial communities" [59].

Within the field, statistical advances can "provide insight to plant breeders on how MATs can be incorporated into cultivar improvement programs to breed for higher quality cultivars that have specific effects on the gut microbiome, ultimately changing the paradigm of current breeding programs" and in the future, ongoing advancements can facilitate the identification and creation of fresh approaches aimed at managing diseases by manipulating the microbiome [59]. In other work studying the effects of microbiome-active traits in Sorghum plants, researchers highlight that analyses of these traits can establish their effects on the gut microbiome and "pave

the way for use of seed traits with major effects on beneficial gut microbes as traits for crop improvement strategies that can have profound outcomes with respect to human health" [157].

Based on work from previous literature and discussions of contributions with collaborators, this work has provided a beneficial method for researchers that has yet to be utilized for similar microbiome experiments. The benefits are also practically relevant given the assistance the polymicrobial traits provide in combination with researchers' newer applications (AiMS phenotyping and MATs (Microbiome Active Traits)). The PM traits can combine with the experimental design from Chapter 2 to reduce and account for spurious variabilities. By incorporating design effects into the analysis and evaluating taxonomic community effects, it is possible to accurately identify the specific impact of dietary changes on gut microbiota and better understand their potential implications for human health. The ability to have pseudo-multivariate GWAS traits (as the polymicrobial traits) demonstrates how multivariate methodology shows similar trends to the univariate cases while emphasizing relationships that the univariate methods could not identify as significant. By using the pseudo-genetic variability matrices (the gencov and gen corr matrices), it is possible to take into account the genetic variability prior to carrying out the GWAS modeling procedures. Overall, to the knowledge of Dr. Van Haute and others assisting in her research, GWAS of the "human gut microbiome phenotypes have not ever been used to study effects on the gut microbiome as a "trait" of food crop plants," and so it has not been completed with these types of polymicrobial traits [59].

3.8 Limitations and Future Work

In order to provide microbiologists with a model that is both statistically accurate and easy to use, certain modeling and analytical decisions were made. These decisions may require further exploration to gain a deeper understanding of how our analyses can be improved as a method to create multivariate phenotypic traits. However, these choices were made deliberately to serve the focus of the current work. There may be opportunities moving forward with this research to modify certain aspects of this work to observe potential improvements or adjustments.

First, an arbitrary cutoff of taxa was selected for inclusion in the multivariate analysis for each subject. Based on conversations with Dr. Van Haute and Dr. Benson, there was already a reduced number of total taxa identified as possible traits of interest for the GWAS models. After that, we further set a cutoff of around an average 90% abundance for the taxa to be included in the final set matrix of responses for the MANOVA. Further research could be done to develop a cutoff for how many or which taxa should be utilized as the "community" of interest to analyze. Here, our focus was on the most abundant genera, but are there other taxa groupings that would be interesting to include in analyses to provide information about other gut microbiotic relationships? Is there a boundary at which there are too many taxa, and how many taxa will make the process of including the polymicrobial traits worth the effort? We could also investigate how variability related to different bean lines and taxa abundances alter the hypothesis and genetic covariance matrices.

The effect of adjusting zero values from the initial data set could also be evaluated for the data collection and transformation procedures. With some of the final taxa, no zeroes appeared, while a few taxa ended up 20 or more observations with zeros. In the univariate methods, a cutoff of 0.0015 was included. With this method,

values of 10E-8 were included, which could be driving some of the loadings toward the values with more zeroes. Adjusting the zero values to 10E-8 could be helpful to identify the effect sparse taxa of interest have on the community structure, or it could be skewing the results towards taxa that may not be valuable to the overall interpretation. To address this issue and enable researchers to prioritize taxa where this method would be most significant, the selection process for taxa in the final statistical analyses can be modified.

Also, with the multivariate models, there has been a discussion about the implications of different MANOVA modeling techniques. First, due to the nature of the current project, MANOVA models fit in PROC GLM dropped some design pieces included initially in Chapter 2 and included an overall simplified analysis focusing on the design batch, plate, and bean line group. To include additional variables of interest, MANOVA models could be extended to account for more attributes, including non-normality and adding fixed effects for rows and columns within plates. Secondly, there may be unaccounted implications of utilizing the MANOVA format of a PROC GLM model with an ANOVA format for the statistical modeling versus a stacked mixed model version of the multivariate format. Using a REML structure to estimate variability could lead to improvements in MANOVAs and PCA models, potentially enhancing the ability of these models to highlight community relationships among response taxa. One method attempted after the fact was to run models with the "stacked" multivariate abundances as a trick to fit the multivariate model using PROC GLIMMIX in SAS. A drawback to this is that computation time and convergence become issues to the modeling process as there is an increase in the number of taxa and parameters included in the model. It is possible that more accurate assessments of variability exist, which could aid in the determination of the "most optimal" principal component analysis by examining the connections between bacterial

abundances.

Selection of the variance methodology to use occurs after the MANOVA fitting. After accounting for different effects and variability within the MANOVA model, all outputs were used to estimate relationships between taxa. Methods utilized include the hypothesis covariance and correlation (for the bean line effect), a pseudo-estimate of the genetic covariance and correlation matrices, alongside two sets of coefficients from canonical discriminant analyses. Within each type of covariance matrix, scores were calculated and used as the polymicrobial trait for GWAS modeling. Since this work was exploratory, all polymicrobial traits were evaluated for use in the final GWAS models. However, more focus in the future could be devoted to assessing the best format across all methods utilized (PCA vs. CDA, covariances vs. correlations). Some of the PCA and CDA components gave similar results, but with more subjects' data or other taxa included, one may be better at identifying significant genetic relationships over the other. Research could evaluate the differences in how the taxa weighting changes between the methods and which are most consistent overall in highlighting scientifically relevant community effects.

Investigation of the differences in analyses across subjects could add another dimension, both statistically and practically, to the overall MANOVA models. Adding additional subjects to the study would provide a more comprehensive understanding of the various microbiomes, potentially leading to a more diverse interpretation of GWAS results that could more practically benefit researchers in the long run. The difficulty of having a subject effect within the analysis is that the community of taxa found in the sequenced genetic output differs widely from one person to another. Based on diet, lifestyle, and background (among other factors), the bacteria in a person's gut changes, and one set of taxa found in a subject could potentially not match someone else. A mismatch of bacteria found in the gut could mean that there

would be a reduction in the number of taxa included in the final response matrix for the MANOVA model. Microbiologists and statisticians would have to work together to determine how the interpretations would change for the results from a GWAS model containing polymicrobial traits calculated across different subjects.

The initial expansion of this work continues in the next chapter, looking at the combination of the design and analysis from Chapters 2 and 3. This work represents an important advancement in integrating experimental design and abundance analysis for studying microbiome activity traits. With more investigation into experimental design configurations, we can find the "best" estimation of variability within the MANOVA to use in the PCA to help find the most informative polymicrobial trait values. These processes have the potential to pave the way toward determining the most suitable design for simulating microbiome samples across well plates to generate highly informative estimates of the taxa's variability linked to the disparities observed between the studied bean lines (or other plant species). The new or more "optimal" designs could lead microbiologists to more informed decisions about where significant associations exist between the bean genome and the communities of bacterial taxa. In turn, these decisions could aid plant breeders in developing bean lines that could benefit gut health.

Chapter 4

Exploration of Optimal Design Strategies in Microbiome Experiments

4.1 Introduction

4.1.1 Background

Chapters 2 and 3 establish the need for updated statistical experimental design and analysis procedures in research examining the correlation between bean genomics and bacterial abundance in gut microbiome communities. We aim to assist researchers studying the subject matter in answering how bean genetics impact the gut microbiota community. This type of community structure can be described via taxa association due to variation among the bean lines. The overarching question becomes: how can researchers design their study to optimize the variability due to the taxa, bean lines, and random noise? Adapting the form of the experimental design could help investigate how the design can best account for the relationships between bean lines and what types of structures appear between all the different taxa. Accounting for various aspects of the data in the design can control extra variation in the experiment to reduce variability affecting the taxa associations. Then, the final analysis would be able to produce the best estimates of the bean line effects across the taxa abundances.

This chapter will look at how adjusting parameters in the design affect the outcomes in a multivariate analysis of the taxa. By obtaining more precise estimates of variabilities from a multivariate mixed model, as demonstrated in Chapter 2, we could improve our ability to quantify the association between changes in the bean genome and gut bacteria.

The statistical methodology of optimal design helps researchers systematically identify the combination of input variables to yield the best possible analysis outcomes while considering any constraints or limitations they may face with their available resources. An optimal design algorithm shares similarities with power analyses in that it aims to determine the optimal sample size by manipulating various design aspects to optimize a specific criterion function. Typically, this criterion function involves transformations of one or more information matrices. The remainder of this chapter details the key parameters that will be adapted and why given the experimental boundaries from Chapter 2. In this chapter, we discuss the algorithm and optimal design procedures used to evaluate the impact of new design structures and changes in variability across the taxa analyzed.

4.1.2 Motivation

This chapter addresses the third objective of this dissertation (from section 1.3): to identify and explore the effects of experimental parameters on statistical optimality criteria for analyzing data from microbiome well-plate experiments. Although the preceding discussions have focused primarily on applied statistical concepts in microbiology, this chapter aims to enhance researchers' experimental design skills by incorporating fundamental principles from statistical methodology. By combining these sources of information, novel methods can be developed to improve collaboration in studying the impact of plant consumption on diet and gut health. Examples in

this chapter primarily draw upon microbiome data, but the methodologies presented can be applied to other fields, including plant or animal breeding, which may also require multivariate models and design optimization techniques.

Objectives one and two highlighted simple and more direct methods of experimental design with Dr. Van Haute's specific aims and goals. Although more straightforward, the methods serve as a valuable starting point for understanding the essential information required for designing and analyzing an experiment. While considering the statistical advantages or disadvantages of a specific design, it is important to consider the ease with which researchers can implement an adaptable design structure. Montgomery wrote that some of the more desired properties to evaluate an experimental design are simplicity, cost-effectiveness, unbiasedness, and precision [97]. We would like this experiment and analysis to be a simple and cost-effective way to incorporate community effects of taxa prior to a GWAS analysis.

One way to enhance the properties of our design is by creating a more generalizable approach that allows for adjustments to be made to the analyses used to generate polymicrobial traits. Statistically, we explore how changes in our data and model can impact the optimality of the design. These adaptations are related to multivariate response models with multiple taxa, the number of bean line replications, variability structures, and the inclusion of genetically correlated effects. An experimental design that considers the correlations between bean lines due to their population structure can assist researchers in comprehending the effect of the chosen bean lines on variability outcomes from the chosen analysis. Equally important in this context is the ability to analyze the data in a multivariate fashion, which gives rise to an optimal design algorithm structure for a multivariate mixed model with variances estimated using REML (restricted maximum likelihood estimation) procedures. This unique exploration of optimal experimental design for multivariate

mixed models with a specific covariance structure, particularly in the context of genetic relationship matrices, constitutes a valuable contribution to both microbiology and statistical literature. Much of the information within the microbiology literature simply highlights the need for new adaptations of statistical design but does not compare or evaluate optimality criteria. While there have been numerous investigations into the optimal design characteristics of univariate models, relatively little attention has been given to the optimal design structures for multivariate models, and even less so for incorporating covariance structure [79].

After initially selecting an experimental design and analysis in the previous chapters, we now assess what improvements can be made to the design to more efficiently evaluate the relationships between the taxa abundances and bean lines. The format of the fixed and random design matrices can be altered to achieve a design with a maximized optimality criterion. We create an algorithm that uses different varieties of input data and design formations to estimate a criterion value to compare design structures. The algorithm uses variability-based criteria influenced by changes in the number of replications across plates and groups of bean lines. By adjusting the replications across bean lines within and across plates, researchers can observe how optimality responds to the different variability and interrelatedness of the bean lines and taxa covariance matrices. Based on initial evaluations of how changes in the data and design affect the selected optimality criteria, statisticians could make recommendations about what microbiologists should consider when designing their experiments and considering different taxa for analysis. As we have stated previously, this work stems from the growing necessity of collaborative work between statisticians and domain researchers. The motivation behind this work stems from the recognition of the importance of incorporating a more comprehensive statistical approach in experimental design selection for fields such as microbiology. Establishing principles for

optimal design in microbiological data analysis may enable a shift in focus towards expanding this work, developing and comparing design strategies based on their capacity to enhance comprehension of the community structure of taxa in relation to genetic changes in the food consumed within the gut. Given the advances in technology and the increasing availability of resources for researchers, there is a need for an updated and broader approach to experimental design that can better accommodate the complex nature of modern experiments.

4.2 Literature review

4.2.1 Overview

Casler used an analogy to convey the importance of creating effective experiments, likening it to the process of using a cookbook to find new recipes that fulfill one's needs while also returning to their tried-and-true favorites [21]. In fields such as microbiology, where the effectiveness of certain designs is not routinely evaluated and where conventional design and analysis approaches are prevalent, we can leverage techniques from other domains to enhance subject-specific methodologies. To create experimental designs that are effective, it is crucial to adhere to the principles of randomization, replication, and blocking. These principles are necessary for ensuring the reliability and accuracy of the experiment's results. Replication allows for estimating experimental error variance and ensures an appropriate number of replicates are used, resulting in precise inferences [98]. To improve our microbiome experiment, we could establish a more comprehensive replication structure within and across the 96-well plates and also collaborate with researchers more openly to customize designs to suit their data analysis needs. The review outlines statistical methodologies for creating optimal experimental designs and provides a backdrop for what has been done in

literature across statistics, plant and animal science, microbiology, and genetics. Simulations later in this chapter utilize design principles from multivariate models based on work completed by Kmail examining design effects on causal models [78, 79]. We will examine relevant literature that describes the selection of various metrics and techniques employed in the methods outlined in section 4.3 and subsection 4.4.2 and how this can help microbiologists design better experiments.

4.2.2 Background of Optimal Experimental Designs

Optimal design's influence extends to many areas of experimental design, and its' ideas have "become well established in the statistical literature as a fundamental tool for comparing designs" [140]. An extended look at the study of optimal design goes back to Kiefer in 1959 with a focus on response surfaces, but some research roots span back to 1918 for one-factor polynomial models [4]. In 1926, Fisher outlined the three critical components of experimental design: replication, randomization, and blocking, and introduced terms such as *factors*, *interaction*, and *confounded* for factorial experiments [116]. Those such as Atkinson, Box, Hotelling, Smith, Williams, Cullis, Yates, Fedorov etc., all contributed to the field and laid the groundwork for research like this to positively adapt the experimental design to answer today's complex research questions [5, 16, 63, 138, 9, 153, 158, 44]. The search for the best design often centers on finding a configuration that is "optimal" under a pre-specified linear model, chosen to closely align with the expected analysis [20]. To guide our search and comparison of designs, we will build upon the analysis completed in Chapter 3, incorporating adaptations that aim to enhance the methodological features for future use.

Optimal design of experiments as a sub-field within statistics involves the development of theory and methodology for constructing designs in various scenarios [12].

To plan an experiment and determine sample size, researchers with basic statistical knowledge may use standard and familiar techniques to calculate the appropriate number of replicate experimental/observational units using power analyses. The field of optimal experimental design goes beyond determining the number of replicates and sample size for statistical significance; it encompasses a range of sophisticated techniques that aim to maximize the efficiency and precision of various metrics derived from different statistical model parameters. The progress of statistical technology has broadened the impact of optimal design theory, prompting a renewed research focus on designs based on various criteria, often functions of a model's covariance matrices [78]. Federov described the two primary areas of experimental design as those "designs of extremal experiments and design of experiments for an elucidation of the mechanism of a phenomenon" [45]. Examples show that the first area focuses on when there is interest in assessing changes in optimality criteria based on changes in experimental conditions and factor levels. The second area focuses more on the domain experts interpreting a mathematical model to describe the factors being investigated.

R. A. Fisher was an early expert in and proponent of randomized designs in agricultural research, and he wrote that "if the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too" [48]. Thus, to acquire new knowledge and achieve meaningful results, it is crucial to employ a sound and suitable experimental design. The experiment conducted with Van Haute and Benson resembles the layout of an experimental design used in field trials. The wells in a plate can be compared to the plots in a field, and the bean line "treatments" are randomly assigned to different "plots". To account for all variability within and across well-plates, we consider them as experimental units, similar to how plots within a field are considered in agriculture.

To determine the optimal design, a set of potential designs must be identified

and evaluated using relevant criteria within the statistical model of interest to compare and select the most effective design. By utilizing this planning approach, the optimal experimental design adapts its structure and size to the specific analysis being performed, which may vary based on the statistical model utilized [119]. From the analysis in Chapter 2, we focus on a multivariate linear mixed model where we account for relatedness between taxa, bean lines, and plates as fixed and random effects. In the next section, background about different strategies for finding an optimal design is provided based on a multivariate framework as an extension of a more basic univariate framework. Section 4.2.3.2 will describe the literature pertaining to the methods utilized for searches conducted to compare experimental designs for adaptations of the design structure from chapter 2.

4.2.3 Optimal Design Methodologies

4.2.3.1 Multivariate Analysis Optimal Design Structure

Estimating community effects with the polymicrobial traits from Chapter 3 would require a search of different design effects on a multivariate model. In light of the previous chapter, it is crucial to shift our attention to multivariate models, given their effectiveness in estimating variability matrices for dimension reduction techniques, as demonstrated by our MANOVA. While the most straightforward approach for analyzing an optimal design involves univariate models, it is important to explore more sophisticated options to ensure robust results. Much of the current literature focused on univariate effects, creating a basis for expanding literature about multivariate models. This section will highlight areas of univariate and multivariate uses of optimal design and what ideas spawned the ideas for the optimal design simulations in this chapter.

Referring back to subsection 4.1.2, Montgomery described the following important aspects of an effective design: simplicity, cost-effectiveness, unbiasedness, and precision [97]. The increasing use of experimental design can be attributed to the absence of theoretical findings on the optimal design of blocked and split-plot experiments, leading to the popularity of computerized design algorithms and the development of algorithms for constructing optimal blocked designs [79]. The techniques discussed in this chapter for determining an optimal design mainly focus on evaluating the model's precision, as this directly impacts the variability considered in statistical analyses. Our objective becomes obtaining a design allowing for precise parameter estimates, "leading to more precise inferential statistics," also avoiding systematic error or bias [78, 79].

This work draws significant inspiration from Dr. Zaher Kmail's dissertation (*Optimal Design for a Causal Structure*) and publication (*D-Optimal Design for a Causal Structure for Completely Randomized and Random Blocked Experiments*). Both were crucial resources, as Kmail's structure for fitting causal inference models outlined the techniques used here to create a multivariate REML model investigation of optimality with varying taxonomic data and well-plate designs. Much of Kmail's work is based on research pertaining to the optimal design of blocked and split-plot designs from those such as Goos, Mylona, Vandebroek, Gilmour, and Jones. These works have introduced differing statistical model procedures, optimal design algorithms, optimality criteria, etc., all for univariate models [100, 99, 74, 28, 55]. This work aims to continue Kmail's expansion of these works in the multivariate field for use with microbiological data.

In 2019, Rodriquez-Diaz and Leon wrote about expanding univariate optimal design techniques to multivariate to account for multiple covariance structures. They aimed to show that specific structures can be added to the multivariate covariances

when considering cases when data is measured on the same experimental units over time (for instance) [123]. This work is similar to what we would like to show, however, the paper focused on analytical properties of the information matrices with the covariance matrices for only fixed effects in a bivariate model, using data examples with data related to processes of non-linear bacterial growth[123]. This work demonstrates the infancy of work on optimal experimental design for multivariate models, especially when considering unique covariance structures. Additional research is needed to address the computational complexity of optimal designs that incorporate comparable extended covariance effects, especially when dealing with intricate covariance structures[123]. Theoretical extensions into multivariate generalized mixed models have been investigated as well, deriving analytic solutions for A- and D-optimal designs with gamma-distributed outcomes [66].

Optimal design issues focusing on multi-response mixed effects models have received minimal attention in the literature [86]. Thus, if we could explore the behavior of simulation analyses in the optimal design of multi-response mixed models concerning microbiome data, we could contribute to statistics and the field of microbiology. This work aims to address the knowledge gap and bridge the divide between optimal statistical design and microbiological experimental design research by developing and examining effective strategies.

4.2.3.2 Optimality Criteria and Search Algorithms

Several optimality criteria exist to evaluate the impact of experimental design variations on specific statistical models of interest. Many criteria focus on the information matrices from fixed and random effects. Many useful optimality criteria in plant and animal studies are A and D optimality. A-optimality "seeks to minimize the average variance of elementary treatment contrasts," and D-optimality maximizes

the determinant of the information matrix [127, 20]. Although an issue with "highly D-efficient designs is that, generally, they do not involve sufficient replicates to allow for pure-error estimates" [99]. For the design of split-plot experiments, Sambo et al. mention that "the I-optimality criterion, which minimizes the average prediction variance, seems to be as appropriate as the D-optimality criterion and in some cases even more appropriate, for generating split-plot response surface designs" because the typical aim is making predictions [127, 75].

To explore and better highlight the differences between optimality criteria can demonstrate some examples of using a design matrix $\mathbf{X}$ on a simple univariate example. For a model

$$\underline{\mathbf{y}} = \mathbf{X}\underline{\beta} + \underline{\epsilon} \tag{4.1}$$

where $\underline{\mathbf{y}}$ is a vector of response values, $\mathbf{X}$ is our design matrix containing information about how the treatment values are arranged in a completely randomized design, β is the parameter coefficient estimates, and ϵ is the experimental error term. The least squares estimates for the parameter matrix β and the covariance matrix are outlined in Equations 4.2 and 4.3, and the information matrix M is found in Equation 4.4.

$$\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y} \tag{4.2}$$

$$Var(\underline{\hat{\beta}}) = (X'X)^{-1}{\sigma_\epsilon}^2 \tag{4.3}$$

$$M = {\sigma_\epsilon}^{-2}X'X \tag{4.4}$$

The following are summaries of different varieties of optimaltiy criteria highlighed by Kmail in 2019 [78]. A D-optimal design will minimize the criteria $|(X'X)^{-1}|$, equivalent to finding the maximum of $|X'X|$. E optimal designs minimize the worst possible variance of any contrast $\min_x[max(\lambda_i)]$. This criterion can be thought of

as the largest possible variance of any contrast based on the least-squares estimator [38]. A-optimality finds the minimum average variance of the parameter estimates $\hat{\beta}$ equivalent to $\min_x[tr(X'X)^{-1}]$ [24].

In recent years, research from combinations of Mylona, Gilmour, Jones, and Goos focus on evaluations of optimal designs with blocking and split plot components, with varying composite optimality criteria evaluating both fixed and random effect variance estimation [99, 100, 127, 75, 54]. In particular, a composite D-optimality criterion is proposed that prioritizes accurate estimation of the fixed treatment effects and the pure-error estimation of the variance components. Kmail discusses these works and their introduction of a composite optimality criteria, placing weights on both the fixed and random effect information matrices [79, 78, 100]. The new criteria is $\Phi = \frac{\alpha}{q} log|M| + \frac{1-\alpha}{2} log|N|$. Within the equation, α is a value between 0 and 1, weighting the "importance" of information from the fixed or random effects, and p is the number of (qualitative or quantitative) fixed effect parameters. For the fixed random effects, the value of M and N represent the fixed and random effect information matrices (respectively) from the REML estimation equations [78, 100]. Lastly, instead of another variable term on the weight of the random effect matrices, from Mylona et al. in 2014, a 2 was included because the model used was $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$ where 2 represents the variance components ${\sigma_\gamma}^2$ and ${\sigma_\epsilon}^2$ [100].

These modeling techniques are based on the univariate model, whereas, Kmail extended these ideas to causal models with multiple dependent endogenous variables, acting more similarly to a multivariate linear mixed model structure. Thus, he adapted this criterion into Equation 4.5[78].

$$\Phi = \frac{1-\alpha}{p+q} log|M(\hat{\delta})| + \frac{\alpha}{h} log|M(\hat{\sigma})| \tag{4.5}$$

Now, $M(\hat{\delta})$ is the information for a combination of exogenous and endogenous parameters, $M(\hat{\sigma})$ is an information matrix for random effect parameters, $p + q$ is the summation of fixed effects for the exogenous and endogenous parameters (with p independent and q dependent variables), with h added into account for the total number of variance components. Between the two criteria, the α weight is flipped, now with the α weight on the variance components, and more interest was on adjusting how much focus was placed on the included variance components. In this work, we employed the adapted version of composite optimality criteria as the primary methodology for determining the optimal design.

Next, we must consider prevalent search algorithms utilized to select a method for creating and comparing experimental designs. Given an exhaustive search of all possible structures for a design assigning treatments would require testing, possibly over millions or billions of entries. Kmail provides a detailed overview of different strategies for selecting a design and algorithmically evaluating a chosen criterion by iteratively modifying the design to identify the most optimal structure for both the fixed and random effects of interest [78]. Designs considered were Kiefer round-off procedure, the Fedorov algorithm, the Wynn-Mitchell algorithm, the Van Schalkwyk algorithm, the Mitchell algorithm, the modified Fedorov algorithm, and the combined Fedorov-Wynn-Mitchell algorithm [78]. However, Cook and Nachtsheim summarized and compared methods, and when efficient designs were of interest, they recommended the modified Fedorov algorithm[25]. Our objective was not to develop a novel algorithmic framework; given the available options and concerns related to computation time, we became interested in more simple and more versatile designs. It was also noted that Cook and Nachtsheim only described the relative performance of the seven algorithms in comparisons and did not make any conclusions on an algorithm's ability to find an optimal design [25].

Though many univariate optimal design algorithms have been developed for univariate mixed models, research is limited on optimal multivariate design [79]. Following the methodology from Kmail, the Modified Federov Algorithm from Cook and Nachtsheim was first investigated. The Modified Federov Algorithm, also known as the Simultaneous Switching algorithm, begins by using an initial design with the same structure and size as the boundaries of the available resources. To assess and compare the optimality of each design, the algorithm iteratively compares the initial design matrices with a secondary design matrix option, with one candidate point exchanged for one corresponding row. Then, the design associated with the better value is kept, and a new candidate point is exchanged into the updated design structure. As the optimality improves to a small change, delta, the design matrix matching the smallest optimality value will be saved as the best design.

Literature into univariate extensions of optimal design search algorithmic procedures for multivariate problems is limited. Even in univariate cases, it is noted that with specific algorithms and computing power, even more sophisticated approaches may be unable to find globally optimal designs [69]. When extending to multivariate cases for this style of microbiome data, the number of observational units in the design increases, larger than many of the studies described in this section and later sections on research for models in agricultural studies. Cook and Nachtsheim outline in their work that for large N "little is gained with the use of an exchange algorithm" [25]. Some have looked at extending these more typical "exchange-point" methods for a larger data set, however, when dealing with extremely large problems, even the use of exchange algorithms can become impractical due to the high computational costs involved [126]. A current challenge in optimal design for larger data sets is the lack of computational optimization methods that can efficiently generate targeted sample sets in a time frame comparable to that of a randomized sampling strategy while

incorporating relevant design criteria [34]. Many exchange algorithms begin by selecting a random design to start the exchange of points. For this work, the algorithm described in section 4.4.2 will utilize a random search methodology to structure the exploration into features design optimality.

4.2.4 Related Applications of Statistical Experimental Design

4.2.4.1 Statistical Methods Accounting for Genetic Effects

The typical multivariate methodology described above may not account for the correlation and relationship between our bean line treatments. From our collaborators, we know kinship information exists for the bean lines, and this genetic information can often be used as additional information in statistical models. We can utilize specific variation estimation techniques to consider these kinship covariances. Our optimal design simulations drew upon a couple of works that provided context for the incorporated relatedness between genetic effects. Over the past two decades, there has been a growing interest in experiment design models that incorporate kinship matrices as correlated random effects, including research on partially replicated (p-rep) designs that account for experiments conducted across different environments [117].

This research approach aims to estimate the variability of covariances among bean lines, experimental errors, and genetic factors in the data rather than focusing on treatment means or differences. These variabilities are useful in the same way as those estimated from MANOVA when analyzing the polymicrobial traits. Our approach to incorporating genetic information into the optimal design framework is based on techniques outlined in *Variance Components* by Searle, Casella, and McCulloch [131]. Although the text is older, it is a valuable resource for future

research. Zhou and Stephens' 2012 paper discusses this text as a foundation for analyzing efficient genome-wide mixed model techniques, as discussed in Chapter 3 [164, 163]. The methodology is widely employed in genetics research, such as for improving the power of GWAS models to assess phenotypic traits in a randomized complete block design using multivariate linear mixed models, predicting microbial temporal dynamics by modeling variability over time, and designing experiments in plant and animal breeding based on available pedigree information [136, 23, 29]. Related works in plant science have presented methods with the same univariate-based LMM methodology aimed at optimal design for arranging genotypes within and across fields based on their kinship information [46].

These resources exemplify how to set up and estimate covariance components in a stacked multivariate model, accounting for correlations between different random factors. Geneticists can integrate population kinship or pedigree covariance components as estimates of these relationships between the genetic effects of interest [131]. The FarmCPU model used by Van Haute fits the total genetic effects as random with included variance and covariance structure identified by the kinship relationship [59, 85]. When analyzing the common bean and gut microbiome data, considering the population-based genetics of the bean lines revealed significant relationships between the landrace and market class of the beans and the active microbiome traits [59]. This finding demonstrates that genetic relationships in the bean lines drive the gut microbiome's bacterial abundance. Inherent connections between the bean lines must be considered in future optimal design and statistical analyses.

Furthermore, for the proposed simulation-based research methodology in section 4.4, we utilize pieces from the following model from Searle et al. [131]. Searle et al. provide an example of the benefit of adding in the genetic structure when analyzing piglets' weight and body length measurements, using a correlation struc-

ture on the random effect to account for relationships between the two related traits. From a univariate perspective, with data from our microbiological background, we can imagine that one bean line is being measured with two taxa being outputted in the final microbiome sample. Additionally, relatedness between multiple bean lines will be included, but a simple example is used to start.

For the variables, y_i represents the two related response measurements, with the overall mean effects μ_i multiplied by an intercept vector of ones. The α terms represent the effect of genetic relationships between the two variables, with a design matrix for the random effects Z, where $\mathbf{Z_1} = \{_{\mathbf{d}}\mathbf{1}_{\mathbf{n_i}}\}_{\mathbf{i=1}}^{\mathbf{a}}$.

$$y_1 = \mu_1 \mathbf{1_N} + \mathbf{Z_1}\alpha_\mathbf{1} + \mathbf{e_1} \tag{4.6}$$

$$y_2 = \mu_2 \mathbf{1_N} + \mathbf{Z_2}\alpha_\mathbf{2} + \mathbf{e_2} \tag{4.7}$$

In the piglet example, this could be a sow effect, for a microbiome example, it is the bean line kinship effect. This model can then be vectorized and rewritten in the following matrix form.

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1_N & 0 \\ 0 & 1_N \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{4.8}$$

$$\rightarrow y = \mathbf{X}\beta + \mathbf{Z}u + e \tag{4.9}$$

Each corresponding vector or matrix is tied to its similar piece in the matrix-style linear model, which could be fit with typical linear mixed model methodology [131, 143]. The corresponding variances can be estimated with G and R matrices for the random effect and error variabilities, in Searle, the random effects are associated

with a letter D rather than G. The residual estimates for the error will be estimated in the typical manner where...

$$var(\mathbf{e}) = \mathbf{R} = \begin{bmatrix} \sigma^2_{e_1}\mathbf{I_N} & \tau_e\mathbf{I_N} \\ \tau_e\mathbf{I_N} & \sigma^2_{e_2}\mathbf{I_N} \end{bmatrix} = \begin{bmatrix} \sigma^2_{e_1} & \tau_e \\ \tau_e & \sigma^2_{e_2} \end{bmatrix} \otimes \mathbf{I_N} \tag{4.10}$$

Now a represents the number of included random effects in Z, and $\tau_\alpha = cov(\alpha_{1i}, \alpha_{2i})$ the covariance that exists between related measurements from the same sow in the piglet example, or the bean lines population covariances from our taxa abundance example. However, instead of utilizing the same basic structure for the model random effects, there is the addition after the arrow of an A kinship relatedness matrix in Equation 4.12.

$$var(\mathbf{u}) = \mathbf{D} = \begin{bmatrix} \sigma^2_{\alpha_1}\mathbf{I_a} & \tau_\alpha\mathbf{I_a} \\ \tau_\alpha\mathbf{I_a} & \sigma^2_{\alpha_2}\mathbf{I_a} \end{bmatrix} = \begin{bmatrix} \sigma^2_{\alpha_1} & \tau_\alpha \\ \tau_\alpha & \sigma^2_{\alpha_2} \end{bmatrix} \otimes \mathbf{I_a} \tag{4.11}$$

$$\rightarrow \begin{bmatrix} \sigma^2_{\alpha_1}\mathbf{A} & \tau_\alpha\mathbf{A} \\ \tau_\alpha\mathbf{A} & \sigma^2_{\alpha_2}\mathbf{A} \end{bmatrix} = \begin{bmatrix} \sigma^2_{\alpha_1} & \tau_\alpha \\ \tau_\alpha & \sigma^2_{\alpha_2} \end{bmatrix} \otimes \mathbf{A} \tag{4.12}$$

Searle introduced this method as a methodological way to utilize genetic covariances "which must be taken into account in estimating the variance component of $\sigma^2_{\alpha_1}$" [131]. The A matrix is a matrix measuring genetic relatedness, and in this case, it is treated as a random effect coming from an overall population [131, 60]. Incorporating the A matrix is an effective approach to account for the population structure of relationships among the bean lines used in the optimal design modeling algorithm. To obtain more precise variance estimates in the final analysis, we plan

to incorporate a multivariate model structure that evaluates community effects for the taxa, taking into account the genetic community relationships within the bean lines as an additional measure of the effects of genetic variability. Searle provides REML modeling formulas for the multivariate model, closely aligning with Kmail's approach. These formulas will be the foundation for developing the optimal design simulation strategy described in section 4.3.

In designing our analysis, we drew upon literature that referenced Searle and explored the development of efficient or optimal designs for plant and animal genetics. Related results from Cullis, Smith, Piepho, Williams, and others establish methods not currently used in microbiological research [27, 138, 113, 112]. However, our work could bridge the gap and extend more statistical thinking about improved designs into microbiome studies. In 2020, Piepho, Vo-Thanh, and Tobias wrote about the processes of generating experimental designs for genetically related treatments, emphasizing agricultural sciences research. The authors note that in the agricultural sciences, classical designs are commonly used, relying on established search methods and statistical software programs to identify appropriate designs, typically based on univariate linear models with blocking and treatment fixed effects, assuming independent errors with constant errors variance [112].

As genetic testing provides kinship information, plant, and animal breeders are increasingly interested in incorporating it into their breeding programs, leading more researchers to explore the use of linear models, including their correlated treatment effects, in their experiments [112]. This is especially true when the statistical methods of interest revolve around genome-wide association analyses focused on best linear unbiased prediction (GBLUPs). Piepho et al. [112] utilizes SAS PROC OPTEX, a SAS procedure specifically designed for creating experimental designs and randomizations, to establish a design generation schema with correlated treatment effects.

To augment their findings, the authors present examples from Bueno Filho's research, which provide additional resources for this work [112, 28].

Bueno Filho and Gilmour offer valuable insights into the impact of genetic relatedness on experimental design, including potential outputs from simulations that can facilitate the comparison of designs. Their work specifically focused on the design and modeling of animal breeding trials, which was a notable contribution to the field, given the limited research on the application of optimal design for breeding and selection at the time of their publication in 2003 [28]. To find efficient methods for selection and breeding, authors modeled genetic breeding values and random effects and included relationship matrices with LMM methodology estimated with REML, similar again to work from Harvile, Searle, and Patterson and Thompson [28, 58, 131, 110]. The modeling techniques and accompanying data offer a distinct perspective on potential comparisons between optimality criteria, particularly as we modify and assess various variability scenarios in our simulations.

This work deals with univariate LMMs, but extensions later will be made to utilize these ideas with a multivariate mixed model structure. The authors use summaries of the established genetic and error variances in terms of heritability, also used in Van Haute's work, to select traits for inclusion in the GWAS models.

$$y = \mathbf{X}\beta + \mathbf{Z}\tau + \epsilon, \tag{4.13}$$

$$\text{where } E(y) = \mu,\ V(\beta) = \mathbf{B},\ V(\epsilon) = \mathbf{R},\ V(\tau) = \mathbf{G}, \tag{4.14}$$

$$\mathbf{R} = \sigma^2\mathbf{I},\ \mathbf{B} = (\sigma^2_\beta)\mathbf{I} \tag{4.15}$$

$$\mathbf{G} = (\sigma^2_\tau)\mathbf{A} = \frac{\sigma^2}{\gamma}\mathbf{A} \tag{4.16}$$

In the model above, $y_{n\,x\,1}$ is a vector of responses, $X_{n\,x\,b}$ is the design matrix for the fixed block effects, $\beta_{b\,x\,1}$ is a vector of unknown block means, $Z_{n\,x\,t}$ is the matrix

for block assignments of treatments, $\tau_{t\ x\ 1}$ is the vector of unknown treatment effects (uncorrelated with β), and $\epsilon_{n\ x\ 1}$ is the vector of random errors. Lastly, $\gamma = \frac{(\sigma^2)}{(\sigma_\tau^2)}$, where $\mathbf{A}$ is a pedigree matrix for the correlations between the genetic treatment random effects. With this model, we can define heritability based on estimates of genetic effects τ as $h^2 = \frac{(\sigma_\tau)^2}{({\sigma_\tau}^2+\sigma^2)}$ [28]. These values will serve as a valuable measure for comparing designs, especially for researchers interested in the genetic effects of treatments. By employing a multivariate extension of the univariate models presented in this study, we will investigate how modifications to the variability matrices impact optimal outcomes, thereby enabling us to make comparisons between designs.

After estimating the optimality of their models with various experimental design structures for six treatments in four blocks of 3, they compared optimality criteria versus different inputs of γ values. Filho and Gilmour compare design optimality criteria values for design structures both known to be optimal and subpar for experiments with unrelated treatments [28]. Through an analysis of design optimality, Filho and Gilmour made the following observation: for a given pedigree structure, the designs that were known to be optimal for fixed treatment structures and criteria such as A-optimality tended to perform better than suboptimal designs for unrelated treatments regardless of heritability; This trend was observed across different values of γ, illustrating its consistency [28]. These outputs allowed authors to make statements about how optimality is affected by the various relationships between the genetic and error variances. For example, they found that for large responses and smaller heritability, "the relative importance of the information on the individual phenotype starts to be smaller" [28]. The authors emphasized the significance of including the genetic relatedness matrix when searching for optimal blocking structures in these cases applicable for selection purposes in plant and animal breeding [28]. These findings could be useful when expanding the multivariate mixed model in our work to

include the genetic relationship matrix between the bean lines as treatments across plates.

Similarly, Paget et al. observed how p-rep designs, like those explored in Chapter 2, could allocate replicates in potato breeding experimental designs [107]. Heritabilities were also utilized within their modeling for design comparisons, and results indicate, based on empirical trials and simulation, that p-rep trials increased genetic gain for yield and overall selection efficiency, and additional advantages are brought by extending experiments across sites [107]. Other work expanding that of Bueno Filho and Gilmour continues to use both broad and narrow sense heritability to evaluate the genetic information within the design of experiments in fieldwork with correlated treatments [20] Related work more recently has found that it is generally better to avoid replicating genotypes within and across trials and instead include as many genotypes as possible in the overall design [121].

4.2.4.2 Applied Multivariate PCA Considerations

Having combined multivariate linear model theory with a principal component aspect in our previous analysis, we can explore different uses of PCA models for evaluating design, power, and replicability. Research in this subject area has evaluated sample size replication, focusing on estimating power to assess relationships between PCA scores derived from phenotypic data analyses and SNPs [76]. Some include a specific PCA model for the genetic variability, called a principal component of heritability[11]. Traditional analysis techniques like Principal Coordinates Analysis (PCoA) or PCA alone are inadequate for handling the unique properties of microbiome data, such as sparsity and heterogeneity[26].

Joint analysis of multiple phenotypes in genetic association studies has gained interest due to the demonstrated power gain, and PCA has been widely used for

dimension reduction in this type of analysis, as it can identify significant top SNPs missed by individual phenotype association tests, as shown in previous studies [88, 89, 3]. Specifically, Liu and Lin have proposed PC-based association tests for multiple phenotype studies using summary statistics from GWASs [88]. Although PCA is commonly used in GWAS, there is limited theoretical analysis on when it is advantageous or disadvantageous to use in analysis, particularly in multiple phenotype settings versus multiple SNP settings, and inappropriately using PCA in such situations could lead to a potential power loss [87].

Liu et al. presented an interesting summary and evaluation of the growing interest among researchers in exploring the potential for increased power in detecting genetic associations with a SNP through combining multiple phenotypes, which aligns with our interest in multivariate modeling techniques [87]. The authors focus on a regression setting based on hypothesis testing scenarios. While they do not use the same statistical techniques, they also showed that "correlation structures either among a group of SNPs or among multiple phenotypes play an important role in the performance of each test"[87]. Although it may not be feasible to incorporate at present into the optimal design outlined in subsection 4.4.2, it is worth noting that in multiple phenotype association studies, higher-order PCs with small eigenvalues are generally favored, while in the SNP-set setting, lower-order PCs with large eigenvalues are preferred concerning power [87]. The connection between this research and our optimal design objectives lies in our belief that the relationships between multiple phenotypes and their corresponding genetic factors will impact the variability produced by statistical analyses. There is continuing interest in the scenario where the association between multiple phenotypes and SNP explanatory variables is sought, but Lin and Liu's modeling is not readily adaptable for this purpose, necessitating further research to enable the simultaneous analysis of multiple phenotypes and SNP

sets [87]. Hence, there is a pressing need in the field to operate within the realm of multivariate phenotypic analyses involving a significant amount of exploratory data, similar to a large number of bean lines in our design.

4.2.4.3 Microbiological Studies

The use of microplates is essential for conducting biomedical experiments, however, the resulting data and quality metric values can be significantly influenced by the specific placement of the samples within the plate [122]. Although several plate layout editors are available, including Brunn, Labfolder, PlateDesigner, and PlateEditor, none can generate effective layouts, despite some creating randomized ones [2, 82, 31, 145]. Rodriguez et al. found that a common data design and collection approach is manually assigning treatments to wells and performing multiple technical and biological replicates. However, these procedures are often associated with higher costs and longer experiments and can lead to a trade-off between the number of samples analyzed and the number of replicates per sample. [122]. If a study's underlying design is flawed due to experimental design problems, such as insufficient statistical power or failure to account for confounding variables, any attempts to test the robustness of a result will also be flawed, posing a risk to replicability [129]. Overall, investigations of the microbiology literature demonstrated that studies about optimizing experimental design for data analysis of microbiome abundance data are rare.

While research in the microbiome has not focused on optimal experimental designs, many within the last several years have begun highlighting current work and best practices for considering the integration of experimental design into microbiome research studies. Numerous collaborative reviews have been published that identify the prevailing methodologies employed in microbiome research and suggest ways to

adapt these approaches for future studies [129, 47, 14, 133, 91, 81, 30, 26, 68]. Despite the potential benefits of using more basic power analyses to assess sample size calculations for replication in -omics research, their utilization remains limited, partly due to the complexity of multivariate data, which poses challenges when implementing such analyses [133, 47]. Outlining values for optimal design or power analyses can be difficult as small variability estimates between similar biological replicates pose a challenge in identifying weak biological signals, especially with small or unknown effect sizes; Small sample sizes may not accurately represent population-based outcomes, thus selecting appropriate sample sizes based on statistical principles is crucial to avoid biases and misinterpretations [14]. To obtain more robust evidence of an effect and determine the optimal amount of a dietary component, Jarett et al. suggest testing multiple levels of the factor of interest with an adequate number of animals, as this approach can produce more convincing results [68]. The concepts discussed in the microbiological literature underscore the importance of updated experimental design standards and procedures. Subject matter scientists would welcome incorporating measures that account for variability and genetic effects to develop more effective designs.

Furthermore, researchers have examined univariate and multivariate analyses to assess the selection accuracy and estimate genetic parameters in common bean. These studies have demonstrated the effectiveness of using several designs of genetic correlations and heritabilities. Although using multiple traits has been commonly employed in animal and forestry crop breeding, this approach has been relatively lacking in annual crops such as common bean [7]. Background information about this work is noteworthy, as it sheds light on the analysis methods employed and the similarities between the data type used and genetic information on the common bean. When evaluating multiple versus one trait analyses, authors noted that when response

traits have smaller genetic correlations, univariate and multivariate analyses yielded comparable results of parameter estimates and predictive accuracy [7]. It is worth emphasizing that adjusting the correlation and relatedness of taxa and bean lines can be altered using optimal design simulation strategies, underscoring the importance of such processes. This work is also notable because they utilized the REML-BLUP analysis framework similar to our extension of Kmail's work and background from Searle and Henderson [131, 60, 78]. A plausible hypothesis is that increasing the number of highly correlated traits may result in greater statistical and practical benefits from adjusting the overall design.

4.3 Application of Multivariate Optimal Design

4.3.1 Overview

Our third primary objective, as outlined in section 1.3, is to evaluate how varying data and experimental parameters impact the optimal statistical design criteria for comparing data and models to analyze microbiome well-plate experiments. To accomplish this, we can consider several factors, including aspects of the experimental design, data variability, and optimality criteria, and use them as parameters for evaluation in an algorithmic optimal design procedure. By modifying these parameters, we aim to evaluate the advantages of adjusting the experimental design and structure for this type of analysis. Furthermore, this study establishes a foundation for comparing the optimality of different designs that consider both fixed and random parameters and account for the genetic relationships between treatments and correlations among multivariate responses.

Due to time constraints, we could not evaluate all possible parameter combinations in this work. However, several design elements were modified and compared to

measure optimality. For instance, our initial experiment involved 12 plates for each of the three subjects, but the number of plates could vary across designs. Additionally, we could modify the arrangement of bean lines within and across plate plates or determine the optimal number of bean lines to test. Although we arranged the bean lines in an incomplete block structure across plates, we could explore whether a complete block design would be more optimal. The optimal design may also change based on the taxa composition in the response matrix and how many taxa are included in the final multivariate response matrix. With the statistical model, the idea is to find the best configuration of the design matrix X based on a specific optimal design criterion, which can often be calculated by "minimizing a function of the variance-covariance matrix of the least – squares estimator" [79]. We will examine the information matrices related to fixed and random effects, providing flexibility in the algorithm based on the researchers' interests. Before finalizing a proposed methodology, the following sections will outline how we set up the statistical methods to conduct simulations estimating optimality criteria and which parameters became a focus for this work.

4.3.2 REML Expansion for Multivariate Models

Following Kmail [78, 79], who utilized notation similar to that of Durbin [36] and Searle [131], we outline the modeling structure for the optimal design of a similar set-up as what was designed in Chapter 2. The analysis structure focuses on a REML univariate structure for stacked multivariate data. The analysis will be simplified to account for the fixed effects of all plates across the experiment, with random effects for bean lines, more similar to what was conducted in Dr. Van Haute's GWAS analyses from Chapter 2.

Zaher begins with the following equations for a multivariate mixed modeling

format.

$$\mathbf{YB} + \mathbf{X}\gamma = \mathbf{ZU} + \mathbf{E} \tag{4.17}$$

where all notation is similar to all linear mixed modeling format, except for the addition of the $\mathbf{B}$ matrix, which can be adapted for contrasts on the endogenous variables in a causal structure, and γ is used as a parameter for the exogenous, fixed, effects. $\mathbf{Y}_{n\,x\,t}$ contains our transformed abundance estimates, with n observations across t taxa. Then $\mathbf{X}_{n\,x\,p}$ is the design matrix for the fixed effects of p plates, while $\gamma_{\mathbf{p\,x\,t}}$ could contain the parameter estimates for the plate effects. $\mathbf{Z}_{n\,x\,u}$ and $\mathbf{U_{u\,x\,t}}$ are the design matrix and parameter estimates for any random effects included, in our case, $u = 1...bl$ where bl will be the number of bean lines included in the design. $E_{n\,x\,t}$ contains all random error effects across all observations and taxa.

If we set $\mathbf{B} = \mathbf{I}$ and $\gamma = -\beta$, our model will begin to resemble a more typical mixed model format, and we can adapt the equations for a multivariate mixed modeling format. By using the rules of transposing matrices, we can reformat the equation in the following manner.

$$\mathbf{YI} - \mathbf{X}\beta = \mathbf{ZU} + \mathbf{E} \tag{4.18}$$

$$\rightarrow \mathbf{I'Y'} - \beta'\mathbf{X'} = \mathbf{U'Z'} + \mathbf{E'} \tag{4.19}$$

$$\rightarrow \mathbf{IY'I} - I\beta'\mathbf{X'} = I\mathbf{U'Z'} + \mathbf{E'} \tag{4.20}$$

Then, to obtain the multivariate model for REML mixed model variance estimation, we used vector notation and Kronecker products to stack the multiple taxa variables into a univariate format.

$$Vec(IY'I) - Vec(I\beta'X') = Vec(IU'Z') + Vec(E') \tag{4.21}$$

Then, using $Vec(ABC) = (C' \otimes A)Vec(B)$, we can say that Equation 4.21 can become Equation 4.22.

$$\rightarrow (I \otimes I)Vec(Y') - (X \otimes I)Vec(\beta') = (Z \otimes I)Vec(U') + Vec(E') \tag{4.22}$$

$$Vec(\mathbf{Y}') = Vec(\beta'\mathbf{X}') + Vec(U'Z') + Vec(E') \tag{4.23}$$

Next, assuming $Vec(AB) = (A \otimes I)Vec(B) = (B' \otimes I)Vec(A)$, we can adapt Equation 4.23 into Equation 4.24.

$$Vec(\mathbf{Y}') = (X \otimes I)Vec(\beta') + (Z \otimes I)Vec(U') + Vec(E') \tag{4.24}$$

Equation 4.25 contains descriptions outlining how each part of the stacked model will follow into a univariate structure.

$$\underbrace{Vec(\mathbf{Y}')}_{\tilde{y^*}} = \underbrace{X \otimes I}_{\tilde{x^*}}\underbrace{Vec(\beta')}_{\tilde{\beta^*}} + \underbrace{Z \otimes I}_{\tilde{Z^*}}\underbrace{Vec(U')}_{\tilde{u^*}} + \underbrace{Vec(E')}_{\tilde{\epsilon^*}} \tag{4.25}$$

Which can be rewritten as our final model outline in Equation 4.26.

$$\tilde{y^*} = \tilde{\mathbf{X}}^*\tilde{\beta^*} + \tilde{\mathbf{Z}}^*\tilde{u^*} + \tilde{\epsilon^*} \tag{4.26}$$

Then if we let $Var[Vec(U')] = A \otimes \Sigma_u$ and $Var[Vec(E')] = I \otimes \Sigma$, the remaining pieces of the model can be addressed. In these variances, A represents our kinship relationship matrix for the bean lines included in our random effects. This is directly tied to what was outlined in the literature from Searle [131]. The multivariate residual format also follows from the format of Kmail and Searle as well.

Therefore,

$$\begin{pmatrix} \tilde{\underline{u}} \\ \tilde{\underline{\epsilon}} \end{pmatrix} \sim N\left(\begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}, \begin{bmatrix} A \otimes \Sigma_u & 0 \\ 0 & I \otimes \Sigma \end{bmatrix} \right) = N\left(\begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix}, \begin{bmatrix} \tilde{G^*} & 0 \\ 0 & \tilde{R^*} \end{bmatrix} \right) \tag{4.27}$$

Where $\tilde{G^*} = A \otimes \Sigma_u$, $\tilde{R^*} = I \otimes \Sigma$, $\Sigma_u = \begin{pmatrix} \sigma^2_{u_{11}} & \sigma_{u_{12}} & \dots & \vdots \\ \sigma_{u_{12}} & \sigma^2_{u_{22}} & \dots & \vdots \\ \vdots & \dots & \ddots & \sigma_{u_{1t}} \\ \sigma^2_{u_{1t}} & \dots & \dots & \sigma^2_{u_{tt}} \end{pmatrix}$,

and $\Sigma = \begin{pmatrix} \sigma^2_{11} & \sigma_{12} & \dots & \sigma_{1t} \\ \sigma_{12} & \sigma^2_{22} & \dots & \vdots \\ \vdots & \dots & \ddots & \sigma_{1t} \\ \sigma^2_{1t} & \dots & \dots & \sigma^2_{tt} \end{pmatrix}$.

Following information from Kmail Chapter 4 discussion, the full model was created with the equations for the P and V matrices and the information matrices for the fixed and random components. In work for evaluating the causal structure, the model has information for a B matrix, which adds contrasts for evaluating relationships in the endogenous variables. In this case, that could be added to our Y response matrix. However, after cutting that out, what results are above, and all of the equations are simplified with B as the identity matrix.

Following the REML structure for a univariate model, $\tilde{y^*} = \tilde{\mathbf{X}}^* \tilde{\beta^*} + \tilde{\mathbf{Z}}^* \tilde{u^*} + \tilde{\epsilon^*}$, we needed to set the remaining model structure to obtain both estimates of the fixed and random information matrices $M(\hat{\delta})$ and $M(\hat{\sigma})$ as described in Equation 4.5. For the fixed effects, $M(\hat{\delta})$ for our model takes the standard format of,

$$M(\hat{\delta}) = \tilde{\mathbf{X}}^{*\prime} \tilde{\mathbf{V}}^{*^{-1}} \tilde{\mathbf{X}}^* \tag{4.28}$$

where $\tilde{\mathbf{V}}^* = Var(\tilde{y^*}) = Var(\tilde{\mathbf{Z}}^*\tilde{G}^*\tilde{\mathbf{Z}}^* + \tilde{\mathbf{R}}^*)$.

Again, following the typical univariate mixed model methods protocol, similar to Stroup, Searle et al., and Kmail [143, 131, 78], we also know that the information for the random effects can be calculated with Equation 4.29. In this format, d and r correspond to the different random effects and residual variance effects to be estimated in the model. These will be tied directly back to the size of our Σ_u and Σ matrices from Equation 4.27. For a three taxa example, the matrix $M(\hat{\sigma})$ will be 12 rows and 12 columns for each of the six variances and six covariance values to be estimated on the upper (or lower) diagonal of the symmetric matrices.

$$M(\hat{\sigma}) = \frac{1}{var(\hat{\sigma})} = \\ = \frac{1}{2}\begin{bmatrix} \{\underset{m}{tr}(\tilde{\mathbf{P}}^*\tilde{\mathbf{Z}}^*\frac{\partial\tilde{\mathbf{G}}^*}{\partial\theta_{di}}\tilde{\mathbf{Z}^*}'\tilde{\mathbf{P}}^*\tilde{\mathbf{Z}}^*\frac{\partial\tilde{\mathbf{G}}^*}{\partial\theta_{d'_i}}\tilde{\mathbf{Z}^*}')\}_{i,i'=1}^{v_d} & \{\underset{m}{tr}(\tilde{\mathbf{P}}^*\tilde{\mathbf{Z}}^*\frac{\partial\tilde{\mathbf{G}}^*}{\partial\theta_{di}}\tilde{\mathbf{Z}^*}'\tilde{\mathbf{P}}^*\frac{\partial\tilde{\mathbf{R}}^*}{\partial\theta_{r_j}})\}_{i=1,\,j=1}^{v_d,\,v_r} \\ \{\underset{m}{tr}(\tilde{\mathbf{P}}^*\frac{\partial\tilde{\mathbf{R}}^*}{\partial\theta_{r_j}}\tilde{\mathbf{P}}^*\tilde{\mathbf{Z}}^*\frac{\partial\tilde{\mathbf{G}}^*}{\partial\theta_{d_i}}\tilde{\mathbf{Z}^*}')\}_{j=1,\,i=1}^{v_r,\,v_d} & \{\underset{m}{tr}(\tilde{\mathbf{P}}^*\frac{\partial\tilde{\mathbf{R}}^*}{\partial\theta_{r_j}}\tilde{\mathbf{P}}^*\frac{\partial\tilde{\mathbf{R}}^*}{\partial\theta_{r_{j'}}})\}_{j,j'=1}^{v_r} \end{bmatrix} \tag{4.29}$$

where $\tilde{\mathbf{P}}^* = \tilde{\mathbf{V}^{*-1}} - \tilde{\mathbf{V}^{*-1}}\tilde{\mathbf{X}}^*M(\hat{\delta})^{-1}\tilde{\mathbf{X}}^{*\prime}\tilde{\mathbf{V}^{*-1}}$, $\tilde{\mathbf{V}}^* = Var(\tilde{\mathbf{Z}}^*\tilde{G}^*\tilde{\mathbf{Z}}^* + \tilde{\mathbf{R}}^*)$, $\tilde{\mathbf{Z}}^* = Z_{n\ x\ b} \otimes I_b$, $\tilde{\mathbf{G}}^* = A_{b\ x\ b} \otimes \Sigma_u$, and $\tilde{\mathbf{r}}^* = I \otimes \Sigma$.

Next, we need to specify the derivative functions for $\tilde{\mathbf{G}}^*$ and $\tilde{\mathbf{R}}^*$. To accomplish this, we will adopt the approach used in Section 4.4 of Kmail's dissertation, which provides examples for the format of the derivatives [78]. The genetic covariance matrix will be incorporated into the derivative functions, and the size of the derivatives will depend on the number of taxa included. To illustrate the process, we will consider the case where there are two responses, as demonstrated in both Kmail and Searle's work[78, 131].

If we let $\tilde{G}^* = A \otimes \Sigma_u = A \otimes \begin{pmatrix} \sigma^2_{u_{11}} & \sigma_{u_{12}} \\ \sigma_{u_{12}} & \sigma^2_{u_{22}} \end{pmatrix}$ and $\tilde{R}^* = I \otimes \Sigma = I \otimes \begin{pmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & {\sigma_2}^2 \end{pmatrix}$ then, the equations below outline the necessary derivatives across the $\tilde{G}^*$ and $\tilde{R}^*$

matrices.

$$\frac{\partial \tilde{\mathbf{G}}^*}{\partial {\sigma_{u_{11}}}^2} = A \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \frac{\partial \tilde{\mathbf{G}}^*}{\partial {\sigma_{u_{22}}}^2} = A \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \frac{\partial \tilde{\mathbf{G}}^*}{\partial \sigma_{u_{12}}} = A \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (4.30)$$

$$\frac{\partial \tilde{\mathbf{R}}^*}{\partial {\sigma_1}^2} = I \otimes \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \frac{\partial \tilde{\mathbf{R}}^*}{\partial {\sigma_2}^2} = I \otimes \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \frac{\partial \tilde{\mathbf{R}}^*}{\partial \sigma_{12}} = I \otimes \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad (4.31)$$

4.4 Proposed Exploratory Optimal Design Methodology

4.4.1 Variables of Interest

The equations are presented as each component is necessary to compute the composite optimality criteria, which was chosen to compare various input data of interest and changes in the design matrices. From Equation 4.5, we adapt Kmail's expansion of Mylona, Goos, and Jones' composite optimality criterion for simulations to evaluate optimality across both the fixed and random effects [78, 100]. For every new design iterated through in the simulation, the following criteria will be calculated and used as a measure of comparison.

$$\Phi = \frac{1-\alpha}{p*t} log|M(\hat{\delta})| + \frac{\alpha}{t*(t+1)} log|M(\hat{\sigma})| \quad (4.32)$$

Within Equation 4.32, $p*t$ finds how many fixed effects parameters there are, including values for p plates across t taxa. Then, $t*(t+1)$ calculates how many random effects should be estimated within Σ and Σ_u. For example, for three taxa cases, the weight on the random effects will be $\frac{\alpha}{3(4)} = \frac{\alpha}{12}$, where 12 comes from the estimates of $\sigma_{u11}, \sigma_{u22}, \sigma_{u33}, \sigma_{u12}, \sigma_{u13}, \sigma_{u23}, \sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}$, and σ_{23}. The matrices are symmetric, so not all of the variabilities will be specifically estimated. To

evaluate the impact of changes in the number of bean line replications in the plates and variability in taxonomic data on statistical design optimality, we conducted a random search simulation-based study. This approach enabled us to assess the effect of varying parameter changes on the overall optimality criterion and examine how optimality changed due to deviations in experimental design matrices and variations across taxa and bean lines.

Overall, to begin the modeling and set the format of the simulation, we need to set values for the following:

- X and Z: These matrices hold information about how many plates are being utilized and how many bean lines exist per plate. Currently, 12 plates are utilized in all simulations (the same as in the study from Chapter 2), and we allow for 87 available wells per plate, again the same as what was allowed in our experiment.

- A: This is for our genetic relationship matrix. We had access to the entire A matrix with covariance relationship parameters between all 297 bean lines from Dr. Van Haute. For this chapter, we have focused on smaller examples with only 50 bean lines.

- Σ_u: This matrix contains information about variability across the taxa included as responses in the model as related to the random effects. In this model, those will be the bean lines of interest. This matrix is used to calculate $\tilde{G}^* = A \otimes \Sigma_u$ where, $Var[Vec(U')] = A \otimes \Sigma_u$, which can be used to find the genetic covariance matrix among the lines due to the covariances of the taxa. This structure is similar to a repeated measures experiment where $Var(\epsilon) = \mathbf{R}\sigma^2$, where R is a matrix containing the covariance structure between time points. In this case, our A matrix acts like a covariance structure associated with the random effects.

In this case, from A and Σ_u, the value $\sigma^2_{uij}\mathbf{A}$ gives an approximate genetic covariance matrix of the lines for the first taxa, using estimates of the variances or covariances based on the values of i and j. We will focus on adaptations of this matrix and making changes in Σ_u to evaluate changes in replications and optimality.

- Σ: information about the residual variability across the taxa from, $\tilde{R}^* = I \otimes \Sigma$, where $Var[Vec(E')] = I \otimes \Sigma$

4.4.1.1 Genetic Relationship A Matrix

As stated in the second bullet point, the A matrix was fully provided to us for all 297 bean lines. The A matrix used in our calculation of $\tilde{G}^* = A \otimes \Sigma_u$ has the properties of a covariance matrix between the bean lines, similar to a covariance structure used in a repeated measures model where the covariances and correlations between time point are included. In our modeling structure, we know that inherently there is a population structure and a covariance structure between the bean lines included in the design. Because there are relationships between the bean lines, we decided it was important to include the structure within our model and our optimal design simulation structure. The A matrix is symmetric and contains 297 rows and 297 columns, where the values inside are estimates of the covariance between each combination of the 297 bean lines. The A matrix was calculated based on the MAD-P, middle American diversity panel of bean cultivars cultivated from the *P. Vulgaris* common bean. The kinship matrix was calculated in TASSEL v5.2.69 from the genotype data to account for population structure in downstream analyses [59]. TASSEL stands for Trait analysis by association, evolution, and linkage used to evaluate trait associations [17]. In total, our A matrix includes 199 Durango lines, as well as 100 Mesoamerican gene

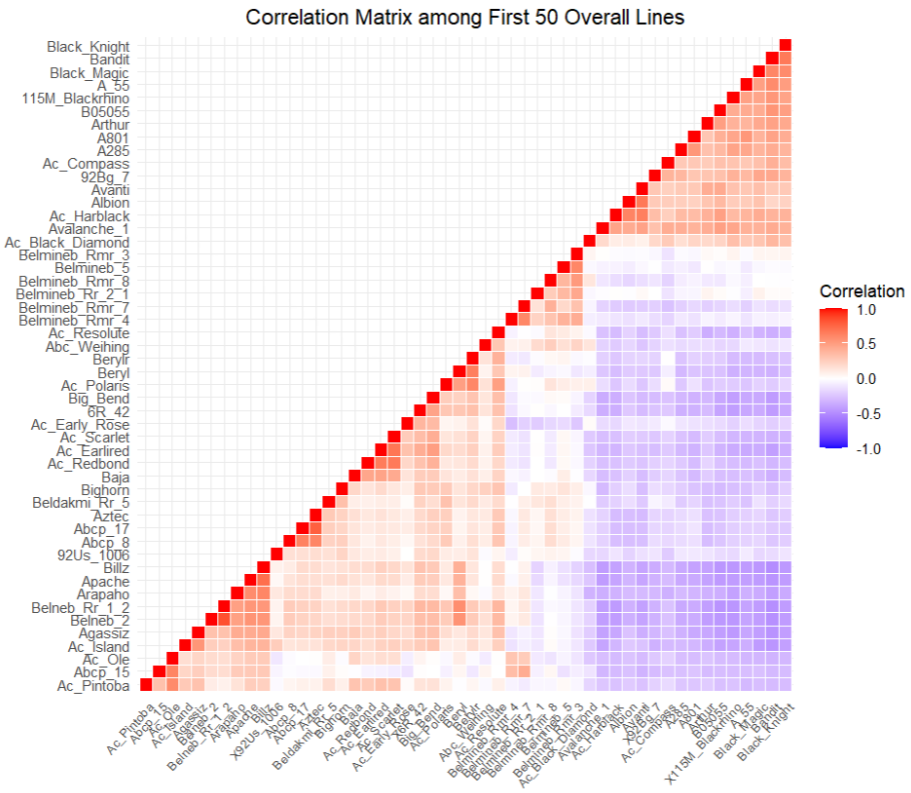


Figure 4.1: Correlation values between first fifty selected bean lines from the Genetic Relationship A Matrix

lines. The kinship matrix was calculated from SNP marker data ($\sim$ 132k positions). Collaborators provided a paper from Endelman and Jannink with specific descriptions of how the values within the A matrix were calculated with provided SNP markers [39].

To reduce the experiment's size and runtime, simulations were conducted using only 50 taxa. For many runs, the first 50 bean lines listed in the A matrix were used as a mix of Mesoamerican and Durango bean lines. An adapted set of 50 bean lines was selected for additional runs, whereas only 50 Mesoamerican lines were selected. The Mesoamerican and Durango lines will be referred to as our *Standard A*, and the only Mesoamerican lines are labeled as *New A*. To visualize the relationships across these two sets of genetic covariance matrices, heat maps of the correlation values

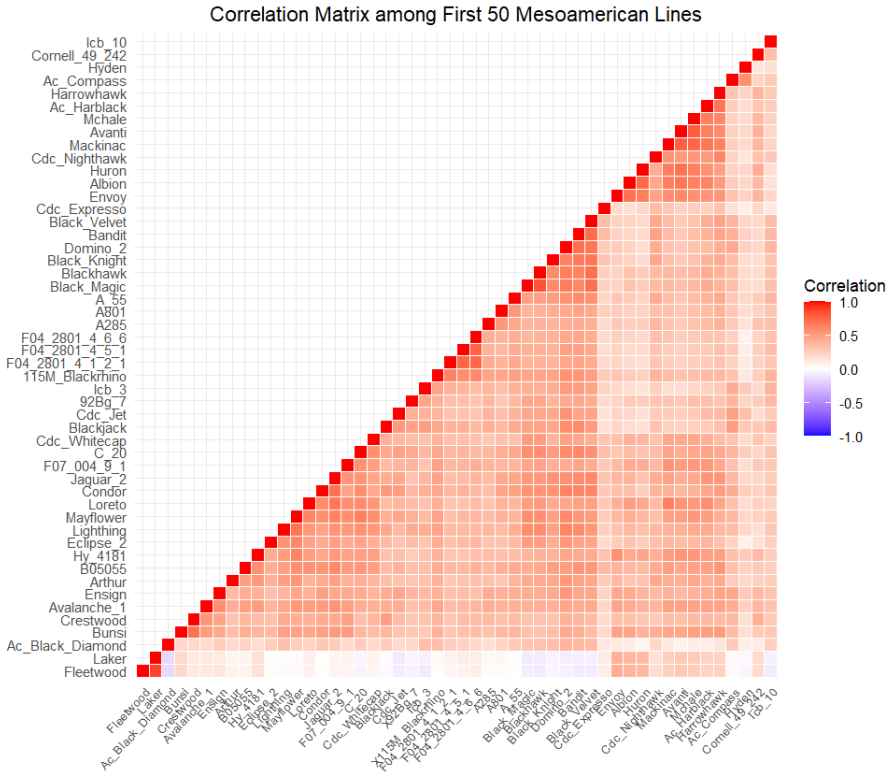


Figure 4.2: Correlation values between first fifty selected Mesoamerican bean lines from the Genetic Relationship A Matrix

have been plotted below. From the first 50 overall bean lines, in Figure 4.1, there is a variety of positive and negative correlations between the Mesoamerican and Durango bean lines. With just the Mesoamerican lines, there are mostly positively correlated bean lines within Figure 4.2.

4.4.1.2 Variability Estimates

The setup of X and Z, as well as the changes in the Z matrix over the simulations, will be described in detail in subsection 4.4.2. For the last two bullet points, both matrices comprise the variabilities and covariances for the bean lines and residuals across the taxa. There are multiple estimates of variabilities for different combinations of bean lines and subjects analyzed in Chapter 3. These estimates were used

to establish boundaries for pilot examples in this chapter. Researchers could obtain pilot information on the Σ and Σ_u matrices to evaluate potential design optimality and variability changes, especially for taxa of interest.

First, for Σ, it may be more challenging to outline exact values of interest. We wanted to see how adaptations of the Σ_u matrix changed the optimality and replications in the design. For all simulations, three basic Σ matrices were utilized for the simulations not based on real pilot data. These three matrices were denoted in the code in Appendix C as

$$RU.1 = \overbrace{\begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}}^{1}, \; RU.2 = \overbrace{\begin{bmatrix} 0.001 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.001 \end{bmatrix}}^{2}, \; RU.3 = \overbrace{\begin{bmatrix} 3 & 0 & 0 \\ 0 & 0.001 & 0 \\ 0 & 0 & 0.001 \end{bmatrix}}^{3};$$

RU.1 represents all large variabilities, RU.2 represents all small variabilities, and RU.3 is a mix of large and two small variabilities. Burgueño et al. compared a range of covariance structures for the genetic effect variability matrices in plant breeding models, incorporating both marker and pedigree information in similar multivariate models. Their findings confirm the superiority of models that utilize both marker and pedigree information over those that rely solely on pedigree information [19]. They highlighted that using both diagonal forms of the covariance matrices for random effects of genetic effects and a diagonal residual structure would be the same as a univariate case for them of single-environment models, and for this work, it would be like fitting individual taxa models [19]. In the future, we could compare cases with all diagonal effects to those cases where we allow for more complicated covariance structures. However, at this time, it was of interest to investigate differences in correlated taxa across the random effects while assuming diagonal, independent taxa in

the residual effects of Σ.

Paired with are a variety of Σ_u matrices. First, types of compound symmetric structures on the Σ_u test matrices were used, where values for variability and one correlation were selected to fill in the following outline.

$$\Sigma_u = \begin{bmatrix} {\sigma_1}^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho \\ . & {\sigma_2}^2 & \sigma_2\sigma_3\rho \\ . & . & {\sigma_3}^2 \end{bmatrix} \tag{4.33}$$

To construct the CS Σ_u matrices, we employed three variance types that corresponded to the variations present in Σ: one with exclusively large variabilities, another with only small variabilities, and a third containing a mixture of one large and two small variances. Then, we paired each with three different correlation structures with varying positive correlations. Specifically, all large, all small, and all mixed variability values were combined with a large correlation value of 0.9, a medium correlation value of 0.5, and a small correlation value of 0.1 were used to create the first group of Σ_u values. From these combinations, nine different structures of this type of covariance matrices were used in combination with each of the three Σ (RU) matrices for 27 combinations based on the CS style Σ_u matrices. The nine styles of these matrices can be seen in the Figure 4.3 below.

To introduce more flexibility in the covariance matrix, we adopted an unstructured style covariance matrix that allows for different correlations between the three taxa, as seen below in Equation 4.34.

$$\Sigma_u = \begin{bmatrix} {\sigma_1}^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} \\ . & {\sigma_2}^2 & \sigma_2\sigma_3\rho_{23} \\ . & . & {\sigma_3}^2 \end{bmatrix} \tag{4.34}$$

	Label: Var, Corr		Variance			Correlation			Final Variance Matrix Used		
1	Large Variance	Large	3	0	0	1	0.9	0.9	3	2.7	2.7
			0	3	0		1	0.9	2.7	3	2.7
			0	0	3			1	2.7	2.7	3
2	Large Variance	Medium	3	0	0	1	0.5	0.5	3	1.5	1.5
			0	3			1	0.5	1.5	3	1.5
			0	0	3			1	1.5	1.5	3
3	Large Variance	Small	3	0	0	1	0.1	0.1	3	0.3	0.3
			0	3			1	0.1	0.3	3	0.3
			0	0	3			1	0.3	0.3	3
4	Small Variance	Large	0.001	0	0	1	0.9	0.9	0.001	0.0009	0.0009
			0	0.001	0		1	0.9	0.0009	0.001	0.0009
			0	0	0.001			1	0.0009	0.0009	0.001
5	Small Variance	Medium	0.001	0	0	1	0.5	0.5	0.001	0.0005	0.0005
			0	0.001	0		1	0.5	0.0005	0.001	0.0005
			0	0	0.001			1	0.0005	0.0005	0.001
6	Small Variance	Small	0.001	0	0	1	0.1	0.1	0.001	0.0001	0.0001
			0	0.001	0		1	0.1	0.0001	0.001	0.0001
			0	0	0.001			1	0.0001	0.0001	0.001
7	Mixed Variance	Large	3	0	0	1	0.9	0.9	3	0.0493	0.0493
			0	0.001	0		1	0.9	0.0493	0.001	0.0009
			0	0	0.001			1	0.0493	0.0009	0.001
8	Mixed Variance	Medium	3	0	0	1	0.5	0.5	3	0.02739	0.02739
			0	0.001	0		1	0.5	0.02739	0.001	0.0005
			0	0	0.001			1	0.02739	0.0005	0.001
9	Mixed Variance	Small	3	0	0	1	0.1	0.1	3	0.00548	0.00548
			0	0.001	0		1	0.1	0.00548	0.001	0.0001
			0	0	0.001			1	0.00548	0.0001	0.001

Figure 4.3: Nine Matrices Utilized as CS-Style Σ_u matrices

We utilized a set of sixteen distinct variance matrices as Σ_u in conjunction with the three Σ matrices specified earlier. These variance matrices follow the same pattern as before, featuring large and small variances. Additionally, two mixed cases are included: one with one large and two small variabilities and another with one small and two large variabilities. We also explored multiple types of correlations, with large (0.9), medium (0.5), and small (0.1) values. We combined each of the four variance cases with four different correlation types: (1) one large, one medium, one small, (2) one large and two small, (3) one large and two medium, and (4) two large and one small. By doing so, we created sixteen new correlation matrices for optimality simulation. The final matrices for these cases can be found in Figure 4.4 below.

	Variance	Corr	Variance			Correlation			Final Matrix		
10	Large	L, M, S	3	0	0	1	0.9	0.5	3	2.7	1.5
			0	3	0		1	0.1	2.7	3	0.3
			0	0	3			1	1.5	0.3	3
11	Small	L, M, S	0.001	0	0	1	0.9	0.5	0.001	0.0009	0.0005
			0	0.001	0		1	0.1	0.0009	0.001	0.0001
			0	0	0.001			1	0.0005	0.0001	0.001
12	Mixed, L S S	L, M, S	3	0	0	1	0.9	0.5	3	0.0493	0.0274
			0	0.001	0		1	0.1	0.0493	0.001	0.0001
			0	0	0.001			1	0.0274	0.0001	0.001
13	Mixed, S L L	L, M, S	0.001	0	0	1	0.9	0.5	0.001	0.0493	0.0274
			0	3	0		1	0.1	0.0493	3	0.3
			0	0	3			1	0.0274	0.3	3
14	Large	1L, 2S	3	0	0	1	0.9	0.1	3	2.7	0.3
			0	3	0		1	0.1	2.7	3	0.3
			0	0	3			1	0.3	0.3	3
15	Small	1L, 2S	0.001	0	0	1	0.9	0.1	0.001	0.0009	0.0001
			0	0.001	0		1	0.1	0.0009	0.001	0.0001
			0	0	0.001			1	0.0001	0.0001	0.001
16	Mixed, L S S	1L, 2S	3	0	0	1	0.9	0.1	3	0.0493	0.0055
			0	0.001	0		1	0.1	0.0493	0.001	0.0001
			0	0	0.001			1	0.0055	0.0001	0.001
17	Mixed, S L L	1L, 2S	0.001	0	0	1	0.9	0.1	0.001	0.0493	0.0055
			0	3	0		1	0.1	0.0493	3	0.3
			0	0	3			1	0.0055	0.3	3
18	Large	1L, 2M	3	0	0	1	0.9	0.5	3	2.7	1.5
			0	3	0		1	0.5	2.7	3	1.5
			0	0	3			1	1.5	1.5	3
19	Small	1L, 2M	0.001	0	0	1	0.9	0.5	0.001	0.0009	0.0005
			0	0.001	0		1	0.5	0.0009	0.001	0.0005
			0	0	0.001			1	0.0005	0.0005	0.001
20	Mixed, L S S	1L, 2M	3	0	0	1	0.9	0.5	3	0.0493	0.0274
			0	0.001	0		1	0.5	0.0493	0.001	0.0005
			0	0	0.001			1	0.0274	0.0005	0.001
21	Mixed, S L L	1L, 2M	0.001	0	0	1	0.9	0.5	0.001	0.0493	0.0274
			0	3	0		1	0.5	0.0493	3	1.5
			0	0	3			1	0.0274	1.5	3
22	Large	2L,1S	3	0	0	1	0.9	0.9	3	2.7	2.7
			0	3	0		1	0.1	2.7	3	0.3
			0	0	3			1	2.7	0.3	3
23	Small	2L,1S	0.001	0	0	1	0.9	0.9	0.001	0.0009	0.0009
			0	0.001	0		1	0.1	0.0009	0.001	0.0001
			0	0	0.001			1	0.0009	0.0001	0.001
24	Mixed, L S S	2L,1S	3	0	0	1	0.9	0.9	3	0.0493	0.0493
			0	0.001	0		1	0.1	0.0493	0.001	0.0001
			0	0	0.001			1	0.0493	0.0001	0.001
25	Mixed, S L L	2L,1S	0.001	0	0	1	0.9	0.9	0.001	0.0493	0.0493
			0	3	0		1	0.1	0.0493	3	0.3
			0	0	3			1	0.0493	0.3	3

Figure 4.4: Additional Variability Cases for Σ_u with Differing Correlations

To integrate the findings of Chapter 3, we included pilot data from nine additional cases using estimated matrices. Three of the most abundant taxa - Bacteroides,

Sutterella, and Bifidobacterium - appeared across all three subjects. To obtain pilot estimates for three additional Σ_u and Σ cases, we ran another MANOVA model using PROC GLM for each subject, focusing only on these three taxa, with the fixed effect of the plate and the random effect of bean line. From the resulting output, we obtained three versions of the Σ_u and Σ covariance matrices. We further decomposed these matrices into their respective variability and correlation components to compare them with the examples generated in our study. The variability, correlations, and final matrices of the pilot data can be found below in Figure 4.5. These matrices are closest to what one might expect to see from the real transformed abundances for the same taxa across subjects, and from these examples as well we can begin to investigate any additional relationships due to negative correlations.

Pilot SigmaU and Sigma Variability Estimates

Case	Variance			Correlation			Final Matrix		
1	0.078	0	0	1	0.33	-0.9	0.078	0.014	-0.109
	0	0.022	0	0.33	1	-0.58	0.014	0.022	-0.037
	0	0	0.189	-0.9	-0.58	1	-0.109	-0.037	0.189
2	0.183	0	0	1	0.83	-0.95	0.183	0.14	-0.423
	0	0.156	0	0.83	1	-0.9	0.14	0.156	-0.37
	0	0	1.085	-0.95	-0.9	1	-0.423	-0.37	1.085
3	0.059	0	0	1	0.65	-0.84	0.059	0.044	-0.111
	0	0.078	0	0.65	1	-0.82	0.044	0.078	-0.124
	0	0	0.294	-0.84	-0.82	1	-0.111	-0.124	0.294

Pilot Residual Variance Matrices

Case	Variance			Correlation			Final Matrix		
1	0.052	0	0	1	0.33	-0.9	0.052	0.01	-0.074
	0	0.017	0	0.33	1	-0.58	0.01	0.017	-0.027
	0	0	0.13	-0.9	-0.58	1	-0.074	-0.027	0.13
2	0.128	0	0	1	0.83	-0.95	0.128	0.111	-0.321
	0	0.139	0	0.83	1	-0.9	0.111	0.139	-0.317
	0	0	0.894	-0.95	-0.9	1	-0.321	-0.317	0.894
3	0.045	0	0	1	0.65	-0.84	0.045	0.038	-0.088
	0	0.074	0	0.65	1	-0.82	0.038	0.074	-0.11
	0	0	0.242	-0.84	-0.82	1	-0.088	-0.11	0.242

Figure 4.5: Pilot Variability Cases for Σ_u with Differing Correlations from Bacteroides, Sutterella, and Bifidobacterium across our three subjects

Our plan was to add more categorical group labels to the plots in order to further categorize the data. Our intention was to use labels in a similar way as we did with the CS and UN structures. However, due to the complexity of these matrices, the summary labels are not as informative as simply examining the results across the different matrix pairings. In some summary graphics, the Σ_{u1} and Σ_{u3} matrices have smaller variances with one larger one, so they will be given a label of "Mixed SSM". Σ_{u2} has two slightly larger variabilities, and one that is greater than one, so this group will be labelled "Mixed MML." We did the same thing in terms of the correlation values for the Σ_u matrices, but they are just labelled with the values of all the correlation values. In matrix form, the final matrices utilized for the 9 pilot combinations are listed below, matching the final values on the right in the figure 4.5 above. A label *2.1*, for example, would correspond to a case with Σ_u matrix 2 and Σ matrix 1.

$$\Sigma_{u1} = \begin{bmatrix} 0.078 & 0.014 & -0.109 \\ 0.014 & 0.022 & -0.037 \\ -0.109 & -0.037 & 0.189 \end{bmatrix}, \ \Sigma_{u2} = \begin{bmatrix} 0.183 & 0.14 & -0.423 \\ 0.14 & 0.156 & -0.37 \\ -0.423 & -0.37 & 1.085 \end{bmatrix}, \ \Sigma_{u3} = \begin{bmatrix} 0.059 & 0.044 & -0.111 \\ 0.044 & 0.078 & -0.124 \\ -0.111 & -0.124 & 0.294 \end{bmatrix};$$

$$\Sigma_1 = \begin{bmatrix} 0.052 & 0.01 & -0.074 \\ 0.01 & 0.017 & -0.027 \\ -0.074 & -0.027 & 0.13 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0.128 & 0.111 & -0.321 \\ 0.111 & 0.139 & -0.317 \\ -0.321 & -0.317 & 0.894 \end{bmatrix}, \ \Sigma_3 = \begin{bmatrix} 0.045 & 0.038 & -0.088 \\ 0.038 & 0.074 & -0.11 \\ -0.088 & -0.11 & 0.242 \end{bmatrix};$$

4.4.2 Setting Up Iterative Algorithm

We used R to code our optimal design simulations, relying on several packages such as pracma, dplyr, data.table, Matrix, spam, and sparseinv [118, 15, 151, 33, 10, 49, 161]. An example of the optimal design simulation code is provided in Appendix C. While R was utilized to run the simulation program, UNL's Holland Computing Center's High-Performance Computing (HPC) clusters. To evaluate the criterion Φ from Equation 4.32, we followed the equations in subsection 4.3.2 and set up all relevant pieces of the model. Our simulation function is designed to provide a comprehensive modeling framework for calculating design optimality criterion values

and keeping track of how these values change as the formation of the Z matrix is altered. The objective of the code is to establish all pieces of the univariate form of a multivariate mixed model and compute our modified variant of Kmail's updated composite criterion (Equation 4.32) that can be tailored to strike a balance of optimality between fixed and random effects. To explore the range of α values, we selected a variety of weight values for the fixed and random effects, ranging from 0.1 to 0.9 (specifically $\alpha = 0.1, 0.25, 0.5, 0.75$, and 0.9). Mylona, Goos, and Jones previously used $\alpha = \frac{1}{2}$, which balanced the two objectives of the composite criterion [100].

We created a function titled *OptDesignThreeTaxa* in R, which fits is capable of fitting various parameter values. The function takes in matrix values of $Sigma = \Sigma_u$, $RU = \Sigma$, and our genetic relationship matrix A. Additionally, it is set up to require values for the number of bean lines, the α weights, the number of plates, the number of wells utilized per plate, and how many Z matrices to iterate through for the search procedure. Our searches all utilize 5000 random generations of the Z matrix. Initial code testing used 2500 replicates and 10,000 random Z matrices. The Crane high-performance clusters within the Holland Computing Center provided resources for running 2500 replicates relatively quickly, so we wanted to run more replicates. However, when we attempted to run 10,000 replicates, the program did not complete within the allotted time on Crane. We reduced the number of replicates to 5,000 and observed a final run time of approximately 3-5 days. The actual time varied based on the number of nodes and computing power specified in the SLURM file, either on our specific resources or the general batch node for all HCC users. We found that the optimality values showed minimal variability across all replicates, indicating that 5,000 replicates were sufficient for all simulations. Despite this, we anticipate that we can increase the number of replicates in future runs with changes in the coding structure and better resource allocation. This would help us determine with greater

certainty whether our sampled Z matrices approach an optimal Φ value.

These values are all set up in a generic framework, where R will build out the X and Z matrices based on what inputs are put into the function. Before running the function, a user is required to set up what particular A relationship matrix will be used. Again, as stated above, we use two cases, one with the first 50 bean lines overall and one with a selected case of the first fifty Mesoamerican bean lines. The A matrix format also tells the function how many bean line treatments will be used within the design. Before running the function, all possible Σ_U and Σ matrices are included in the code. From this list, an individual combination of Σ and Σ_U is selected for use within the function. Since these values are the only ones that change throughout most of the code, we have included only one example in Appendix C. The one thing overall that is not easily adaptable is how many taxa can be included with this specific function. All code formulations for the derivative functions of $\frac{\partial \tilde{\mathbf{G}}^*}{\partial \sigma_{u_{ij}}}$ and $\frac{\partial \tilde{\mathbf{R}}^*}{\partial \sigma_{ij}}$ are specifically created for three taxa within this function.

Inside the function, descriptions for R of how the program needs to classify all the individual values and matrices are set up at the beginning, and then two for-loops are utilized within the function to literature between the values of α weights and between random formations of our Z design matrix. We begin by setting a start time to see how long each simulation runs and establish all values that do not change as the Z matrix changes. This reduces run time as R will not need to recalculate certain values across the iterations. Values that do not change as we iterate through different designs include the A matrix, the number of taxa, the number of fixed and random effects (utilized within the Φ function), the number of plates, and any formulas used within our modeling structure dealing with our X matrix. We also set up the matrices of 0s and 1s necessary later within our calculations of the derivative matrices $\frac{\partial \tilde{\mathbf{G}}^*}{\partial \sigma_{u_{ij}}}$ and $\frac{\partial \tilde{\mathbf{R}}^*}{\partial \sigma_{ij}}$.

Our outer loop runs through five iterations of the simulation for each level of α. As the values of α change, the weights on the fixed and random portions of Φ are calculated and remain the same for each value of alpha. Next, the inner loop iterates through values of Z. Random rows of Z are filled with values of 1 in columns corresponding to individual bean lines. For instance, our Z matrices result in a 1044 x 50 matrix, where each well is filled with a random selection of the 50 bean lines. Next, all values necessary to compute the information matrices are determined, including Z, $\tilde{Z}^*$, G, R, V, V^{-1}, $M(\delta)$, and $M(\delta)^{-1}$, where $M(\delta)$ represents the information for the fixed effects. These values are used to evaluate the Kronecker products for the corresponding Σ_u and Σ values. Then, P and all matrix pieces for the $M(\sigma)$ matrix are calculated. To obtain the $M(\sigma)$ matrix for three taxa, a 12 x 12 matrix is generated, and a small loop is utilized to calculate the values on the diagonal and lower diagonal of the matrix. As the matrix is symmetric, these values are used to complete the full $M(\sigma)$ matrix. The computation time required for these calculations is extensive, particularly for larger matrices. In the case of three taxa, running the code can take between two to five days, depending on the capacity of UNL's Holland Computing Center (HCC). However, as more taxa are included, the matrices increase in size, making the current computational approach impractical to run within the 7-day time limit on HCC's CRANE server.

Once we have the $M(\sigma)$ matrix, we can compute the information criteria labeled *InfMat* in the code. To finish the calculation of Φ from Equation 4.32, we take the logarithm of the determinants of both information matrices and multiply the values by the corresponding weights. The resulting values for the optimality criterion and each fixed and random component are stored in a matrix as we iterate through each Z matrix format. If the Modified Federov Switching algorithm is included in the code, it is necessary to check and compare the criteria values as we loop through the different

Z values. This structure remains in the code, and the most optimal Z matrix and optimal criteria values are stored during the simulation. Although keeping every Z matrix is impractical, a summary of each matrix is retained. We summarize each Z matrix into the column sums for how many replications of each bean line appear within the matrix. After each run of the full simulation, the *TotalReps.ALL.Zs* matrix contains 50 rows by 5000 columns of how many times each bean line was replicated in all Z variations. We use this matrix to explore how the optimality may change based on the variability of the number of replicates across all the bean lines. In addition, we will use this information in our results to explore the variability and spread in the bean line replicates to examine if they differ between the "best" and "worst" designs based on the optimality criteria. Although not evaluated currently, we determined the number of replicates for each bean line across the 12 plates within the experiment for the Z matrix with the largest optimality criterion. This feature adds flexibility to the code, enabling it to evaluate where the lines are allocated within a plate. In the future, these results could be used to adapt the X matrix as well for the fixed effects of plates if researchers would be interested in looking at adapting the fixed effects as well.

The function generates several significant data sets saved in .csv format. Five files are created for each corresponding α level containing the criterion values for each of the 5000 Z matrices. Next, for each level of α, a data set is outputted containing the summary of how many times each bean line was replicated across all 5000 simulated Z matrices. Lastly, the function in R directly outputs the time taken to run and the largest optimality criterion value by α. Additionally, optimal Z matrices and the total reps by plates for each optimal Z matrix are outputted, although they were not used in the analysis. [18]

4.4.2.1 Outputs

The following information was outputted and calculated into a final data set for graphics and summaries of the simulation runs produced thus far. We have results for every level of α tested $(0.1, 0.25, 0.5, 0.75, 0.9)$. Then, we have indicator variables labeling what variance and correlation combination was utilized for Σ_u and Σ for each set of the simulation results. A variable labeled "*Group*" also identifies which version of the A matrix was used for the calculations. Then, summaries are evaluated across the 5000 replicated Z matrices, including the minimum, upper and lower quartile, mean and median, and maximum Φ composite optimality criteria values. In addition to these statistical summaries, we calculated the variance in the number of replications for the most and least optimal design of Z, determined by which designs created the maximum and minimum Φ outputted across all 5000 calculated optimality criteria. These variances were also used to find coefficients of variation $(\frac{\sigma}{mu})$ for designs with the largest and smallest Φ values. For this calculation, the σ is the square root of the variance of the number of replications for each bean line in the selected design by each value of α, and μ is the average number of replications across all the bean lines. This gives an idea of the spread of the available replications across the bean lines. For the case where we use 50 bean lines, there are more places possible to replicate each bean line. However, in cases included in the results where 297 bean lines are included, the number of total replications across each bean line is much more restrictive.

Lastly, similar to Bueno Filho and Gilmour [28], we wanted to compare the optimality values across quantitative summaries of the combinations of Σ and Σ_u. Heritability is defined as $h^2 = \frac{{\sigma^2}_u}{{\sigma^2}_u + \sigma^2}$ for univariate modeling components. In a multivariate setting, we aim to compare similar summaries of the Σ and Σ_u matrices across variance and covariance components to analyze the relationships between taxa.

To do this, from every combination of Σ and Σ_u we calculated and focused on variables titled, h^2, *det2*, and *eigen.ratio*, from the formulas in Equation 4.35 to Equation 4.37.

$$h^2 = \frac{tr(\Sigma_u)}{tr(\Sigma_u) + tr(\Sigma)}; \tag{4.35}$$

$$det2 = \frac{|\Sigma_u|}{|\Sigma_u| + |\Sigma|}; \tag{4.36}$$

$$Eigen.Ratio = \frac{\lambda_{1_{\Sigma_u}}}{\lambda_{1_{\Sigma_u}} + \lambda_{1_{\Sigma}}}; \tag{4.37}$$

Additionally, equations like $g = \gamma = \frac{tr(\Sigma)}{tr(\Sigma_u)}$; and a form of this γ equation with the determinants, $det1 = \frac{|\Sigma|}{|\Sigma_u|}$;, could also be used to compare optimal designs.

Ge et al. utilized Equation 4.35 as the heritability of a multidimensional trait, so we have used a similar notation for the calculations of both h^2 and γ [52]. We use these two values as a multivariate extension to how designs were compared by Bueno Filho and Gilmour [28]. Det1 is also a determinant expansion of the g value, which only includes the trace. Next, to compare the overall values of the Σ and Σ_u matrices, we calculated the determinants of the individual matrices and compared them using similar ratios as the heritability calculations. Lastly, some authors have utilized the largest eigenvalues from the genetic matrices of interest in univariate or multivariate model frameworks [73, 35, 7, 77, 52]. We have also utilized another ratio similar to the heritability but using the largest eigenvalue, λ_1, calculated from Σ and Σ_u. Regarding the results, it was found that the relationships in γ and $det1$ did not yield any significant findings beyond what was observed for the other three ratios. Therefore, these cases were not further analyzed in this study.

4.5 Results

We present our results in separate summary graphics, highlighting the optimality criterion and variability in the number of replications for each group's most and least optimal designs. This approach will enable us to effectively compare and contrast the performance of different designs across multiple groups. Additionally, we observe how the optimality and replicate variability changes for summary combination values of Σ_u and Σ, including values from the summaries h^2, g, $det1$, $det2$, and the *eigen.ratio*. By providing these summaries, we can gain insights into the types of information that domain researchers should consider when using a similarly structured optimality search. This information will be valuable not only for designing studies of this nature, but also for informing future adaptations of such designs.

4.5.1 CS Variance Structures

First, Figure 4.6 evaluates how the maximum Φ values changed across different levels of α weights on the x-axis. The plot is divided into panels, each corresponding to a different Σ_u matrix used in the first 27 sets of variability matrices. The legend on the bottom indicates that the shapes of the points correspond to which of the three Σ residual variances was used, and the color matches with what size correlation was used between all three taxa. The points are labeled with values for which values of Σ_U and Σ were combined. In our table within Figure 4.3, values are numbered by Σ_u and Σ. For example, the point labelled 7.1, is the seventh Σ_u style, with a mixed variance (one large and two small) and all large correlations, with the first residual diagonal matrix with all large values.

In the points, we noted where points were overlapping and where there were areas where some designs within similar variance structures began to separate, with

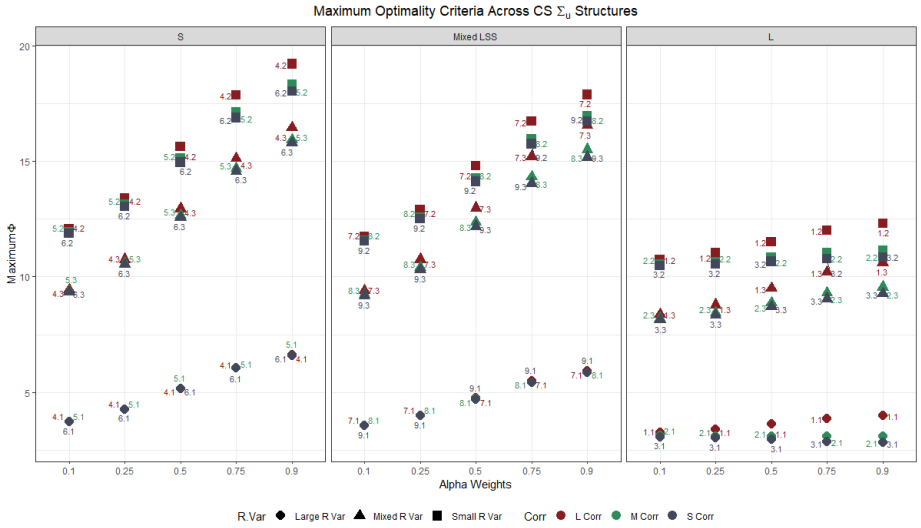


Figure 4.6: Maximum Φ values across CS Σ_u Structures. Points are labeled by numeric descriptions of which Σ_u and Σ matrix combination were utilized for the simulation. For example, if a point is labeled 7.1, this is the seventh Σ_u matrix and the first Σ residual matrix

which designs were optimal. Across all three panels of Σ_u varieties, the small variability cases tend to have more optimal designs than those with the mixed and large cases. Within the large panel of cases on the right, there is a smaller gap of differences in optimality criteria across α weights. Also, within these cases, the circle points at the bottom of the pane for the large residual values show more separation in optimal design for the mixed and small Σ_u values, where the points all seem to overlap. The separation shows that the largely correlated taxa begin to have the ability to produce a more optimal design.

These same separations show up in the other Σ_u cases, and the gap between large, medium, and small correlations between taxa become more pronounced as we put more weights on the random effects. Gaps between small and medium correlation values of 0.1 and 0.5 remain smaller. As the researchers become more interested in weighting their design optimality on the random bean line effects with highly correlated taxa, we are beginning to see trends that show that the style of design plays

a larger role based on the variances and correlations of the taxa and residuals. We have observed expected trends in the variabilities, with certain groups demonstrating greater optimality. Notably, we have observed larger differences in Φ magnitudes within the Σ_u variability options across the residual variabilities, as opposed to the relatively smaller differences observed within the same Σ values across the panes.

Next, we can see how these differences in optimality correlated to differences in the variability of number of replications across the bean lines. The idea here is to understand how the balance of lines across wells is related to optimality.

For each of the optimal designs, we calculated the variance of the number of replications for each of the 50 bean lines in the final design. A plot was originally created looking at the changes in the variability are observed across levels of α weights for each of the three types of Σ_u matrices. There were not many overall trends easily evaluated across the changes in the combinations of the variabilities and correlation values. There may be a bit more clustering of the variability for the smaller α weights, but overall, these changes not too noticeable in passing glances on the graphic. This plot was not included, but then we focused on three specific α weights.

To better observe the points, we could focus on what happens with the 0.1, 0.5, and 0.9 levels of α, with the values in the middle and at the extremes, in Figure 4.7. With this closer look, we can see individual α level plots trends across the three types of Σ_u matrices. Across the panels in the rows for L correlation cases, for α values 0.1 and 0.5, as the optimality criteria values increase, there are increases in the amount of variability in the replication variance. There is a much greater increase with the large 0.9 α weight. However, for the small and mixed variability cases in the top two rows, there is not much change in the variability for the 0.1 and 0.5 α cases, but more drastic declines in variability for the 0.9 weights. Overall, the variabilities begin to decrease in the places with large Φ values for optimal designs. There also

does not appear to be one matching structure of what happens between the same variability cases for differing correlations. We are still seeing the trend of the larger correlated taxa having higher optimality values, but it does not often translate to other relationships of the variance in the replications. With small and mixed residual variabilities we are seeing more balanced designs than with the larger cases.

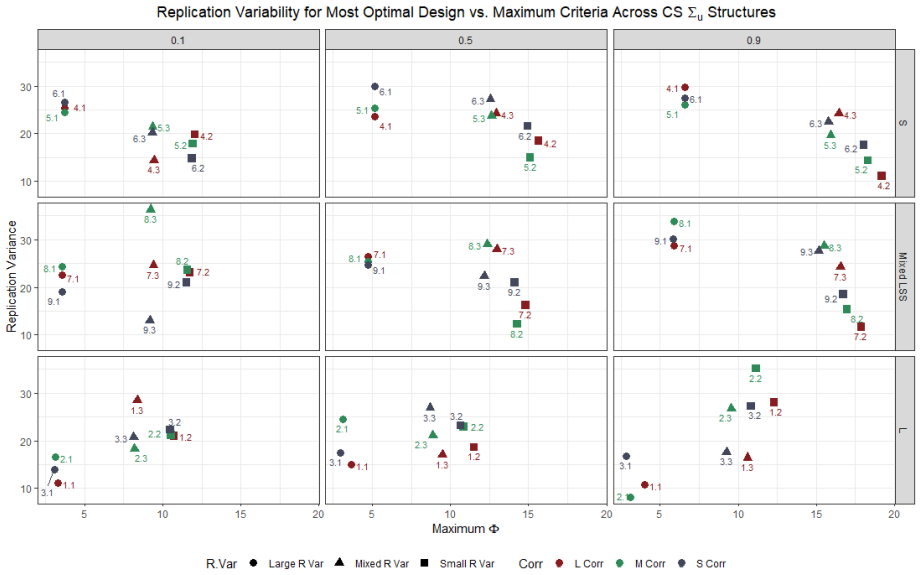


Figure 4.7: Maximum Variability vs. Maximum Φ Values across $\alpha = 0.1, 0.5, 0.9$. Plot panels have the values of α in the columns, and varieties of Σ_u in the rows. Shapes of the points are the Σ matrix variety, and colored based on the correlations.

While the figures offer visualizations of how variability in the number of replications affects different variance structures, we can also gain insights by examining the differences between the most and least optimal designs. These designs are from the Z matrices with the largest (most optimal) and smallest (least optimal) values of Φ. By doing so, we can identify the factors that contribute to the optimal performance of certain designs, and if there are changes between the designs that the optimal criteria values determine are the most and least optimal. Since variance is not a unit-less value, we utilize the coefficient of variation as a standardized value to compare variabilities. CV values were calculated for the most and least optimal designs, and then

the difference was taken between the least optimal design's coefficient of variation and the most optimal design.

In plots for the difference in coefficient of variation, a horizontal line is added at a zero value for the difference. When the difference is zero, the coefficients of variation are close to the same, and the variability between the most and least optimal designs are similar with respect to their overall number of replications across the bean lines. When taking the difference in the coefficient of variation values across the designs, we can see how much change there is in the most and least optimal designs, the more positive the value, the more variability exists in the least optimal design.

Originally trends were outlined across all levels of α, however, the overall implications of these findings for determining an optimal design are not straightforward. Even so, our graphics for the CS matrices can begin to reveal distinct changes in optimality and design structure across different combinations of Σ_u and Σ matrices. By considering these variations, we can gain a more comprehensive understanding of the factors that contribute to the effectiveness of certain design choices. Rather than observing trends across all α levels, we focused on selected cases.

In this last, Figure 4.8, only the α values 0.1, 0.5, and 0.9 are plotted to look closer at any trends in the change of CV. Similarly cases across all levels of α, the coefficient of variation values for smaller weights are closely clustered together, while the spread between points becomes more pronounced for weights of 0.9 α. This trend highlights the increased variability in performance among designs with higher weights, underscoring the importance of carefully selecting the optimal weight for the given scenario. Researchers can adjust the value of α to prioritize either the fixed or random effects, depending on their specific design or model of interest. Our plots demonstrate that this flexibility can impact the overall optimal design, particularly in terms of how the replicates are distributed across treatments.

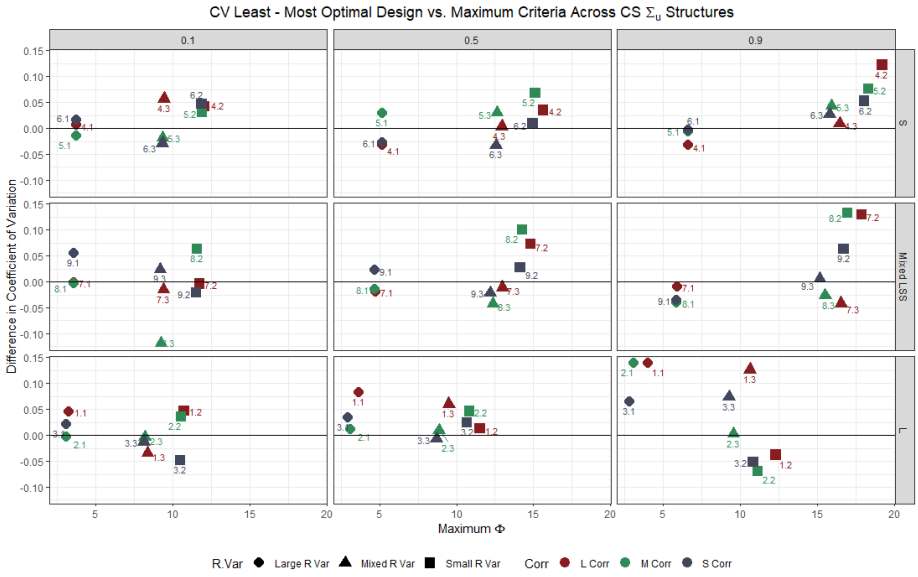


Figure 4.8: CV Most - Least Optimal Design vs. MaximumΦ Values across $\alpha = 0.1, 0.5, 0.9$. Plot panels have the values of α in the columns, and varieties of Σ_u in the rows. Shapes of the points are the Σ matrix variety, and colored based on the correlations.

4.5.2 Unstructured Variance Structures

Moving forward, we examine scenarios where the correlations between the three taxa within the Σ_u matrices vary. By analyzing the same graphics and values as in the CS structures, we can assess how distinct correlations among taxa responses may impact the optimality criteria values and the variability of replications across all relevant cases. For all cases with the UN structures, there were some combinations of the Σ and Σ_u matrices that had noninvertible information matrices for the fixed or random effects, either that were not positive definite, or they created extreme values where R could not take the log of the determinant and NAs were produced. Several different functions in R were tested to produce the inverse, however, errors were not able to be fixed in these particular cases. The missing cases for all UN results will be 22.1, 22.2, 22.3, 23.2, 24.2, 24.3, and 25.2. All 22 cases for Σ_u contained the large variabilities, 23.2 is a case where both Σ and Σ_u have the smallest variabilities. These

are cases where perhaps the variances were too extreme given the other data. The Σ_u matrices for cases 22 through 25 also all include two large correlations and one small. This could also be lending to the extreme nature of the cases producing nonpositive definite matrices.

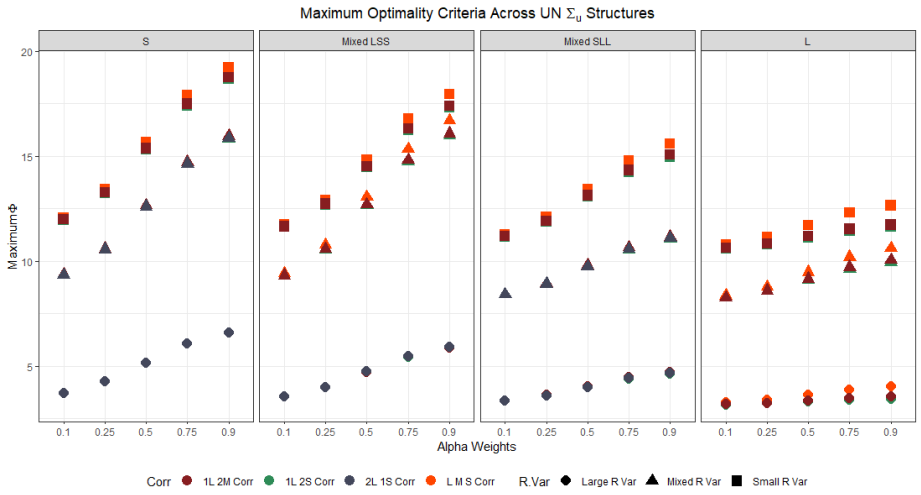


Figure 4.9: Maximum Φ values across UN Σ_u Structures. Panels represent the different varieties of Σ_u, points are colored by the correlation structure, and the shape of the points

When examining the changes in optimality criteria values across the selected UN cases for varying α weights, we observe that Φ values exhibit much steeper changes for small variance values of Σ_u than for larger variance values, within Figure 4.9. This suggests that the performance of the selected designs is more sensitive to changes in the composite weights when the variance values in Σ_u are low. Points are unlabelled in these cases, as the plot would get too busy. Many of the points are close enough in Φ where they overlap, but those points that stick out are those of the large, medium, and small correlation cases (in orange) and the 1 large and 2 medium cases in the maroon color. Regarding the shape of the points and the Σ matrix values, we observe that smaller variance values yield the most optimal designs, similar to the mixed case. The circle points associated with large Σ variances are once again found at the

lower end. Moreover, we notice that the rate of change in optimality across the Σ_u variabilities remains consistent when the variances are small.

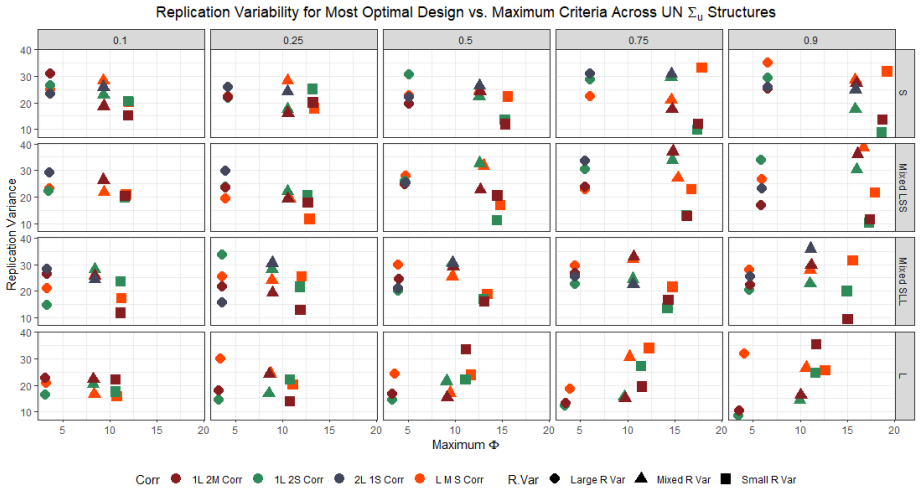


Figure 4.10: Maximum Variability vs. Maximum Φ Values across all α. Plot panels have the values of α in the columns and varieties of Unstructured Σ_u in the rows. Shapes of the points are the Σ matrix variety and are colored based on the correlations.

Again, in plots Figure 4.10 and Figure 4.11 values are plotted between the Φ values and the replication variability for all and select α weights. In comparison to the CS cases previously, there are even fewer trends in these points. However, there are still very flat relationships between the criteria and the replication variance for the smaller levels of α. But, then the increases for the 0.75 and 0.9 α weights are less pronounced. No particular cases overall are sticking out as more or less variable.

For the particular cases in Figure 4.11, there are some slightly increasing points, as Φ goes up, so is the amount of variability, with the largest points being for the small variabilities in maroon with the 1 large and 2 medium correlations. Circle points for the large residual variabilities have closer variabilities with more or equal fixed effect weights, whereas, with the 0.9 weights, those cases are more spread out. But, the relationships are not consistent with how the variances change as the types of correlation change. The LMS correlation type tends to have larger variabilities,

but also, for the α 0.9 weight column of plots, the 1L2S grouping also tends to have large replication variability as well. Again, these are not the most interesting notes to make, as we are not seeing consistent patterns in these particular hypothetical examples.

In standardizing the variabilities and observing changes in the coefficients of variation between the most and least optimal designs, there is again a tendency towards closer CV values between the most and least optimal designs for the weights more on the fixed effects. For the larger weights, there is more of a gap in what the variabilities are in the designs with the largest and smallest Φ values. An additional figure was created to narrow in on only three values of α, but not many new insights can be gained from focusing on these particular cases.

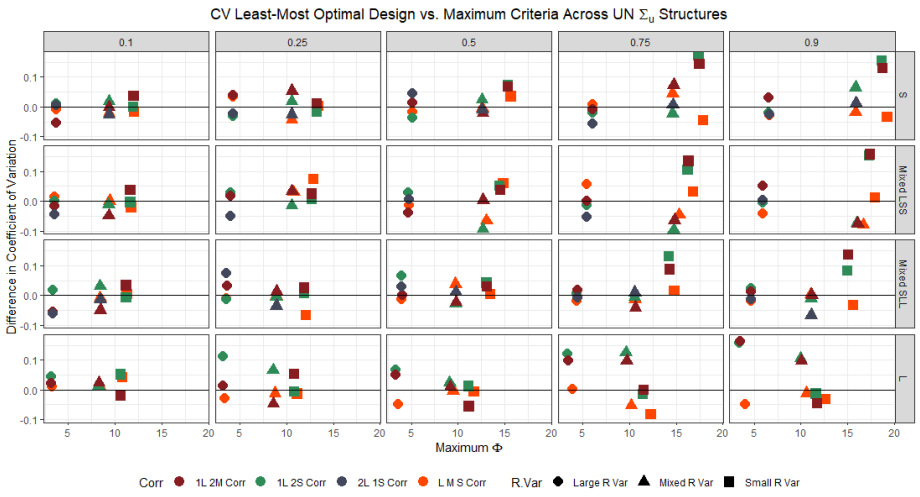


Figure 4.11: CV Most - Least Optimal Design vs. Maximum Φ Criteria Across UN Structures. Plot panels have the values of α in the columns and varieties of Unstructured Σ_u in the rows. Shapes of the points are the Σ matrix variety and color based on the correlations.

4.5.3 Pilot Data

The pilot data contains values from nine simulations across three Σ_u and Σ matrices created based on the data from Chapter 3 from taxa Bacteroides, Sutterella,

and Bifidobacterium. These cases estimated from the three subjects utilized in Chapter 3 will allow us to observe what might be recommendations for researchers talking to a statistician about their real data before setting up a similar experiment. First, we have cases like the ones above where we use a reduced plotting format because now there are three cases of each Σ_u and Σ matrices, but there is a less direct structure to the format of the matrices.

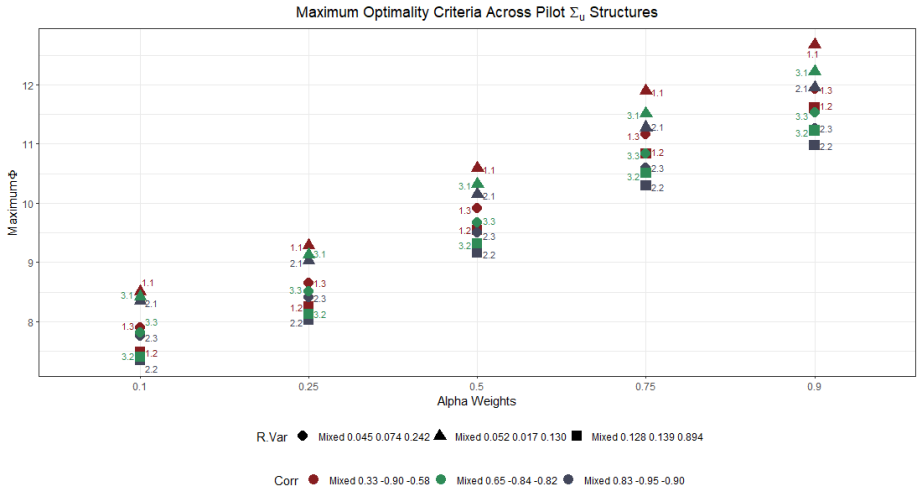


Figure 4.12: Maximum Φ values across pilot Σ_u Structures. Points are labeled by combinations of Σ_u and Σ, with shapes based on the residual pilot matrix and colors based on the correlation of the Σ_u matrix.

Points are labeled within Figure 4.12 and other pilot plots according to the numbering on the matrices from Figure 4.5 and the following matrices, where for example, a value of 3.2 would be from the third pilot Σ_u matrix and the second Σ matrix. These matrices are outlined in subsubsection 4.4.1.2 Again, the pattern in optimality values was closer for the simulations that were weighted more on the fixed effects, and there were larger gaps in Φ for the 0.9 weight. Across all α values, the *1.1* case with the Σ_u matrix $= \begin{bmatrix} 0.078 & 0.014 & -0.109 \\ 0.014 & 0.022 & -0.037 \\ -0.109 & -0.037 & 0.189 \end{bmatrix}$, with taxa correlated with the values of 0.33, -0.90, and -0.58, and the triangle case of the residual variability, which overall has the smaller of the variance estimates out of the three groupings.

Since this is the case of the pilot data that ties to the first Σ_u matrix and the first residual matrix, this pilot information goes back to the subject one grouping. Across all points, the triangle residual points with the mixed variance 0.052, 0.017, and 0.130 cases (from the first pilot residual matrix) have the largest optimality, followed by the circle variance cases (from subject's threes residual pilot matrix).

The cases with the square Σ points tend to have lower optimality. The square case for Σ had the largest of the three variabilities, so it makes sense that these cases fall more toward the bottom, within Figure 4.12. Finally, points of the same shape within the α values are often ordered from largest to smallest optimality as maroon, green, to blue from larger to smaller Φ values. The maroon points have smaller correlated values, followed by the green, than the blue. This demonstrates that values with higher correlation generally achieve less optimal designs compared to their less correlated counterparts. However, this relationship is more intricate than what we observed in the CS matrices, where cases featuring predominantly large correlations exhibited greater optimality. In the CS matrices, correlations were identical; however, the maroon cases in Figure 4.12, which exhibit the larger Φ values, display the widest differences in correlation values, potentially influencing changes in optimality. As the correlation values get closer together in magnitude, the optimality decreases, which is similar to trends we saw for the UN style cases. Overall, we observe that intricate interactions between taxa correlations and variances can impact optimality, and these pilot cases demonstrate trends that more closely resemble those found in actual data.

Instead of visualizing the replication variabilities, Figure 4.13 displays the coefficients of variation (CV) for the design with the highest Φ value in each case. By examining the CV values, we can identify trends and draw more meaningful conclusions than by considering variances alone. Notably, the CV values tend to be higher for the 0.1, 0.25, and 0.5 cases and then decrease and become more variable for the

0.75 and 0.9 weights. As some of the cases have increased optimality across α, the replicate variability decreases, and the more optimal designs have more consistent numbers of replicates across the 50 bean lines.

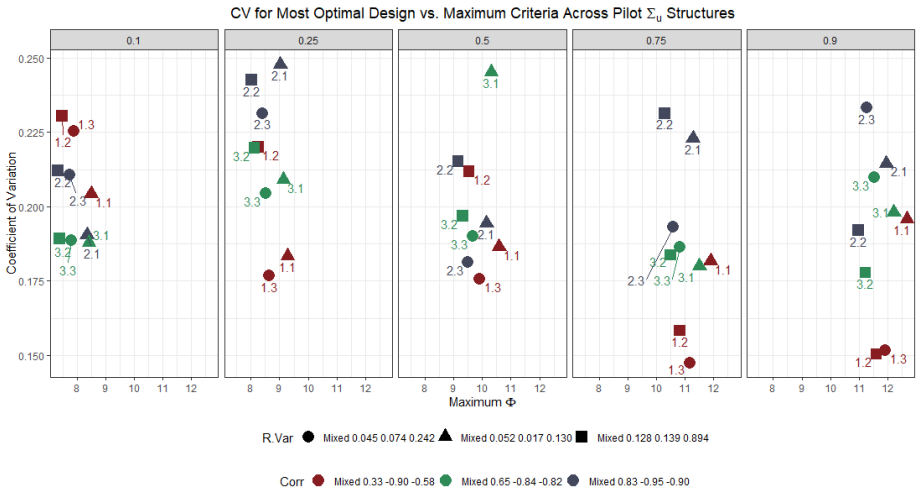


Figure 4.13: Coefficient of Variation vs. Maximum Φ Values across all α for most optimal designs. Plot panels have the values of α in the columns. Shapes of the points are the Σ matrix variety and are colored based on the correlations.

As was done for the CS and UN cases, Figure 4.14 illustrates the difference in coefficients of variation for the number of bean line replications between the least and most optimal designs from the 5000 simulated Z matrices. This plot shows a more pronounced pattern than the differences in CV values observed for the CS and UN cases. As we shift our attention towards the fixed effect information matrices, the variability among the replicates of bean lines decreases between designs with the highest and lowest Φ values. However, when researchers prioritize the optimality criteria of the random effects, they would observe more significant differences between the types of designs that yield optimal values for the Φ criteria. The largest differences between the least and most optimal designs are for the correlation case with the values of 0.33, -0.9, and -0.58 with the maroon coloring, followed by the green cases and the blue ones last. The maroon case has smaller correlation values, including the 0.33

and -0.58 values, so wider differences between bean line replications for the most and least optimal designs occur for cases with smaller correlated values vs. those with two or more correlations closer to 1 (or negative 1).

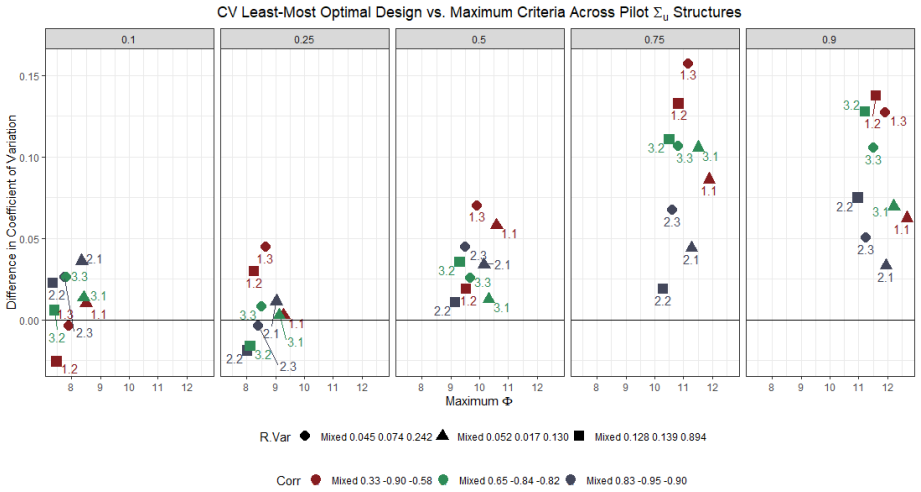


Figure 4.14: CV Most - Least Optimal Design vs. Maximum Φ Criteria Across CS Structures. Plot panels have the values of α in the columns. Shapes of the points are the Σ matrix variety and are colored based on the correlations of the pilot Σ_u matrices.

4.5.4 Optimality Across Variation Ratios of Interest

Rather than dividing the variabilities and correlations of the compound symmetric, unstructured, and pilot-style Σ_u and Σ matrices into segments, we can examine how our formulas outlined in subsubsection 4.4.2.1, summarizing the Σ_u and Σ matrices, affect optimality differences. Summaries were based on the values from Equation 4.35 to Equation 4.37. Although the plots may depict numerous potential relationships, analyzing the trends across all variations of Σ and Σ_u, featuring distinct variances and correlations, will be more accessible using these ratios. This is because the amalgamation of all versions from above provides a clearer understanding of the behavior of the relevant values of interest, namely h^2, $det2$, and $eigen.ratio$. While possible to display many different combinations of values, we found that not all were

visually interesting to compare. As such, we focused on all data overall for some cases, and for others, only pilot data covariance matrices may offer more noteworthy cases.

Over the larger set of combinations we used for our simulations, there are some interesting trends to note. All of these cases will be the ones from the sections above that used the same first 50 bean lines from the A genetic relationship matrix. Initially, by examining the ratios of h, $det2$, and the $eigenratio$ against the maximum optimality criteria Φ values for all cases, we can identify certain relationships. To explore these relationships in greater detail, we will use additional graphics with a more targeted focus. Some plots may not indicate as many groupings in both the Σ and Σ_u matrices because the purpose of incorporating values like heritability is to provide a comprehensive summary of the variability in both matrices.

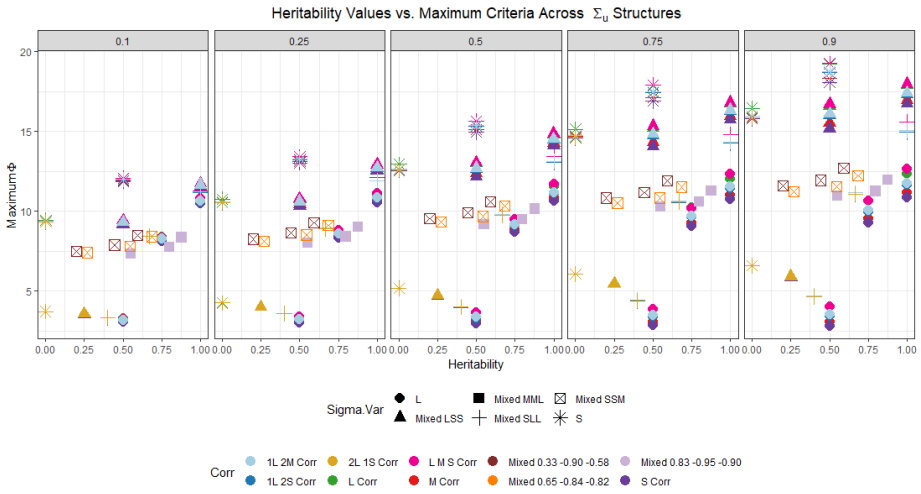


Figure 4.15: Heritability (equation 4.35) Summaries vs. Maximum Optimality Criteria Φ. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

Across the values of heritability in Figure 4.15, there are some groupings of points where, as the values of heritability increase, the design optimality values do as well. The residual matrix format, which was excluded from this plot's graphical

summaries, contributes to the separation between the points in different groups. The clustering of points at the bottom of the plot (with optimality values in the range of five) signifies that as the heritabilities increase, the optimality values decrease more sharply across the range of α values. This phenomenon is predominantly noticed in regions where the group's residual variability is defined by large variance values. In these cases, the variation in the points across their Σ_U categories leads to a reduction in the Φ value, with the highest Φ being associated with smaller Σ_u variances that decrease towards higher variabilities around the circular points on the heritability values of 0.5. Although the overlapping points make it challenging to interpret, the standout color of some points indicates that for the 0.5 heritability points with high Σ and Σ_u variances, groups with lower correlation values exhibit smaller Φ values, whereas groups with higher correlation values, such as the pink points for the LMS correlation group, display larger optimality criteria values. By generating multiple Σ and Σ_u matrices at different levels of heritability, researchers can gain insights into how the design optimality may vary when adapting models to account for diverse taxa and their varying relationships, as well as different levels of unaccounted residual variability in the model.

It is less interesting to identify the residual Σ relationships for the pilot cases, but these points can be seen in the middle of the optimality values with the boxed shape points, with x's through the box and the filled-in square. When α is weighted at 0.1, the cases are all quite similar to each other. However, as the Φ criteria emphasizes the importance of the random effects, the differences between the most optimal design values become more dissimilar. More focused cases using the heritability values will be presented to highlight pilot cases for specific levels of α.

There are similar relationships when the whole determinant of the Σ and Σ_u matrices are considered instead of simply the trace of the values, which is the basis of

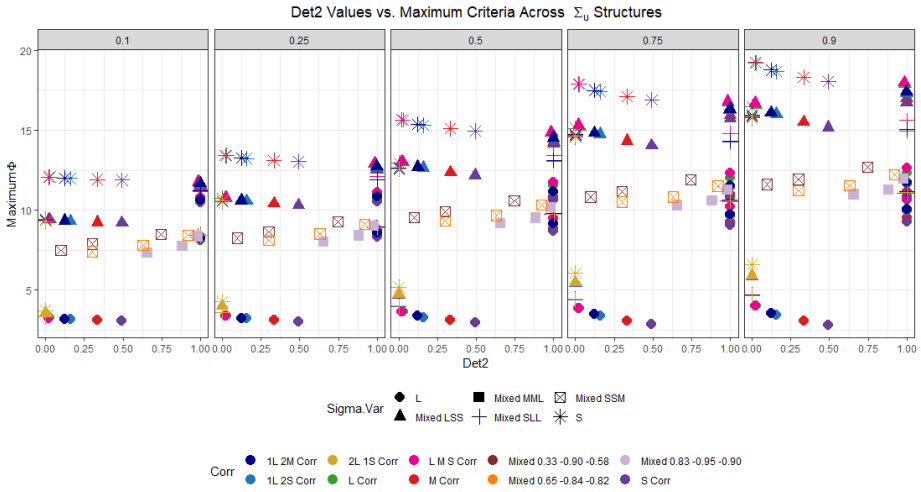


Figure 4.16: Det2 (equation 4.36) Ratio Summaries vs. Maximum Optimality Criteria Φ. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

the heritability calculations, in Figure 4.16. Because typical heritability calculations focus on the variance terms, the determinant ratio allows researchers to summarize the overall magnitudes of the taxa variance matrices. In our cases, the Φ values exhibit no change and maintain a constant level with lower criteria weights. However, the general pattern suggests that as the determinant ratio rises, the Φ value decreases, and by analyzing the color and shape of the points, researchers can identify the additional input components that contribute to this decline. In the pilot cases, the trends indicate that as the determinant ratio rises, so does the level of optimality. The box points, positioned between the 7.5 and 10 values of Φ, reveal that the points with identical shapes and colors have increasing determinant ratios, depending on which of the three residual matrices was employed in the simulation. The optimality criteria for the three points differed, with the second Σ matrix producing the smallest values, followed by the third and the first matrix having the largest. The second Σ matrix had larger variances and more correlated taxa compared to the other two. The third matrix had medium correlations and variances, while the first one had even

smaller variabilities and correlations. Therefore, it appears that in the pilot cases, the relationship between the determinant and Φ values increased due to a decrease in variability and correlations.

The eigenvalue ratio (Equation 4.37) values consider a slightly different comparison of the two variance matrices, and we can see different relationships within some of the data groupings in our simulations. Researchers who perform an analysis akin to that of Chapter 3, which centers on polymicrobial traits and the eigen analysis of variability matrices, may find these values valuable and may wish to base their design structure on the most optimal design for their pilot data of interest. Again, there are separate trends in our made-up examples versus our pilot examples. Across all variations of the variability summaries, many of our pilot data examples will provide a helpful real-life mirror for what a researcher may bring into a statistician when trying to design their experiments.

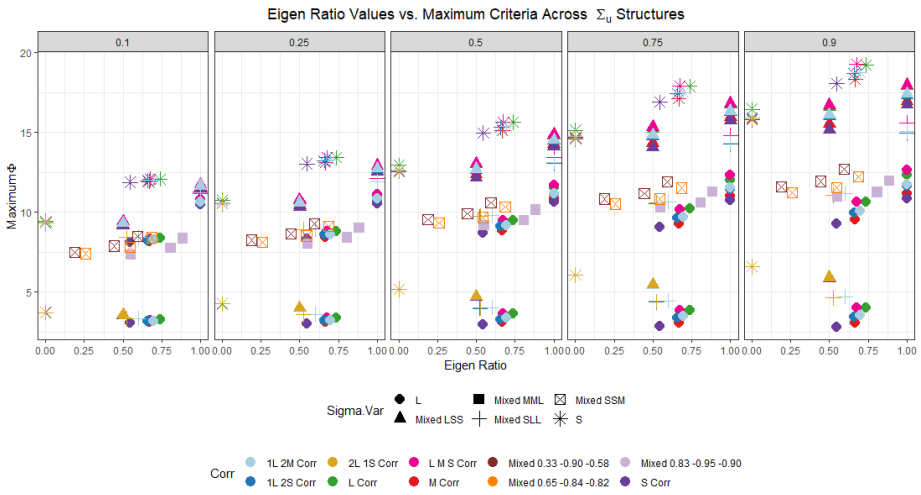


Figure 4.17: Eigen ratio (equation 4.37) Summaries vs. Maximum Optimality Criteria Φ. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

The plot in Figure 4.17 reveals that several cases are closely clustered together, and across different groups of cases, it is evident that as the eigenvalue ratios increase,

so do the corresponding Φ values. There are smaller clusters, such as the ones at the bottom from the large residual variance values. In this cluster, the optimality criteria for the 0.5, 0.75, and 0.9 weights exhibit a decline as the eigenvalue ratio increases. However, within the circle of points having large Σ_u variability, the Φ values begin to increase once again. The circle points for the large Σ_u variability group in the middle of the Φ values (with the pilot information) are a small group of cases with the mixed Σ case with one large and two small variances. The values of Φ for these cases are generally in the middle range. When the correlations within a group change, those with predominantly large correlations or correlations with larger differing magnitudes tend to yield larger optimality criteria compared to scenarios where the correlations are mostly small or a mix of small and medium values (in both cases with positive and negative correlations). For instance, as depicted in Figure 4.16, cases exhibiting all large correlations, as well as those with a mix of large, medium, and small correlations, generally display similar and higher optimality values.

Aside from examining the correlation between changes in the optimal criteria Φ values, we can also focus on how the alterations in optimality lead to changes in the structure of the Z design. To accomplish this, the plot will showcase how the variability of bean line replications changes based on variations in the ratio values. We will concentrate on the pilot cases to illustrate how statisticians can generate and present this information to researchers who want to determine more optimal replication structures for their 50 bean lines. As the variables are more interpretable versions of the variance of the number of replications across bean lines, we focus on plots with coefficient of variability values, and how ratios affect differences in the coefficients of variation between the least and most optimal designs. Colors for the new plots tie into the same correlation colors for the pilot cases as in the previous graphics, but the shapes have changed so that we can tie each case to its corresponding

pilot Σ_u case in Figure 4.5.

First, in Figure 4.18, Figure 4.19, and Figure 4.20, there are the plots looking at the coefficient of variation values for the pilot data cases vs. ratio values of h, g, $det2$, $det1$, and $eigenratio$. Within the plots, shapes of the values match which pilot version of the Σ_u matrix was used from Figure 4.5.

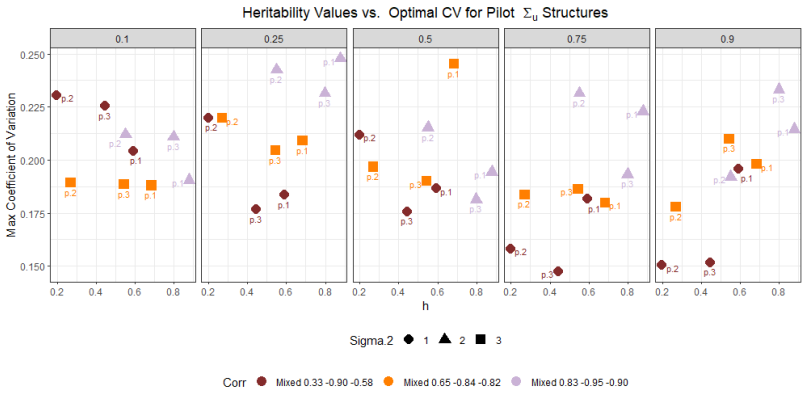


Figure 4.18: Heritability Summaries vs. Most Optimal CV. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

For the heritability, determinant2, and eigenvalue ratios, based on the way the ratios are taken, with the Σ_u values in the numerator of the ratio, these trends are the same in that as the ratio value goes up, the variation in the number of replications across the bean lines are going up as well. When the variation in the Σ_u matrix surpasses the overall variability and magnitude of the eigenvalues, the optimal design necessitates greater variation and spread in the number of replications across the bean lines. The most optimal designs would involve bean lines with varying levels of replication, corresponding to a larger coefficient of variation in the number of replicates.

However, again across figures 4.18, 4.19, and 4.20, at the smaller α levels, there are fewer changes in how the designs are structured across the corresponding x-axis value. The steeper descents across h, $det2$, and the $eigenratio$ occur for the 0.75

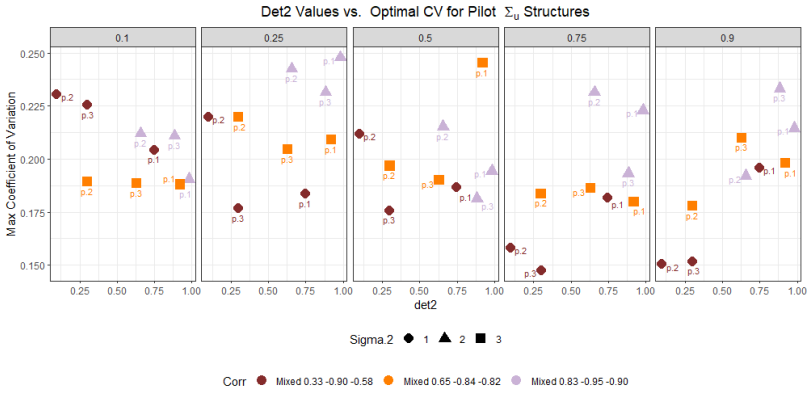


Figure 4.19: Det2 Ratio Summaries vs. Most Optimal CV. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

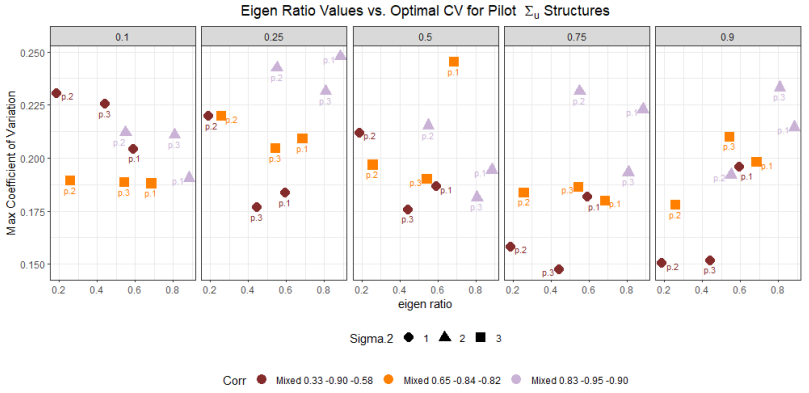


Figure 4.20: Eigenvalue Ratio Summaries vs. Most Optimal CV. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

and 0.9 α weights. Similarly, for these three ratios, the relationships among the CV outcomes exhibit similarity across cases. This suggests that evaluating all the ratios may not be necessary to compare differences in designs. Instead, researchers can choose a specific format based on their requirements and the factors they wish to assess when evaluating their designs.

Then, after looking at the variabilities on their own, the differences in the least and most optimal coefficient of variation values are plotted in figures 4.21 through 4.23. These plots highlight the impact of prioritizing optimality in the random effects calculation over the fixed effects. Notably, there is little change in the variability of

the Z design matrices that our simulation identifies as the least and most optimal when there is more weight on the fixed effects. For the equally weighted cases, there is more variance with larger CV values for the less optimal design. The difference values are all much more separated out above 0 for the 0.75 and 0.9 cases, but the trends remain similar for the heritability, $det2$, and $eigenvalue$ groups.

Upon examining the individual cases based on the colors, shapes, and labels of the points, several patterns emerged. However, not all of these patterns are consistent across each ratio or even across the various values of α. The first and third varieties of the residual matrix (points labeled p.1 and p.3) with the purple triangles tend to be closer together, with larger values of h, $det2$, and the $eigenvalue$ ratios, with the smaller differences in CV between the least and most optimal design. These are the cases, in purple, with the more largely correlated taxa in Σ_u and the triangles, corresponding to the variabilities of 0.183, 0.156, and 1.085, which are the largest of the variances out of the three pilot Σ_u matrices. The cases towards the other ends of the patterns tend to be the p.2 residual matrix, which has the largest residual variabilities out of the three options. The larger of the two p.2 cases is often the maroon case from Σ_u correlations with the smaller overall values, which edges out in most of the plots the orange case for the more mixed correlation values of 0.65, -0.84 and -0.82.

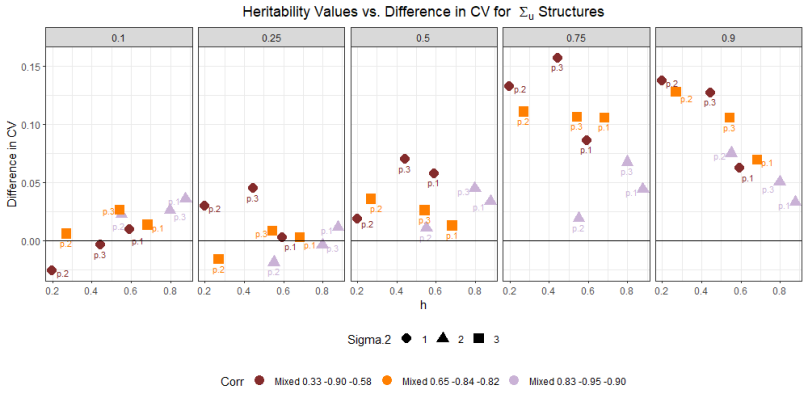


Figure 4.21: Heritability Summaries vs. Difference in CV. Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

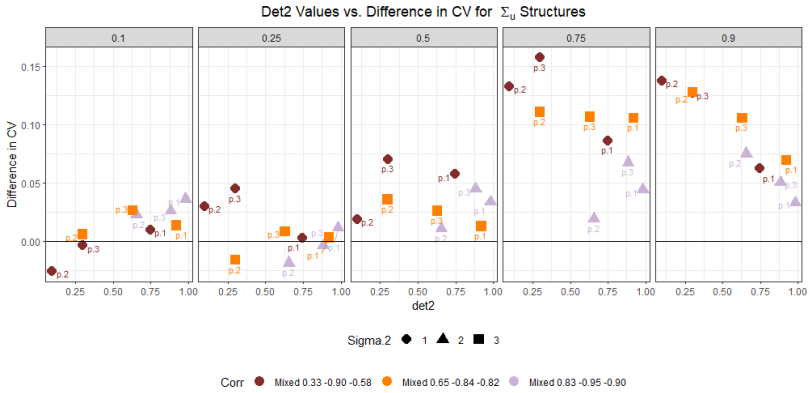


Figure 4.22: Det2 Ratio Summaries vs. Difference in CV (Least-Most Optimal). Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

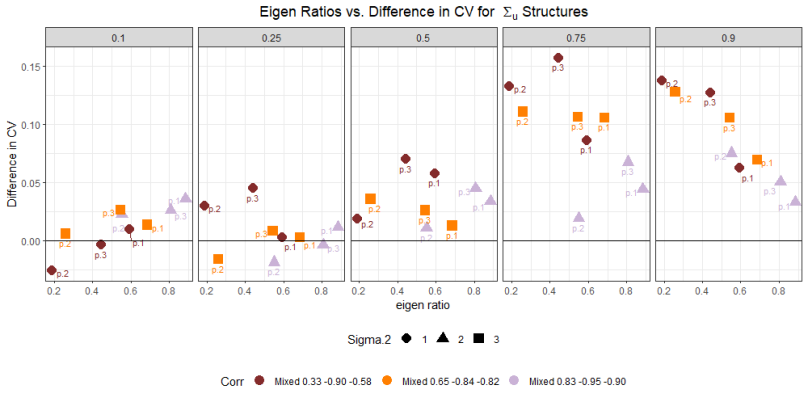


Figure 4.23: Eigenvalue Ratio Summaries vs. Difference in CV (Least-Most Optimal). Plots are paneled across values of α weights, with point shapes for the version of Σ_u and colored based on correlation.

4.5.5 Variance Structures Compared Across Two Genetic Relationship Matrices

For some cases, simulations were altered to look at a selection of bean lines containing only the first 50 Mesoamerican bean lines. These cases will be compared to those containing the first 50 overall lines, which were made up of a mix of Mesoamerican and Durango bean lines. Cases were not initially run for the CS variance cases but were adapted for the UN and pilot cases. For this section, we will show general results in changes in optimality for the unstructured cases but will focus on the changes from the pilot variance matrices. These provide a more representative picture of what would happen if researchers changed the bean lines selected for the experiment and then used pilot cases to see how the adaptations would change the optimality of their designs.

The plot in Figure 4.24 shows the relationship between the differences in the maximum optimality criteria Φ values across values of α for the unstructured Σ_u matrices with the standard three residual variance matrices. Points around zero would have designs that were equal in the optimality criteria produced between the two A matrices, and positive values show designs that were more optimal with the standard A genetic relationship matrix with a mix of both Mesoamerican and Durango lines. The mix of lines had a mix of correlation values, with small and large values. However, from Figure 4.2, we know that when the bean lines only contained Mesoamerican lines, as in the new A format, the lines were all a lot more highly correlated with each other. Note that there may be some trends that are missing, as for both styles of the A matrix the same cases were missing due to not being able to invert the random effects information matrices of interest.

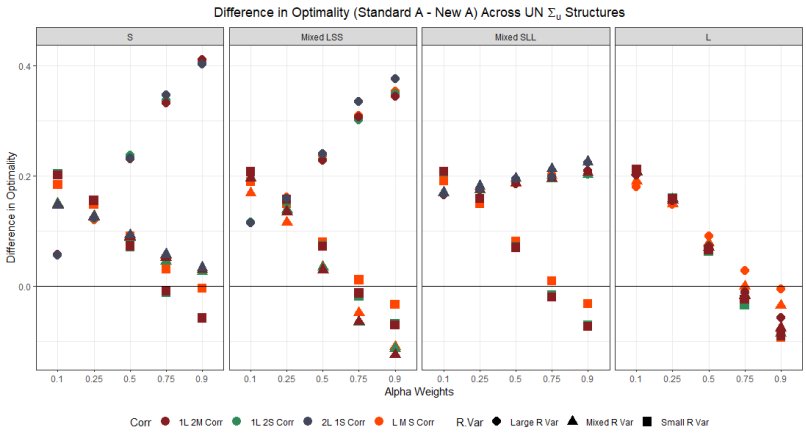


Figure 4.24: Difference in Maximum Φ Value, Standard - New A groups, across α weights for the UN cases. Plots are paneled across types of Σ_u matrices, with point shapes for the version of Σ and colored based on correlation.

Across the values of α, there are changing relationships between which combinations have increasing or decreasing differences between the optimality for the two genetic relationship matrices. Also, across the panels, the Σ_u matrices greatly affect how the groups respond optimally to the change in the two A matrices. For Σ_u matrices containing all, or primarily small, variance components, there are many more cases where the large residual variability cases increase in optimality as more weight is put on the random effects. For these cases, the standard A matrix creates larger Φ values by a difference of between 0.2 and 0.4. The relationship flattens more when more large variability is introduced, and for all cases when all large variances are utilized for Σ_u, the trend goes towards the new A becoming more optimal as we put more weight on the random effects. For the large Σ_u case, with all correlations and a weight of 0.9, the new A was more optimal by about 0.065 units on average. These values are still very close to zero, so as the relationship goes more toward all Mesoamerican lines having larger optimality, the two cases are still fairly close on average. Additionally, the points having the increasing relationship tend to be the maroon, blue, and green circle points, with the larger correlations from the large

residual variance group. Orange points, with a mix of a large, medium, and small correlation, always trend down toward the new A having larger Φ values.

In most cases, combining the Mesoamerican and Durango lines leads to the creation of more optimal designs. However, in some scenarios with smaller residual matrices (square points), the new A matrix may produce slightly more optimal designs, although the difference in optimality is often negligible. Increasing the α values places more emphasis on the information matrix containing the A matrix. As a result, these relationships suggest that as the optimality focuses more on the bean lines themselves, the Φ values will be relatively similar, regardless of whether the bean lines have a mixture of correlations or are predominantly correlated with each other. These findings also imply that a highly correlated A matrix with all Mesoamerican lines can lead to less optimal designs, particularly when equal or less weight is given to the random effects. Therefore, when using lines with this structure, greater attention should be paid to the design process to increase the likelihood of producing optimal designs.

The same comparison, but for the pilot structures, is similar to that for the decreasing trends across the UN cases. In Figure 4.25, all three panels across varieties of Σ_u are showing decreasing trends, whereas we put more weight on the random effects, the designs are more closely related to one another.

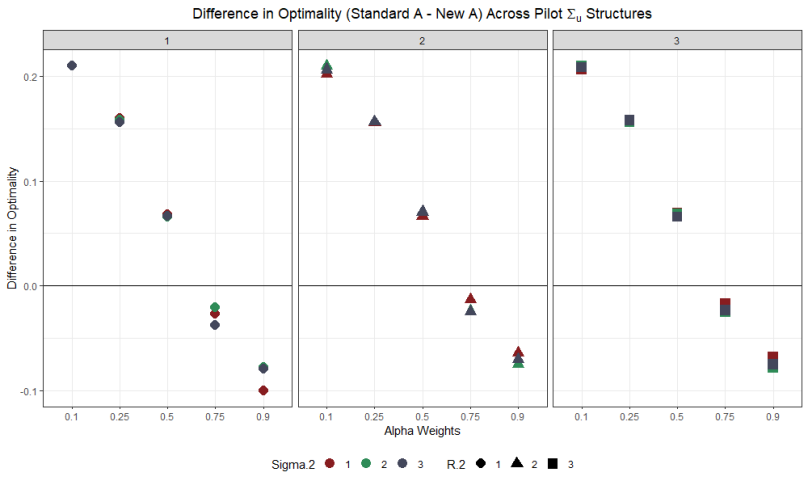


Figure 4.25: Difference in Maximum Φ Value, Standard - New A groups, across α weights for the nine pilot cases. Plots are paneled across types of Σ_u matrices, with point shapes for the version of Σ and colored based on correlation.

For the 0.1, 0.25, and 0.5 α weights, the standard A matrix produces a more optimal design versus the new A matrix for the remaining two cases. The Mesoamerican A produces the largest difference in optimality for the first Σ_u matrix with the red point for the first residual matrix. This case specifically is that of the first subject, as it is matrix one and one. From Figure 4.5, this grouping of the first of each style of the matrix has large, medium, and small correlations, with the smallest residual variance values of the three pilot Σ matrices. This grouping closely resembles a similar case to the orange square points in the UN plot above. Therefore, we can observe similar trends across both real-life and hypothetical examples, despite the pilot group having negative correlations for two out of the three values.

Figure 4.26 shows how close many of the cases Φ values are and where some of the larger differences occur across the 9 groupings of Σ_u and Σ matrices. The three types of Σ_u matrices are up and down across the rows and the columns represent the three Σ structures. The point labels take the optimality for the triangle standard A points from the mix of Mesoamerican and Durango bean lines and subtract the Φ

from the circle point from the all Mesoamerican bean line matrix.

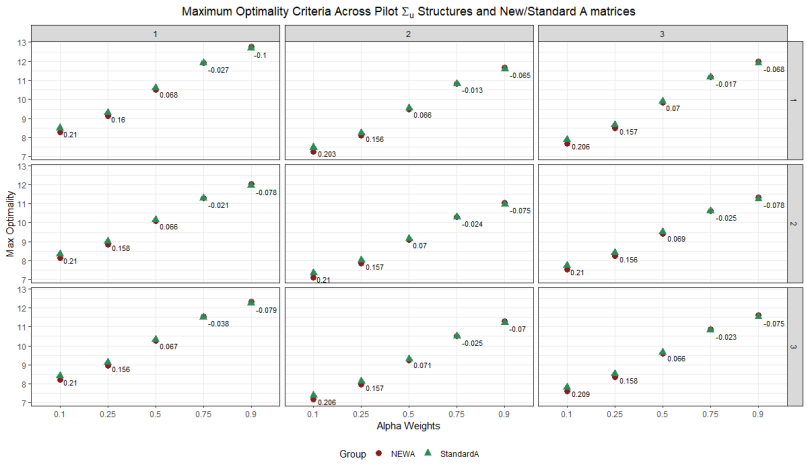


Figure 4.26: Maximum Φ Values, standard - new A groups, across α weights for the nine pilot cases. Plots are paneled in rows across types of Σ_u matrices and columns across the Σ structures. Point shapes and colors identify which A matrix was used. Labels on the points are the size of the difference between the optimality for the Standard - New A.

We can again see all the same relationships as in the previous plot but with more visualization of the sizes of each of the differences between Φ. There are some slight differences between groupings, but overall, trends across our pilot examples did not change much versus their closely related cases in the unstructured simulations. It seems that the type of A matrix and the bean lines employed will have a more significant impact on the optimality for extreme cases (like those additional cases in the UN structure) rather than on cases with relatively similar characteristics, as evidenced by the lack of variations across our nine similar pilot cases.

For the Pilot examples below, many of the cases are labeled based on their combination of the pilot Σ_u and Σ matrices. As a note, here are the matrices utilized for the 9 pilot combinations. A label *2.1*, for example, would correspond to a case with Σ_u matrix 2 and Σ matrix 1.

$$\Sigma_{u1} = \begin{bmatrix} 0.078 & 0.014 & -0.109 \\ 0.014 & 0.022 & -0.037 \\ -0.109 & -0.037 & 0.189 \end{bmatrix}, \; \Sigma_{u2} = \begin{bmatrix} 0.183 & 0.14 & -0.423 \\ 0.14 & 0.156 & -0.37 \\ -0.423 & -0.37 & 1.085 \end{bmatrix}, \; \Sigma_{u3} = \begin{bmatrix} 0.059 & 0.044 & -0.111 \\ 0.044 & 0.078 & -0.124 \\ -0.111 & -0.124 & 0.294 \end{bmatrix};$$

$$\Sigma_1 = \begin{bmatrix} 0.052 & 0.01 & -0.074 \\ 0.01 & 0.017 & -0.027 \\ -0.074 & -0.027 & 0.13 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0.128 & 0.111 & -0.321 \\ 0.111 & 0.139 & -0.317 \\ -0.321 & -0.317 & 0.894 \end{bmatrix}, \ \Sigma_3 = \begin{bmatrix} 0.045 & 0.038 & -0.088 \\ 0.038 & 0.074 & -0.11 \\ -0.088 & -0.11 & 0.242 \end{bmatrix};$$

Then figures 4.27 through 4.29 show the different ratios between the Σ_u and Σ matrices on the x-axes versus the Φ values for the optimal designs in the pilot cases, where the shape of the points now represent the style of A matrix used within the simulation. Across the panels are the varying α weights, and as the values increase, the all Mesoamerican A matrix becomes more optimal, and the gaps between Φ increase as we have seen before. Overall, these graphics are demonstrating very similar results to what we have seen before in other sections for the "standard" A matrix and reiterate what we know about the differences in A matrices from the overall plots earlier in this section. The pilot cases again do not show much variation across how the change in A affects overall optimality.

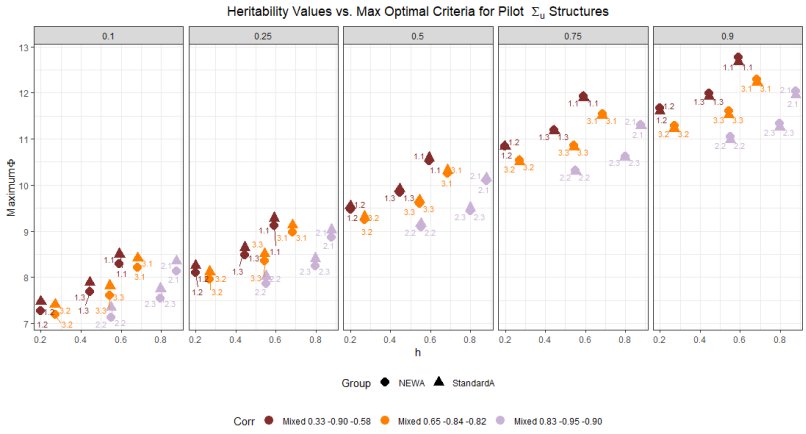


Figure 4.27: Heritability Summaries vs. Most Optimal Φ. Plots are paneled across values of α weights, with point shapes for the version of A and colored based on correlation.

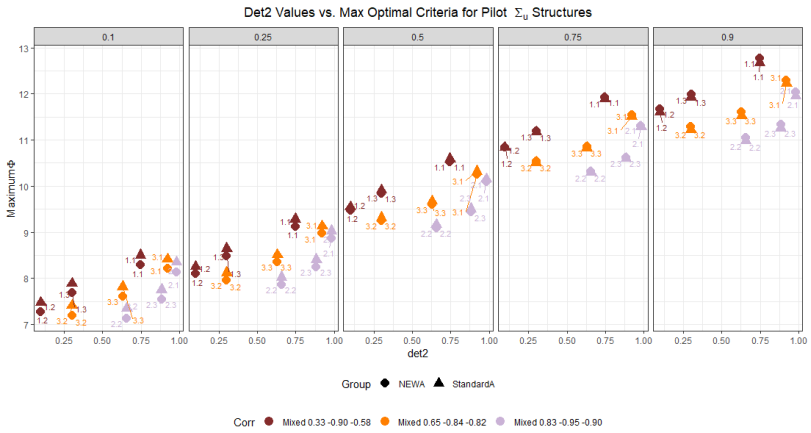


Figure 4.28: Det2 Ratio Summaries vs. Most Optimal Φ. Plots are paneled across values of α weights, with point shapes for the version of A and colored based on correlation.

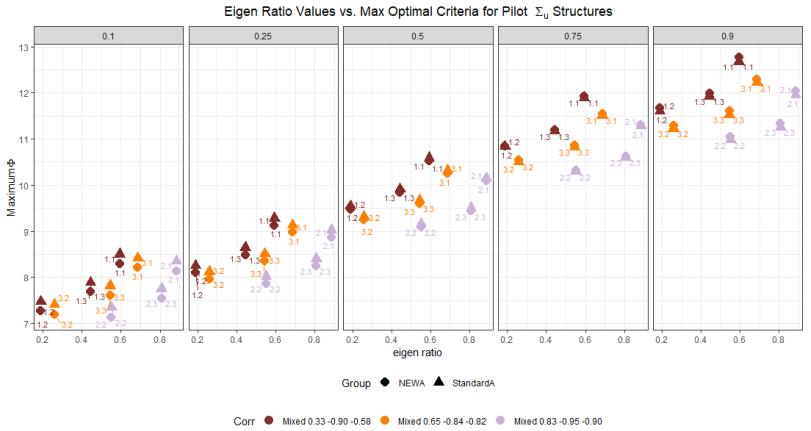


Figure 4.29: Eigenvalue Ratio Summaries vs. Most Optimal Φ. Plots are paneled across values of α weights, with point shapes for the version of A and colored based on correlation.

Figures 4.30 through 4.32 illustrate how the replication variability across the 50 bean lines changes based on which lines are included in the design structure. The graphics display the Σ_u and Σ ratios against the coefficient of variation for each of the designs, with colors representing the correlations. The circle points represent the Mesoamerican-only A groups, while the triangle points indicate the A matrix cases across a mixture of Mesoamerican and Durango bean lines. When following the points along the x-axis, cases from the same variability groups have identical ratio values.

The different A points are positioned either above or below each other, depending on which group had a higher coefficient of variation value.

In all plotted ratio values, the new A cases with all Mesoamerican lines exhibit higher coefficient of variation values in most scenarios. The optimal designs seem to be those with greater replication variances among the bean lines in groups with all Mesoamerican lines, even when the last set of plots indicates that highly correlated lines do not always result in optimality values that differ from those with mixed lines. The trends in the coefficient of variation (CV) values for all Mesoamerican bean lines, represented by circle points, are generally consistent with the triangle points, but they exhibit a less pronounced ascent or descent across the x-axis ratio values. For the size of the gap between the two A matrix cases, points are closer together with more weight on the fixed effects, which again mirrors what we had seen before for the Φ values.

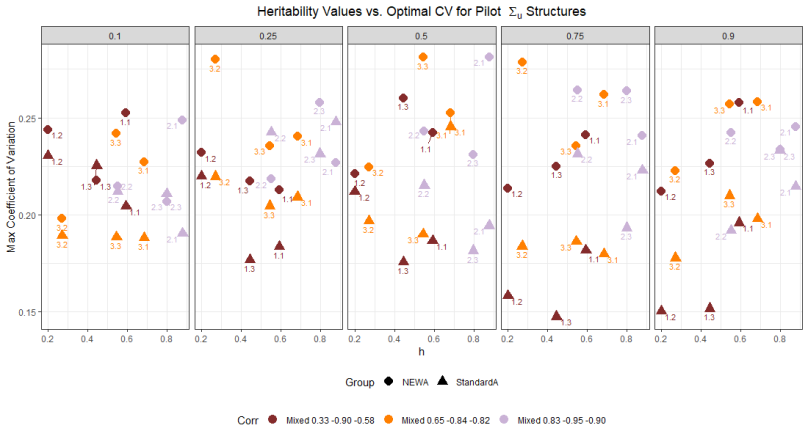


Figure 4.30: Heritability Summaries vs. Most Optimal CV. Plots are paneled across values of α weights, with point shapes for the version of A and colored based on correlation.

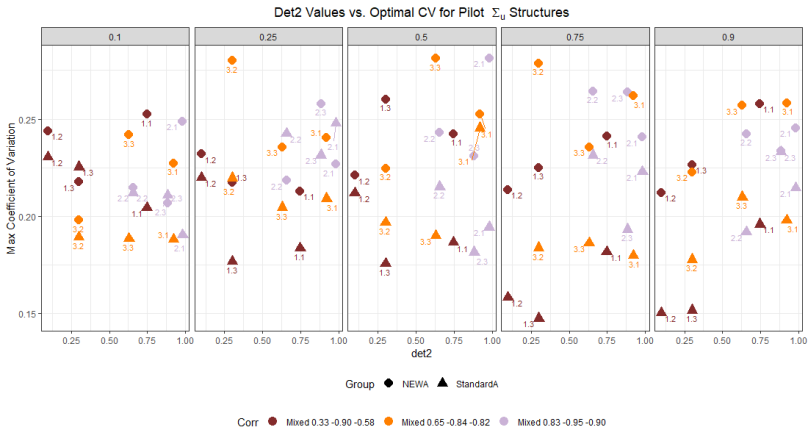


Figure 4.31: Det2 Ratio Summaries vs. Most Optimal CV. Plots are paneled across values of α weights, with point shapes for the version of A and colored based on correlation.

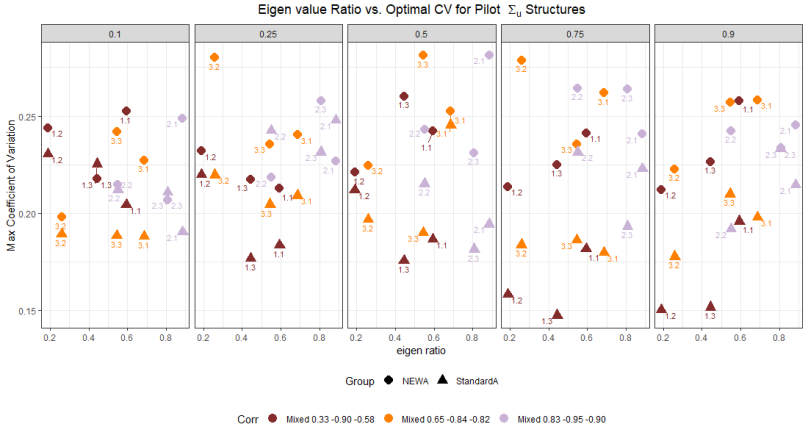


Figure 4.32: Eigenvalue Ratio Summaries vs. Most Optimal CV. Plots are paneled across values of α weights, with point shapes for the version of A and colored based on correlation.

Then plots in figures 4.33 through 4.35 demonstrate how the different A matrices alter the difference in how the coefficient of variation values change between the least and most optimal design. The difference in CV values is plotted on the y-axis. Then we can observe if the gap between the least and most optimal designs changes between the A bean lines from the mix of Mesoamerican and Durango lines in the standard A matrix vs. from the only Mesoamerican lines in the new A matrix.

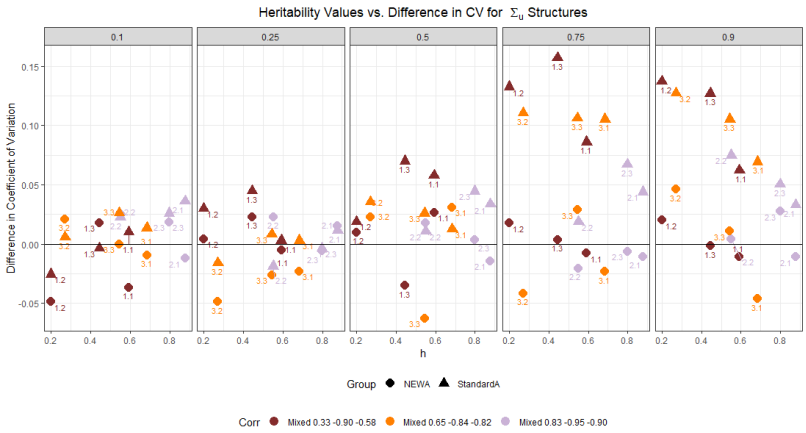


Figure 4.33: Heritability Summaries vs. Difference in CV. Plots are paneled across values of α weights, with point shapes for the versions of A and colored based on correlation.

The relationship changed from the plots of the CV values, where now the triangle points for the standard bean lines are larger than for the new A matrix. This demonstrates that the disparity in coefficients of variation for the least versus most optimal designs is larger for the mixed bean line version of A, where the least optimal designs tend to be more variable with larger CVs. The circle points fall more closely to the line around zero, where there is not much change between the variability in the two designs with the smallest and largest values of Φ. Points that fall below zero indicate that the most optimal designs with the largest Φ value had a larger coefficient of variation versus the least optimal design. Many of these negative differences are still fairly small, so the designs with circle points are relatively close to each other. Additionally, as the values of the α weights increase, all designs across the new and standard versions of A are raising above the line where the most and least optimal designs have the same CV value. This demonstrates that there are larger gaps between the "best" and "worst" designs based on Φ the more weight is put on estimating the random effects across the information matrices.

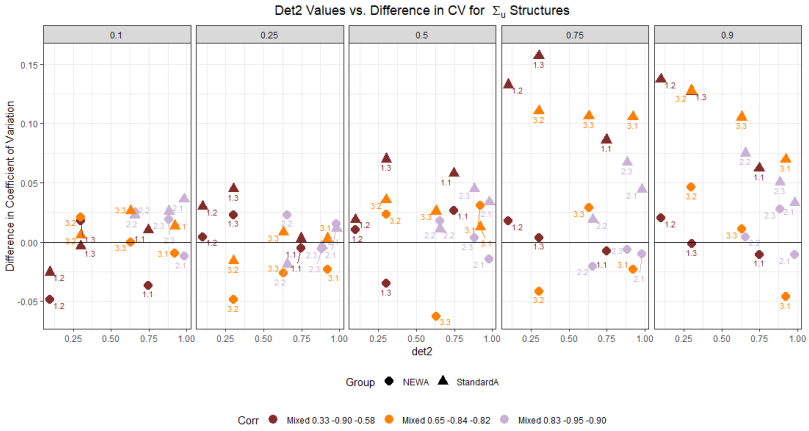


Figure 4.34: Det2 Ratio Summaries vs. Difference in CV (Least-Most Optimal). Plots are paneled across values of α weights, with point shapes for the versions of A and colored based on correlation.

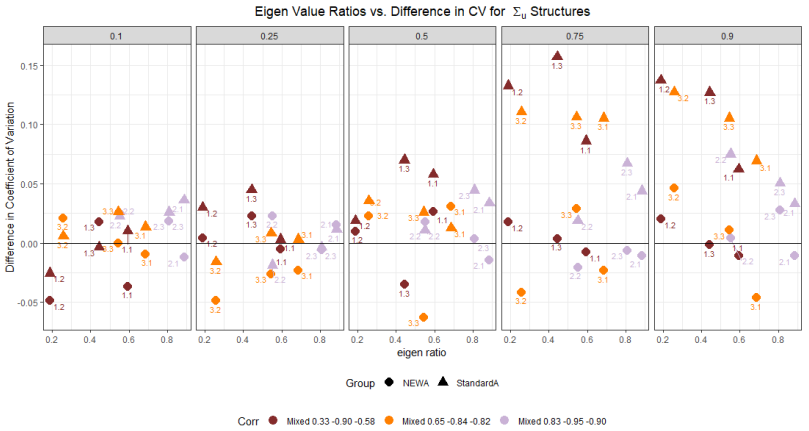


Figure 4.35: Eigenvalue Ratio Summaries vs. Difference in CV (Least-Most Optimal). Plots are paneled across values of α weights, with point shapes for the versions of A and colored based on correlation.

4.5.6 Simulated Comparisons with Z Matrices from Real Experiment

After analyzing the bean line cases across various hypothetical and pilot structures, we hypothesized whether these findings would remain consistent when using the entire A matrix in an expanded experiment of the same size as that conducted for the three subjects in Chapters 2 and 3. In these instances, the optimality of the experimental design developed in Chapter 2 for each subject could be compared to

that of the best design resulting from the 5000 simulated Z matrices. To create these cases, the code used for all other simulations was adjusted to include the entire A matrix and all 297 bean lines spread out among 12 plates with 87 open wells per plate. Now on average, bean lines could have 3.52 replicates per plate. In our experiment from Chapter 2, all bean lines were replicated at least three times, with the selected "check" lines replicated at most seven times. These were the lines that were replicated once in each of the three design reps and then once again in each plate within one particular design rep.

The X and Z matrices for each subject's final design was generated in R using the data from the output data set used for data analysis in Chapter 3. Instead of simulating Z matrices, this was used as the Z matrix in our function. The optimality of this Z matrix was computed for each subject at each level of α. To compare optimality values from our real Z matrices to randomly generated cases, we ran new simulation codes to determine the most optimal Z matrix as we did previously for the pilot cases. We employed the same methodology as before, but this time with the entire A matrix. Plots 4.36, 4.37, and 4.37 provide a comparison of the difference in Φ values between the best design obtained from the simulated Z cases and the optimality derived by using the Z matrix for each of the three subjects, across each of the nine pilot variance combinations. The specific combinations tied to each subject can be seen below.

$$Subject\ 1 \rightarrow \Sigma_{u1} = \begin{bmatrix} 0.078 & 0.014 & -0.109 \\ 0.014 & 0.022 & -0.037 \\ -0.109 & -0.037 & 0.189 \end{bmatrix},\ \Sigma_1 = \begin{bmatrix} 0.052 & 0.01 & -0.074 \\ 0.01 & 0.017 & -0.027 \\ -0.074 & -0.027 & 0.13 \end{bmatrix};$$

$$Subject\ 2 \rightarrow \Sigma_{u2} = \begin{bmatrix} 0.183 & 0.14 & -0.423 \\ 0.14 & 0.156 & -0.37 \\ -0.423 & -0.37 & 1.085 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 0.128 & 0.111 & -0.321 \\ 0.111 & 0.139 & -0.317 \\ -0.321 & -0.317 & 0.894 \end{bmatrix};$$

$$Subject\ 3 \rightarrow \Sigma_{u3} = \begin{bmatrix} 0.059 & 0.044 & -0.111 \\ 0.044 & 0.078 & -0.124 \\ -0.111 & -0.124 & 0.294 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.045 & 0.038 & -0.088 \\ 0.038 & 0.074 & -0.11 \\ -0.088 & -0.11 & 0.242 \end{bmatrix};$$

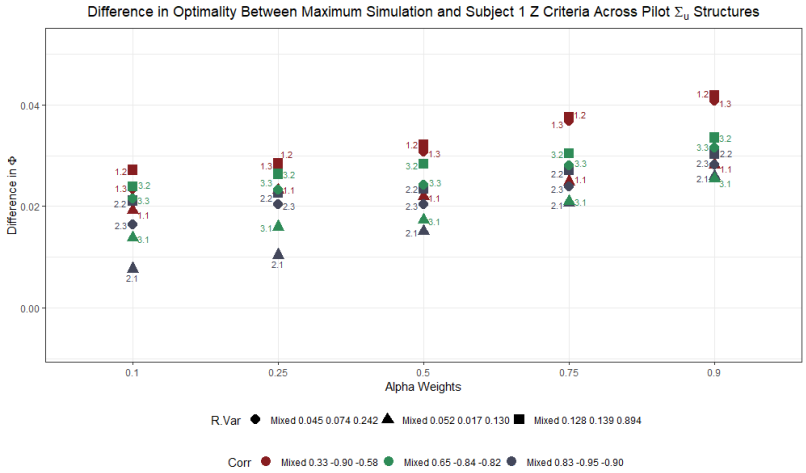


Figure 4.36: $\Phi_{Opt} - \Phi_{Actual}$ difference in optimality for optimal designs from simulated pilot examples with all 297 bean lines compared with the Z design matrix from Subject 1's experiment. All points are labeled with their $\Sigma_u.\Sigma$ values for the variance groups, where Subject 1's estimated variability pilot case is labeled as *1.1*. Shapes correspond to Σ variability and colored based on correlations.

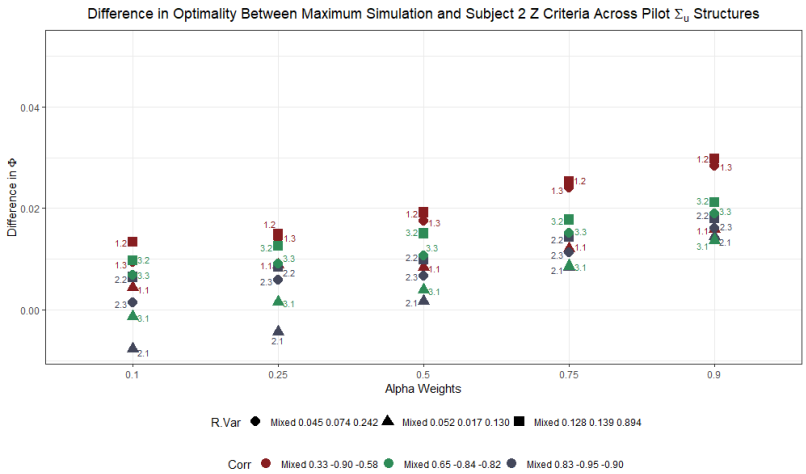


Figure 4.37: $\Phi_{Opt} - \Phi_{Actual}$ difference in optimality for optimal designs from simulated pilot examples with all 297 bean lines compared with the Z design matrix from Subject 2's experiment. All points are labeled with their $\Sigma_u.\Sigma$ values for the variance groups, where Subject 2's estimated variability pilot case is labeled as *2.2*. Shapes correspond to Σ variability and colored based on correlations.

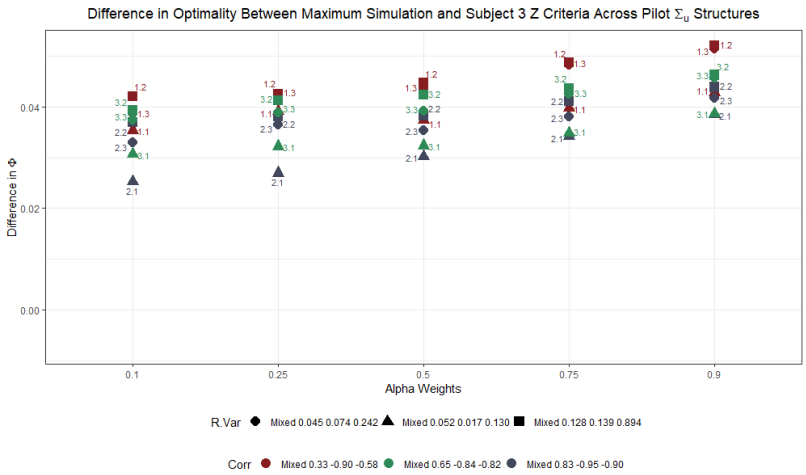


Figure 4.38: $\Phi_{Opt} - \Phi_{Actual}$ difference in optimality for optimal designs from simulated pilot examples with all 297 bean lines compared with the Z design matrix from Subject 3's experiment. All points are labeled with their $\Sigma_u.\Sigma$ values for the variance groups, where Subject 3's estimated variability pilot case is labeled as *3.3*. Shapes correspond to Σ variability and colored based on correlations.

Each plot corresponds to a different subject's Z matrix. Upon examining the relationships in the three plots, we observe increasing trends for all subjects across the range of alpha weights. As the α weight increases, the cases become more similar in how different the randomly selected designs are from the ones used for subjects 1, 2, and three in our experiments. All three of the different plots use the same y-axis boundaries, and comparing the three plots, the design from subject 1 was in the middle range of the values for how different the designs were in optimality. Then, subject 2 had the smallest differences between optimality values, even having a few cases where our design was more optimal than the best matrix out of the randomly generated cases. Lastly, the third subject's Z design was the one with the most different values in optimality, where the Φ values were larger for the randomly generated best design.

This ordering matches with the overall average number of replicates in each of the three experimental designs, where subject 3's design had the smallest number of average replications per bean line, with 3.33 wells per bean line on average, followed by

subject 1 with 3.38, and subject 2 with 3.42. For the randomly generated cases, they have 3.52 replications per bean line on average, equal to taking the number of wells and dividing by the number of bean lines, taking $\frac{Total\ Lines}{Number\ of\ Bean\ Lines} = \frac{1044}{297} = 3.52$. Based on plots 4.36, 4.37, and 4.38 it appears that the designs become more similar to the best randomly selected Zs when the number of replications on average per bean line was the largest for subject 2 (in Figure 4.37), where subject 2's design had 3.42 replicates per bean line on average. Thus, the simulated best cases have slightly larger Φ values, it appears that more balanced cases across the replications of the best lines are potentially better designs. Throughout all the plots, the trends in how the various combinations of Σ and Σ_u respond to design changes remain consistent. The maroon square cases, representing the smallest correlations with the larger residual variabilities, are located towards the top and are the combinations that differ the most between the two design types. Conversely, many of the cases with the second Σ_u structures, represented by the blue triangles for the mixed residuals with smaller variabilities, lead to closer optimality values between the best selected random design and the subject's structure 1, 2, or 3.

Lastly, there is more of a spread in the differences between the designs when more weight is put on the fixed effects. This makes sense compared to only our randomly generated cases because now, the subjects' designs have different X matrices. This is not due to our randomization, but rather because in the real-life experiments, not every one of the 87 wells was measured due to missing observations. So, when we focus on the fixed effects, there are differences with the randomly generated design maximum. We opted to use the real-life Z matrix derived from the data analysis to assess the optimality of our design, given the actual variability observed in these types of experiments. From the originally created designs, all wells possible were not filled as there were three of the 300 initial lines removed. So, from the initial designs

provided to the researchers, there were already differences between how many wells were filled in the design, thus, using the Z matrices from our final data sets across the three subjects gives us an appropriate representation of what our designs are as utilized in our experiments. As we will address in the discussion of limitations and future work, numerous other types of comparisons could be conducted between actual real-world design examples and the experimental design selected as the best outcome from random or other search algorithms.

Ultimately, if we would like to assess how structure in the Σ_u matrices affects the design optimality, we would want to try provide intuition behind these relationships. In this case, we are seeing that the cases with the first Σ_u structure has the largest optimality values, followed by case 3, than 1. These are the same relationships we were seeing when only 50 of the bean lines were used. Looking at the final matrices, we could compare their forms. Below is a repeat of a previous graphic with the formations across all pilot cases. In Figure 4.39, we see that the case one of Σ_u in the top set of tables, has the smallest variabilities, with the largest spread in correlation values, across 0.33, -0.9, and -0.58. Then, the case that provides the smallest optimality values, case 2, has the largest and most similar correlation values, 0.83, -0.95, and -0.9, along with the largest variability values. Together these combinatiosn of variance and correlations are still contributing the magnitude of the optimality values, similar to what we had seen in previous pilot and UN cases.

Pilot SigmaU and Sigma Variability Estimates

Case	Variance			Correlation			Final Matrix		
1	0.078	0	0	1	0.33	-0.9	0.078	0.014	-0.109
	0	0.022	0	0.33	1	-0.58	0.014	0.022	-0.037
	0	0	0.189	-0.9	-0.58	1	-0.109	-0.037	0.189
2	0.183	0	0	1	0.83	-0.95	0.183	0.14	-0.423
	0	0.156	0	0.83	1	-0.9	0.14	0.156	-0.37
	0	0	1.085	-0.95	-0.9	1	-0.423	-0.37	1.085
3	0.059	0	0	1	0.65	-0.84	0.059	0.044	-0.111
	0	0.078	0	0.65	1	-0.82	0.044	0.078	-0.124
	0	0	0.294	-0.84	-0.82	1	-0.111	-0.124	0.294

Pilot Residual Variance Matrices

Case	Variance			Correlation			Final Matrix		
1	0.052	0	0	1	0.33	-0.9	0.052	0.01	-0.074
	0	0.017	0	0.33	1	-0.58	0.01	0.017	-0.027
	0	0	0.13	-0.9	-0.58	1	-0.074	-0.027	0.13
2	0.128	0	0	1	0.83	-0.95	0.128	0.111	-0.321
	0	0.139	0	0.83	1	-0.9	0.111	0.139	-0.317
	0	0	0.894	-0.95	-0.9	1	-0.321	-0.317	0.894
3	0.045	0	0	1	0.65	-0.84	0.045	0.038	-0.088
	0	0.074	0	0.65	1	-0.82	0.038	0.074	-0.11
	0	0	0.242	-0.84	-0.82	1	-0.088	-0.11	0.242

Figure 4.39: Pilot Variability Cases for Σ_u with Differing Correlations from Bacteroides, Sutterella, and Bifidobacterium across our three subjects

Finally, we computed the variability and average replication per bean line for our real cases, allowing us to compare each subject's design using a coefficient of variation value. The combinations 1.1, 2.2, and 3.3 of Σ_u and Σ in the pilot cases corresponded to the variability estimated from the associated multivariate models fit with the three selected taxa (Bacteroides, Sutterella, and Bifidobacterium) for each subject (for example, case 1.1 corresponds to the variability from subject 1, 2.2, for subject 2). Both the replication variance and coefficient of variation values were larger than those from the real experiments at the bottom in yellow. Once again, it appears that the randomly generated designs identified as optimal have greater variability in the number of replicates per bean line than those utilized in our data analysis. This is understandable, considering that in Chapter 2, we intentionally incorporated approximate balance within each line of our designs. Moving forward, it would be valuable to investigate these patterns further to determine whether the

approach adopted in Chapter 2 was essential or if a less balanced design, with varying degrees of replication across lines, would be more suitable.

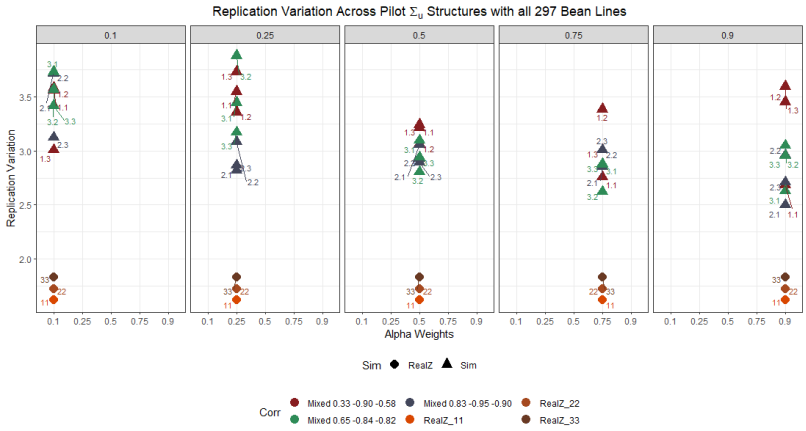


Figure 4.40: Replication variability compared between nine pilot examples compared to real Z matrices used in experiments in Chapters 2 and 3. Points are labeled with their $\Sigma_u.\Sigma$ values for the variance groups, with shapes corresponding to whether the Z matrices were real or simulated and colored based on correlations and real Z combinations.

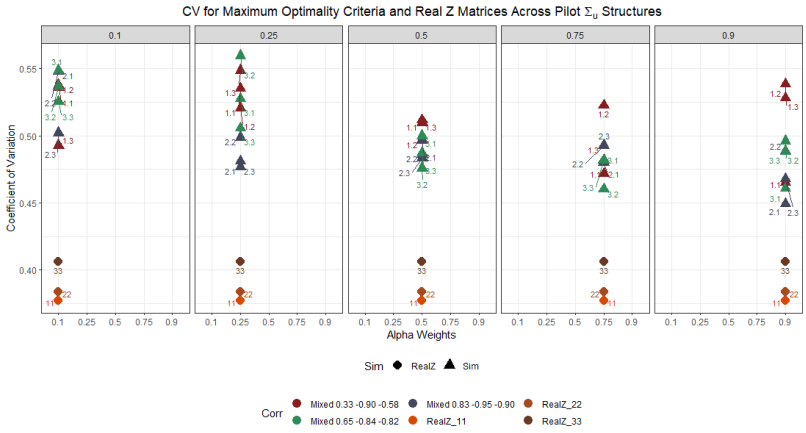


Figure 4.41: Coefficient of variation between nine pilot examples compared to real Z matrices used in experiments in Chapters 2 and 3. Points are labeled with their $\Sigma_u.\Sigma$ values for the variance groups, with shapes corresponding to whether the Z matrices were real or simulated and colored based on correlations and real Z combinations.

4.6 Conclusions

4.6.1 Summary and Discussion

Throughout this chapter, the presented examples have laid the groundwork for the versatility and adaptability of the simulations developed to explore the application of experimental optimal design methodology in genetic microbiome experiments. This objective aimed to identify and explore the impact of various experimental parameters on statistical optimality criteria when analyzing data from microbiome well-plate experiments. The literature reveals a knowledge gap in optimal experiment design for multivariate mixed models incorporating genetic effects, particularly in microbiology and studying bacterial community changes in the human gut. This gap is evident in both statistics and microbiology, suggesting a need for further research to bridge this gap and provide a more comprehensive understanding of the experimental design strategies for multivariate mixed models in microbiological research.

The results presented in this chapter highlight the significant impact of variability inputs, such as Σ, Σ_u, and A, along with the framework of the Z matrices, on the composite optimality criteria Φ. Understanding how these factors interplay can provide crucial insights for researchers who want to optimize their experimental designs. Additionally, the values of α as weights on the optimality criteria can directly affect the optimization process, particularly concerning fixed or random effects of interest to the researchers. Thus, considering the effects of both the variability inputs and α values can provide a more comprehensive understanding of the factors influencing the effectiveness of experimental design choices.

The simulation results, which included combinations of Σ_u, Σ, A, and Z, revealed numerous relationships between these values and the ultimate optimality of Φ. The analysis of the covariance structures revealed that the experimental designs for

cases with small Σ variability tend to be more optimal compared to those with mixed and large variance. Furthermore, it was observed that when the three taxa were all highly positively correlated could yield more optimal designs within the groupings of varying Σ variance structures. We saw across all examples, the differing correlation and variability structures of Σ_u affected the Φ values. Looking ahead, it would be worthwhile to consult with microbiologists and investigate the feasibility of gathering information about an individual's dietary and lifestyle habits to predict the correlation patterns of their gut bacteria. If such an approach were viable, it could potentially inform a selective subject recruitment strategy to optimize experimental design. The changes in optimality criteria values across the selected unstructured Σ_u cases for varying α weights indicate that Φ values exhibit much steeper changes for small variance values of Σ_u than for larger variances. Results demonstrated that the performance of the selected designs is more sensitive to changes in the composite Φ criterion weights when the variance values in Σ_u are low. In general, there were no consistent patterns in the results for these cases, and the relationships between the criteria and the replication variance are inconsistent with how the variances change as the types of correlation change. However, there is a tendency toward closer coefficient of variation values between the most and least optimal designs for the weights more on the fixed effects. In the UN cases, we observed that discerning optimality and replication variability changes were more challenging for cases with similar overall characteristics, such as varying correlations and mixed variabilities.

Our pilot data included combinations from three taxa (Bacteroides, Sutterella, and Bifidobacterium) Σ_u and Σ matrices created based on the data from our real experiment. Pilot data can assist researchers in determining optimal protocols and procedures when establishing a comparable experiment. By prioritizing the optimality criteria of the random effects, researchers can more clearly distinguish be-

tween the types of designs that produce optimal values for the Φ criteria, revealing larger differences between them. Again, wider gaps in Φ were observed for larger α weights. Regardless of the weights, residual Σ variability continues to play a role in which cases produce the most optimal designs. Although challenging to manage, researchers should take steps to mitigate extraneous noise in the data and account for variables that contribute significantly to variability in the statistical model of interest. This thought also highlights the benefit of integrating a selective subject recruitment strategy to pick subjects that may provide for a more optimal design and helpful information for analysis. In instances with identical Σ matrices, the correlation type and Σ_u variability consistently maintained a similar ordering. When the Φ values increase for some cases across α, the replicate variability decreases, and the more optimal designs have more consistent numbers of replicates across the 50 bean lines. Our recommendations for researchers to achieve the most optimal design should primarily emphasize the importance of reducing residual variability and that it is crucial to consider the taxa being included and how their variability is affected by the bean lines. Additionally, we found that the most significant differences between optimal and suboptimal designs were in cases where there was a large variation in the number of replications across bean lines.

Although these recommendations are not novel and mainly focus on reducing overall variability, which is already a goal of researchers using various methods such as adjusting lab structure, experimental design, and selecting appropriate analytical tools, they remain valuable insights despite being based on general cases. To try and make more guided recommendations, we used various ratios to summarize the matrices and evaluate their impact on design optimality. Bueno Filho and Gilmour emphasized that incorporating information on relatives could enhance analysis, resulting in several proposed extension models; However, no research has yet been

conducted to investigate the impact of genetic relatedness on the efficacy of block designs in different treatments [28].In the future, we could look into if there are better A matrix structures that facilitate more optimal designs.

We examined the impact of multivariate genetic extensions on the relationships between Σ_u and Σ on the optimality and composition of the most effective Z design matrices, drawing on our findings and relevant literature. We found that as heritability and determinant ratios increase, there were mixed relationships in the variability matrices that caused Φ to increase or decrease. Pilot cases across tended to have increased Φ values as the ratios increased. For example, though, within cases with a particular Σ and Σ_u, variability tends to decrease in optimality as heritability and det2 ratios increase in value. For the comparisons across all of the Σ and Σ_u cases for the eigenvalue ratios, some relationships tend to be much flatter and then increase for most cases in the larger α weights. This is an important contradiction to note based on using a function of Fisher information as our optimality criterion. This criterion seems to not recognize the importance of a large genetic variation relative to the overall error variance. Thus, this might lead us to try different optimality criteria in the future, like the function of heritability we used in Equation 4.35. Then we could find designs that lead to the largest values of heritability, where the genetic variation accounts for a large proportion of the overall variation.

For the pilot cases, we examined how the ratios affect the structure of the design by comparing the variation in the number of replications across bean lines. There was little change in the variability of the design matrices when there was more weight on the fixed effects. However, for α weights of 0.75 and 0.9 (with more weight on the random effects), as the ratios of heritability, det2, and the eigenvalue ratio increase, the coefficients of variation also increase in the optimal designs. As a researcher becomes more interested in cases where Σ_u variability explains a larger

proportion of the overall variation, it becomes necessary to increase the variability in the number of replications across bean lines to achieve an optimal experimental design. In conclusion, we observed that when comparing designs with the smallest and largest values of Φ, the similarity between the designs was higher for smaller α weights, when researchers would be more concerned with the fixed effects, and for larger values of heritability, det2, and eigenvalue ratios. For researchers, designs with larger values of these metrics may be more relevant, and thus the selection of the optimal design may have fewer implications.

Overall, this section provides insight into the potential impact of the choice of covariance matrix ratio on the design matrices' variability and the experiment's overall effectiveness. However, more research is needed to determine the optimal ratio of covariance matrices for a given experimental design. These explorations provide insights into how the design optimality may vary when adapting models to account for diverse taxa and their varying relationships and different levels of unaccounted residual variability in the model.

Comparisons were also made to evaluate how different genetic relationship matrices (A) impact the optimality of experimental designs. Specifically, the simulations compare using an A that mixes Mesoamerican and Durango bean lines with a new A that contains only Mesoamerican bean lines. The results suggest that, in general, the mixed A leads to more optimal designs, but the difference in optimality is often negligible. However, the new A may produce less optimal designs in extreme cases with highly correlated Mesoamerican bean lines. The simulations also demonstrate that as more weight is placed on estimating random effects, the optimality of the two As becomes more closely related. Additionally, the simulations show that designs with greater replication variances among the bean lines generally lead to more optimal designs, even when using the new Mesoamerican-only A. Overall, the findings

suggest that the choice of A and the characteristics of the bean lines used can impact the optimality of experimental designs, and greater attention should be paid to the design process to increase the likelihood of producing optimal designs. In the future, we could attempt a similar strategy with A matrices as with Σ_u where a variety of different covariance structures are used for A and the resulting changes in the optimality criteria are observed.

Lastly, we compared the optimality of an experimental design developed in Chapter 2 with the best design resulting from simulated Z matrices. The expansion to use the entire A matrix and all 297 bean lines allowed us to compare the optimality of our design to what was selected from the simulations. As the value of α increased, there was a rising trend for all subjects, with the randomly generated optimal design exhibiting optimality values that progressively approached those of our design. Across α, designs became more similar to the best randomly generated Z when the number of replications on average per bean line was the largest. The trends in how the various combinations of Σ and Σ_u respond to design changes remained consistent. The randomly generated designs identified as optimal have greater variability in the number of replicates per bean line than those utilized in the data analysis. In conclusion, the experiment showed that the optimality of the experimental design developed in Chapter 2 was not too much different relative to the best design resulting from the simulated Z matrices. The results also suggest that the number of replications on average per bean line and the combination of Σ and Σ_u impact the design optimality. The study provides insight into the design of experiments and highlights the importance of considering various factors when designing experiments.

4.6.2 Contributions

As no significant research has been conducted in the realm of optimal experimental design within microbiology, this work provides a comprehensive overview of the potential for enhancing statistical modeling by improving how researchers plan and design their microbiome plates. Overall, this work has provided insight into the factors that contribute to the effectiveness of certain design choices and can inform future adaptations of similar designs. However, the implications of these findings for determining an optimal design are not straightforward. Considering these variations can provide a more comprehensive understanding of the factors contributing to specific design choices' effectiveness.

To ensure reliable and accurate research outcomes, it is important to establish an optimal experimental design structure with the flexibility to optimize a criterion of interest for researchers while observing changes in the genetic covariance parameters across different taxa and bean lines. This work proposes such a design, which could significantly improve measurements and enhance the quality of research. The study reviews optimal design changes based on Σ_u and Σ matrices directly linked to calculating polymicrobial traits from Chapter 3 analyses. In the future, while using these models for analysis, researchers could find the design that best maximizes the variability available in the $\tilde{G}^* = A \otimes \Sigma_u$. Then the entire $\tilde{G}^*$ matrix containing the genetic covariances could be used in a PCA to calculate polymicrobial traits. By optimizing the experimental design based on these matrices, researchers can obtain better variability estimates and identify polymicrobial traits that provide more information about the community relationships in the gut. This, in turn, could help identify significant GWAS peaks and inform us about genetic traits associated with the taxa community relationships in bean lines. Working collaboratively with researchers, we

could focus on finding the most optimal Z designs, where optimality is maximized for typical cases of variability, and better explain the statistical relationships between a person's diet and the community of bacteria in the gut.

4.7 Future Work

4.7.1 Parameter Adaptations in the Optimal Design Algorithm

While the compound symmetric, unstructured, and pilot variability matrices were valuable starting points, researchers may encounter many other combinations of Σ, Σ_u, and A matrices in their experiments. To provide a more comprehensive representation of the information that researchers might observe, it would be valuable to generate additional combinations of these matrices based on prior knowledge, as was initially demonstrated in the pilot cases. Our main question of interest is, as the combinations of Σ, Σ_u, and A change, how do the adaptations affect overall optimal design criteria? This approach could also involve creating variability estimates that are more applicable to researchers in the field, which would foster greater collaboration between statisticians, microbiologists, and plant breeders. Microbiologists can contribute their knowledge about bacterial taxa in the gut, while plant breeders can provide information about typical structures of the A matrix. Together, these collaborations can lead to more effective experimental designs that better reflect the complexity of real-world data. By establishing a statistical foundation for exploring optimal design for multivariate linear mixed models with genetic associations in the random effects, this work lays the groundwork for future research in this area.

In addition to adjustments to the variability matrices used in the simulation, incorporating more structure into the simulation could enhance the calculation of optimality for predetermined structures of the X and Z design matrices for the fixed

and random effects. For instance, researchers could explore whether a complete or incomplete block design would be more optimal and whether there are important effects to be accounted for across plate rows and columns. This could be achieved by adapting the code to allow for the optimality to be calculated for different statistical models. The models could account for different fixed or random effects of plates and different groups of plates as batches. The effect of these changes on the optimality of different designs could then be measured. Boundaries could also be set within the R code to generate specific formats for the design matrices. For instance, in our simulations, the Z matrix required a bean line to be present in every well while allowing any number of replications, including zero, for each bean line. However, it can be modified such that at least one instance of each bean line is required or a specific number of replicates is mandated for a particular bean line.

In addition to the optimality criteria, supplementary calculations can be performed within the design process to facilitate design comparisons. For instance, from a statistical perspective, researchers may seek designs that minimize the standard errors of genetic covariance parameters across bean lines. These values can be computed using the information matrix for random effects, $M(\sigma^2)$, obtained from our simulations. During design iterations, we can calculate $\sqrt{diag(M(\sigma^2)^{-1})}$ for each Z matrix to determine the standard errors of the variance components. When discussing with researchers, this enables us to compare standard errors among designs with similar optimality Φ values. We incorporated this value into some of our simulations. However, it was unclear which of the six random variance components to compare and in what context. For example, when considering scenarios using pilot data matrices and all 297 bean lines, the average difference between the random variance component estimates $(\sigma_u 11, \sigma_u 22, \sigma_u 33)$ for the most optimal design of Z (with the largest Φ value) was smaller than those from the least optimal design (with the smallest Φ

value) by -0.000141513 units. As this value is negative, the variance estimates are, on average, larger for designs with the smallest Φ value. We observed minimal changes in standard errors across many cases. This is generally consistent with our expectations, as our responses are centered log-ratio (CLR) transformations of abundance values. These values are small and therefore associated with small errors.

Expanding the simulation in these ways would provide researchers with a more comprehensive understanding of the factors that contribute to the effectiveness of certain design choices and could inform future adaptations of similar designs. Similar explorations have been common in different domains, including plant and animal breeding for univariate cases. However, the expansions across multivariate relationships are much newer, especially in microbiology. Even though new, this work is pertinent to assisting microbiologists in researching relationships across the effect of diet on the community effects of bacteria in the gut.

4.7.2 Changes to the Optimal Design Algorithm

The R function was developed to perform all required matrix algebra for a multivariate mixed model with three response variables, including a fixed plate effect and incorporating genetic relationships between random effects. Collaborating with researchers who have diverse R coding (or other platforms) expertise and styles could facilitate adapting the code to accommodate additional taxa. Due to technological limitations, we were unable to evaluate more than three taxa across all desired cases in a reasonable time frame, but working with others can aid in scaling the function to handle more complex experimental designs. The size of the matrices necessary to multiply and invert in the code expands greatly as more multivariate taxa are added, so research could be done to evaluate better methods for including more responses. For example, in Chapter 3, for subject 1, 17 taxa were included to calculate the

polymicrobial traits. With the function and code from this work, that would exceed the seven days allotted computation time allowed on Crane and would take over two weeks to run with 5,000 replications. An adaptation to assist with running larger taxa cases could be running more code chucks

The current version of the function only allows researchers to iterate over different values of α and randomly generated Z matrices. However, in practice, researchers may only be interested in one specific value of α while wanting to adjust the inclusion of different styles of X and Z design matrices. To address this, the code could be modified after the adjustments are made to reduce computation time when dealing with more than three taxa. After optimizing the code, the focus could shift to implementing other search algorithms, as discussed in the literature review. Although random search is a common starting point for many algorithms, an iterative method could be introduced to make incremental adjustments to the X or Z matrices while storing changes to the optimality Φ values throughout the simulation. Upon completion of the simulation, the researcher would be provided with the final optimality values, as well as the corresponding design matrices for the optimal design.

Chapter 5

Summary, Conclusions, Future Work

5.1 Summary and Conclusions

The gut microbiome plays a crucial role in human health, and by working collaboratively with microbiologists, we hope to further our understanding of the gut microbiome and its impact on human health. To promote a diverse microbiome, it is essential to identify ways to expand and diversify the community of gut bacteria. Many authors in the microbiological literature reviewed in this work underscore the importance of collaborative efforts that prioritize best practices for statistical design and analysis in microbiome studies. To ensure the validity and accuracy of the results, involving a statistician in the entire process, starting from experimental design and layout to statistical analysis and interpretation, is crucial [129]. We collaborated with Dr. Mallory Van Haute and Dr. Andrew Benson in their work investigating the link between the genetic features of bean lines and the human gut microbiome. By doing so, we acquired a deeper understanding of microbiologists' unique requirements when devising experiments and scrutinizing data to emphasize the multivariate community influences spanning various taxa. Chapter 2 of this dissertation presents an experimental design tailored for Dr. Van Haute. In Chapter 3, we examined our analytical approaches and collaboratively pinpointed associations between taxonomic

abundance data and particular regions within the *P. vulgaris* bean genome. Finally, in Chapter 4, we developed a simulation method for calculating a composite optimality criterion for multivariate linear mixed models with a covariance structure on the genetic random effects. We used this method to observe trends in how changes in the bean line and bacterial taxa information affect the optimal experimental design.

Through each chapter in this work, we aimed to address each of the following three objectives:

1. To develop a new experimental design for next-generation sequencing 96 well-plate lab experiments, specifically dealing with the human microbiome.

2. To use multivariate analysis methods to analyze data outputted from the experiment designed from objective 1.

3. To identify and explore the effects of experimental parameters on statistical optimality criteria for analyzing data from microbiome well-plate experiments.

We collaborated with Dr. Van Haute to devise a new experimental design approach for arranging bean lines across well plates, taking into account the resources typically employed in microbiome research. We did not utilize novel statistical techniques but rather particular experimental design strategies not commonly utilized within microbiology research using well plates. We created an incomplete block design overlaid with a partially replicated (p-rep) α lattice design to randomly assign bean line treatments to wells across sets of 96-well micro-plates. The statistical program, GENDEX, was used to create the randomizations utilizing univariate optimal design methodology to assign treatments to wells within and across plates. At the time, this process aimed to assist researchers in creating the best and simplest design for assigning treatments while obtaining the most informative statistical information

from all samples. Moreover, this chapter emphasizes the significance of integrating the incomplete block design structure into experiments as a means of addressing the substantial variability observed in taxonomic abundances within individuals' gut microbiomes and among distinct bacterial taxa within the same subject's gut. We found that incorporating the IBD structure was up to 35% more efficient than using a design with just the replicates. However, the efficiency gain varied depending on the subjects and taxa. We summarize that because the relative efficiency is quite large in some cases, there is a benefit to using more intricate design structures, as researchers cannot know which taxa will be found in a subject's gut or how the variability changes between subjects.

While working with Dr. Van Haute on the design process, we gained insight into the specific needs of microbiologists when designing and analyzing experiments focusing on the relationships between treatments and bacterial taxa. We utilized multivariate statistical methods to analyze the data generated from the design in Chapter 1 and create new polymicrobial traits. These traits act as multivariate phenotypes, accounting for abundance information across bacterial taxa in the simulated microbiome samples. This helped Dr. Van Haute outline where community effects were associated with areas in the *P. Vulgaris* common bean genome in her GWAS models. We outputted different variance decompositions from out MANOVA models to use as covariance matrices in principal component and canonical discriminant analyses. We utilized these methods to explore the relationship between genetic variation in a range of common bean lines and the composition of gut microbial communities in three subjects. Dimension reduction analyses were used to create weighted combinations of the community of taxa within each microbiome. Biplots were produced, demonstrating a variety of community relationships within each subject, and specific taxa driving the values of the PM traits were highlighted.

Forty-one of our total 72 polymicrobial traits had large enough heritability to be included within the GWAS models. All included polymicrobial traits had significant associations across each of the selected seven Major Effect Loci (MEL) areas on the chromosome. Individually, each of the 41 traits had significant associations with between four and nine chromosomes. Overall, out of 72 calculated polymicrobial traits, 57% met the heritability cutoff and were included in the genome-wide association study (GWAS) modeling. Of the 41 traits included, 18, 11, and 12 were significant within each subject, respectively. Significant associations were observed between each MEL and one or more polymicrobial traits, particularly those representing the first PC. This trend makes sense as with PCA we expect the first PC to account for the most variation. We found numerous overlapping associations within a given microbiome, where the same SNP was significantly associated with a polymicrobial trait as well as the individual taxa that were the main drivers of that polymicrobial trait.

For instance, overlapping associations of individual and polymicrobial traits from S776 to MEL-C demonstrated hits of HypCorrPC2 (Faecalibacterium, Coprococcus3, LachnospiraceaeUCG004) as well as the individual taxonomic traits of Faecalibacterium and Lachnospiraceae. These findings, together with other types of microbiome traits that were significantly associated with the same SNPs or MEL from multiple microbiomes, provide further evidence of causality in regions of the common bean genome that appear to impact the microbiome. Symbols were added to the trait names to indicate different category labels, including whether a trait had overlapping taxa with univariate traits. Of the significant polymicrobial traits used in the GWAS models, 54% had either a direct or related overlap with a univariate taxon on the same multivariate effect locus (MEL). Only 12.2% of significant polymicrobial traits did not have any largely weighted taxa overlapping with univariate taxa. Finally, about 34% of the significant polymicrobial traits were found to indicate areas in

which the genome exhibited a statistically significant association with a more diverse community of taxa, which were identified as highly abundant in the microbiome of each subject.

Specifically, we also highlighted a result from Dr. Van Haute that had particular applicability within the relationships between gut taxa and human health. An interesting relationship was identified between traits and a specific SNP location within MEL C. The high pleiotropy of this locus provided strong evidence that genuine variation at this locus has a major effect on the microbiome [59]. The study identified genes related to glycyrrhizinate biosynthesis that were related to significant PM traits, mainly driven by Lachnospiraceae, Faecalibacterium, and Clostridium, which can produce butyrate, a microbial metabolite with known health benefits [59]. In Chapter 3, we highlighted a biplot (Figure 3.5) comparing these traits in the GenCov PC1 loadings versus the HypCorr PC2 loadings. In their loadings, we can see high values for this taxa, with additional weights across the remaining taxa. Thus we have some specific and overarching evidence that these multivariate methods can help identify important candidate genomic regions and support the conclusion that these MELs define areas with the most significant effects on the human gut microbiome. The findings underscore the importance of examining the interactions among different microbes, as these relationships can significantly influence community structure.

Thus, the study highlights the need to shift from an exclusive focus on individual microbes to a more holistic approach considering polymicrobial traits. The polymicrobial traits provided a beneficial method for researchers, reducing and accounting for spurious variabilities and assisting with identifying the effects of dietary changes on the gut microbiota. By decomposing various correlation and covariance matrices from the bean line and genetic effects and weighting the abundances across the taxa that explain the most variation in the selected structures, we were able to create many

options for the pseudo-multivariate community-type traits. Across all options, over 50% of the selected traits showed significant associations across multiple areas of the *P. Vulgaris* bean genome. This research has the potential to advance strategies for investigating microbiome active traits as complex traits of crop plants, thus providing a foundation for food scientists to integrate health-related traits into crop enhancement programs. Furthermore, this work may serve as a catalyst for statisticians to devise innovative methods for accommodating community effects without the need to modify or adapt conventional GWAS techniques.

Our final objective was to investigate how experimental parameters affect the statistical optimality criteria used to analyze data from microbiome well-plate experiments. We developed a function that calculates the composite optimality of fixed and random effects for a multivariate mixed model. This function emphasizes relationships between values of Φ, Σ_u, Σ, the genetic relationship matrix A, and the random effect design matrix Z. The objective of Chapter 4 is to develop a simulation with the capabilities to compare design optimality for varying input values pertinent in the field of microbiology and genetics. Overall, we would like to balance statistical accuracy and ease of use for domain specific researchers. To achieve this, we employed the multivariate mixed model framework originally proposed by Kmail, and extended the model by incorporating the genetic relationship matrix A into the random effect variability structure.

The work represents an important advancement in integrating experimental design and abundance analysis for studying microbiome activity traits. In these methods, there is potential for determining the most suitable design for simulating microbiome samples across wellplates to generate highly informative estimates of the taxa's variability linked to the variability between bean lines based on their genetic population structure. We worked to adapt our model to meet the needs of researchers

like Dr. Van Haute, who aim to investigate the variability of multiple responses (i.e., taxa) resulting from genetic changes from foods (like beans) in the diet. Specifically, we have optimized the model's design structure to accurately reflect the covariance between the population structure of the bean lines in A. This modification allows for a more precise analysis of the genetic effects and their impact on the different taxa, thereby enhancing the model's overall effectiveness.

The model is built around a stacked form of a multivariate model, using REML estimation procedures to find the information matrices for the fixed and random effects. Genetic information is frequently considered a random effect; therefore, in this model, we refine this aspect in comparison to the analysis of abundance carried out in Chapter 3, which facilitated the calculation of polymicrobial characteristics. In utilizing the Φ optimality criteria split between the fixed and random effects, we offer flexibility to researchers who would like to make sure they are finding a design to emphasize their interest in the variability of the bean line treatments. Researchers using our methodology to find their design matrices would then have more optimal designs to perhaps use this multivariate mixed model to create new versions of the polymicrobial traits, only specifically using the $\tilde{\mathbf{G}}^* = A \otimes \Sigma_u$ matrix instead of having to decompose multiple matrices to calculate a genetic effect matrix as we did in Chapter 3 with the GenCov and GenCorr formats.

After creating the multivariate mixed model framework within our simulations, we compared Φ optimality values and variability in the number of replications contained in Z for different variabilities and A matrix structures. For these selected combinations of matrices, we also framed the results in terms of different variance ratios that may be of interest to subject matter researchers. For example, heritability was defined as a value that compared variability in the Σ_u matrix to the overall variability for the added effects of Σ and Σ_u. The Σ_u matrix represents the contri-

bution of random effects in explaining the variation in taxa, specifically how changes in bean lines' random effects account for the observed variability. Similar to what was done by Bueno Filho and Gilmour, we know that researchers may be interested in optimal designs for traits that have large heritability. Additionally, we created a ratio of interest for the largest eigenvalue in the Σ_u matrix divided by the sum of the largest eigenvalues from Σ and Σ_u. Eigenvalues would be of interest as well for the creation of polymicrobial traits, as we used PCA to observe the variability around similar matrices in Chapter 3. Lastly, The Z matrix structures were also compared for the most optimal design out of 5000 randomly selected Z matrices versus the design structures implemented for subjects 1, 2, and three created in Chapter Two. For researchers, designs with larger values of these metrics may be more relevant, and thus the selection of the optimal design may have fewer implications.

In the hypothetical cases of variability, we found more distinct trends in the gaps between optimality criteria for the more extreme combinations across the CS and UN structures. Smaller residual variability leads to more optimal designs but with smaller differences as more weight is put on the random bean line effects in the designs. When using specific pilot structures, we were able to see trends across negative correlations between the taxa, with the more highly negatively correlated taxa often having slightly less optimal designs, especially when combined with larger residual variabilities as well. The pilot cases also had larger disparities between correlations between the three taxa, with the cases with large correlations of 0.83, -0.95, and -0.9 having less optimal designs than cases with correlations of 0.33, -0.9, and 0.58.

Researchers may not care necessarily about the optimality values specifically but rather what are commonalities in the designs that are produced that are more optimal. For all cases, we then compared the variability in the number of replicates

across bean lines for what designs were created as most optimal. The pilot cases provided the most information about how the optimality changes and what are the most optimal designs. For the coefficients of variation across the pilot cases, we found that for the optimality more weighted on the fixed effects, there are larger disparities in the variability in the number of replications produced within the design with the larger values of Φ, versus when more weights are on the fixed effects. Also, as α increases, the gaps are increased between what are identified as the best and worst designs. Furthermore, the choice of covariance matrix ratio and genetic relationship matrices (A) can impact the optimality of experimental designs. Our analysis shows that when the values of our most informative ratios - namely, heritability, det2, and eigenvalue ratio - increase, the best design found from a random sample of 5000 designs leads to more optimal designs in our pilot cases. However, for other variability cases, we observe decreasing trends. For example, in cases with large residual variation, as Σ_u changed from small to large variability, the optimality decreased. This trend makes sense because heritability will get larger as the values in the Σ_u matrix get larger, and optimality values decrease as there is more variability in the Σ and Σ_u matrices. Differences were again smaller for the smaller α weights, with more gaps between cases when more weight was put on the random effects.

In the pilot cases where the A matrix contained 50 bean lines, we observed that for larger α weights, there was a steeper increase in Φ values and variability in the number of replications as the ratios of heritability, det2, and the eigenvalues increased. When we plotted the same comparisons for cases with all 297 bean lines, we observed the same relationship for Φ, i.e., as the ratios increased, so did the optimality. In contrast to the 50-line case, we did not observe the same trend in replication variability for the nine simulations run with all 297 bean lines. Instead, as the ratio values increased, the coefficient of variation for the number of replications

for each bean line decreased. For clarity, we have added plots for the 297 line cases in 5.1, 5.2, and 5.3 for heritability, det2, and the eigenvalue ratio, to demonstrate the decreasing trends. To compare, we can return to Figure 4.18, and see the increasing trend in CV across heritability for the 50-line case.

In selecting an optimal design for cases with 50 lines, it appears that as Φ increases, there is a wider gap in the number of replications across bean lines. We hypothesize that this could be because, with fewer lines relative to the number of wells, there are more opportunities across the wells to replicate the different lines. Thus, perhaps the information within the A matrix for the genetic relationships between the bean lines may lead the optimal design to replicate certain bean lines more than others. However, for cases with all the lines, there can be fewer replications on average for cases with many more bean lines. The more optimal designs may have less variability between the number of replications so that more lines can be replicated. In the future, we could investigate if there are commonalities between which lines are selected more or less often to be replicated.

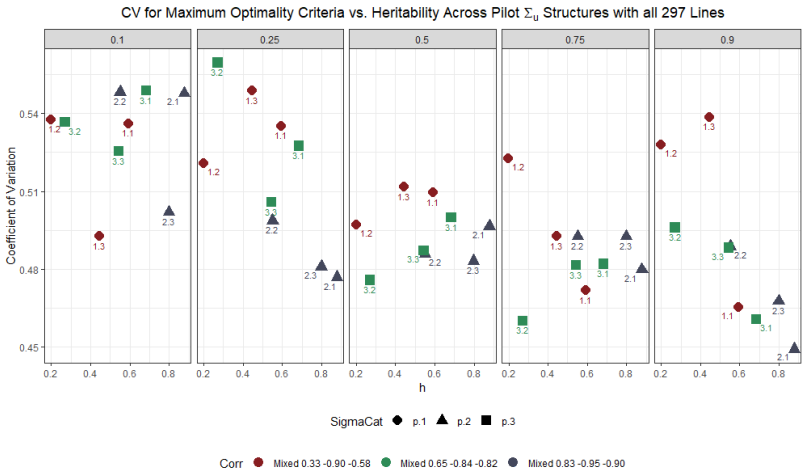


Figure 5.1: Maximum CV vs. Heritability across Pilot Σ_u Structures using all 297 bean lines. Panels represent the different values of α, points are colored by the correlation structure, and the shape of the points correspond to the variety of Σ_u

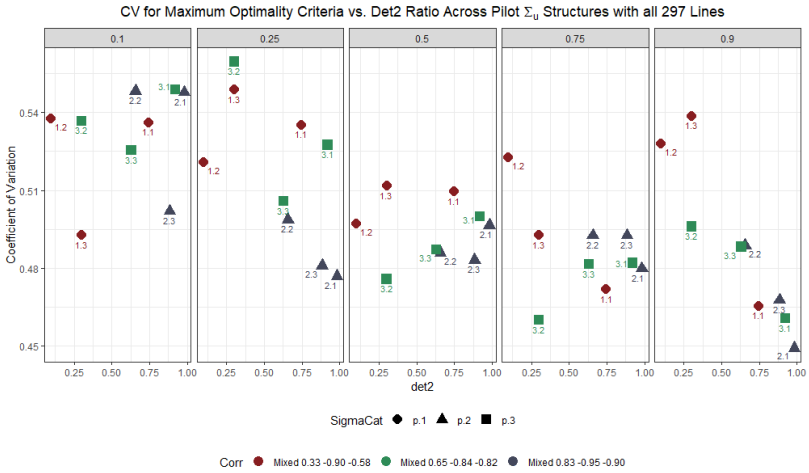


Figure 5.2: Maximum CV vs. Det2 ratio across Pilot Σ_u Structures using all 297 bean lines. Panels represent the different values of α, points are colored by the correlation structure, and the shape of the points corresponds to the variety of Σ_u

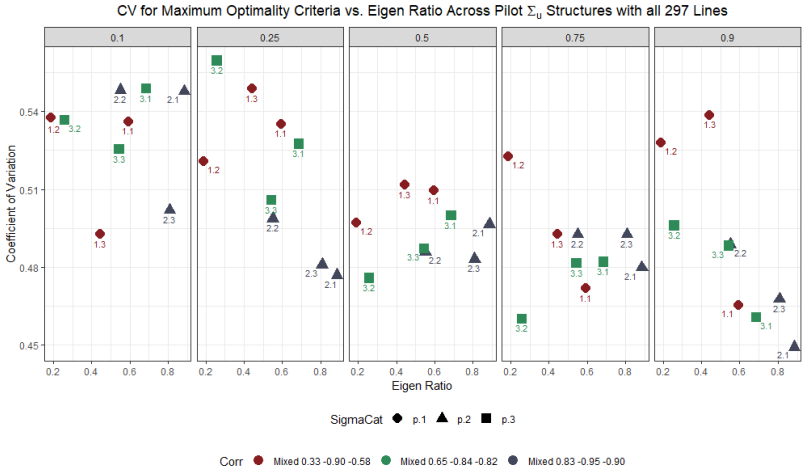


Figure 5.3: Maximum CV vs. Eigenvalue ratio across Pilot Σ_u Structures using all 297 bean lines. Panels represent the different values of α, points are colored by the correlation structure, and the shape of the points corresponds to the variety of Σ_u

When comparing simulations with the two different types of genetic relationship matrices, we found that the results suggest that, in general, the mixed A leads to more optimal designs, but the difference in optimality is often negligible. However, the new A may produce less optimal designs in extreme cases with highly correlated Mesoamerican bean lines. Thus, researchers may want to discuss with a statistician

the bean lines or treatments of interest that are being added to the design, as their effects on the optimality could change the final type of design. We found that the optimal designs for the more correlated bean lines lead to a larger amount of variability for the number of replications across the 50 bean lines as compared to the set of the mixed Mesoamerican and Durango bean lines.

Lastly, the optimality of the experimental design developed in Chapter 2 was compared to the best designs resulting from 5000 simulated Z matrices. Often the difference in optimality was very small, but the magnitude and which design was more optimal changed depending on the design of which subject was utilized. The random search produced slightly more optimal designs, with more variability in the replications of the bean lines included within the Z matrix. As highlighted in the Chapter 4 results, our design filled every well, and bean lines typically received anywhere from 3 to 7 replications at the end of the study after data was collected. The Z matrices from the simulation allow bean lines to have no replication and put no max on the number of replicates possible. Thus, even though the random search is finding more optimal designs, it might not be the most practical of designs. However, this type of structure may be useful for observing in the future if there is ever a commonality with which bean lines receive the smallest number of replications.

The designs we've presented have the potential to help us better understand the structure of taxa communities, allowing for the designs used within this field to find the best design for a multivariate mixed model considering the genetic population structure of treatments and the covariance structure among the taxa. While we need to conduct further research to fully comprehend how to enhance our analyses, the choices we made were intentional and aligned with the goals of this current exploratory work. Our work contributes to the field of microbiology by shedding light on the factors that influence the effectiveness of experimental design choices,

especially when studying bacterial community changes in the human gut. We still have much to do to identify common trends and collaborate with microbiology researchers to adapt this methodology to their needs. Through our optimal design function, we investigated the application of optimal design methodology in genetic microbiome experiments, thereby addressing a knowledge gap regarding the optimal design of experiments for multivariate mixed models that incorporate genetic population structure. The modeling techniques used throughout Chapter 4 can be tied back to finding the best design to estimate variability matrices to calculate polymicrobial traits. Our findings demonstrate the impact of input variability and the framework of the Z matrices on the composite optimality criteria Φ. By observing changes in these factors, researchers can optimize their experimental designs to obtain better variability estimates, find better polymicrobial trait estimates, and gain deeper insights into the relationships within the gut community. In summary, our research provides a foundation for researchers to optimize their experimental designs based on the discussed factors, ultimately improving the quality of their research.

5.2 Future Work

Much of what should be addressed in future work pertaining to the specific topics in Chapters 2, 3, and 4 are addressed in sections 2.6, 3.8, and 4.7. The following two subsections delve into potential future endeavors that may yield significant contributions to the fields of quantitative genetics and microbiology. This exploration aims to provide researchers with valuable insights and guidance for advancing their respective areas of study.

5.2.1 Adapting the Optimal Design Methodology for use by Domain Researchers

To maximize the usefulness of the method to stakeholders with similar types of data, it is important to maintain the current direction of the work and adapt it accordingly. To ensure that our work is aligned with the goals of researchers, we need to take into account the questions that microbiologists are asking and how our findings can best support their research objectives. For example, researchers may be interested in creating a design focusing on evaluating how the genetic variation across bean lines affect the community structure of the taxa in the gut. This outcome could be directly related to the calculations of the polymicrobial traits from Chapter 3. Researchers may also ask questions similar to those in a power analysis, such as how many replications are best across and within plates. Given our current objectives, we could investigate further what structure in the A matrix has effect on the optimality criteria, and on the structure of the most optimal designs. In the future to answer both of these types of questions, we can further adapt what we calculate in our simulation to evaluate different types of optimality criteria.

For the calculations of polymicrobial traits, we could adapt our optimal design function to pick a design based on optimization of the variability within the $\tilde{\mathbf{G}}^* = A \otimes \Sigma_u$ matrix. If values like heritability are of particular interest, so we could focus our simulation on maximizing heritability $= \frac{\hat{\Sigma}_u}{\hat{\Sigma}+\hat{\Sigma}_u}$ where the variability matrices will be estimated from the model within the function. For each, we would add in full calculations of these values into the optimal design function, and for each different structure of the X and Z design matrices, the new values of the variances and heritability could be calculated. Then, we could make note of which note has the most informative variability estimates, or largest heritability values. As we saw

in Chapter 3, only the trait values with the largest heritabilities were included in the GWAS models. Thus, to provide the best PM traits for GWAS, we would want to make sure there is large enough heritability values. As a whole, to really drive the contribution of this work, the future work on the simulation will focus in on the direct needs of researchers who would like to further how to integrate better designs into their experiments.

5.2.2 Design and Analysis Protocols, Optimal Design Web Application

Overarchingly, in addition to the methodology presented in this dissertation, a significant contribution in the future would be the development of a set of standardized protocols for designing and analyzing microbiome data. This would be an overall incorporation of all the future work information. These protocols could serve as a guide for researchers, providing step-by-step instructions and recommendations for setting up an experiment, from the randomization of treatments to the analysis and interpretation of results. Since the concepts presented in this dissertation are relatively new to the field, researchers would benefit from a comprehensive resource that includes detailed descriptions of the experimental design and analysis methods proposed in Chapters 2 and 3.

The protocols could also address potential challenges that may arise during the experiment and provide solutions for dealing with them. Ultimately, these protocols would help ensure the reproducibility and accuracy of experimental results and facilitate the advancement of knowledge in this area of research. Again, the work could be a culmination of work from many different types of scientists who contribute to similar projects in genetics, microbiology, plant breeding, etc. Furthermore, by building upon the investigative findings of Chapter 4, we can conduct more precise comparisons between designs, enabling us to draw broader conclusions about their

optimality. We could expand on some more of the univariate optimal designs of interest and see how they perform across different combinations of the Σ_u and Σ matrices in ratios such as the heritability so the results would be applicable to those across the fields in genetic sciences.

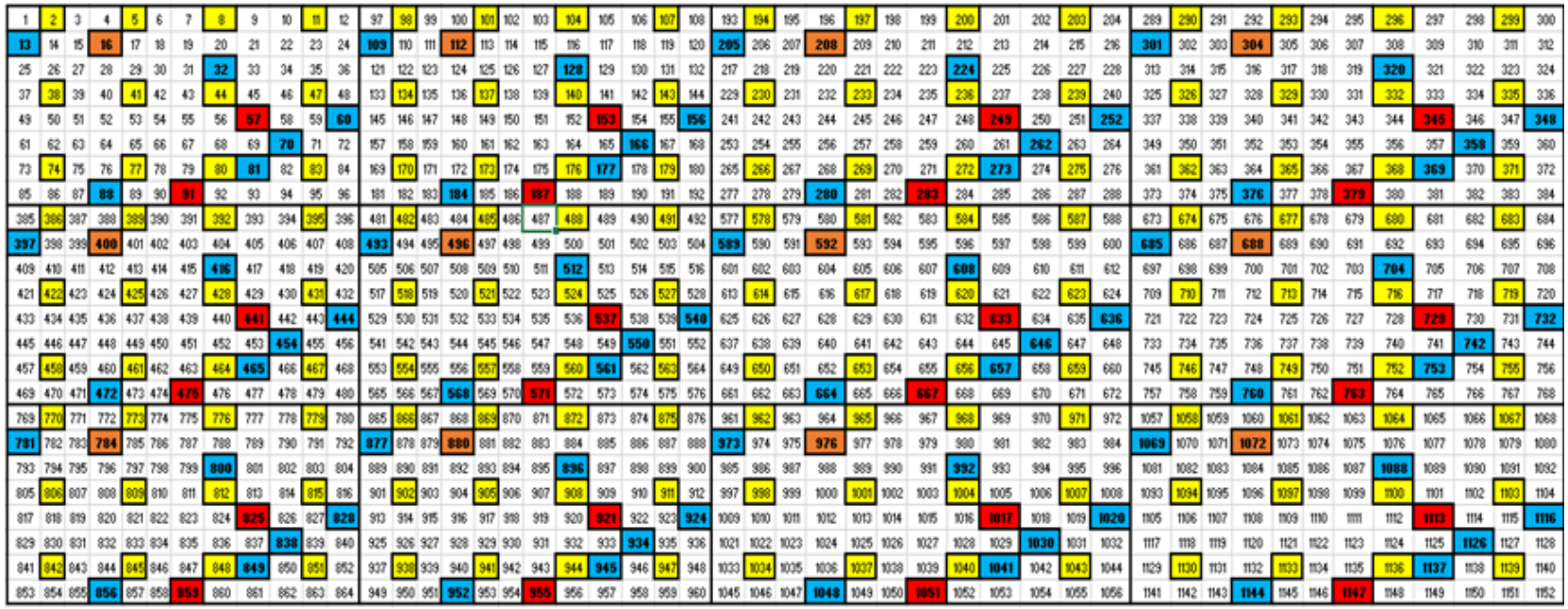
Improving the linkage between the design of the experiments and the methods of analysis used can yield dual benefits. First, advanced experimental design techniques may be easier to implement if researchers better understand how the design is closely related to the analysis. This is the tie-in between the expansion of the information in Chapters 3 and 4. We wanted to focus on a more flexible statistical analysis model in Chapter 4, closer to what researchers might want to include in Chapter 3 for calculations of the polymicrobial traits. The REML multivariate expansion needs to be adjusted to account for more taxa and more flexible capabilities to change the variables included in the model, however, we have laid the groundwork to demonstrate the usefulness of being able to compare design through an optimal experimental design framework. Often the design can be used to specifically define what analysis should be used, and when a researcher cannot utilize a standard set of statistical analyses, it "becomes tempting to employ progressively complex data-transformation, data subsetting, or analytical approaches, in the hope that one or more will identify an effect" [6]. Utilizing methods that are overly complex could potentially be helpful in the short term, but in the long run, could cause more harm than good in the attempt to find results that accurately represent what occurred in the experiment.

Eventually, efforts could set up a protocol and procedural framework similar to the PlateDesigner web application that Suprun and Suarez-Farinas proposed in 2018 [145] and other web-based well-plate experimental design applications described in the review of the literature. To the best of their knowledge, the developers specified that PlateDesigner is the first and only web-based application available to researchers that

can generate randomization schemes for microplate experiments. So, an expansion on their proposals and establish something similar but for the design set up in Chapter 2 of this dissertation. Given that the function for the optimal design is currently implemented in R and can be executed on high-powered computing clusters, it is possible to create an application such as an R shiny webpage. This would allow researchers with an interest in comparing different experimental designs to easily compare optimal design criteria given their particular pilot data of interest. The function can be customized with a specific statistical model of interest. Researchers can then input their preferred levels of variability and design matrix constraints, allowing them to obtain an optimality value for a specific design or conduct simulations across multiple matrices to identify the most optimal design.

Appendix A

Figure A.1: Full Randomization Summary with Labelled Wells



Appendix B

Figure B.1: Subject 1 (S770) Polymicrobial Trait Loadings

Subject 1 (S770)																		
PM Trait	Component	Prevotella 9	Succinivibrio	Dialister	Enterobacteriaceae	Bacteroides	Coprococcus3	Anaerostipes	Sutterella	Prevotellaceae	Coprococcus1	Bifidobacterium	Lachnospiraceae_NA	Dorea	Blautia	Roseburia	Fournierella	Faecalibacterium
Genetic Correlation	1	-0.27	-0.27	-0.31	-0.26	-0.29	0.3	0.27	-0.29	-0.08	-0.29	-0.08	-0.1	0.24	0.25	-0.04	0.27	0.25
	2	-0.01	0.28	-0.05	-0.27	-0.07	-0.14	0.07	0.11	-0.26	-0.18	**0.57**	-0.08	-0.19	0.07	**-0.5**	-0.2	0.21
	3	0.05	0.14	0.11	-0.23	-0.12	-0.07	0.21	0.08	**0.6**	0.1	0.12	**-0.64**	0.15	-0.1	0.03	0.06	-0.11
	4	-0.15	-0.08	0	0.06	0.31	0.03	0.05	0.1	0.35	-0.25	-0.26	0.15	-0.14	**-0.42**	**-0.52**	0.24	0.24
Genetic Covariance	1	-0.14	-0.32	-0.15	-0.36	-0.18	0.23	0.17	-0.13	-0.04	-0.2	-0.04	-0.03	0.12	0.31	-0.04	0.25	0.61
	2	-0.09	**-0.47**	-0.03	0.27	0.01	0.25	0.03	-0.08	0.04	0.04	-0.39	0	0.22	0.2	0.22	0.31	-0.5
	3	-0.04	0.25	-0.02	-0.35	-0.12	0.03	0.11	0	-0.08	0	0.34	-0.14	0.15	**0.62**	-0.07	-0.14	-0.46
	4	0.02	0.09	0.05	**-0.46**	-0.19	0.01	0.2	-0.01	**0.42**	0.1	0.06	-0.21	0.2	**-0.5**	0.03	0.36	-0.2
Hypothesis Correlation	1	0.23	0.3	0.33	0.28	0.31	-0.29	-0.27	0.31	0.11	0.22	0.07	0.15	-0.27	-0.26	-0.04	-0.22	-0.22
	2	0.29	-0.24	0.05	0.06	0.06	0.12	-0.09	-0.16	0.25	0.34	**-0.45**	-0.2	0.1	-0.15	**0.46**	0.29	-0.22
	3	0.25	0.27	0.13	-0.37	-0.23	-0.1	0.29	0.16	0.15	0.08	0.38	**-0.55**	0.09	0.02	0.11	-0.05	-0.18
	4	0.08	-0.03	0.03	-0.28	-0.04	-0.01	0.1	0.1	**0.47**	-0.24	-0.17	0.01	-0.4	-0.4	-0.13	0.17	0.48
Hypothesis Covariance	1	0.13	0.32	0.15	0.33	0.18	-0.19	-0.14	0.15	0.05	0.16	0.09	0.07	-0.12	-0.28	0.01	-0.34	-0.62
	2	0.08	-0.28	-0.02	0	-0.01	0.14	0.01	-0.09	0.07	0.16	-0.36	-0.1	0.12	-0.03	**0.45**	0.51	-0.49
	3	0.1	0	0.08	0.15	0.12	-0.12	-0.16	0.04	0.09	0.09	-0.32	0.08	-0.23	**-0.67**	0.09	0.18	0.49
	4	0.09	0.39	0.05	**-0.58**	-0.15	-0.13	0.13	0.1	0.06	0.04	0.37	-0.21	-0.02	-0.3	-0.04	0.38	-0.11
Raw CDA	1	-0.05	**3.71**	-0.24	2.16	1.22	-1.36	-0.03	-0.47	0.48	1.7	-0.35	-1.77	1.66	0.15	0.54	0.61	0.22
	2	**-4.46**	**-3.9**	-0.48	0.26	-2.2	-1.37	-2.16	2	-2.39	**-2.97**	**-2.72**	**-4.52**	-0.86	-2.02	-1.7	-1.52	-2.18
	3	**5.15**	-1.68	-0.16	1.74	-0.44	0.65	0.42	**3.53**	**-3.46**	0.82	1.48	**2.6**	-0.84	1.21	0.77	0.77	0.88
	4	1.83	**6.59**	**6.78**	**5.42**	**7.62**	**4.25**	**5.65**	**6.29**	**5.32**	**4.22**	**5.54**	**8.07**	**6.83**	**6.17**	**6**	5.81	5.41
Standardized CDA	1	-0.01	1.89	-0.05	1.09	0.37	-0.58	-0.01	-0.13	0.13	0.53	-0.19	-0.58	0.54	0.09	0.35	0.5	0.21
	2	-1.19	-1.99	-0.11	0.13	-0.67	-0.59	-0.7	0.55	-0.65	-0.93	-1.53	-1.48	-0.28	-1.24	-1.09	-1.24	-2.05
	3	1.38	-0.85	-0.04	0.88	-0.13	0.28	0.14	0.97	-0.94	0.26	0.83	0.85	-0.27	0.74	0.49	0.63	0.83
	4	0.49	**3.35**	1.51	**2.74**	2.31	1.82	1.83	1.73	1.45	1.32	**3.11**	**2.65**	2.22	**3.77**	**3.83**	4.75	5.09

Figure B.2: Subject 2 (S776) Polymicrobial Trait Loadings

Subject 2 (S776)																
PM Trait	Component	Bacteroides	Phascolarcto_ bacterium	Faecaliba cterium	Clostridium sensu stricto 1	Sutterella	Bifidobact erium	Coprococcus3	Dorea	Alistipes	Roseburia	Lachnospiraceae_ UCG004	Blautia	Parabacter oides	Eubacterium halli	Coprococcus1
Genetic Correlation	1	-0.31	-0.32	0.25	-0.3	-0.29	0.23	-0.28	0.06	-0.31	-0.03	0.21	0.2	-0.32	0.3	0.24
	2	0.06	0.05	0.33	-0.1	0.18	-0.26	-0.23	**-0.58**	0.16	-0.18	**0.41**	-0.36	0.09	0.07	0.14
	3	-0.19	-0.18	-0.07	0.12	-0.25	**-0.41**	0.25	-0.31	-0.12	**0.63**	-0.18	-0.1	0.01	0.13	0.23
	4	-0.08	-0.06	-0.15	0.18	-0.17	**-0.47**	-0.06	-0.22	-0.12	-0.39	-0.02	**0.41**	-0.06	0.24	**-0.49**
Genetic Covariance	1	-0.16	-0.14	0.22	**-0.53**	-0.12	0.07	-0.11	0	-0.13	-0.01	0.14	0.17	-0.17	**0.69**	0.14
	2	-0.1	-0.08	**-0.41**	0.3	-0.16	0.02	0.09	0.29	-0.14	0.07	-0.33	**0.63**	-0.11	0.19	-0.16
	3	-0.07	-0.03	-0.03	**0.56**	-0.07	-0.35	0.03	-0.29	-0.01	0.16	-0.06	-0.36	0.03	**0.55**	0.06
	4	0.16	0.15	0.11	0.15	0.14	-0.15	-0.17	-0.13	0.1	**-0.65**	0.26	0.23	0.04	0.16	**-0.51**
Hypothesis Correlation	1	0.38	0.37	0.01	0.2	0.36	-0.16	-0.01	-0.22	0.37	-0.05	0	-0.27	0.37	-0.31	-0.18
	2	0.01	0.02	**-0.5**	0.37	-0.08	0.09	**0.45**	0.31	-0.06	0.13	**-0.45**	0.1	0.02	-0.17	-0.19
	3	0.06	0.02	-0.09	-0.03	0.08	0.33	-0.24	0.09	-0.02	**-0.72**	-0.04	0.21	-0.01	0.11	**-0.48**
	4	0.13	-0.07	0.05	-0.38	0.21	**0.73**	0.03	0.15	0.13	0.12	-0.13	-0.06	0.07	-0.31	0.3
Hypothesis Covariance	1	0.18	0.17	-0.06	0.37	0.16	-0.08	0.04	-0.07	0.16	0.01	-0.04	-0.26	0.19	**-0.79**	-0.09
	2	0.12	0.1	**0.44**	**-0.58**	0.17	-0.09	-0.18	-0.28	0.15	-0.05	0.31	-0.4	0.11	-0.04	0.11
	3	0.02	0.1	0.01	**0.52**	0	-0.37	-0.02	-0.18	0.02	-0.06	0.04	**-0.51**	0.05	**0.52**	-0.11
	4	0.06	0.08	0.12	0.04	0.05	**-0.64**	-0.14	-0.27	0.07	-0.05	0.13	**0.65**	0.07	-0.04	-0.14
Raw CDA	1	2.5	1.12	-0.88	1.34	0.62	0.98	2.15	0.92	2.39	1.25	1.36	1.25	1.85	1.12	1.16
	2	**-2.63**	1.53	2.26	2.4	**4.41**	2.13	**2.92**	1.78	**3.38**	**2.73**	2.33	2.01	**5.05**	2.19	**2.68**
	3	**5.42**	1.49	**3.81**	2.16	-0.09	2.18	2.49	2.31	**2.97**	**3.05**	0.81	2.02	1.64	2.17	1.52
	4	0.47	1.72	**3.26**	**2.63**	**3.96**	2.24	**2.74**	**3.08**	**5.06**	1.3	2.17	2.31	1.1	2.08	1.69
Standardized CDA	1	1.98	0.92	-0.74	1.9	0.51	1.09	1.36	0.95	1.83	0.82	1.07	1.98	1.56	**2.75**	0.83
	2	-2.08	1.25	1.9	**3.4**	**3.69**	2.37	1.86	1.84	**2.59**	1.79	1.84	**3.2**	**4.27**	**5.37**	1.93
	3	**4.29**	1.22	**3.2**	**3.07**	-0.08	2.42	1.58	2.37	2.28	2.01	0.64	**3.21**	1.39	**5.33**	1.09
	4	0.37	1.41	**2.74**	**3.73**	**3.31**	2.49	1.74	**3.17**	**3.89**	0.85	1.71	**3.67**	0.93	**5.1**	1.21

Figure B.3: Subject 3 (S768) Polymicrobial Trait Loadings

Subject 3 (S768)										
PM Trait	**Component**	**Prevotella9**	**Megasphaera**	**Acidaminococcus**	**Escherichia**	**Bacteroides**	**Clostridium sensu stricto 1**	**Phascolarcto_ bacterium**	**Sutterella**	**Faecalibacterium**
Genetic Correlation	1	0.05	0.27	**-0.45**	-0.24	-0.39	-0.34	-0.31	-0.35	**0.41**
	2	**-0.44**	**-0.55**	-0.07	**-0.47**	-0.19	0.22	0.33	-0.26	0.09
	3	**0.83**	-0.06	-0.15	-0.3	-0.12	0.26	0.32	-0.02	-0.07
	4	0.09	-0.21	0.21	**-0.59**	0.2	-0.31	-0.34	**0.53**	0.15
Genetic Covariance	1	0	0.05	-0.11	-0.11	-0.14	-0.23	-0.19	-0.12	**0.92**
	2	-0.16	-0.29	-0.07	**-0.41**	-0.15	**0.51**	**0.6**	-0.19	0.17
	3	0.02	0.08	-0.02	-0.05	0.1	**0.76**	**-0.63**	0.01	0.06
	4	-0.29	-0.32	0.28	**-0.55**	0.32	-0.19	-0.18	**0.51**	0.02
Hypothesis Correlation	1	0.29	0.17	0.35	0.35	**0.43**	0.23	0.21	0.39	**-0.47**
	2	**0.42**	**0.45**	-0.24	0.3	0.02	**-0.44**	**-0.51**	0.06	0.09
	3	0.06	**-0.61**	-0.39	-0.02	0.38	-0.32	0.05	**0.45**	0.11
	4	0.33	0.22	0.24	**-0.54**	-0.17	**-0.47**	**0.45**	0.19	0.03
Hypothesis Covariance	1	0.07	0.04	0.12	0.11	0.12	0.15	0.15	0.12	**-0.95**
	2	0.3	0.33	-0.1	0.33	0.11	**-0.49**	**-0.64**	0.14	-0.08
	3	0.11	-0.26	-0.07	-0.08	0.22	**-0.68**	**0.52**	0.37	0.02
	4	0.1	**0.6**	0.33	-0.17	**-0.42**	-0.32	0.32	-0.35	-0.04
Raw CDA	1	**-6.43**	-1.19	0.14	0.41	1.64	-1.22	-1.63	0.21	-1.06
	2	**2.92**	1.82	**3.27**	**8.66**	-0.67	**2.7**	**3.48**	1.94	**2.66**
	3	**2.73**	-1.08	-0.87	0.1	-0.86	2.23	0.96	0.12	0.18
	4	**2.58**	1.72	**3.16**	1.01	**3.78**	**3.01**	-0.07	2.42	1.97
Standardized CDA	1	-2.37	-0.5	0.07	0.17	0.63	-0.72	-1.09	0.09	-2.46
	2	1.07	0.77	1.59	**3.57**	-0.26	1.61	2.34	0.85	**6.17**
	3	1	-0.46	-0.42	0.04	-0.33	1.33	0.64	0.05	0.43
	4	0.95	0.73	1.54	0.42	1.45	1.8	-0.05	1.06	**4.57**

Table B.1: Taxa with Largest Weights from Polymicrobial trait loadings. (Cutoff values: PCA loadings < -0.4 and > 0.4; N/A was assigned to PCs of polymicrobial trait methods that had loadings all smaller than the cutoffs.

Polymicrobial Trait Method	**PC**	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**
Genetic Correlation	**PC1**	Acidaminococcus, Faecalibacterium	n/a	n/a
	PC2	Prevotella9, Megasphaera, Escherichia/Shigella	Bifidobacterium, Roseburia	Dorea, LachnospiraceaeUCG004
	PC3	Prevotella9	Prevotellaceae, Lachnospiraceae_NA	Bifidobacterium, Roseburia
	PC4	Escherichia/Shigella, Sutterella	Blautia, Roseburia	Bifidobacterium, Blautia, Coprococcus1
Genetic Covariance	**PC1**	Faecalibacterium	Faecalibacterium	Clostridium sensu stricto, Eubacterium hallii
	PC2	Phascolarctobacterium, Clostridiumsensustricto, Escherichia/Shigella	Succinivibrio, Faecalibacterium	Faecalibacterium, Blautia
	PC3	Phascolarctobacterium, Clostridiumsensustricto	Blautia, Faecalibacterium	Clostridium sensu stricto, Eubacterium hallii
	PC4	Escherichia/Shigella, Sutterella	Enterobacteriaceae, Prevotellaceae, Blautia	Roseburia, Coprococcus1
Hypothesis Correlation	**PC1**	Bacteroides, Faecalibacterium	n/a	n/a
	PC2	Prevotella9, Megasphaera, Phascolarctobacterium, Clostridiumsensustricto	Bifidobacterium, Roseburia	Faecalibacterium, Coprococcus3, LachnospiraceaeUCG004
	PC3	Megasphaera, Sutterella	Lachnospiraceae_NA	Roseburia, Coprococcus1
	PC4	Clostridiumsensustricto, Escherichia/Shigella, Phascolarctobacterium	Prevotellaceae, Dorea, Faecalibacterium	Bifidobacterium

Table B.1: Taxa with Largest Weights from Polymicrobial trait loadings. (Cutoff values: PCA loadings < -0.4 and > 0.4; N/A was assigned to PCs of polymicrobial trait methods that had loadings all smaller than the cutoffs.

Polymicrobial Trait Method	**PC**	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**
Hypothesis Covariance	**PC1**	Faecalibacterium	Faecalibacterium	Eubacteriumhallii
	PC2	Phascolarctobacterium, Clostridiumsensustricto	Roseburia, Fournierella, Faecalibacterium	Faecalibacterium, Clostridium sensu stricto
	PC3	Phascolarctobacterium, Clostridiumsensustricto	Blautia, Faecalibacterium	Clostridium sensu stricto, Blautia, Eubacterium hallii
	PC4	Bacteroides, Megasphaera	Enterobacteriaceae	Bifidobacterium, Blautia

Table B.2: Taxa with Largest Weights from Polymicrobial trait loadings. (Cutoff values: CDA loadings < -2.5 and > 2.5. N/A was assigned to CDs of polymicrobial trait methods that had loadings all smaller than the cutoffs.

Polymicrobial Trait Method	**CD**	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**
Raw CDA	**RawCD1**	Prevotella9	Succinivibrio	n/a
	RawCD2	Prevotella9, Acidaminococcus, Escherichia/Shigella, Phascolarctobacterium, Clostridiumsensustricto, Faecalibacterium	Prevotella9, Succinivibrio, Coprococcus1, Bifidobacterium, Lachnospiraceae_NA	Bacteroides, Sutterella, Coprococcus3, Alistipes, Roseburia, Parabacteroides, Coprococcus1
	RawCD3	Prevotella9,	Prevotella9, Sutterella, Prevotellaceae, Lachnospiraceae_NA	Bacteroides, Faecalibacterium, Alistipes, Roseburia,
	RawCD4	Prevotella9, Acidaminococcus, Bacteroides, Clostridiumsensustricto	Succinivibrio, Dialister, Enterobacteriaceae, Bacteroides, Coprococcus3, Anaerostipes, Sutterella, Prevotellaceae, Coprococcus1, Bifidobacterium, Lachnospiraceae_NA, Dorea, Blautia, Roseburia, Fournierella, Faecalibacterium	Faecalibacterium, Clostridium sensu stricto, Sutterella, Coprococcus3, Dorea, Alistipes,
Standardized CDA	**StdCD1**	n/a	n/a	Eubacteriumhallii
	StdCD2	Escherichia/Shigella, Faecalibacterium	n/a	Clostridium sensu stricto, Sutterella, Alistipes, Blautia, Parabacteroides, Eubacteriumhallii
	StdCD3	n/a	n/a	Bacteroides, Faecalibacterium, Clostridium sensu stricto, Blautia, Eubacteriumhallii
	StdCD4	Faecalibacterium	Succinivibrio, Enterobacteriaceae, Bifidobacterium, Lachnospiraceae_NA, Blautia, Roseburia, Fournierella, Faecalibacterium	Faecalibacterium, Clostridium sensu stricto, Sutterella, Dorea, Alistipes, Blautia, Eubacteriumhallii

Table B.3: Significant Associations within MEL A

Chr1 – MEL A (Pv01_ 40,625,364 – 41,296,564 bp)				
Position	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**	**# Sig**
41,025,998	ASV219_ Phascolarcto -bacterium	-	-	1
41,030,018	-	-	ASV58_Eubacterium Coprococcus1	2
41,074,103	-	-	ASV10_Clostridium ASV69_Faecalibacterium ASV246_Faecalibacterium Clostridiumsensustricto1 **raw.can.RCan1 (*n/a*)** Clostridiaceae1	6
41,093,387	-	Enterobacteriaceae	-	1
41,152,814	-	ASV5_Prevotella	ASV70_Lachnospiraceae LachnospiraceaeUCG004	3
41,155,494	-	-	ASV23_Coprococcus ASV95_Eubacterium **gen.corr.V1 (*n/a*)** $gen.cov.V1^{+}$ **(*Clostridiumsensustricto, Eubacteriumhallii*)** $hyp.cov.Comp.1^{\sim}$ **(*Eubacteriumhallii*)**	5
41,208,518	-	-	ASV33_Lachnospiraceae Lachnospiraceae_NA	2
41,263,233	-	-	ASV249_Clostridium	1

Table B.4: Significant Associations within MEL B

Chr3 – MEL B (Pv03_ 51,239,343 – 52,368,862 bp)				
Position	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**	**# Signif**
51,484,051	ASV7_Coprococcus ASV216_Megasphaera	Lachnospiraceae_fam ASV11_Dorea ASV142_Succinivibrio $raw.can.RCan1^{+}$ (***Succinivibrio***) Dorea Succinivibrio Ruminococcaceae hyp.corr.Comp.1 (***n/a***) gen.corr.V1 (***n/a***) ASV7_Coprococcus Coprococcus3 $gen.cov.V1^{-}$ (***Faecalibacterium***)	-	14
51,562,786	-	-	Enterobacteriaceae	1
51,606,348	-	-	EscherichiaShigella	1
51,773,201	-	Agathobacter	-	1
51,885,792	-	-	Veillonella	1
51,886,054	-	Prevotellaceae	-	1
51,902,743	-	ASV44_Blautia ASV52_Agathobacter Blautia Shannon	-	4
51,934,691	-	ASV39_Blautia	-	1
51,934,861	-	Succinivibrionaceae	-	1
51,966,148	Megasphaera	-	Acetate Ruminococcaceae_uncul	3
51,977,930	-	pielou_e	-	1
52,306,019	-	Prevotellaceae	-	1

Table B.5: Significant Associations within MEL C

Chr5 – MEL C (Pv05_9,107,694 – 10,041,586 bp)

Position	Subject 3: S768	Subject 1: S770	Subject 2: S776	# Signif
9206068	-	-	Butyricicoccus	1
9547397	-	-	ASV10_Clostridium	1
9614950	ASV50_Bacteroides	-	ASV33_Lachnospiraceae ASV49_Lachnospiraceae Lachnospiraceae_NA *hyp.corr.Comp.*2$^{\sim}$ (***Faecalibacterium,*** ***Coprococcus3,*** ***LachnospiraceaeUCG004***)	5
9615001	-	-	ASV9_Parasutterella ASV40_Faecalibacterium ASV58_Eubacterium ASV246_Faecalibacterium Agathobacter Parasutterella	6
9615026	Propionate	ASV145_Enterobacteriaceae ASV155_Prevotella ASV2_Bacteroides ASV23_Coprococcus ASV44_Blautia ASV56_Anaerostipes ASV6_Faecalibacterium Bacteroides Coprococcus1 Faecalibacterium Prevotella2	Acetate ASV256_Ruminococcaceae ASV70_Lachnospiraceae Eubacteriumhallii *gen.cov.V*1^{+} (***Clostridiumsensustricto,*** ***Eubacteriumhallii***) LachnospiraceaeUCG004 Ruminococcaceae_uncultured Ruminococcaceae RuminococcaceaeUCG002	21

Table B.6: Significant Associations within MEL D

Chr6 – MEL D (Pv06_7,267 - 2,570,337 bp)

Position	Subject 3: S768	Subject 1: S770	Subject 2: S776	# Sig
17,140	-	Acetate	-	1
145,568	-	ASV160_Lachnospiraceae	-	1
206,596	-	Shannon	-	1
583,639	-	*hyp.cov.Comp.*4^{-} (***Enterobacteriaceae***)	-	1
587,596	-	Ruminococcaceae	-	1
818,805	-	-	-	1
852,965	-	-	RuminococcaceaeUCG002	1
916,250	-	Succinivibrionaceae	-	1
1,101,942	-	-	ASV49_Lachnospiraceae	1
1,145,031	-	-	ASV255_Bacteroides	1
1,317,217	-	ASV6_Faecalibacterium Erysipelotrichaceae	-	2
1,318,580	-	-	Butyricicoccus	1
1,319,423	*hyp.cov.Comp.*2^{-} (***Phascolarctobacterium, Clostridiumsensustricto***)	-	-	1
1,319,499	-	-	ASV256_Ruminococcaceae	1
1,329,862	-	ASV52_Agathobacter	-	1

Table B.6: Significant Associations within MEL D

Chr6 – MEL D (Pv06_7,267 - 2,570,337 bp)				
Position	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**	**# Sig**
1,531,686	*gen.cov.V*1^{+} (***Faecalibacterium***) *hyp.corr.Comp*.1^{+} (***Bacteroides, Faecalibacterium***) *hyp.cov.Comp*.1^{+} (***Faecalibacterium***) *std.can.SCan*2^{+} (***Escherichia, Faecalibacterium***) *std.can.SCan*4^{+} (***Faecalibacterium***)	ASV44_Blautia ASV72_Eubacterium Faecalibacterium gen.corr.V1 (n/a) *gen.cov.V*1^{+} (***Faecalibacterium***) hyp.corr.Comp.1 (***n/a***) *hyp.cov.Comp*.1^{+} (***Faecalibacterium***) std.can.SCan1 (***n/a***) std.can.SCan3 (***n/a***) *std.can.SCan*4^{+} (***Succinivibrio, Enterobacteriaceae, Bifidobacterium, Lachnospiraceae_NA, Blautia, Roseburia, Fournierella, Faecalibacterium***)	Agathobacter Faecalibacterium	17
1,661,151	ASV220_Prevotella	-	-	1
1,709,844	-	-	*hyp.corr.Comp*.2^{+} (***Faecalibacterium, Coprococcus3, Lachnospiraceae-UCG004***)	1

Table B.6: Significant Associations within MEL D

Chr6 – MEL D (Pv06_7,267 - 2,570,337 bp)

Position	Subject 3: S768	Subject 1: S770	Subject 2: S776	# Sig
1,822,853	-	Anaerostipes	-	1
1,824,924	-	ASV78_Desulfovibrio	-	1
1,832,386	Coprococcus1	-	-	1
1,912,742	$gen.corr.V1^{+}$ (***Acidaminococcus, Faecalibacterium***)	-	-	1
1,986,819	Faecalibacterium	-	observed_ASVs	2
1,997,097	ASV6_Faecalibacterium	-	-	1
2,391,824	Parabacteroides ASV19_Parabacteroides Tannerellaceae	-	-	3
2,480,413	Ruminococcaceae	ASV7_Coprococcus ASV142_Succinivibrio Coprococcus3 Fournierella		5

Table B.7: Significant Associations within MEL E

Chr7 – MEL E (Pv07_ 9,511,250 - 10,149,594 bp)

Position	Subject 3: S768	Subject 1: S770	Subject 2: S776	# Signif
9,621,160	-	-	ASV8_Escherichia	1
9,628,646	Faecalibacterium	ASV104_Ruminococcaeae RuminococcaceaeUCG005	-	3
9,640,574	-	-	ASV58_Eubacterium	1
9,640,692	-	Butyrate ASV6_Faecalibacterium Faecalibacterium *gen.cov.V*1^{+} (***Faecalibacterium***) std.can.SCan2 (***n/a***) std.can.SCan3 (***n/a***) *std.can.SCan*4^{+} (***Succinivibrio,Enterobacter., Bifidobact., Lachnospiraceae_NA, Blautia,Roseburia, Fournierella, Faecali.)*** Ruminococcaceae	-	8
9,656,023	Acidaminococcus ASV29_Acidaminococcus	ASV2_Bacteroides ASV72_Eubacterium Enterobacteriaceae	ASV9_Parasutterella ASV70_Lachnospiraceae Parasutterella raw.can.RCan1 (***n/a***)	9
9,656,074	-	Agathobacter	-	1
9,661,392	-	*hyp.corr.Comp.*4^{+} (***Prevotellaceae, Dorea, Faecalibacterium***)	-	1
9,825,333	-	Acetate	-	1
9,900,516	-	*HypCorrPC*3^{-} (***Lachnospiraceae_NA***)	-	1

Table B.8: Significant Associations within MEL F

Chr7 – MEL F (Pv07_ 16,443,493 - 22,231,725)

Position	Subject 3: S768	Subject 1: S770	Subject 2: S776	# Signif
9,621,160	**-**	-	ASV8_Escherichia	1
9,628,646	Faecalibact.	ASV104_Ruminococcaeae RuminococcaceaeUCG005	-	3
9,640,574	**-**	-	ASV58_Eubacterium	1
9,640,692	**-**	Butyrate ASV6_Faecalibact. Faecalibact. *gen.cov.V*1^{+} (***Faecalibacterium***) std.can.SCan2 (***n/a***) std.can.SCan3 (***n/a***) *std.can.SCan*4^{+} (***Succinivibrio, Enterobacteriaceae, Bifidobacterium, Lachnospiraceae_NA, Blautia, Roseburia, Fournierella, Faecalibact.***) Ruminococcaceae	-	8
9,656,023	Acidaminococcus ASV29_ Acidaminococcus	ASV2_Bacteroides ASV72_Eubacterium Enterobacteriaceae	ASV9_Parasutterella ASV70_Lachnospiraceae Parasutterella raw.can.RCan1 (***n/a***)	9
9,656,074	**-**	Agathobacter	-	1
9,661,392	**-**	*hyp.corr.Comp.*4^{+} (***Prevotellaceae, Dorea, Faecalibact.***)	-	1
9,825,333	**-**	Acetate	-	1

Table B.8: Significant Associations within MEL F

Chr7 – MEL F (Pv07_ 16,443,493 - 22,231,725)				
Position	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**	**# Signif**
9,900,516	**-**	$HypCorrPC3^{-}$ (***Lachnospiraceae_NA***)	-	1

Table B.9: Significant Associations within MEL G

Chr8 – MEL G (Pv08_ 3,658,283 – 4,029,365 bp)				
Position	**Subject 3: S768**	**Subject 1: S770**	**Subject 2: S776**	**# Signif**
3,777,523	Megasphaera	-	-	1
3,792,888	Acidaminococcus Bacteroides Prevotella9 ASV2_Bacteroides ASV4_Prevotella ASV29_ Acidaminococcus ASV50_Bacteroides pielou_e Shannon Simpson $raw.can.RCan1^{+}$ (***Prevotella9***) Acidaminococcaceae Bacteroidaceae Prevotellaceae	ASV13_ Bacteroides	-	15
3,838,416	-	-	ASV81_ Fusicatenibacter	1

Appendix C

Three Taxa Example Simulation Code

```
load_pkg <- rlang::quos(pracma, dplyr, data.table, Matrix, spam, sparseinv)
invisible(lapply(lapply(load_pkg, rlang::quo_name),
library, character.only = TRUE))

#### Data
my_data <- read.delim("/work/eskridge/kkarnik/KinshipClean.txt", row.names = 1)
cov <- as.matrix(my_data)
A = as.matrix(cov[1:50, 1:50])
bl = nrow(A)
##########################################
##########################################

Opt_Design_ThreeTaxa <- function(sigma, RU , bl, A, alpha_options,
plates, ObsPerPlate, randomZs){# Set Up beginning values for loops, alpha, etc.
  start_time <- Sys.time()

  # Set Up beginning values for loops, alpha, etc.
  val = 1;

  alpha_total <- length(alpha_options)

  A = as.matrix(A)
  CanZ <- diag(bl)
  numtaxa = nrow(RU)
  I.taxa = diag(numtaxa)
  rand_eff = numtaxa*(numtaxa+1);
```

```
## Setting up places to store outputs

leng <- randomZs*length(alpha_options)
TotalReps.ALL.Zs <- as.data.frame(matrix(data = NA,
nrow = bl, ncol = leng, byrow = FALSE))
values <- vector(length = leng)
all.values <- matrix(data = NA, nrow = leng,
ncol = 3, byrow = FALSE)
variances <- matrix(data=NA, nrow=leng,
ncol = numtaxa*12+12, byrow=FALSE)
## 12 parameters from NRAN matrix (6 for sigma u var and covar,
##and 6 from sigma),
## then  ##numtaxa*12 for plate variances for fixed effects
Z_list <- list()
MaxAlpha <- matrix(data=NA, nrow = length(alpha_options), ncol=1,byrow=TRUE)

## Plate information, NOW standard with 12 plates
plates = plates
fixed = plates * numtaxa
ident <- diag(plates)
replicates <- as.vector(rep(1, ObsPerPlate))
X = kronecker(ident, replicates)
Xstar = kronecker(X, I.taxa)
XstarP=t(Xstar)
nobs <- plates*ObsPerPlate
diag.nobs <- Matrix(diag(nobs))

##################################################
## Setting up Derivatives
dG1dSu1  <-  matrix(NA,3,3); dG1dSu12 <-  matrix(NA,3,3); dG1dSu13 <-  matrix(NA,3,3);
dG1dSu2  <-  matrix(NA,3,3); dG1dSu23 <-  matrix(NA,3,3);
dG1dSu3  <-  matrix(NA,3,3);

dG1dSu1[lower.tri(dG1dSu1  ,  diag=TRUE)] <- c(1,rep(0,5))
dG1dSu12[lower.tri(dG1dSu12 , diag=TRUE)] <- c(rep(0,1),1,rep(0,4))
dG1dSu13[lower.tri(dG1dSu13 , diag=TRUE)] <- c(rep(0,2),1,rep(0,3))

dG1dSu2[lower.tri(dG1dSu2,diag=TRUE)]     <- c(rep(0,3),1,rep(0,2))
dG1dSu23[lower.tri(dG1dSu23,diag=TRUE)]   <- c(rep(0,4),1,rep(0,1))
```

```
dG1dSu3[lower.tri(dG1dSu3,diag=TRUE)]     <- c(rep(0,5),1)

makeSymm <- function(m) {
  m[upper.tri(m)] <- t(m)[upper.tri(m)]
  return(m)
}
makeSymm2 <- function(m) {
  m[lower.tri(m)] <- t(m)[lower.tri(m)]
  return(m)
}

list.of.deriv <- list(dG1dSu1,dG1dSu12,dG1dSu13,
                      dG1dSu2,dG1dSu23,
                      dG1dSu3)

dG.dSu <- lapply(list.of.deriv, makeSymm)

#Loop 1, Changed from Plates to alpha

for (i in 1:alpha_total){
  #   i = 1
  print(i)

  D = -10;  D1 = 0; D2 = -2; diff = 10;

  alpha <- alpha_options[i]
  infmat.fixed  <- (1-alpha)/fixed
  infmat.random <- alpha/rand_eff

  for (j in 1:randomZs){

    print(j)

    set.seed(NULL)
    Z <- Matrix(0, nobs, bl, sparse = TRUE)
    Z[cbind(1:nobs, sample(ncol(Z), nobs, TRUE))] <- 1
    Zstar = kronecker(Z, I.taxa)
    ZstarP=t(Zstar)
```

```
G = kronecker(A, sigma)
R = kronecker(diag.nobs, RU)
v = Zstar%*%G%*%ZstarP + R
#vinv = inv(v)
#vinv = solve(v)
vinv = chol2inv(chol(v))

M=XstarP%*%vinv%*%Xstar;

InfMatFixed = M;

if (det(M) == 0){
  Minv = matrix(0, fixed, fixed)}else{
    #Minv = inv(M)
    #Minv = solve(M)
    Minv = chol2inv(chol(as.matrix(M)))
  }

####################################

dD1dSu1   = kronecker(A, dG.dSu[[1]]);
dD1dSu12  = kronecker(A, dG.dSu[[2]]);
dD1dSu13  = kronecker(A, dG.dSu[[3]]);

dD1dSu2   = kronecker(A, dG.dSu[[4]]);
dD1dSu23  = kronecker(A, dG.dSu[[5]]);

dD1dSu3   = kronecker(A, dG.dSu[[6]]);

##############################

dR1dS1   = kronecker(diag.nobs, dG.dSu[[1]]);
dR1dS12  = kronecker(diag.nobs, dG.dSu[[2]]);
dR1dS13  = kronecker(diag.nobs, dG.dSu[[3]]);

dR1dS2   = kronecker(diag.nobs, dG.dSu[[4]]);
dR1dS23  = kronecker(diag.nobs, dG.dSu[[5]]);

dR1dS3   = kronecker(diag.nobs, dG.dSu[[6]]);
```

```
################################################
vinv.Xstar <- vinv%*%Xstar
XstarP.vinv <- XstarP%*%vinv

P = vinv-(vinv.Xstar%*%(Minv%*%XstarP.vinv))
P.Z <- P%*%Zstar
################################################

P.Z.D1.1.ZP  <- P.Z%*%(dD1dSu1%*%ZstarP );
P.Z.D1.2.ZP  <- P.Z%*%(dD1dSu2%*%ZstarP );
P.Z.D1.12.ZP <- P.Z%*%(dD1dSu12%*%ZstarP);
P.Z.D1.23.ZP <- P.Z%*%(dD1dSu23%*%ZstarP);
P.Z.D1.13.ZP <- P.Z%*%(dD1dSu13%*%ZstarP);
P.Z.D1.3.ZP  <- P.Z%*%(dD1dSu3%*%ZstarP );

################################################

PR1.1  <- P%*%dR1dS1;  PR1.12 <- P%*%dR1dS12; PR1.13 <- P%*%dR1dS13;
PR1.2  <- P%*%dR1dS2;  PR1.23 <- P%*%dR1dS23;
PR1.3  <- P%*%dR1dS3;

list.values <- list(P.Z.D1.1.ZP,  P.Z.D1.12.ZP,   P.Z.D1.13.ZP,
                    P.Z.D1.2.ZP,  P.Z.D1.23.ZP,
                    P.Z.D1.3.ZP,
                    PR1.1,        PR1.12, PR1.13,
                    PR1.2,        PR1.23,
                    PR1.3)

NRAN <- matrix(NA, 12, 12)
r.c <- 1
for (l in 1:length(list.values)){
  for (m in r.c:length(list.values)){
    mat.l <- list.values[[l]]
    mat.m <- list.values[[m]]
    nran.l.m <- crossprod(mat.l, mat.m)
    diagonal <- diag(nran.l.m)
    NRAN[l,m] <- 0.5*sum(diagonal)
  }
```

```
  r.c <- r.c+1
}
NRAN <- makeSymm2(NRAN)

H1=det(InfMatFixed);
H2=det(NRAN);

FixedOptCrit=log(H1);
RandOptCrit=log(H2);

if(H2<0){H2 = 10e-8}

InfMat= infmat.fixed*log(H1) + infmat.random*log(H2);

# Standard errors calculated using fisher info, for fixed and random effects
FixedEffects.SE <- sqrt(diag(solve(InfMatFixed)))
RandomEffects.SE <- sqrt(diag(solve(NRAN)))

all.values[val,] <- c(InfMat, FixedOptCrit, RandOptCrit)
variances[val,] <- c(FixedEffects.SE, RandomEffects.SE)

## Total reps and how many reps for all Z matrices along the way

TotalReps.ALL.Zs[,val] <- as.data.frame(colSums(Z));

if (InfMat > D){
  DM <- Z
  D <- InfMat
}

end_time <- Sys.time()
totaltime <- as.numeric(end_time - start_time, units="hours")
print(totaltime)

# Z=DM;
diff=D-D1;
D1=D;

values[val] <- D1
```

```
    val = val+1

  } #end random Z reps

  #  if (D1 > D2){
  #
  #    zfinalx <- Z
  #    finalx <- X
  #    D2 <- D1
  #  }

  zfinalx <- DM
  Z_list[[i]] <- zfinalx
  MaxAlpha[i,] <- D1

  # end i loop
}

TotalReps.ALL.Zs <<- TotalReps.ALL.Zs
#D2_1 <<- D2;
Z  <<- zfinalx;
Full.Opt.ZList <<- Z_list
Alpha_Opt_Values <<- MaxAlpha

dataframeZ <- as.data.frame(as.matrix(Z))
dataframeZ$PlateGroups <- as.factor(rep(1:12, each=87))
ByPlate <- dataframeZ %>% group_by(PlateGroups)
Split <- group_split(ByPlate)

plate1  <- Split[[1]][,-51];
plate2  <- Split[[2]][,-51];
plate3  <- Split[[3]][,-51];
plate4  <- Split[[4]][,-51];
plate5  <- Split[[5]][,-51];
plate6  <- Split[[6]][,-51];
plate7  <- Split[[7]][,-51];
plate8  <- Split[[8]][,-51];
plate9  <- Split[[9]][,-51];
plate10 <- Split[[10]][,-51];
```

```
plate11 <- Split[[11]][,-51];
plate12 <- Split[[12]][,-51];

TotalReps <- as.data.frame(colSums(Z));

TotalReps$Plate1Reps  <- colSums(plate1);
TotalReps$Plate2Reps   <- colSums(plate2);
TotalReps$Plate3Reps  <- colSums(plate3);
TotalReps$Plate4Reps   <- colSums(plate4);

TotalReps$Plate5Reps  <-colSums(plate5);
TotalReps$Plate6Reps  <- colSums(plate6);
TotalReps$Plate7Reps  <-colSums(plate7);
TotalReps$Plate8Reps  <- colSums(plate8);

TotalReps$Plate9Reps  <-colSums(plate9);
TotalReps$Plate10Reps <- colSums(plate10);
TotalReps$Plate11Reps <-colSums(plate11);
TotalReps$Plate12Reps <- colSums(plate12);

TotalReps <<- TotalReps
# X <<- finalx;

FinalAllCriteria <- as.data.frame(all.values)
FinalValues <- as.data.frame(values)
FinalAllVariances <- as.data.frame(variances)

HowManyLevelsofD <<- levels(as.factor(FinalValues$values))
FinalValues$DiffChanges <- FinalValues %>%
  mutate(diff = values - lag(values, default = first(values)))

FinalValues <<- FinalValues
FinalAllCriteria <<- FinalAllCriteria
FinalAllVariances <<- FinalAllVariances

end_time <- Sys.time()

time = end_time-start_time
```

```
  finaltime <<- time

  return(list(Alpha_Opt_Values, finaltime))

  # end of the function
}
###################################
## Variabilty estimates
###################################

sigma.1 = as.matrix(rbind(c(3,2.7,2.7),c(2.7,3,2.7),c(2.7,2.7,3)))
sigma.2 = as.matrix(rbind(c(3,1.5,1.5),c(1.5,3,1.5),c( 1.5,1.5,3)))
sigma.3 = as.matrix(rbind(c(3,0.3,0.3),c(0.3,3,0.3),c( 0.3,0.3,3)))
sigma.4 = as.matrix(rbind(c(0.001,0.0009,0.0009),
c( 0.0009,0.001,0.0009),c(0.0009,0.0009,0.001)))
sigma.5 = as.matrix(rbind(c(0.001,0.0005,0.0005),
c(0.0005,0.001,0.0005),c(0.0005,0.0005,0.001)))
sigma.6 = as.matrix(rbind(c(0.001,0.0001,0.0001),
c(0.0001,0.001,0.0001),c(0.0001,0.0001,0.001)))
sigma.7 = as.matrix(rbind(c(3,0.0493,
0.0493), c(0.0493,0.001,0.0009),
c(0.0493,0.0009,0.001)))
sigma.8 = as.matrix(rbind(c(3,0.0274,0.0274),
c(0.0274,0.001,0.0005),
c(0.0274,0.0005,0.001)))
sigma.9 = as.matrix(rbind(c(3,0.0055,0.0055),
c(0.0055,0.001,0.0001), c(0.0055,0.0001,0.001)))
sigma.10 = as.matrix(rbind(c(3,2.7,1.5),
c(2.7,3,0.3),c(1.5,0.3,3)))
sigma.11 = as.matrix(rbind(c(0.001,0.0009,0.0005),
c(0.0009,0.001,0.0001), c(0.0005,0.0001,0.001)))
sigma.12 = as.matrix(rbind(c(3,0.0493,0.0274),
c(0.0493,0.001,0.0001), c(0.0274,0.0001,0.001)))
sigma.13 = as.matrix(rbind(c(0.001,0.0493,0.0274),
c(0.0493,3,0.3), c(0.0274,0.3,3)))
sigma.14 = as.matrix(rbind(c(3,2.7,0.3),
c(2.7,3,0.3), c(0.3,0.3,3)))
sigma.15 = as.matrix(rbind(c(0.001,0.0009,0.0001),
c(0.0009,0.001,0.0001), c(0.0001,0.0001,0.001)))
```

```
sigma.16 = as.matrix(rbind(c(3,0.0493,0.0055),
c(0.0493,0.001,0.0001), c(0.0055,0.0001,0.001)))
sigma.17 = as.matrix(rbind(c(0.001,0.0493,0.0055),
c(0.0493,3,0.3), c(0.0055,0.3,3)))
sigma.18 = as.matrix(rbind(c(3,2.7,1.5),
c(2.7,3,1.5), c(1.5,1.5,3)))
sigma.19 = as.matrix(rbind(c(0.001,0.0009,0.0005),
c(0.0009,0.001,0.0005), c(0.0005,0.0005,0.001)))
sigma.20 = as.matrix(rbind(c(3,0.0493,0.0274),
c(0.0493,0.001,0.0005), c(0.0274,0.0005,0.001)))
sigma.21 = as.matrix(rbind(c(0.001,0.0493,0.0274),
c(0.0493,3,1.5), c(0.0274,1.5,3)))
sigma.22 = as.matrix(rbind(c(3,2.7,2.7),
c(2.7,3,0.3), c(2.7,0.3,3)))
sigma.23 = as.matrix(rbind(c(0.001,0.0009,0.0009),
c(0.0009,0.001,0.0001), c(0.0009,0.0001,0.001)))
sigma.24 = as.matrix(rbind(c(3,0.0493,0.0493),
c(0.0493,0.001,0.0001), c(0.0493,0.0001,0.001)))
sigma.25 = as.matrix(rbind(c(0.001,0.0493,0.0493), c(0.0493,3,0.3), c(0.0493,0.3,3)))

RU.1 = as.matrix(rbind(c(3,0,0), c(0,3,0), c(0,0,3)))
RU.2 = as.matrix(rbind(c(0.001,0,0), c(0,0.001,0), c(0,0,0.001)))
RU.3 =as.matrix(rbind(c(3,0,0), c(0,0.001,0), c(0,0,0.001)))

sigma <- list(sigma.1, sigma.2, sigma.3, sigma.4,
              sigma.5, sigma.6, sigma.7, sigma.8,
              sigma.9, sigma.10, sigma.11, sigma.12,
              sigma.13, sigma.14, sigma.15, sigma.16,
              sigma.17, sigma.18, sigma.19, sigma.20,
              sigma.21, sigma.22, sigma.23, sigma.24,
              sigma.25)
RU <- list(RU.1, RU.2, RU.3)

sigma = sigma[[16]]
RU = RU[[2]]

#####################################################################

alpha_options <- c(0.1,0.25,0.5,0.75,0.9)
```

```
randomZs <- 5000
plates <- 12

Design_Out <- Opt_Design_ThreeTaxa(sigma, RU=RU, bl, A,
alpha_options, plates, ObsPerPlate = 87, randomZs = randomZs)

Design_Out

#################################################################

### Summaries of All Values

Alpha_0.1_optimalityvalues  <- FinalAllCriteria[1:randomZs,]
Alpha_0.25_optimalityvalues <- FinalAllCriteria[(randomZs+1):(randomZs*2),]
Alpha_0.5_optimalityvalues  <- FinalAllCriteria[(randomZs*2+1):(randomZs*3),]
Alpha_0.75_optimalityvalues <- FinalAllCriteria[(randomZs*3+1):(randomZs*4),]
Alpha_0.9_optimalityvalues  <- FinalAllCriteria[(randomZs*4+1):(randomZs*5),]

summary(Alpha_0.1_optimalityvalues$V1)
summary(Alpha_0.25_optimalityvalues$V1)
summary(Alpha_0.5_optimalityvalues$V1)
summary(Alpha_0.75_optimalityvalues$V1)
summary(Alpha_0.9_optimalityvalues$V1)

#####################################################################

dataframeZ.Alpha_0.1  <- as.data.frame(as.matrix(Full.Opt.ZList[[1]]))
dataframeZ.Alpha_0.25 <- as.data.frame(as.matrix(Full.Opt.ZList[[2]]))
dataframeZ.Alpha_0.5  <- as.data.frame(as.matrix(Full.Opt.ZList[[3]]))
dataframeZ.Alpha_0.75 <- as.data.frame(as.matrix(Full.Opt.ZList[[4]]))
dataframeZ.Alpha_0.9  <- as.data.frame(as.matrix(Full.Opt.ZList[[5]]))

# Looking for where changes happened in the optimality values
# FinalValues$test <- FinalValues %>%
# mutate(diff = values - lag(values, default = first(values)))

# all.values --> FinalAllCriteria -->
##  overall criteria, fixed log determ, random log determ
write.csv(FinalAllCriteria)
```

```
write.csv(FinalAllVariances)
write.csv(FinalValues)

write.csv(dataframeZ.Alpha_0.1)
write.csv(dataframeZ.Alpha_0.25)
write.csv(dataframeZ.Alpha_0.5)
write.csv(dataframeZ.Alpha_0.75)
write.csv(dataframeZ.Alpha_0.9)

# TotalReps
write.csv(TotalReps)
write.csv(TotalReps.ALL.Zs)

##########################################################################
```

Bibliography

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[2] J. Alvarsson, C. Andersson, O. Spjuth, R. Larsson, and J. E. Wikberg. Brunn: An open source laboratory information system for microplates with a graphical plate layout design process. *BMC bioinformatics*, 12:1–8, 2011.

[3] H. Aschard, B. J. Vilhjálmsson, N. Greliche, P.-E. Morange, D.-A. Trégouët, and P. Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676, May 2014.

[4] A. C. Atkinson. The usefulness of optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):59–76, Jan. 1996.

[5] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Oxford Statistical Science Series. Clarendon Press, Oxford, England, Aug. 1992.

[6] M. Bailey, A. Thomas, O. Francis, C. Stokes, and H. Smidt. The dark side of technological advances in analysis of microbial ecosystems. *Journal of Animal Science and Biotechnology*, 10(1), June 2019.

[7] M. Balestre, P. P. Torga, R. G. V. Pinho, and J. B. dos Santos. Applications of multi-trait selection in common bean using real and simulated experiments. *Euphytica*, 189(2):225–238, Sept. 2012.

[8] K. Banerjee, J. Chen, and X. Zhan. Adaptive and powerful microbiome multivariate association analysis via feature selection. *NAR Genomics and Bioinformatics*, 4(1), Jan. 2022.

[9] K. Basford, E. Williams, B. R. Cullis, A. Gilmour, and G. Hammer. Experimental design and analysis for variety trials. *Plant adaptation and crop improvement*, pages 125–138, 1996.

[10] D. Bates, M. Maechler, and M. Jagan. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2022. R package version 1.5-1.

[11] J. T. Bensen, L. A. Lange, C. D. Langefeld, B.-L. Chang, E. R. Bleecker, D. A. Meyers, and J. Xu. Exploring pleiotropy using principal components. *BMC Genetics*, 4(Suppl 1):S53, 2003.

[12] M. P. Berger and W.-K. Wong. *An introduction to optimal designs for social and biomedical research*. John Wiley & Sons, 2009.

[13] M. P. F. Berger and W. K. Wong. *An Introduction to Optimal Designs for Social and Biomedical Research*. John Wiley & Sons, Ltd, May 2009.

[14] R. Bharti and D. G. Grimm. Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, 22(1):178–193, Dec. 2019.

[15] H. W. Borchers. *pracma: Practical Numerical Math Functions*, 2022. R package version 2.4.2.

[16] G. E. P. Box and D. W. Behnken. Some new three level designs for the study of quantitative variables. *Technometrics*, 2(4):455–475, Nov. 1960.

[17] P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635, June 2007.

[18] J. Burgueño, J. Crossa, F. Rodríguez, K. M. Yeater, B. Glaz, and K. M. Yeater. Chapter 13: Augmented designs-experimental designs in which all treatments are not replicated. In *ACSESS Publications*. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Inc., 2018.

[19] J. Burgueño, G. de los Campos, K. Weigel, and J. Crossa. Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, 52(2):707–719, Mar. 2012.

[20] D. G. Butler, A. B. Smith, and B. R. Cullis. On the design of field experiments with correlated treatment effects. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(4):539–555, Dec. 2014.

[21] M. D. Casler. Fundamentals of experimental design: Guidelines for designing successful experiments. *Agronomy Journal*, 107(2):692–705, Mar. 2015.

[22] E. Z. Chen and H. Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, May 2016.

[23] Y. Chen, H. Wu, W. Yang, W. Zhao, and C. Tong. Multivariate linear mixed model enhanced the power of identifying genome-wide association to poplar tree

heights in a randomized complete block design. *G3 Genes|Genomes|Genetics*, 11(2), Jan. 2021.

[24] H. Chernoff and N. Divinsky. The computation of maximum-likelihood estimates of linear structural equations. *Studies in econometric method*, 14:236–302, 1953.

[25] R. D. Cook and C. J. Nachtrheim. A comparison of algorithms for constructing exact d-optimal designs. *Technometrics*, 22(3):315–324, 1980.

[26] C. M. Cullen, K. K. Aneja, S. Beyhan, C. E. Cho, S. Woloszynek, M. Convertino, S. J. McCoy, Y. Zhang, M. Z. Anderson, D. Alvarez-Ponce, E. Smirnova, L. Karstens, P. C. Dorrestein, H. Li, A. S. Gupta, K. Cheung, J. G. Powers, Z. Zhao, and G. L. Rosen. Emerging priorities for microbiome research. *Frontiers in Microbiology*, 11, Feb. 2020.

[27] B. R. Cullis, A. B. Smith, and N. E. Coombes. On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4):381–393, Dec. 2006.

[28] J. S. de S. Bueno Filho and S. G. Gilmour. Planning incomplete block experiments when treatments are genetically related. *Biometrics*, 59(2):375–381, June 2003.

[29] J. S. de S. Bueno Filho and S. G. Gilmour. Planning incomplete block experiments when treatments are genetically related. *Biometrics*, 59(2):375–381, 2003.

[30] J. Debelius, S. J. Song, Y. Vazquez-Baeza, Z. Z. Xu, A. Gonzalez, and R. Knight. Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biology*, 17(1), Oct. 2016.

[31] V. Delorme, M. Woo, V. C. de Almeida Falcão, and C. Wood. PlateEditor: A web-based application for the management of multi-well plate layouts and associated data. *PLOS ONE*, 16(5):e0252488, May 2021.

[32] T. Ding and P. D. Schloss. Dynamics and associations of microbial community types across the human body. *Nature*, 509(7500):357–360, April 2014.

[33] M. Dowle and A. Srinivasan. *data.table: Extension of 'data.frame'*, 2022. R package version 1.14.6.

[34] C. C. Drovandi, C. C. Holmes, J. M. McGree, K. Mengersen, S. Richardson, and E. G. Ryan. Principles of experimental design for big data analysis. *Statistical Science*, 32(3), Aug. 2017.

[35] R. J. Dugand, J. D. Aguirre, E. Hine, M. W. Blows, and K. McGuigan. The contribution of mutation and selection to multivariate quantitative genetic variance in an outbred population of *Drosophila serrata*. *Proceedings of the National Academy of Sciences*, 118(31), July 2021.

[36] J. Durbin. Maximum likelihood estimation of the parameters of a system of simultaneous regression equations. *Econometric Theory*, 4(1):159–170, Apr. 1988.

[37] A. Eetemadi, N. Rai, B. M. P. Pereira, M. Kim, H. Schmitz, and I. Tagkopoulos. The computational diet: A review of computational methods across diet, microbiome, and health. *Frontiers in Microbiology*, 11, Apr. 2020.

[38] S. Ehrenfeld. On the efficiency of experimental designs. *The Annals of Mathematical Statistics*, 26(2):247–255, June 1955.

[39] J. B. Endelman and J.-L. Jannink. Shrinkage estimation of the realized relationship matrix. *G3 Genes|Genomes|Genetics*, 2(11):1405–1413, Nov. 2012.

[40] S. Fatumo, T. Carstensen, O. Nashiru, D. Gurdasani, M. Sandhu, and P. Kaleebu. Complimentary methods for multivariate genome-wide association study identify new susceptibility genes for blood cell traits. *Frontiers in Genetics*, 10, April 2019.

[41] W. T. Federer. Augmented (or hoonuiaku) designs. In *Biometrics Unit Technical Reports; Number BU-74-M*, 1956.

[42] W. T. Federer. Augmented designs with one-way elimination of heterogeneity. *Biometrics*, 17(3):447, Sept. 1961.

[43] W. T. Federer and D. Raghavarao. On augmented designs. *Biometrics*, 31(1):29, Mar. 1975.

[44] V. Fedorov. Optimal experimental design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):581–589, 2010.

[45] V. V. Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

[46] V. Feoktistov, S. Pietravalle, and N. Heslot. Optimal experimental design of field trials using differential evolution. In *2017 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, June 2017.

[47] I. Ferrocino, K. Rantsiou, R. McClure, T. Kostic, R. S. C. de Souza, L. Lange, J. FitzGerald, A. Kriaa, P. Cotter, E. Maguin, B. Schelkle, M. Schloter, G. Berg,

A. Sessitsch, and L. C. and. The need for an integrated multi-OMICs approach in microbiome science in the food system. *Comprehensive Reviews in Food Science and Food Safety*, 22(2):1082–1103, Jan. 2023.

[48] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

[49] F. R. Furrer, R. and F. Gerber. *spam: SPArse Matrix*, 2022. R package version 2.9-1.

[50] A. H. Gaike, D. Paul, S. Bhute, D. P. Dhotre, P. Pande, S. Upadhyaya, Y. Reddy, R. Sampath, D. Ghosh, D. Chandraprabha, J. Acharya, G. Banerjee, V. P. Sinkar, S. S. Ghaskadbi, and Y. S. Shouche. The gut microbial diversity of newly diagnosed diabetics but not of prediabetics is significantly different from that of healthy nondiabetics. *mSystems*, 5(2), Apr. 2020.

[51] J. Galloway-Peña and B. Hanson. Tools for analysis of the microbiome. *Digestive Diseases and Sciences*, 65(3):674–685, Jan. 2020.

[52] T. Ge, M. Reuter, A. M. Winkler, A. J. Holmes, P. H. Lee, L. S. Tirrell, J. L. Roffman, R. L. Buckner, J. W. Smoller, and M. R. Sabuncu. Multidimensional heritability analysis of neuroanatomical shape. *Nature Communications*, 7(1), Nov. 2016.

[53] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8, Nov. 2017.

[54] P. Goos. *The Optimal Design of Blocked and Split-Plot Experiments*. Springer New York, 2002.

[55] P. Goos and M. Vanderbroek. D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics*, 45(3):235–245, 2003.

[56] N. S. Grantham, Y. Guan, B. J. Reich, E. T. Borer, and K. Gross. MIMIX: A bayesian mixed-effects model for microbiome data from designed experiments. *Journal of the American Statistical Association*, 115(530):599–609, July 2019.

[57] S. Hackinger and E. Zeggini. Statistical methods to detect pleiotropy in human complex traits. *Open Biology*, 7(11):170125, Nov. 2017.

[58] D. Harville. a.(1997). matrix algebra from a statistician's perspective. *TechnometricsVolume*, 40:164–164, 1974.

[59] M. J. S. V. Haute. *Genetic Analysis in Common Bean for Variation Affecting the Human Gut Microbiome*. PhD thesis, University of Nebraska Lincoln, 2021. Unpublished Doctoral Dissertation.

[60] C. Henderson and R. Quaas. Multiple trait evaluation using relatives' records. *Journal of animal science*, 43(6):1188–1197, 1976.

[61] C. R. Henderson. *Applications of linear models in animal breeding*. University of Guelph, 1994.

[62] J. Hooton and V. Paetkau. Random assignment of treatments in a 96-well (8 $\times$ 12) microtiter plate a practical method. *Journal of Immunological Methods*, 94(1-2):81–89, Nov. 1986.

[63] H. Hotelling. Experimental determination of the maximum of a function. *The Annals of mathematical statistics*, 12(1):20–45, 1941.

[64] B. E. Huang, K. L. Verbyla, A. P. Verbyla, C. Raghavan, V. K. Singh, P. Gaur, H. Leung, R. K. Varshney, and C. R. Cavanagh. MAGIC populations in

crops: current status and future prospects. *Theoretical and Applied Genetics*, 128(6):999–1017, Apr. 2015.

[65] C. Huttenhower, R. Knight, C. T. Brown, J. G. Caporaso, J. C. Clemente, D. Gevers, E. A. Franzosa, S. T. Kelley, D. Knights, R. E. Ley, A. Mahurkar, J. Ravel, and O. White. Advancing the microbiome research community. *Cell*, 159(2):227–230, Oct. 2014.

[66] O. Idais. Locally optimal designs for multivariate generalized linear models. *Journal of Multivariate Analysis*, 180:104663, Nov. 2020.

[67] S. Institute, Feb 2019.

[68] J. K. Jarett, D. D. Kingsbury, K. E. Dahlhausen, and H. H. Ganz. Best practices for microbiome study design in companion animal research. *Frontiers in Veterinary Science*, 8, Apr. 2021.

[69] W. A. Jensen. Open problems and issues in optimal design. *Quality Engineering*, 30(4):583–593, Oct. 2018.

[70] C. Jiang and Z. B. Zeng. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics*, 140(3):1111–1127, July 1995.

[71] J. A. John and E. R. Williams. *Cyclic and Computer Generated Designs*. Springer US, 1995.

[72] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, Apr. 2016.

[73] A. G. Jones, S. J. Arnold, and R. Bürger. Stability of the g-matrix in a population experiencing pleiotropic mutation, stabilizing selection, and genetic drift. *Evolution*, 57(8):1747–1760, Aug. 2003.

[74] B. Jones and P. Goos. A candidate-set-free algorithm for generating d-optimal split-plot designs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(3):347–364, May 2007.

[75] B. Jones and P. Goos. I-optimal versus d-optimal split-plot response surface designs. *Journal of Quality Technology*, 44(2):85–101, Apr. 2012.

[76] L. Klei, D. Luca, B. Devlin, and K. Roeder. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*, 32(1):9–19, Jan. 2008.

[77] C. P. Klingenberg and L. J. Leamy. Quantitative genetics of geometric shape in the mouse mandible. *Evolution*, 55(11):2342–2352, Nov. 2001.

[78] Z. Kmail. *Optimal Design for a Causal Structure*. PhD thesis, University of Nebraska Lincoln, 2019. Unpublished Doctoral Dissertation.

[79] Z. Kmail and K. Eskridge. D-optimal design for a causal structure for completely randomized and random blocked experiments. *Journal of Probability and Statistics*, 2022:1–15, Aug. 2022.

[80] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dor-

restein. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, May 2018.

[81] J. Kuczynski, C. L. Lauber, W. A. Walters, L. W. Parfrey, J. C. Clemente, D. Gevers, and R. Knight. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*, 13(1):47–58, Dec. 2011.

[82] LabForward. Well plate templates, Aug 2022.

[83] H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, Apr. 2015.

[84] H.-J. Liu and J. Yan. Crop genome-wide association study: a harvest of biological relevance. *The Plant Journal*, 97(1):8–18, Dec. 2018.

[85] X. Liu, M. Huang, B. Fan, E. S. Buckler, and Z. Zhang. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLOS Genetics*, 12(2):e1005767, Feb. 2016.

[86] X. Liu, R.-X. Yue, and W. K. Wong. D-optimal designs for multi-response linear mixed models. *Metrika*, 82(1):87–98, Sept. 2018.

[87] Z. Liu, I. Barnett, and X. Lin. A comparison of principal component methods between multiple phenotype regression and multiple SNP regression in genetic association studies. *The Annals of Applied Statistics*, 14(1), mar 2020.

[88] Z. Liu and X. Lin. Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics*, 74(1):165–175, June 2017.

[89] Z. Liu and X. Lin. A geometric perspective on the power of principal component association tests in multiple phenotype studies. *Journal of the American Statistical Association*, 114(527):975–990, Feb. 2019.

[90] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower. The healthy human microbiome. *Genome Medicine*, 8(1), Apr. 2016.

[91] H. Mallick, S. Ma, E. A. Franzosa, T. Vatanen, X. C. Morgan, and C. Huttenhower. Experimental design and quantitative analysis of microbial community multiomics. *Genome Biology*, 18(1), Nov. 2017.

[92] C. Matthew, C. R. O. Lawoko, C. J. Korte, and D. Smith. Application of canonical discriminant analysis, principal component analysis, and canonical correlation analysis as tools for evaluating differences in pasture botanical composition. *New Zealand Journal of Agricultural Research*, 37(4):509–520, Dec. 1994.

[93] B. H. McArdle and M. J. Anderson. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, Jan. 2001.

[94] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*, 17(4):628–641, Mar. 2016.

[95] M. M. Mensack, V. K. Fitzgerald, E. P. Ryan, M. R. Lewis, H. J. Thompson, and M. A. Brick. Evaluation of diversity among common beans (phaseolus vulgaris l.) from two centers of domestication using 'omics' technologies. *BMC Genomics*, 11(1), Dec. 2010.

[96] J. J. Minich, J. G. Sanders, A. Amir, G. Humphrey, J. A. Gilbert, and R. Knight. Quantifying and understanding well-to-well contamination in microbiome research. *MSystems*, 4(4), Aug. 2019.

[97] D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, NY, 2001.

[98] L. Mramba, G. Peter, V. Whitaker, and S. Gezan. Generating improved experimental designs with spatially and genetically correlated observations using mixed models. *Agronomy*, 8(4):40, Mar. 2018.

[99] K. Mylona, S. G. Gilmour, and P. Goos. Optimal blocked and split-plot designs ensuring precise pure-error estimation of the variance components. *Technometrics*, 62(1):57–70, June 2019.

[100] K. Mylona, P. Goos, and B. Jones. Optimal design of blocked and split-plot experiments for fixed effects and variance component estimation. *Technometrics*, 56(2):132–144, Apr. 2014.

[101] N.-K. Nguyen. Gendex DOE Toolkit 8.0 FAQ. `http://www.designcomputing.net/gendex/index.html`.

[102] N.-K. Nguyen. An algorithm for constructing optimal resolvable incomplete block designs. *Communications in Statistics - Simulation and Computation*, 22(3):911–923, Jan. 1993.

[103] N.-K. Nguyen. Construction of optimal block designs by computer. *Technometrics*, 36(3):300–307, Aug. 1994.

[104] N.-K. NGUYEN. A modified cyclic-coordinate exchange algorithm as illustrated by the construction of minimum-point second-order designs. In *Advances in Statistics, Combinatorics and Related Areas*. WORLD SCIENTIFIC, Dec. 2002.

[105] N.-K. Nguyen and E. Williams. An algorithm for constructing optimal resolvable row-column designs. *Australian Journal of Statistics*, 35(3):363–370, Sept. 1993.

[106] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. M. Coin. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, 7(5):e34861, May 2012.

[107] M. F. Paget, P. A. Alspach, J. A. D. Anderson, R. A. Genet, W. F. Braam, and L. A. Apiolaza. Replicate allocation to improve selection efficiency in the early stages of a potato breeding scheme. *Euphytica*, 213(9), Sept. 2017.

[108] O. Paliy and V. Shankar. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, 25(5):1032–1057, Feb. 2016.

[109] O. Paliy and V. Shankar. Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, 25(5):1032–1057, feb 2016.

[110] H. D. PATTERSON and R. THOMPSON. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[111] H. D. PATTERSON and E. R. WILLIAMS. A new class of resolvable incomplete block designs. *Biometrika*, 63(1):83–92, 1976.

[112] H.-P. Piepho, N. Vo-Thanh, and R. Tobias. Generating experimental designs for estimation of genetically related treatment effects using SAS. *Agronomy Journal*, 112(5):3929–3940, Aug. 2020.

[113] H.-P. Piepho and E. R. Williams. Augmented row-column designs for a small number of checks. *Agronomy Journal*, 108(6):2256–2262, Nov. 2016.

[114] Z. Ping, Z. Yang, Q. Cheng, Z. Liwei, Z. Ruyang, G. Jianwei, L. Jin, L. Liya, and C. Feng. Statistical analysis for genome-wide association study. *The Journal of Biomedical Research*, 29(4):285, 2015.

[115] R. A. Power, J. Parkhill, and T. de Oliveira. Microbial genome-wide association studies: lessons from human GWAS. *Nature Reviews Genetics*, 18(1):41–50, Nov. 2016.

[116] D. A. Preece. R.A. Fisher and experimental design: A review. *Biometrics*, 46(4):925, Dec. 1990.

[117] M. Prus and H.-P. Piepho. Optimizing the allocation of trials to sub-regions in multi-environment crop variety testing. *Journal of Agricultural, Biological and Environmental Statistics*, 26(2):267–288, Jan. 2021.

[118] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[119] D. Rasch, J. Pilz, L. R. Verdooren, and A. Gebhardt. *Optimal Experimental Design with R*. Whittles Publishing, Caithness, UK, May 2011.

[120] M. Ringnér. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, Mar. 2008.

[121] S. Rio, D. Akdemir, T. Carvalho, and J. I. y Sánchez. Assessment of genomic prediction reliability and optimization of experimental designs in multi-environment trials. *Theoretical and Applied Genetics*, 135(2):405–419, Nov. 2021.

[122] M. A. F. Rodríguez, J. C. Puigvert, and O. Spjuth. Designing microplate layouts using artificial intelligence. *bioRxiv*, Apr. 2022.

[123] J. Rodríguez-Díaz and G. Sánchez-León. Optimal designs for multiresponse models with double covariance structure. *Chemometrics and Intelligent Laboratory Systems*, 189:1–7, June 2019.

[124] P. S. L. Rosa, J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS ONE*, 7(12):e52078, Dec. 2012.

[125] C. Roselle, T. Verch, and M. Shank-Retzlaff. Mitigation of microtiter plate positioning effects using a block randomization scheme. *Analytical and Bioanalytical Chemistry*, 408(15):3969–3979, Apr. 2016.

[126] J. Royle. Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, 100(2):121–134, 2002.

[127] F. Sambo, M. Borrotti, and K. Mylona. A coordinate-exchange two-phase local search algorithm for the d- and i-optimal designs of split-plot experiments. *Computational Statistics & Data Analysis*, 71:1193–1207, Mar. 2014.

[128] D. J. Schaid, X. Tong, A. Batzler, J. P. Sinnwell, J. Qing, and J. M. Biernacka. Multivariate generalized linear model for genetic pleiotropy. *Biostatistics*, December 2017.

[129] P. D. Schloss. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio*, 9(3), July 2018.

[130] P. D. Schloss. Amplicon sequence variants artificially split bacterial genomes into separate clusters. *mSphere*, 6(4), Aug. 2021.

[131] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance components*. John Wiley & Sons, 1992.

[132] R. M. Shah, E. J. McKenzie, M. T. Rosin, S. R. Jadhav, S. V. Gondalia, D. Rosendale, and D. J. Beale. An integrated multi-disciplinary perspective for addressing challenges of the human gut microbiome. *Metabolites*, 10(3):94, Mar. 2020.

[133] E. R. Shanahan, J. J. McMaster, and H. M. Staudacher. Conducting research on diet–microbiome interactions: A review of current challenges, essential methodological principles, and recommendations for best practice in study design. *Journal of Human Nutrition and Dietetics*, 34(4):631–644, Feb. 2021.

[134] J. Shankar. Insights into study design and statistical analyses in translational microbiome studies. *Annals of Translational Medicine*, 5(12):249–249, July 2017.

[135] A. Sharma, J. S. Lee, C. G. Dang, P. Sudrajad, H. C. Kim, S. H. Yeon, H. S. Kang, and S.-H. Lee. Stories and challenges of genome-wide association studies in livestock — a review. *Asian-Australasian Journal of Animal Sciences*, 28(10):1371–1379, Mar. 2015.

[136] L. Shenhav, O. Furman, L. Briscoe, M. Thompson, J. D. Silverman, I. Mizrahi, and E. Halperin. Modeling the temporal dynamics of the gut microbial community in adults and infants. *PLOS Computational Biology*, 15(6):e1006960, June 2019.

[137] N.-R. Shin, T. W. Whon, and J.-W. Bae. Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in Biotechnology*, 33(9):496–503, Sept. 2015.

[138] A. B. Smith, P. Lim, and B. R. Cullis. The design and analysis of multi-phase plant breeding experiments. *The Journal of Agricultural Science*, 144(5):393–409, Sept. 2006.

[139] F. W. Stearns. One hundred years of pleiotropy: A retrospective. *Genetics*, 186(3):767–773, Nov. 2010.

[140] D. M. Steinberg and W. G. Hunter. Experimental design: Review and comment. *Technometrics*, 26(2):71–97, May 1984.

[141] B. Stemshorn, D. Buckley, G. Amour, C. Lin, and J. Duncan. A computer-interfaced photometer and systematic spacing of duplicates to control within-plate enzyme-immunoassay variation. *Journal of Immunological Methods*, 61(3):367–375, July 1983.

[142] M. Stephens. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*, 8(7):e65245, July 2013.

[143] W. W. Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

[144] H. Sun, Y. Wang, Z. Xiao, X. Huang, H. Wang, T. He, and X. Jiang. multiMiAT: An optimal microbiome-based association test for multicategory phenotypes. *bioRxiv*, July 2022.

[145] M. Suprun and M. Suárez-Fariñas. PlateDesigner: A web-based application for the design of microplate experiments. *Bioinformatics*, 35(9):1605–1607, Oct. 2018.

[146] SAS Institute Inc. *SAS/STAT Software, Version 9.4*. Cary, NC, 2015.

[147] P. J. Turnbaugh, V. K. Ridaura, J. J. Faith, F. E. Rey, R. Knight, and J. I. Gordon. The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Science Translational Medicine*, 1(6), Nov. 2009.

[148] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), Aug. 2021.

[149] A. W. Walker. A lot on your plate? well-to-well contamination as an additional confounder in microbiome sequence analyses. *MSystems*, 4(4), Aug. 2019.

[150] Q. Wang, K. Wang, W. Wu, E. Giannoulatou, J. W. K. Ho, and L. Li. Host and microbiome multi-omics integration: applications and methodologies. *Biophysical Reviews*, 11(1):55–65, January 2019.

[151] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan. *dplyr: A Grammar of Data Manipulation*, 2023. R package version 1.1.0.

[152] E. Williams, H.-P. Piepho, and D. Whitaker. Augmented p-rep designs. *Biometrical Journal*, 53(1):19–27, Nov. 2010.

[153] E. R. Williams. Row and columan designs with contiguous replicates. *Australian Journal of Statistics*, 28(2):154–163, June 1986.

[154] Y. Xia and J. Sun. Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases*, 4(3):138–148, Sept. 2017.

[155] Y. Xia, J. Sun, and D.-G. Chen. *Statistical Analysis of Microbiome Data with R.* Springer Singapore, 2018.

[156] J. J. Yang, J. Li, L. K. Williams, and A. Buu. An efficient genome-wide association test for multivariate phenotypes based on the fisher combination function. *BMC Bioinformatics*, 17(1), Jan. 2016.

[157] Q. Yang, M. V. Haute, N. Korth, S. E. Sattler, J. Toy, D. J. Rose, J. C. Schnable, and A. K. Benson. Genetic analysis of seed traits in sorghum bicolor that affect the human gut microbiome. *Nature Communications*, 13(1), Sept. 2022.

[158] F. Yates. *The design and analysis of factorial experiments.* Imperial Bureau of Soil Science, 1937.

[159] K. M. Yeater, S. E. Duke, and W. E. Riedell. Multivariate analysis: Greater insights into complex systems. *Agronomy Journal*, 107(2):799–810, Mar. 2015.

[160] W. J. Youden. Use of incomplete block replications in estimating tobacco-mosaic virus. *Journal of Quality Technology*, 4(1):50–57, Jan. 1972.

[161] A. Zammit-Mangion. *sparseinv: Computation of the Sparse Inverse Subset*, 2018. R package version 0.1.3.

[162] W. Zhang, B. Hu, C. Liu, H. Hua, Y. Guo, Y. Cheng, W. Yao, and H. Qian. Comprehensive analysis of sparassis crispa polysaccharide characteristics during the in vitro digestion and fermentation model. *Food Research International*, 154:111005, Apr. 2022.

[163] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, June 2012.

[164] X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409, February 2014.

[165] H. Zhu. *Statistical methods for analyzing multivariate phenotypes and detecting rare variant associations*. PhD thesis, Michigan Technological University, 2021.