MINE: Maximally Informative Next Experiment - Genetics application and novel computational methodology

by

Isaac Manuel Torres Bermeo

(Under the Direction of Jonathan Arnold)

Abstract

The computational methodology of Genome Wide Association Studies (GWAS) currently has several limitations: 1) the number of observations (rows) on a quantitative trait tends to be smaller than the number of single nucleotide polymorphisms (SNPs) (columns) in the design matrix; 2) each SNP is usually modeled separately, failing to acknowledge interaction between each other; 3) there is implicit linkage disequilibrium (LD) between neighboring SNPs. To overcome these issues, we developed a tool that uses ensemble methods to fit mixed linear models into GWAS, and these ensemble methods include the development of a new experimental design approach in GWAS which uses the resultant models and data to select the next informative experiment over time. This new adaptive approach for GWAS experimental design was developed and tested in a 3 year adaptive model-guided discovery experiment.

Index words: [Bioinformatics, simulation, MCMC, sorghum, GWAS]

MINE: Maximally Informative Next Experiment - Genetics application and novel computational methodology

by

Isaac Manuel Torres Bermeo

B.S., Escuela Superior Politecnica del Litoral, 2015

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2024

MINE: Maximally Informative Next Experiment - Genetics application and novel computational methodology

by

Isaac Manuel Torres Bermeo

Major Professor: Jonathan Arnold

Committee: Bernd Schuttler
Katrien Devos
Alexander Bucksch

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2024

Dedication

To God, yours is the glory and honor.
To my parents, Dr. Isaac Torres and Dr. Ana Bermeo, for your unconditional support along these years.
To uncle Freddy, Pablo and Antonio; aunt Gladys and Elena, for your help and support during my studies in the United States.
To Dr. Jonathan Arnold and Dr. Bernd Schuttler for your amazing patience to teach and guide me during this PhD program as well as for your support.
To my previous advisors, Dr. Frank Drews, Dr. Daniel Ochoa and Prof. Carlos Jordan, for teaching the necessary skills to get to this PhD program.

Acknowledgments

This bioinformatics journey started by chance in December 2012 at Prof. Carlos Jordan office when I realized that my computer science skills could be applied to help solving biology/medical problems; I wanted to come up with a significant undergraduate thesis, and Prof. Jordan showed me an interesting systems biology project. I am deeply grateful to Prof. Jordan for showing me this interesting bioinformatics path, and his support during those early research years.

I cannot forget mentioning and being thankful to Dr. Daniel Ochoa, who collaborated with Prof. Jordan, and offered me staying working in research projects after graduation.

The beginning of my graduate studies was not easy, however, thanks to Dr. Frank Drews everything stabilized, and I am profoundly grateful for his teaching and recommendation.

My time at the Institute of Bioinformatics has been amazing from the first day when I communicated with April and Sandra about onboarding details, I want to thank all IOB staff and leadership, they play such an important role in students well-being.

I am deeply grateful to my current advisors, Dr. Jonathan Arnold and Dr. Bernd Schuttler, for their support, guidance, and all the time invested in checking my work, giving me feedback and teaching; I hope to follow your steps wherever my next professional chapter is.

Thanks to my PhD committee for all the feedback received during these years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Genome Wide Association Studies (GWAS) have been a means to obtain insights about the relation between Single nucleotide polymorphisms (SNPs) and phenotypes, such as height or certain diseases [5]; The ultimate goal of GWAS is to identify genomic regions that control a trait. One important part is the experiment design, which involves selecting the genotypes for the GWAS experiment; this work presents an approach to select the most informative genotypes in a series of annual adaptive GWAS on *Sorghum bicolor*; the method is called MINE which stands for Maximally Informative Next Experiment, and it is specially useful when resources are limited and large amount of plants cannot be considered. The MINE approach is also accompanied by novel GWAS computational methodology using ensemble methods [6] to fit mixed linear models. This approach addresses a few limitations of regular regression models such as having a number of observations lower than the number of SNPs, as well as the selection of final genomic regions. The usual way to address the small n (observations), large p (SNPs) problem is by feature selection, thereby rendering the statistical analysis addressable by standard approaches. The problem with doing this is that most of the data are thrown away. An alternative approach is to use ensemble methods to address this problem [7]. Ensemble approaches arise from statistical physics to specifically address this problem and have only been introduced recently into the biological domain [6]. Another feature presented here is the incorporation of all SNPs into the mixed linear modeling phase by forming new chromosomal regions of 50 KB to avoid linkage disequilibrium; this represents an implicit interaction between SNPs instead of generating an isolated model per SNP.

This work is illustrated with a field experiment done at the University of Georgia's Wellbrook Farm located in Watkinsville, Georgia; each year 81 genotypes were selected, and seeds were ordered at GRIN website; plants were grown

at the University of Georgia greenhouse for 2 weeks and then transplanted to the farm; harvesting took place 3 months later, and the following traits were collected: height, dry weight and disease. To obtain dry weight, the plants were chopped and bagged, and put into ovens for around 1 week. The SNP data used to explain various complext traits were taken from Morris work [8] and generated in [2]. The genotypes were selected yearly using MINE in years 2 and 3 of the 3 year adaptive GWAS guided by MINE.

Chapter 2

MINE: a new way to design genetics experiments for discovery

2.1 Introduction

The classic approach to experimental design was developed by R. A. Fisher for linear models in 1935 and had a profound effect on all of science[9][10]. Growing out of his work at the Rothamsted Experiment Station, he introduced widely the notion of precision of an experiment, randomization, ways of controlling heterogeneity through blocking and by the use of covariates, and the vast subject of experimental design in the context of linear models. Now the subject of design is permanently associated with the mathematics of Latin squares, Graeco-Latin Squares, factorial designs, response surfaces, partial factorial designs, and incomplete block designs[11].

The focus of all of these efforts was not on discovery per se. Rather, the end goal was the precision of estimates and the power to test effects in a controlled experiment with the proper randomization and blocking practices in place. The number of replicates was such that the number of observations (n) was typically much greater than the number of effects (p) being estimated in the model. Unfortunately this is no longer the typical situation of an omics experiment[12], such as a Genome Wide Association Study (or GWAS) or quantitative trait locus (QTL) mapping of plant biomass, height, photoperiod sensitivity, and tillering as examples. Instead, there may be only n=1943 samples of sorghum accessions but over p=400,000 potential effects of single nucleotide polymorphisms (or SNPs) on the complex trait of interest for an agronomic crop. The problem has only grown more complicated as more variables are added to the

mix, such as the microbiomes of plants[13] or the expression of quantitative trait loci (eQTLs)[14]. The new variables in the mix add to the complexity because they are inter-related in unknown ways. Geneticists are faced with the challenge of designing very costly omics experiments in which the number of variables (p) measured on each sample, e.g., plant accession, vastly exceeds the number of samples (n). Classical design is not well-equipped for this situation, and new design approaches are needed to "find the needles in the haystack", i.e., the few variables in potentially millions of variables measured that really matter in systems biology[15].

The nature of experimental design has also fundamentally changed. While the original goal was the precision of estimates[10], the new goal is discovery. The reason that it is less important because the data in large omic studies are no longer capable of precise measurements of individual parameters, but rather the goal of such studies has shifted to the discovery of relationships in the data. The focus on precision of effects can only be addressed in followup studies when the relevant variables in the experiment have been identified and related. We wish to discover the relation of plant functional traits to SNPs in the nuclear genome or the assemblage of fungal symbionts in the microbiome most beneficial for plant growth[16] - [17] using models drawn from systems and population ecology [18]. We desire to discover the appropriate nonlinear kinetic models that underly the biological clock at the molecular level[1] as we carry out very expensive transcriptomic experiments. How can the classic linear models[19] and newer models of systems biology[1] guide a discovery process, in which as many as 232,303 SNPs are available for triangulation in the nuclear genome but only a small number have a profound effect on traits of interest, such as biomass[2]. Some GWAS experiments have even millions of SNPS, such as in a human height study recently [20].

Here a new approach to the design of large genomics experiments is introduced, one in which model guided discovery is used to find the variables that matter, considering a system in which the number of potential effects in the system (p) far exceeds the number of observations (n) on the system. The methodological approach to solving such problems utilizes ensemble methods[6] drawn from statistical physics[21] and, ultimately, Boltzmann's 19nth Century work[22]. The particular ensemble method explored here for model-guided discovery is called MINE, which stands for maximally informative next experiment[23].

This chapter is a review of MINE methods. In the next section 2.2 a simple example is used to illustrate ensemble methods that underly the design tool MINE. The ensemble is a collection of models that are consistent with the available data, and an ensemble method is used to identify the relevant models and

make predictions about future experiments[6]. In the third section the ensemble is used to select the maximally informative next experiment (MINE)[1][19], which is illustrated with a simple example drawn from mixture experiments as a particular class of linear models[24]. The predictions made by each member of the ensemble are used to guide the choice of the next experiment to discover as much as we can about the biological system. In section IV some of the properties of MINE are described, providing the rationale for its use in model-guided discovery, such as its consistency in finding the true model[23]. In sections VI.-IX. the use of MINE is illustrated as a discovery tool in genetics experiments and associated field trials. The review finishes with some concluding remarks.

2.2 Ensemble methods

Ensemble methods were developed in the 19^{th} century by Boltzmann to describe the motion of an ideal gas[22]. In this situation there is an Avogadro number (A) of particles in a one liter box, but only 3 measurements are made: temperature, pressure, and volume. How is the motion of the particles in the box described? How is the motion of A particles described with 6A degrees of freedom each with only three measurements?

With so little data, the data did not strongly support just one model. Boltzmann's solution was to give up on identifying one best model, but rather to make predictions from an ensemble of models. Omics experiments face exactly the same problem[6], but the paucity of data with respect to the complexity of the model is not as severe as in the problem Boltzmann faced. The interest may be in identifying the dynamics of genes and their products in carbon metabolism[6] or the biological clock[25], but the genetics dictates that only a limited number of samples at different time points can be made to identify the system, while there are many parameters required to describe the system. For example, the number of measurements at different time points on the biological clock may be on the order of 60,000 measurements (n), but there are over 90,000 rate constants and initial conditions (p) in the model that must be estimated[26][27]. Much as for an ideal gas, averaging over the ensemble allows for detailed predictions about complex biological systems, such as the clock. A simple example is used to illustrate the approach of ensemble methods. The first step is writing down the model specification for the measurements. There are n measurements $Y = (y_1, ..., y_n)$ drawn from an unknown distribution parameterized by p parameters in $\theta = (\theta_1, ..., \theta_p)$, and some of these parameters are ancillary parameters, such as α, in which there is less interest. In addition, the variables $U = (u_1, ..., u_n)$ describe the experimental conditions. For example,

the list U might specify the SNPs used in a GWAS field trial. Then the model specification would take the form:

$$P(Y|\theta, U) = C(Y)e^{-H(Y|\theta,U)} \tag{2.1}$$

where C is a normalization constant chosen to make the integral over the data Y to equal 1. The quantity $H(Y|\theta, U)$ is known as the Hamiltonian. Ideally the distribution would be observed directly, but in practice what is available are sample moments of the data Y.

Since the goal is to identify a model θ supported by the data Y, a change of viewpoint is needed. As in the method of maximum likelihood[28], the model specification $P(Y|\theta, U)$ is viewed as a function of the model parameters θ, and the data Y and experimental conditions U are taken as fixed:

$$Q(\theta|Y, U) = \Omega^{-1}P(Y|\theta, U) \tag{2.2}$$

where Ω is a normalization constant chosen to make the integral over all parameters θ in the parameter space equal to 1. This normalization constant is only a function of the data Y. The magnitude of the ensemble $Q(\theta|Y, U)$, or $Q(\theta)$ for short, is larger when the model θ is more supported by the data Y. It may be useful to think of the ensemble $Q(\theta|Y, U)$ as a posterior distribution to the model specification $P(Y|\theta, U)$ with the two functions, $P(Y|\theta, U)$ and $Q(\theta|Y, U)$, connected by Bayes Theorem[29].

The ensemble $Q(\theta|Y, U)$, or $Q(\theta)$ for short, is the collection of models θ consistent with the available data Y. Model-averaging with respect to the model ensemble $Q(\theta)$ allows predictions about the system's behavior. Instead of identifying one model θ, a distribution of models $Q(\theta)$ is identified. With the number of parameters p being vastly greater than the number of data points n, predictions can still be made and tested with respect to averages computed from the ensemble $Q(\theta)$.

Monte Carlo Methods are used to identify the ensemble $Q(\theta)$[7] because the model specifications are complicated[25][30]. A simple example will illustrate how this is done. Take the Hamiltonian viewed as a function of θ as having the following simple form:

$$H(\theta) = \beta[-\theta^2 + \alpha\theta^4] \tag{2.3}$$

The model parameter θ is the one we are truly interested in, and the remaining parameters α and β are ancillary. A graph of the ensemble $Q(\theta) = e^{-H(\theta)}$ is shown in Figure 2.1d. There are two maxima in the ensemble or equivalently,

two minima in the Hamiltonian. The goal is to reconstruct the ensemble $Q(\theta)$ by Monte Carlo for prediction.

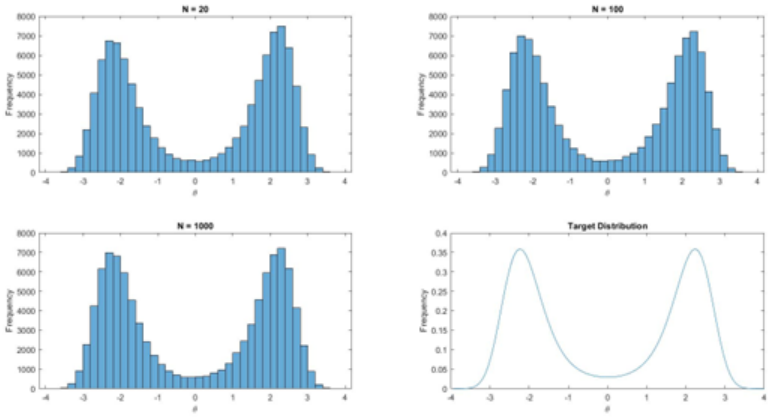

Figure 2.1: Convergence of an ensemble to the target distribution: (a) ensemble after 20 moves; (b) ensemble after 100 moves; (c) ensemble after 1000 moves; (d) true ensemble or "target distribution". In panel (d) is shown the target distribution as the ensemble converges to the true target distribution under a Monte Carlo Experiment (a-c). The starting guess at to the parameter θ was 3. After each of 20, 100, and 1000 moves, 10,000 samples of θ from the resulting distribution were drawn to characterize the ensemble. The ancillary parameters were $\alpha = 0.1$ and $\beta = 1.0$. The plots were created in Matlab R2018B.

In this example, we are in the perfect world in which the ensemble, or equivalently the Hamiltonian, is observed from 10,000 values after each move in the Monte Carlo experiment. To reconstruct the ensemble $Q(\theta)$ by Monte Carlo at each move a new model parameter $\theta^{'}$ is drawn from the ensemble $Q(\theta)$ when the current proposal is the model parameter θ. The goal is to move into a region of the parameter space which is well supported by the ensemble $Q(\theta)$ in the equilibration phase. Once equilibrated many 1,000s or 10,000s of models are accumulated that are well supported to reconstruct the ensemble from the sample histogram of these θ-values[25]. The question remains how to choose the well-supported θ-values.

One greedy approach to moving in the parameter space is to draw a model parameter $\theta^{'}$ and proceed up hill using some procedure like steepest ascent to climb the hill(s) in the ensemble. As shown in Figure 2.1, this might lead to a local maximum. In fact in Figure 2.1 there are 2 such maxima. To avoid local maxima, a model parameter $\theta^{'}$ is drawn randomly from the ensemble $Q(\theta)$, being greedy when there is an improvement in the ensemble probability, i.e., $Q(theta^{'}) > Q(\theta)$ or equivalently $H(\theta^{'}) < H(\theta)$, but occasionally when

$Q(\theta') < Q(\theta)$, move downhill anyway. The occasional downhill move may allow escape from a local maximum. In practice in systems biology it may be more appropriate to think of the ensemble surface as gently rolling hills as on a golf course because they data are limited (n < p).

Metropolis and colleagues[31] developed a stochastic search procedure in statistical physics for this and now many other optimization problems[7]. The probability of a move is:

$$p = min(1, \frac{Q(\theta')}{Q(\theta)}) \tag{2.4}$$

The probability p of a move from $\theta \rightarrow \theta'$ occurs with probability 1 if the proposed move takes us up hill, but if the proposed move takes us downhill, then the probability of a move downhill decreases with the amount of drop from $Q(\theta) \rightarrow Q(\theta')$. The sequence of moves are made 10,000 or more times to move into a region of the parameter space well supported by the data during the equilibration phase. This sequence of moves is known as a Markov Chain.

In the equilibration phase the inferred ensemble converges to the true ensemble known as the target distribution (Figure 2.1). The Monte Carlo search for this simple model is successful in the reconstruction in less than a 1,000 moves. Once equilibration is achieved, another sequence of Monte Carlo moves called the accumulation phase is used to build the target distribution. In this simple example only a 1,000 moves are needed to carry the ensemble identification into the accumulation phase.

In practice a sweep is introduced to describe the number of moves taken to visit each model parameter on average once. If there were 10 parameters, then a sweep would consist of 10 moves. The standard length of an equilibration run is 40,000 sweeps, which will vary in practice with the complexity of the model; likewise, the standard length of an accumulation run is also 40,000 sweeps[25][26].

2.3 MINE

Once an ensemble method produces a collection of models supported by the data, then it is possible to make predictions from the ensemble distribution about the next experiment. By averaging some variable of interest over the models in the ensemble distribution $Q(\theta|Y, U)$, a prediction can be made, given the current data Y and the experimental conditions U. For example, Y might be plant biomasses measured in year 1 of a 5 year GWAS experiment to identify SNPs to predict biomass in sorghum with a certain collection of SNPs from

the Bioenergy Accession Panel (BAP)[2]. The question is what SNPs are to be used in year 2. The SNPs to be chosen in year 2 are specified by the design U. In year 1, 79 accessions are measured in the GWAS and can be used to inform the accession choice in year 2. What BAP accessions in year 2 should be selected to give us the most new information about the underlying QTLs for biomass?

One way to make this choice is to select experimental conditions permitting us to distinguish the models in the ensemble $Q(\theta|Y,U)$ identified from Year 1. The best way to distinguish experimentally two models randomly chosen from the model ensemble is if the predictions $F(\theta,Y)$ of each model ($\theta = \theta_1 or \theta_2$) are orthogonal as shown in Figure 2.2. For experiment 1 on the left, the predictions of the two selected models θ_1 and θ_2 are correlated and are harder to distinguish under experimental conditions U_1. The same two models under experimental conditions U_2 are easier to distinguish - model θ_1 is easily tested again model θ_2. The goal of a MINE criterion is then to support making "the angle" between the two predictions of a random pair in the ensemble as large as possible on average in year 2 as a function of the experimental conditions U and current ensemble $Q(\theta|Y,U)$ identified from the data in year 1.

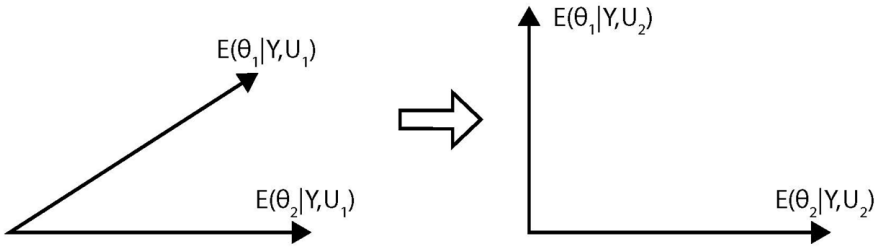

Figure 2.2: Two models can be better distinguished by their predictions in the next experiment if their predictions are less correlated. The predictions of model θ_1 and θ_2 under experimental condition U_1 are the expectations $f_1 = E(\theta_1|Y,U_1)$ and $f_2 = E(\theta_2|Y,U_1)$, respectively. If the two models θ_1 and θ_2 are chosen independently from the model ensemble $Q(\theta|Y,U)$, the expectations are calculated with respect to the product density $Q(\theta_1|Y,U)Q(\theta_2|Y,U)$, where $U = U_1 or U_2$

There are two standard ways to measure the correlations between the predictions[1]. One is by the covariances between the components of the data Y (MINE by Covariance Ellipsoid Volume); the other is by the correlations between the components of the data Y (MINE by Correlation Ellipsoid Volume). There are a variety of reasons for advocating the use of MINE by Correlation Ellipsoid Volume[1]. One of the main reasons is that when there are a large number (p »n) of almost linearly dependent observations as found in practice, it would

be highly desirable to emphasize the new directions in the data Y as done by Correlation Ellipsoid Volume. Denote by E the correlation matrix between the components of Y. The MINE Correlation Ellipsoid Volume is then a determinant (det):

$$V(U) = det(E(U)) \tag{2.5}$$

When the predictions are on average highly correlated (Figure 2.2A), the determinant is nearly zero. When the predictions are nearly orthogonal (going in new directions) (Figure 2.2B), the determinant is nearly 1.

A microscope analogy[19] provides insights on how MINE works (Figure 2.3). MINE is highly analogous to a microscope and its optics. The object in the microscope field described by the data Y is the observed system. MINE, like the optics of the microscope, picks up each component of Y through the prediction $F(\theta, Y)$ about the system. For example, $F(\theta, Y)$ could be the list of predictions of plant biomasses in a GWAS study. The optics $(F(\theta, Y))$ and likewise, MINE, then magnify the predictions to create the image or model of the system (Figure 2.3).

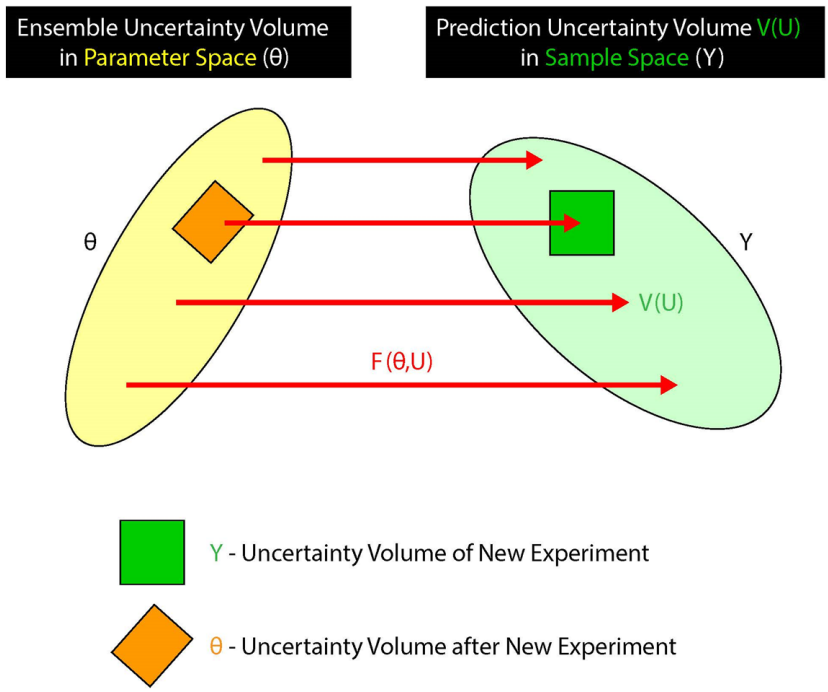

Figure 2.3: MINE is analogous in function to the optics on a microscope. The data Y are the objects in the field of view. The models θ are in the image. The MINE criterion with the predictions $F(\theta, U)$ is the optics. The Uncertainty Volume in the image is the magnification measured by the MINE criterion, V(u) = det(E(U)).

The microscope has a field of view of the object, which we refer to as the Uncertainty Volume of the new experiment Y. The uncertainty in the observations on the field of view comes from our uncertainty about the optics controlled by θ and in the measurements Y on the object. The optics (predictions) then translate the Uncertainty Volume V(U) in the sample space into an image, the Uncertainty Volume in the parameter space. The result is that an Uncertainty Volume in the sample space (object) is mapped by the optics $F(\theta, Y)$ to the Uncertainty Volume in the parameter space (image).

The magnification applied to the object is adjusted to reduce the Uncertainty Volume in the parameter space (image). Another interpretation of the image quality is given by the determinant det(E(U)). The determinant is a measure of the volume of a parallelepiped defined by the Uncertainty Volume in the Sample Space[32]. The determinant is also a measure of the Uncertainty volume in the parameter space (inside the ensemble). As the magnification knob is twiddled, the clarity of the image (model parameters) is increased and uncertainty, reduced (Figure 2.3). The choice of experiments under MINE is designed

to yield maximum clarity about nature and our view of it through the models in the ensemble. If the parallelepiped is squashed in the parameter space, less detail from the observations Y in the sample space are being retained in imaging (i.e., model fitting). MINE is doing the focusing and representing the object in higher clarity in the image constructed by the observer using MINE.

A mixed linear model is used for predicting hyphal extension colonization by arbuscular mycorrhizal fungi (AMF) in plant roots to illustrate MINE. In mixture experiments the design matrix is used to specify the proportion of different treatments, in this case of different AMF species; they are used in population genetics[33] and science and engineering in general[34]. Mixture experiments are examples of linear models that are the focus of experimental design[10]. Mixture experiments can be used to study how AMF affect the health of the plant through colonization of the root system. The assembly of the AMF biome in plant roots is a product of choices imposed by the plant genotype[35][36], competition between AMF, ecological drift[17], historical contingency[37], abiotic factors such as Phosphorous (P) and Nitrogen (N) in the soil[38][39], and other factors. Consider three AMF species, S1, S2 and S3, competing for colonization area in the plant roots of sorghum[17] of ONE plant genotype. These AMF are potential partners with the plant in one of the oldest symbioses on the planet[40]. Potentially the plant provides carbon, and in return potentially the AMF provide P and N. The success of this partnership is measured in part by AMF hyphal extension in the roots and the resulting biomass of the plant host. The AMF hyphal extension determines the access to soil nitrogen and phosphorous for the plant. To study this symbiosis the experimenter inoculates sorghum with a mixed population at least 10 percent of the conidial cells being S1, at least 15 percent are S2 and at least 5 percent of the conidial cells being S3. The inoculum is a co-culture in the plant root cells. Denoting respective spore percentages by u_1, u_2 and u_3, respectively, u_1, u_2 and u_3, are thus constrained by lower limits,

$$u_1 \geq u_1^{(lo)} = 0.10, u_2 \geq u_2^{(lo)} = 0.15, u_3 \geq u_3^{(lo)} = 0.05 \tag{2.6}$$

and by the normalization condition

$$u_1 + u_2 + u_3 = 1 \tag{2.7}$$

Given eq. 2.7, only two of the three species fraction values can be freely chosen. In the following, we will use proportions u_1 and u_2 as those two free variables, with u_3 then being determined via eq. 2.7. Furthermore, the proportions u_1 and u_2 are then subject to upper and lower bounds, resulting from eq.

2.6 and eq. 2.7. When referring, below, to experimenters freely choosing (u_1, u_2 , u_3), it should be understood that these choices must be within the constraints imposed by conditions eq. 2.6 and eq. 2.7.

Assume that, by setting appropriate experimental conditions, the experimenter can construct an inoculum with a constant total spore population size, N_c, and constant species fractions, u_1, u_2 and u_3. Assume also that, subject to the foregoing constraints eq. 2.6 and eq. 2.7, the experimenter can precisely set the values of u_1, u_2 , u_3 and N_c.

Each of the three AMF taxa can increase its rate of occupancy of the root space in the plant, denoted by θ_1, θ_2 and θ_3, for species, S_1, S_2 and S_3, respectively. The experimenter wishes to determine, or at least impose constraints on, the values of these rates in percent area increase, θ_1, θ_2 and θ_3, by performing a sequence of time series experiments wherein the linear filament extension in a root image, denoted by y(t), is measured as a function of time, t, at certain time points, $t_1, t_2, \ldots, t_K$. Here, K is the total number of experimental observation time points. Each experiment thus produces a series of observed filament extension amounts, $y(t_k)$ for $k = 1, 2, \ldots, K$, denoted by $y_1, y_2, \ldots, y_K$. That is, y_k is value of y(t) observed at time t_k, with $k = 1, 2, \ldots, K$ labeling the different observation time points. Each of these experiments is to be performed on a conidial population begun with a different combination, (u_1, u_2, u_3), of AMF inoculation fractions. For simplicity assume, however, that the values of the rates of hyphal extension, θ_1, θ_2 and θ_3, remain the same throughout all these experiments, i.e., assume that the hyphal extension rates, θ_1, θ_2 and θ_3, do not change when the experimenter changes the population composition (u_1, u_2, u_3) from one experiment to the next as in a race tube experiment[1][41]. For simplicity we will refer to θ_1, θ_2 and θ_3 as the rates of colonization success.

The extraction of any information about the success rate in root colonization, θ_1, θ_2 and θ_3, from the experimental time series data, y_k, requires, a mathematical model which treats the rates θ_1, θ_2 and θ_3, as well as the known experimental control parameters, u_1, u_2 and u_3, as input parameters. The model must then use these input parameters to provide a predicted value for each experimental observation, y_k, the hyphal extension colonized in a plant root. For a given experimental data point, y_k, we denote the corresponding value predicted by the model by f_k. Obviously, whatever the model predicts depends on the model input parameters, θ_1, θ_2 , θ_3, u_1, u_2, and u_3, that were used to make the prediction. We will therefore often write f_k as a function of these input parameters, i.e., as $f_k(\theta_1, \theta_2, \theta_3, u_1, u_2, u_3)$, to make it explicit that f_k is dependent on the assumed values of the rate parameters θ_1, θ_2 and θ_3, and on the given values of the control parameters, u_1, u_2, and u_3, set by the experimenter.

For the scenario assumed here, i.e., for a mixed population of cells from three AMF species jointly producing percent root colonization, X, at constant rates per conidia cell type, a simple mathematical model for f_k is easy to construct. Assume that the mixed cell population is established, and starts producing colonization, at time t=0, with no initial colonization length X being present at that time. Then the total percent colonization X produced by the entire AMF population in the roots, by observation time t_k, is given by:

$$X = f_k(\theta_1, \theta_2, \theta_3, u_1, u_2, u_3) = N_c u_1 \theta_1 t_k + N_c u_2 \theta_2 t_k + N_c u_3 \theta_3 t_k \quad (2.8)$$

The percent colonization X can be measured in roots by bright field microscopy[42] - [43]. To understand this linear model, which is linear in the model parameters θ, recall here that N_c is the total number of AMF in the inoculum, and hence $N_c u_1$ is the number S_1-cells in the inoculum. Hence, $N_c u_1 \theta_1$ is the rate of increase by all S_1-cells combined producing percentage root area X-contribution. Each spore produces a hypopodium by which to colonize the root cortex. Recall now that

(Rate of increase in colonization length) x (Time) = (total length colonized)

Hence, the length colonized, by all AMF S_1-cells combined, by time t_k, is $N_c u_1 \theta_1 t_k$. Likewise, the length colonized of the roots produced by all AMF S2-cells and by all S_3-cells, by time t_k, are $N_c u_2 \theta_2 t_k$ and $N_c u_3 \theta_3 t_k$, respectively. We then obtain f_k , i.e., the predicted total amount of hyphal extension colonized X produced by all cells until time t_k, by simply adding up the foregoing three X-contributions from all three AMF species. The result is eq. 2.8. The same model structure would arise if X and θ are in terms of root area occupied instead of hyphal extension.

Suppose we have performed multiple experiments, to be labeled by an "experiment index" $l = 1, 2, \ldots L$, where L is the total number of experiments. In each experiment, a different AMF species composition (u_1, u_2, u_3) was used. To distinguish these u_1, u_2 and u_3 , used in the different experiments, we therefore have to label them with the additional index l, as $u_1^{(l)}, u_2^{(l)} and u_3^{(l)}$, for l=1,2,...L. Consequently, a different time series of X-data, $y_1, y_2, \ldots, y_K$, was observed in each experiment, and we therefore also have to label the observed data, $y_1, y_2, \ldots, y_K$ with the additional index l, as $y_1^{(l)}, y_2^{(l)}, \ldots, y_K^{(l)}$, for l=1,2,...L. Also assume that each data point, $y_k^{(l)}$, has been measured with some experimental uncertainty, quantified by an experimental standard deviation $\sigma_k^{(l)}$. The χ^2-function (or by another name, the Hamiltonian) is then given by:

$$\chi^2(\theta, U) = \sum_{l=1}^{L} \sum_{k=1}^{K} \frac{1}{(\sigma_k^{(l)})^2} [y_k^{(l)} - f_k(\theta, u^{(l)})]^2 \tag{2.9}$$

To simplify and compactify the notation, we have introduced here the following abbreviations:

$$\theta := (\theta_1, \theta_2, \theta_3) \tag{2.10}$$

$$u^{(l)} := (u_1^{(l)}, u_2^{(l)}, u_3^{(l)}) \quad for \quad l = 1, 2, ...L \tag{2.11}$$

$$U := (u^{(1)}, u^{(2)}, ...u^{(L)}) = (u_1^{(1)}, u_2^{(1)}, u_3^{(1)}, u_1^{(2)}, u_2^{(2)}, u_3^{(2)}, ...u_1^{(L)}, u_2^{(L)}, u_3^{(L)}) \tag{2.12}$$

That is, θ (without subscript) is shorthand for a vector which comprises the rates of colonization θ_1, θ_2 and θ_3. The $u^{(l)}$ (without subscript) denotes the vector of the three AMF species inoculation fractions used in experiment number l, and U is the vector comprising the species fractions from all experiments combined. Note that θ does not have an (l)-superscript here because θ_1, θ_2 and θ_3 are assumed to have the same values in all experiments.

Note that $f_k(\theta, u^{(l)})$ is the model prediction of hyphal extension, from eq. 2.8, for $y_k^{(l)}$, i.e., for the k_{th} time series data point for percent root area colonized observed in the l_{th} experiment. The square of the so-called residual, on the right-hand side of eq. 2.9,

$$r_k^{(l)}(\theta, u^{(l)}) := y_k^{(l)} - f_k(\theta, u^{(l)}) \tag{2.13}$$

thus measures the deviation of the model prediction $f_k(\theta, u^{(l)})$ from the experimental observation of hyphal extension $y_k^{(l)}$: The larger $(r_k^{(l)})^2$, the worse, i.e., greater, is the deviation of the model prediction, $f_k(\theta, u^{(l)})$, from the observed data point, $y_k^{(l)}$. By taking the sum of all squared residuals, the χ^2- function in eq. 2.9 thus provides a composite measure of the overall deviation of the model predictions from the data, for all data points on hyphal length colonized combined. In the least-squares fitting approach the "best possible" choice of model parameters is then obtained by finding a parameter combination, $(\theta_1, \theta_2, \theta_3)$, which minimizes this deviation, i.e., by minimizing $\chi^2(\theta, U)$ with respect to θ_1, θ_2 and θ_3. In the following, let $\theta^{(b)} = (\theta_1^{(b)}, \theta_2^{(b)}, \theta_3^{(b)})$ denote that best possible parameter combination which minimizes $\chi^2(\theta, U)$.

Note, in passing, that the squared residuals entering into $\chi^2(\theta, U)$ in eq. 2.9 are weighted by the reciprocals of the variances, $(\sigma_k^{(l)})^2$. This means that

experimental data points with larger experimental uncertainties carry less weight and have less of an effect on the choice of the optimal, "best match" parameter combination, $\theta^{(b)}$, than data points with smaller experimental uncertainties. In that sense, $\theta^{(b)}$ can be regarded as a 'weighted compromise" between all data points, $y_k^{(l)}$: most likely, neither one of the $y_k^{(l)}$ will be perfectly matched by the model prediction $f_k(\theta, u^{(l)})$, but each data point will be matched as best as possible, in such a way that the overall mismatch, i.e., $\chi^2(\theta, U)$, is minimized. Each data point "gets a vote" in this compromise, but the vote from a very uncertain data point, having a large $(\sigma_k^{(l)})^2$, carries less weight than the vote from a less uncertain data point, having a lower $(\sigma_k^{(l)})^2$.

While there are, in principle, many different ways to define an ensemble probability distribution function having these general characteristics, an obvious, simple choice, supported by statistical theory[19], is given by:

$$Q(\theta|U) = \frac{1}{\Omega} e^{-\chi^2(\theta,U)/2} \tag{2.14}$$

The $\frac{1}{\Omega}$ -factor in eq. 2.14 is a normalization factor, chosen to ensure that the ensemble PDF integrates to a probability of 1. That is, for our model for a mixture experiment with $\theta = (\theta_1, \theta_2, \theta_3)$, the Ω is chosen such that

$$\int_{\theta_{Lo}}^{\theta_{Hi}} \int_{\theta_{Lo}}^{\theta_{Hi}} \int_{\theta_{Lo}}^{\theta_{Hi}} Q(\theta_1, \theta_2, \theta_3|U) d\theta_1 d\theta_2 d\theta_3 = 1 \tag{2.15}$$

Here, θ_{Lo} and θ_{Hi} denote, respectively, a reasonable lower and upper limit imposed on θ_1, θ_2 and θ_3. Eq. 2.14 is then to be understood to hold only when $theta_1$, θ_2 and θ_3 each falls within the interval between θ_{Lo} and θ_{Hi}; if θ_1 or θ_2 or θ_3 lies outside of this interval we set $Q(\theta|U) = 0$.

Notice that $Q(\theta|U)$ in eq. 2.14 has the desired general characteristics: For very large values of $\chi^2(\theta, U)$, the exponential function $e^{(-\chi^2(\theta,U)/2)}$, and hence $Q(\theta|U)$, becomes very small; for smaller values of $\chi^2(\theta, U)$, $Q(\theta|U)$ becomes larger. Hence, θ-choices whose model predictions agree poorly with the experimental data, will have a low probability of being drawn from $Q(\theta|U)$; θ-choices whose model predictions agree well with the experimental data, will have a higher probability of being drawn from $Q(\theta|U)$.

Given $Q(\theta|U)$, we can now calculate, for example, expectation values, variances and histograms of any observable quantity, $A(\theta)$, which the model allows us to predict as a function of θ. Specifically for the expectation value, $E[\dots]$, and variance, $\sigma^2[\dots]$, of such an "observable" $A(\theta)$, we need to calculate:

$$E[A(.)] = \int_{\theta_{Lo}}^{\theta_{Hi}} \int_{\theta_{Lo}}^{\theta_{Hi}} \int_{\theta_{Lo}}^{\theta_{Hi}} A(\theta) Q(\theta|U) d^3\theta \tag{2.16}$$

with $\theta = (\theta_1, \theta_2, \theta_3)$ and $d^3\theta = d\theta_1 d\theta_2 d\theta_3$ for short, and then

$$\sigma^2[A(.)] = E[(A(.))^2] - (E[A(.)])^2 \tag{2.17}$$

Here, $E[(A(.))^2]$ is obtained, analogous to $E[A(.)]$, with $A(\theta)$ in eq.2.16 replaced by $(A(\theta))^2$.

Within the ensemble approach, $E[A(.)]$ can serve as a prediction of a representative value of $A(\theta)$, given the experimental control parameters U and prior experimental data, $y_k^{(l)}$ for all l and all k. However, the ensemble approach also allows us to evaluate the uncertainty of that prediction, by way of $\sigma[A(.)]$. Furthermore, with similar expectation value calculations, we can also analyze in more detail the random distribution of $A(\theta)$ by way of histograms of all possible A-values. This would tell us, for example, if the values of $A(\theta)$ have a uni- or a multi-modal distribution, for random θs drawn from the ensemble $Q(\theta)$.

These are just a few examples of what kinds of data analyses and model predictions the ensemble approach itself allows us to implement. In the context of the MINE approach of experiment design, we will have to evaluate certain correlations between pairs of observables, $A(\theta)$ and $B(\theta)$, say. This will require the calculation of expectation values of the general form $E[A(.)B(.)]$, with $A(\theta)$ in eq.2.16 replaced by the product $A(\theta)B(\theta)$.

The evaluation of all the foregoing expectation values usually requires numerical techniques to carry out the θ-integrations as in eq. 2.16. In general, the θ-space is very high-dimensional, far greater than the θ-dimension of $dim(\theta) = 3$ in our simple model here. Markov chain Monte Carlo methods are then the only approach available to perform the required expectation value calculations efficiently for omics experiments and field studies[21]. The basics of the Markov chain Monte Carlo approach are illustrated in more detail in section II of this article. In Figure 2.4 is a simulation of the mixture experiment with the application of the ensemble method to the simulated data. The ensemble method converges quite well to the true colonization rates θ.

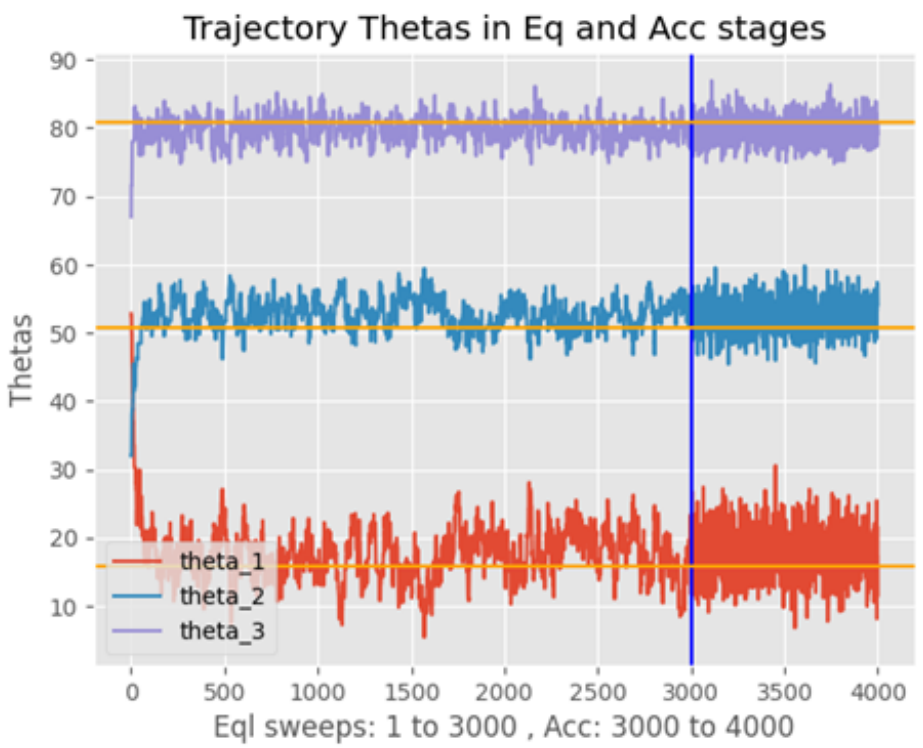

Figure 2.4: An ensemble method to identify the mixture experiment's hyphal extension rates θ is carried out on simulated data from the mixture experiment, and MINE is used to choose the next mixture experiment with inoculation proportions u_1 and u_2. The figure illustrates the ensemble method on simulated data for a mixture experiment of AMF colonizers. The orange lines are the true colonization rates θ. In the Monte Carlo experiment the estimated rates are plotted as a function of sweep, a visit on average once to each of the three rates θ. In the first 3000 sweeps the Monte Carlo experiment is equilibrated to get in the neighborhood of parameters θ that fit the simulated data. In the accumulation phase (last 1000 sweeps) the estimates of θ are accumulated to form the ensemble estimate.

Assume L prior experiments have already been performed, with experimental control parameter vectors $u^{(l)}$, as defined in eq. 2.11, and observed values $y_k^{(l)}$, for l=1,2,...L and k=1,2,...K. The experimental data points, $y_k^{(l)}$, combined with the corresponding model predictions, $f_k(\theta, u^{(l)})$, from eq.2.8, define an ensemble PDF, $Q(\theta|U)$, via eqs. 2.9 and 2.16. The $Q(\theta|U)$, in turn, will determine V(U), the predicted uncertain volume of the observables, to be measured in the new experiment(s), as follows:

As a simplest case, assume that we want to design just one new experiment, with a new experimental control parameter vector $u = (u_1, u_2, u_3)$. The MINE objective is then to choose to input inoculation proportions u so as

to maximize the information content of the new experiment about the rates of production of colonization by hyphal extension (X), by maximizing the predicted uncertainty volume of the observables to be measured. In our simple mixture experiment example, there are K such observables in any experiment: the hyphal extension X-amounts to be measured at times t_k, for k=1,2,...K. The predicted values for these observed hyphal extensions X are then $f_k(\theta, u)$, as defined by the model eq. 2.8, for given θ and u. These predicted values for these K observations can be thought of as the components of a vector in a K-dimensional space, the so-called observation space (Figure 2.3). For a given θ and u, this vector of predicted observations, is in the following denoted by $f(\theta, u)$, and given by

$$f(\theta, u) := (f_1(\theta, u), f_2(\theta, u), ..., f_k(\theta, u)) \tag{2.18}$$

We can now use the ensemble PDF, $Q(\theta, U)$, to define, in some way, a volume of likely θs in θ-space. If we let θ sweep over that finite volume then, by eq. 2.18, $f(\theta, u)$ will sweep over some corresponding finite volume (or hypersurface) in the observation space: the uncertainty volume of the predicted observation vector, to be denoted by V(u), for given u, and illustrated in Figure 3. There is of course no precise prescription of how to define a volume of likely-θ in θ-space, or a corresponding uncertainty volume, V(u), in observation space. That definition is not unique: it requires some arbitrary, but reasonable choices to be made. In the following, two specific possible choices for V(u) will be discussed. In both of these choices, one actually defines V(u) directly in the observation space, in terms of correlations and variances of the observable prediction vector, $f(\theta, u)$,without first defining an underlying volume of likely θs in θ-space.

2.3.1 MINE by Covariance Ellipsoid Volume

In the covariance matrix approach, we define V(u) in terms of the uncertainty ellipsoid, constructed from the covariances of the K observable predictions, $f_1(\theta, u), f_2(\theta, u), \ldots f_K(\theta, u)$, subject to the ensemble PDF $Q(\theta, U)$. Let $D_{kj}(u)$ denote those covariance matrix elements, i.e., for k,j=1,2,...K, let

$$D_{kj}(u) := E[f_k(., u)f_j(., u)] - E[f_k(., u)]E[f_j(., u)] \tag{2.19}$$

with expectation values E[...] defined as in eq.2.16. On general mathematical grounds, the corresponding K×K covariance matrix, D(u), is symmetric and positive semi-definite. Therefore, D has K real, non-negative eigenvalues, λ_ν;

and it has an orthonormal basis of corresponding K-dimensional eigenvectors, $e^{(\nu)}$, with $\nu = 1, 2, \ldots K$. That is, for $k, j, \nu, \mu = 1, 2, \ldots K$, we have:

$$\sum_{j=1}^{K} D_{kj}(u) e_j^{(v)} = \lambda_\nu e_k^{(\nu)} \tag{2.20}$$

$$\lambda_\nu \geq 0 \tag{2.21}$$

$$\sum_{j=1}^{K} e_j^{(\nu)} e_j^{(\mu)} = \delta_{\nu,\mu} \tag{2.22}$$

$$\sum_{\nu=1}^{K} e_k^{(\nu)} e_j^{(\nu)} = \delta_{k,j} \tag{2.23}$$

The eigenvalues and eigen vectors are of course dependent upon u, but for notational simplicity we have suppressed that functional dependence, i.e., $\lambda_\nu(u)$ and $e_j^{(\nu)}(u)$, in eqs. 2.19-2.23. By eq. 2.22, the eigenvectors, $e^{(\nu)}$, are orthogonal, i.e., pairwise perpendicular to each other. The eigenvalues, λ_ν, are the variances of the predicted observation vector, $f(\theta, u)$, along the corresponding eigenvector directions. That is, if we take the projection of the vector $f(\theta, u)$ onto $e^{(\nu)}(u)$, i.e., let $p^{(\nu)}(\theta, u)$ denote that projection, with

$$p^{(\nu)}(\theta, u) := e^{(\nu)}(u) f(\theta, u) = \sum_{k=1}^{K} e_k^{(\nu)}(u) f_k(\theta, u) \tag{2.24}$$

then λ_ν is the variance of that projected $f(\theta, u)$–vector:

$$\sigma^2[p^{(\nu)}(., u)] = \lambda_\nu \tag{2.25}$$

where $\sigma^2[\ldots]$ is defined as in eq.2.17. The eigenvalues and eigenvectors of D(u) define the so-called "covariance ellipsoid" or "error ellipsoid" of the predicted observation vector, $f(\theta, u)$, in the K-dimensional observation space: The eigenvectors, $e^{(\nu)}$, can be thought of as the orientations of the principal axes of the ellipsoid; the standard deviations of the projections $p^{(\nu)}(\theta, u)$, i.e., $\sigma[p^{(\nu)}(., u)] = \sqrt{\lambda_\nu}$, are the lengths of the principal semi-axes along the $e^{(\nu)}$-direction. This ellipsoid serves as our uncertainty volume, and V(u) is given by the product of the semi-axis lengths,

$$V(u) = C_K \sqrt{\lambda_1(u)\lambda_2(u)...\lambda_K(u)} \tag{2.26}$$

where C_K is an unimportant geometrical prefactor,

$$C_K = \frac{2\pi^{n/2}}{n\Gamma(\frac{n}{2})} \tag{2.27}$$

with $\Gamma(x)$ denoting Euler's gamma function. Eq. 2.26 can also be written in terms of the determinant of the D-matrix:

$$V(u) = C_K\sqrt{det(D(u))} \tag{2.28}$$

2.3.2 MINE by Correlation Ellipsoid Volume

In the correlation matrix approach, we define V(u) in terms of an uncertainty ellipsoid constructed from the Pearson correlations of the K observable predictions, $f_1(\theta, u), f_2(\theta, u), \ldots f_K(\theta, u)$, subject to the ensemble PDF $Q(\theta, U)$. The Pearson correlation matrix elements, denoted by $E_{kj}(u)$, are related to the covariance matrix elements, $D_{kj}(u)$, from eq.2.19, by

$$E_{kj}(u) := \frac{D_{kj}(u)}{\sqrt{D_{kk}(u)D_{jj}(u)}} \tag{2.29}$$

Note that $E_{kj}(u)$ can also be written as the covariance matrix of the predicted observations, $f_k(\theta, u)$, normalized by their standard deviations:

$$E_{kj}(u) := E[g_k(., u)g_j(., u)] - E[g_k(., u)]E[g_j(., u)] \tag{2.30}$$

where

$$g_k(\theta, u) := \frac{1}{\sigma[f_k(., u)]} f_k(\theta, u) \tag{2.31}$$

Therefore, the correlation matrix E has the same mathematical properties of symmetry and semi-positivity as the covariance matrix D. Analogous, to the covariance ellipsoid constructed from D, we can therefore construct a correlation ellipsoid from the eigenvalues and orthonormal eigenvectors of the correlation matrix E. Using the volume of the correlation ellipsoid as the uncertainty volume of the predicted observables, we then have, analogous to eq.2.26,

$$V(u) = C_K\sqrt{K_1(u)K_2(u)...K_K(u)} \tag{2.32}$$

where $K_1(u), K_2(u), \ldots K_K(u)$ are the eigenvalues of the correlation matrix E. Analogous to eq.2.28 we can also write this as

$$V(u) = C_K\sqrt{det(E(u))} \tag{2.33}$$

The surface of the MINE criterion in a contour plot is shown for the next mixture experiment (Figure 2.5). The MINE experiment involves using ~ 0.25 of AMF1 in the inoculum and ~ 0.40 of AMF2 in the inoculum to characterize the rates of hyphal extension in the next experiment. While this simple linear model captures some features of the competition between AMF, more elaborate nonlinear models will be considered in Section IX that describe the competition between AMF. Also, these mixture experiments only reveal part of the story – the success of plant root colonization depends on the genotype of the plant colonized as considered in sections VII and VIII. As a final note some theorems about properties of MINE have been established for the class of linear models, such as the mixture experiments[19].

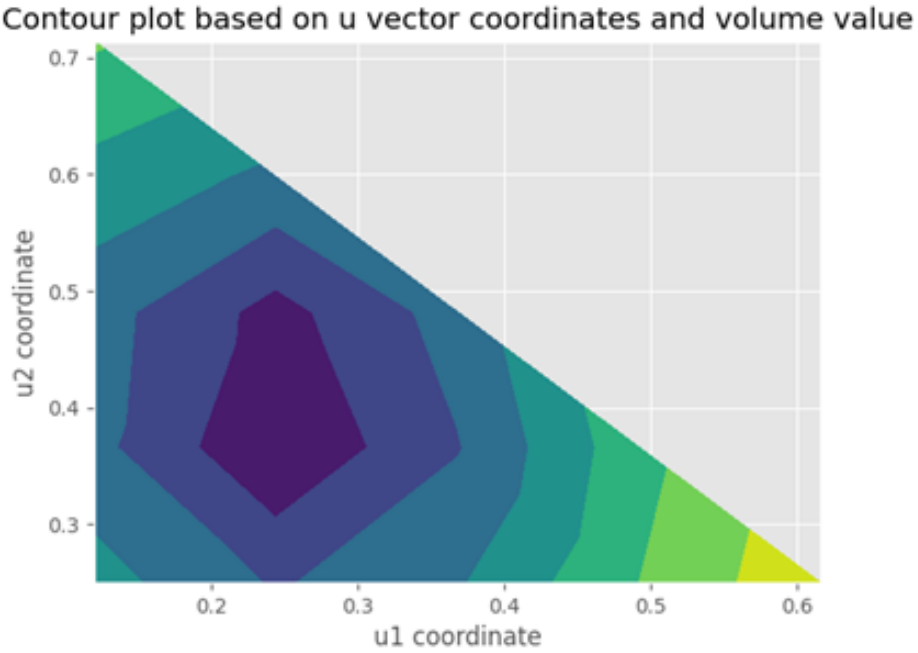

Figure 2.5: An ensemble method to identify the mixture experiment's hyphal extension rates θ is carried out on simulated data from the mixture experiment, and MINE is used to choose the next mixture experiment with inoculation proportions u_1 and u_2. The figure presents the next MINE mixture experiment recommended. The contour plot is of the MINE criterion det(E) as a function of the mixture inoculum proportions u_1 and u_2.

2.4 Application of MINE to RNA profiling experiments coupled to genetic networks

One of the earliest developments in functional genomics was the use of RNA profiling to characterize the response of yeast to a diauxic shift from anaerobic to aerobic conditions, a process that has been of interest for 1,000s of years in the production of fermented products[44]. MINE was developed specifically for this kind of transcriptomics problem and used to close the loop in the computing life cycle (Figure 2.6) proposed by Hood and Abersold[15]. Transcriptomic experiments have a limited number of time points n, but have many 1,000s of genes (and hence parameters (p)) to be identified in the process[1]. While both MINE criteria using the Covariance by Ellipsoid Volume and Correlation by Ellipsoid Volume, only the Correlation by Ellipsoid Volume was reported in the end in designing the experiments[1].

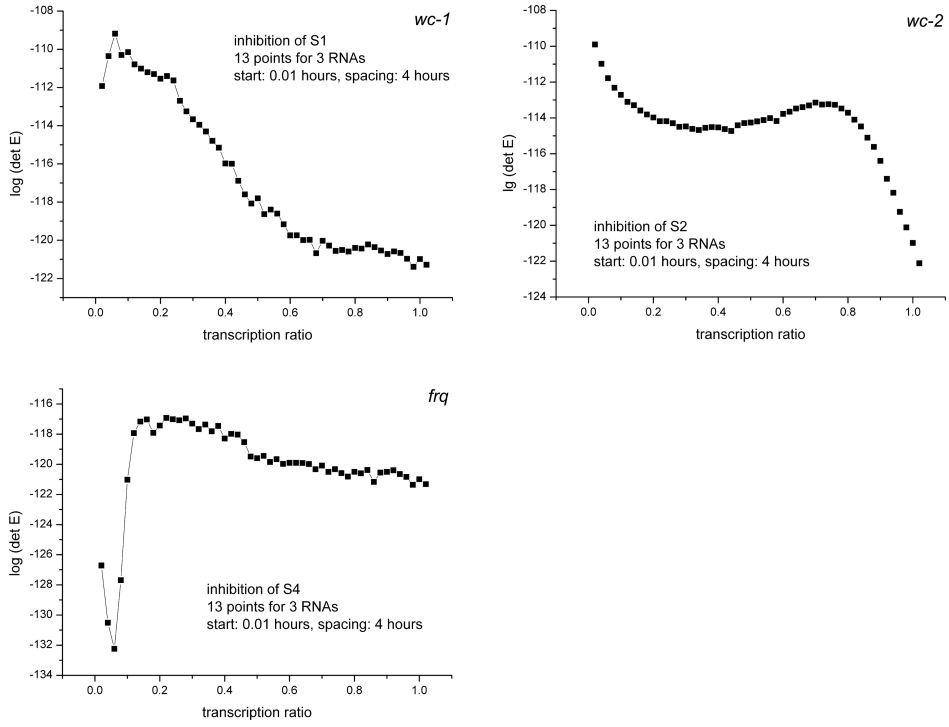

Figure 2.6: The MINE experiment is a 90 percent knockdown of the wc-1 gene. The MINE criterion displayed is the correlation ellipsoid volume det(E(U)), which is graphed as a function of the remaining activity of the three clock mechanism genes. The predictions F are of the log base 10 concentrations over time of frq, wc-1, and wc-2 mRNAs over time from the RNA profiling experiments. The mRNA levels were measured at 14 time points over an 8 hour window. The drawing is taken from [1].

The design problem was very simple. Transcriptomics was to be used to explore the mechanism of the biological clock in one of the most well studied model systems [45], the filamentous fungus, *Neurospora crassa*. Three major components of the clock mechanism were: (1) frequency (frq), the gene encoding the oscillator of the system and a negative regulator; (2) white-collar-1 (wc-1), the gene encoding the light response element and a positive transcriptional activator for the system; (3) and white-collar -2(wc-2), a second positive transcriptional activator for the system. Together wc-1 and wc-2 encode WC-1 and WC-2 proteins that act as positive elements in the clock through the dimeric complex WCC=WC-1/WC-2, while frq encodes a protein FRQ, which acts as the negative regulator for the system. The FRQ protein provides negative feedback to wc-1 and wc-2. The beauty of this system is all three of these elements appear in single copy in the *Neurospora* genome, but they have homologs in fly and mammalian systems[46][47].

What genetic experiments could be done to discover the most new information about the clock mechanism given that an ensemble of models had already been successfully fitted to the data in the literature[25] Different experiments provide more or less new information about a system. As an illustrative example, suppose $y = 1 - (x - 1)^2$, but this relation is unknown to the experimenter. For a given known experimental condition x, the measurement y might be taken. If the experimenter only chose to make measurements y for conditions x from 0 to 1, the experimenter might conclude there is an increasing relation between y and x. If the experimenter chose instead to measure y for x from 1 to 2, the experimenter might conclude there is a decreasing relation between y and x. Only when observations are taken over an interval from 0 to 2 would it become clearer there is a quadratic relation. Each of these three experiments provide different information about the system.

Exactly this problem has arisen in studying the quadratic relation between AMF fungal biomass or equivalently benefit to plant (y) and the level of phosphorous (x) in the soil[48]. Most measurements of phosphorous were done on the high end of phosphorous (high x) to the right of the bump in the unsuspected quadratic relation between fungal biomass (y) and phosphorous (x)[49]. The conventional wisdom in studies in the northern hemisphere was increasing phosphorous would decrease fungal biomass because most measurements of fungal biomass were done under high phosphorus conditions. The problem was in the tropics scientists were seeing the opposite relation where increasing phosphorous increased fungal biomass. Treseder and Allen[48] hypothesized that in a low phosphorous environment adding phosphorous could help both the plant and fungus and that there might be an optimum in phosphorous

(x). Not until the work of Propster and Johnson on the Serengeti[49] was it clarified what was happening at the low and high ends of phosphorous (x) to fungal biomass (y). Thus, different experiments led to different amounts of new information about an ecosystem.

Now consider a series of RNA profiling experiments which were conducted, guided by MINE to choose an informative sequence of experiments. The last in a series of three adaptive experiments guided by MINE involved a choice of whether or not to do a knockdown or overexpression experiment on: (1) frq; (2) wc-1; (3) or wc-2. The conventional wisdom was to mutate the oscillator gene frq. Each of these three experiments cost about 250,000 dollars, thus the need for a rational choice of design.

The first step in the MINE application is to make predictions for various mutations in the clock mechanism genes using an available ensemble. The RNA profiles of all 11,000 genes were measured at each of 14 timepoints. Unlike previous models so far considered, the clock model is a nonlinear model (in the parameters describing the model), which specifies a genetic network of nonlinear ordinary differential equations describing the time course of the genes, their cognate RNAs and proteins[25]. The model ensemble was used to predict RNA profiles of frq, wc-1, and wc-2 under different possible experiments and their correlations as shown in Figure 2.2.

With the correlation matrix in hand for the predictions, the MINE criterion based on the correlation volume ellipsoid was calculated as a function of the degree of knockdown of the three clock genes (Figure 2.6). Similar calculations were done for mutants involving overexpression, but were not competitive relative to the experiments considered in the MINE calculations in Figure 2.6 with respect to MINE score (and not shown). The result was surprising. A knockdown of wc-1 was performed as the MINE experiment and used to identify 2,323 genes responding to the knockdown12. The second surprise from this model-guided discovery process was that ribosome biogenesis was under clock control. This was later confirmed in mammalian systems[50].

A better way to do this MINE calculation would have been to use all of the transcriptomic data on 11,000 N. crassa genes instead of just the three clock genes driving the system. The ensemble methods now exist for the whole genome-scale network with all of its 1,000s of genes and ensembles now exist for the entire clock network[26][27][51]. Using Graphical Processing Units (GPUs) ensemble methods, such as MINE, can now be implemented on a genomic scale with unknown network structure[26][51].

2.5 Application of MINE to QTL mapping using RILs for AMF/Sorghum project

A simple example of the use of MINE arises upon consideration of a hybrid cross between *Sorghum bicolor* and *Sorghum propinquum*[52]. Several quantitative traits are of interest including plant biomass. One of the main methods of mapping Quantitative Trait Loci (QTLs) is interval mapping[53]. For simplicity, consider a dihybrid cross between *S. bicolor* and *S. propinquum*. A QTL of interest is heterozygous for alleles Q1 and Q2. There are neighboring markers with alleles M1 and M2 and N1 and N2. The alleles Q1, M1, and N1 are characteristic of *S. bicolor* parent (P1); the alleles Q2, M2, and N2 are characteristic of *S. propinquum* parent (P2). Classical Mendelian genetics allows us to write down a model for the inheritance of the loci in the F2 progeny of a dihybrid cross (Table 2.1). In this table there are 3 possible orderings of the marker and QTL locus represented by 3 separate columns in Table 2.1. While the QTL genotype is unknown without a known position, the marker genotypes are observable at their known positions in the genome. The markers sort the F2 offspring into 9 marker classes labeled 1-9 in Table 2.1. The recombination distance as measured by the recombination fraction between each of the three loci is denoted by rMQ, rNQ, and rMN, the last of which are known by mapping. Once these recombination distances are known, then the position and hence the identity of the QTL is known. The map is assumed dense enough so that there is only at most one recombination event between these loci. Within each marker class there are three genotypic classes of the QTL mixed together. Knowing the QTL genotype leads to a specification of the quantitative trait developed by R. A. Fisher[54][55]. If the QTL genotype is Q_1/Q_1 characteristic of the S. bicolor parent (P1), then the mean of biomass is μ_1. If the QTL genotype is Q_2/Q_2 characterisitic of the S. propinquum parent (P2), then the mean of biomass is μ_2. If the QTL genotype is Q_1/Q_2 of the F_1 hybrid, then the mean biomass is μ_{12}. The model assumptions are that each genotype produces a normally distributed trait with appropriate genotypic mean and the same variance σ^2.

The columns sum to one when multiplied by the factor in column 3, which represents the chance of independent segregation of offspring of the selfing F_1 cross. The final model specification is then a product of a 3-component normal mixture density for each marker class. The mixing distribution for each marker class can also be computed from Table 2.1 by calculating the conditional probability of a particular QTL genotype within a marker class given that the individual is drawn from that marker class. As an example, for marker class 1, the conditional probability of Q_1/Q_1 is:

Table 2.1: Mendelian model for quantitative trait with one QTL and two adjacent markers M and N.

Marker Class	Genotype	Factor	QMN	MQN*	MNQ
1	$M_1N_1Q_1/M_1N_1Q_1$	1/4	$(1-r_{MQ})^2(1-r_{MN})^2$	$(1-r_{MQ})^2(1-r_{NQ})^2$	$(1-r_{MN})^2(1-r_{NQ})^2$
$M_1M_1N_1N_1$	$M_1N_1Q_1/M_1N_1Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$(1-r_{MN})^2(1-r_{NQ})r_{NQ}$
	$M_1N_1Q_2/M_1N_1Q_2$	1/4	$r_{MQ}^2(1-r_{MN})^2$	$r_{MQ}^2r_{NQ}^2$	$(1-r_{MN})^2r_{NQ}^2$
2	$M_1N_1Q_1/M_1N_2Q_1$	1/2	$(1-r_{MQ})^2(1-r_{MN})r_{MN}$	$(1-r_{MQ})^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
$M_1M_1N_1N_2$	$M_1N_1Q_1/M_1N_2Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})^2$	$(1-r_{MN})r_{MN}(1-r_{NQ})^2$
	$M_1N_1Q_2/M_1N_2Q_1$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}r_{NQ}^2$	$(1-r_{MN})r_{MN}r_{NQ}^2$
	$M_1N_1Q_2/M_1N_2Q_2$	1/2	$r_{MQ}^2(1-r_{MN})r_{MN}$	$r_{MQ}^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
3	$M_1N_2Q_1/M_1N_2Q_1$	1/4	$(1-r_{MQ})^2r_{MN}^2$	$(1-r_{MQ})^2r_{NQ}^2$	$r_{MN}^2r_{NQ}^2$
$M_1M_1N_2N_2$	$M_1N_2Q_1/M_1N_2Q_2$	1/2	$(1-r_{MQ})r_{MQ}r_{MN}^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$r_{MN}^2(1-r_{NQ})r_{NQ}$
	$M_1N_2Q_2/M_1N_2Q_2$	1/4	$r_{MQ}^2r_{MN}^2$	$r_{MQ}^2(1-r_{NQ})^2$	$r_{MN}^2(1-r_{NQ})^2$
4	$M_1N_1Q_1/M_2N_1Q_1$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})^2$	$(1-r_{MN})r_{MN}(1-r_{NQ})^2$
$M_1M_2N_1N_1$	$M_1N_1Q_1/M_2N_1Q_2$	1/2	$(1-r_{MQ})^2(1-r_{MN})r_{MN}$	$(1-r_{MQ})^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
	$M_1N_1Q_2/M_2N_1Q_1$	1/2	$r_{MQ}^2(1-r_{MN})r_{MN}$	$r_{MQ}^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
	$M_1N_1Q_2/M_2N_1Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}r_{NQ}^2$	$(1-r_{MN})r_{MN}r_{NQ}^2$
5	$M_1N_1Q_1/M_2N_2Q_1$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$(1-r_{MN})^2(1-r_{NQ})r_{NQ}$
	$M_1M_2N_1N_2M_1N_2Q_1/M_2N_1Q_1$	1/2	$(1-r_{MQ})r_{MQ}r_{MN}^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$r_{MN}^2(1-r_{NQ})r_{NQ}$
	$M_1N_1Q_1/M_2N_2Q_2$	1/2	$(1-r_{MQ})^2(1-r_{MN})^2$	$(1-r_{MQ})^2(1-r_{NQ})^2$	$(1-r_{MN})^2(1-r_{NQ})^2$
	$M_1N_1Q_2/M_2N_2Q_1$	1/2	$r_{MQ}^2(1-r_{MN})^2$	$r_{MQ}^2r_{NQ}^2$	$(1-r_{MN})^2r_{NQ}^2$
	$M_1N_2Q_1/M_2N_1Q_2$	1/2	$(1-r_{MQ})^2r_{MN}^2$	$(1-r_{MQ})^2r_{NQ}^2$	$r_{MN}^2r_{NQ}^2$
	$M_1N_2Q_2/M_2N_1Q_1$	1/2	$r_{MQ}^2r_{MN}^2$	$r_{MQ}^2(1-r_{NQ})^2$	$r_{MN}^2(1-r_{NQ})^2$
	$M_1N_1Q_2/M_2N_2Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$(1-r_{MN})^2(1-r_{NQ})r_{NQ}$
	$M_1N_2Q_2/M_2N_1Q_2$	1/2	$(1-r_{MQ})r_{MQ}r_{MN}^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$r_{MN}^2(1-r_{NQ})r_{NQ}$
6	$M_1N_2Q_1/M_2N_2Q_1$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}r_{NQ}^2$	$(1-r_{MN})r_{MN}r_{NQ}^2$
$M_1M_2N_2N_2$	$M_1N_2Q_1/M_2N_2Q_2$	1/2	$r_{MQ}^2(1-r_{MN})r_{MN}$	$(1-r_{MQ})^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
	$M_1N_2Q_2/M_2N_2Q_1$	1/2	$(1-r_{MQ})^2(1-r_{MN})r_{MN}$	$r_{MQ}^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
	$M_1N_2Q_2/M_2N_2Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})^2$	$(1-r_{MN})r_{MN}(1-r_{NQ})^2$
7	$M_2N_1Q_1/M_2N_1Q_1$	1/4	$r_{MQ}^2r_{MN}^2$	$r_{MQ}^2(1-r_{NQ})^2$	$r_{MN}^2(1-r_{NQ})^2$
$M_2M_2N_1N_1$	$M_2N_1Q_1/M_2N_1Q_2$	1/2	$(1-r_{MQ})r_{MQ}r_{MN}^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$r_{MN}^2(1-r_{NQ})r_{NQ}$
	$M_2N_1Q_2/M_2N_1Q_2$	1/4	$(1-r_{MQ})^2r_{MN}^2$	$(1-r_{MQ})^2r_{NQ}^2$	$r_{MN}^2r_{NQ}^2$
8	$M_2N_2Q_1/M_2N_1Q_1$	1/2	$r_{MQ}^2(1-r_{MN})r_{MN}$	$r_{MQ}^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
$M_2M_2N_1N_2$	$M_2N_2Q_1/M_2N_1Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}r_{NQ}^2$	$(1-r_{MN})r_{MN}r_{NQ}^2$
	$M_2N_2Q_2/M_2N_1Q_1$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})r_{MN}$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})^2$	$(1-r_{MN})r_{MN}(1-r_{NQ})^2$
	$M_2N_2Q_2/M_2N_1Q_2$	1/2	$(1-r_{MQ})^2(1-r_{MN})r_{MN}$	$(1-r_{MQ})^2(1-r_{NQ})r_{NQ}$	$(1-r_{MN})r_{MN}(1-r_{NQ})r_{NQ}$
9	$M_2N_2Q_1/M_2N_2Q_1$	1/4	$r_{MQ}^2(1-r_{MN})^2$	$r_{MQ}^2r_{NQ}^2$	$(1-r_{MN})^2r_{NQ}^2$
$M_2M_2N_2N_2$	$M_2N_2Q_1/M_2N_2Q_2$	1/2	$(1-r_{MQ})r_{MQ}(1-r_{MN})^2$	$(1-r_{MQ})r_{MQ}(1-r_{NQ})r_{NQ}$	$(1-r_{MN})^2(1-r_{NQ})r_{NQ}$
	$M_2N_2Q_2/M_2N_2Q_2$	1/4	$(1-r_{MQ})^2(1-r_{MN})^2$	$(1-r_{MQ})^2(1-r_{NQ})^2$	$(1-r_{MN})^2(1-r_{NQ})^2$

$$(1-r_{MQ})^2(1-r_{MN})^2/[(1-r_{MQ})^2(1-r_{MN})^2+(1-rMQ)r_{MQ}(1-r_{MQ})r_{NQ}+r_{MQ}^2r_{NQ}^2] \tag{2.34}$$

This specifies the model for each marker interval. We then simply take the product over all marker intervals to obtain the model specification and hence the Hamiltonian for the problem. The MINE problem is to choose the accessions to best inform how genes control a particular complex trait, such as height or percent colonization by AMF.

2.6 Application of MINE to GWAS field studies for AMF/Sorghum project

Consider a GWAS study composed by 343 plant accessions from Sorghum bicolor BAP Panel[2]; its objective is to understand relevant genes for biomass and AMF colonization. There was a previous study where *S. bicolor* genetic information from different sources was compiled and put together into variant

call files[8], and as a result there are 232,303 SNPs available. To characterize biomass, dry weight was measured. Neural networks were used to measure AMF colonization in roots [56]. Initial dry weight data was taken from the panel to estimate a model.

The GWAS study is running for 5 years, and in each year MINE is used to select the most informative 79 accessions; these accessions will be planted in a randomized block design, 3 blocks will be set up, each block will have 12 replicates of each accession. The relation between AMF percent colonization and AMF community composition will be addressed, as well as SNPs in the plant host. The effect of AMF on plant health is also studied by the relation of plant biomass and SNPs.

A few features will be measured during the MINE field experiment: plant genotype, plant eQTLs, the microbiome, Phosphorus (P), Nitrogen (N), time of harvest, and other variables relevant to plant health such as biomass (Figure 2.7). These variables are combined into a mixed linear model to predict biomass. The resulting model is a special case of a structural equation model [57].; these structural equation models have been successfully used in field studies of AMF [17][30][58]. The difference between structural equation models and regression models in Section 3 is that the independent variables, such as eQTLs and accessions, are random and not fixed, since accessions come from a worldwide population of sorghum[12]. The standard model for GWAS is the mixed linear model, representing a special case of the structural equation model, in which some variables are random with mean zero [59]. A mixed linear model is presented below, and will be used in an adaptive GWAS experiment underway at Wellbrook farm, Athens, GA.

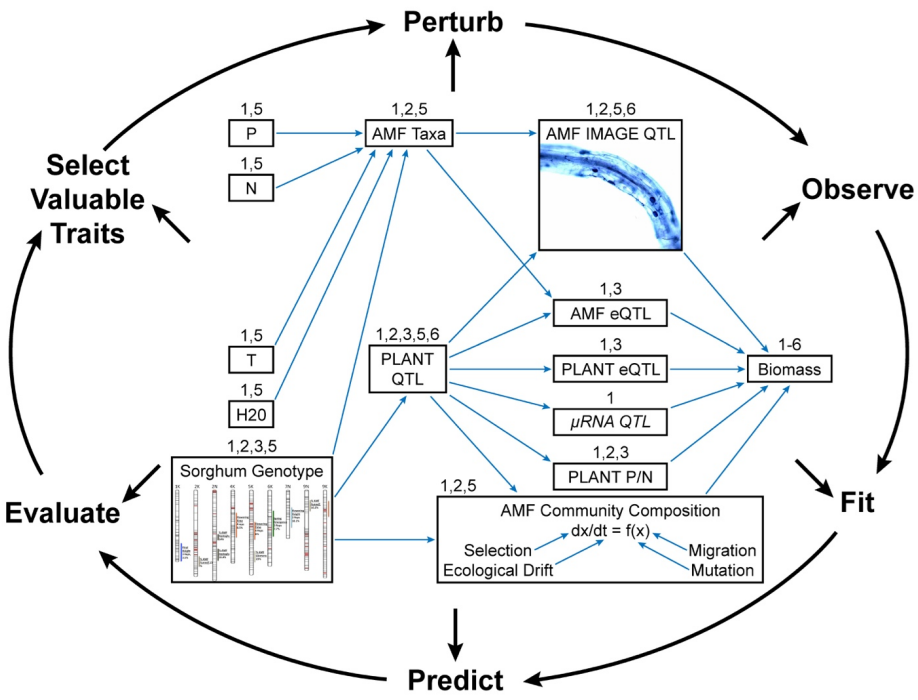

Figure 2.7: A sequence of MINE experiments are to be used in a 5 year GWAS experiment to examine the relation between biomass and SNPs in Sorghum bicolor using the BAP accessions[2]. MINE is used to select the BAP accessions to be used in each year in order to map AMF colonization and biomass to the sorghum genetic map in a GWAS study. Multi-scale structural equation model (SEM) for the project (center boxes and arrows). Lotka-Volterra community models are nested within the SEM and predict associations that affect biomass. The dependent variable is biomass, and the arrows in the diagram denote causal relationships between independent variables in the SEM. The labels on each box index the subproject(s) involved in characterizing the properties of the plant-AMF-microbiome-abiotic environment interaction depicted in that box. In this model, sorghum genotype is the primary independent variable that correlates with the remaining variables. This conceptual model will evolve continuously using the model guided discovery process of maximally informative next experiment (MINE; outer ring)[1].

Two sets of inputs were collected in year 1 at Wellbrook Farm, Athens, GA: 1) fixed variables in the design matrix X, such as block number, harvest time of each plant, N level, P level, and the fixed effect of each BAP accession; 2) random effects Z for the genotype of each BAP accession. The additive genetic variation in plant genotype was captured by binning the number of alleles in a given genomic region different from the reference genome using the sum method [60]. The number of SNPs in each chromosomal region was adjusted

to ensure that at least each genomic regions was 50 kb in size. The result was that the number of genomic regions was 2748 in the sorghum genome, each containing an average of 10-12 genes. The number of such alleles in a region is treated as a continuous random variable with mean 0 and variance component σ_i^2 for the i^{th} accession and summed over regions to obtain a random effect for an accession. The fact that each random effect is the sum of 2748 small random effects of chromosomal regions makes it plausible that the random effect of accession is normally distributed. The fixed effects of block number and BAP genotype are denoted by the vector β and the random effects, by u.

These fixed and random effects are used to predict some measure of biomass, such as log dry weight. There are a total of $n \sim 606$ plants in the field, which is lower than p = 2748 fixed effects + 79 variance components. At harvest time it was feasible to dig up at least 2 plants per row for measurement in a field setup as a randomized block design with 3 blocks and 12 replicate plants in a plot (i.e., row). The measurements to be predicted are summarized in the n x 1 vector Y.

The error in biomass is also considered in the model, it is denoted by ϵ_i for the i^{th} plant in the experiment. The error measurements are summarized in a n x 1 column vector Y.

In year 1 of the adaptive GWAS experiment, no block effect was found, and N and P applied were not varied. A randomized block design was used to plant the 79 BAP accessions selected; a total of three blocks were set up with 79 plots in each block, 1 genotype per plot (i.e., a row), and 12 replicates per plot. The mixed linear model for the experiment was the following:

$$Y = X\beta + Zu + \epsilon \tag{2.35}$$

where Y is a n x 1 vector of observations on biomass. The X variable is the design matrix of size n x p, p represents the number of bins representing the chromosomal regions. β represents the parameters to be estimated, its size is p x 1, each parameter is a fixed effect of the corresponding chromosomal region in X. The matrix Z of size n x r represents the number of alleles in each accession on n plants in the field, for this experiment r = 79. u is a vector of size r x 1 representing the random effects of each accession on biomass. The errors in the dependent variable Y are compiled in the vector ϵ of size n x 1. Three assumptions of this model are: 1) the random effects u are independent of the biomass errors ϵ ; 2) the errors ϵ are normally distributed with mean 0 and variance σ^2; 3) the random effects u are normally distributed with mean 0 and variance $\sigma^2_{j(i)}$ plant I with accession j(i). That is, the assumptions are that the random effects u and errors ϵ are independent and normally distributed with mean 0 and variance-covariance matrix $I\sigma_i^2$ and $I\sigma^2$, respectively.

Under this model the prediction of biomass is:

$$E(Y) = X\beta \tag{2.36}$$

The variance components and heritability are used to calculate the variance-covariance matrix V of the biomass measurements Y:

$$V = VAR(Y) = \sum_{i=1}^{n} Z_i' Z_i \sigma^2_{j(i)} + \sigma^2 I = \sum_{i=1}^{n} X_i' X_i \sigma^2_{j(i)} + \sigma^2 I \tag{2.37}$$

where Z_i and X_i are the i^{th} row vectors of Z and X, respectively. Each observation Y_i describes a corresponding X_i row vector. Each term $X_i' X_i \sigma_i^2$ is an n x n block. The variance-covariance matrix is diagonal with p blocks each with the same diagonal elements $\sigma^2_{j(i)}$. The index j(i) is a lookup that returns the variance component of the i^{th} observation as determined by accession j. Plant i has an assigned accession j.

For the mixed linear model, the ensemble Q can be written down as a multivariate normal with the θ-vector consisting of the fixed effects β and the variance components:

$$Q(\beta) = \frac{e^{-1/2(Y-X\beta)' V^{-1}(Y-X\beta)}}{(2\pi)^{n/2}|V|^{1/2}} \tag{2.38}$$

No fertilizers were used in the first year of MINE field experiment in 2021 at Wellbrook Farm, Athens, GA. The model only considered fixed effects with the number of alleles in a bin (chromosomal region) as the set of independent variables using the sum method[60]. The problem is different from the classical design[19] because of the "big p, little n" problem characteristic of genomic experiments[19]. The variance components were estimated from the replicates on each accession.

The ensemble method was used to estimate the models from the published dry weight data in 3 years from 2013-2015 collected in Florence, South Carolina[2] to make predictions in the use of MINE. Typically in a omics experiment there is prior published data available, and this should be used when available [1] to initialize the MINE sequence.

A total of 100,000 equilibration sweeps were done, and then 1000 sets of model parameters were accumulated, each model parameter was separated by 100 decorrelation sweeps. The chi-squared per data point was 6.12 with n = 606 dry weight measurements. As a control the ensemble run was repeated with the only change being 1000 decorrelation sweeps.

MINE was then applied to the model estimated from the data from Florence, South Carolina to select 80 accessions for use in 2022 for planting at Wellbrook Farm. The result is shown graphically (Figure 2.8). The MINE criterion was optimized by evaluating det(D) on all $\binom{300}{80}$ tuples drawn from 300 BAP accessions at USDA GRIN in Griffin, GA. Details of calculating det(D) follow the directions in the introduction to MINE in section III.

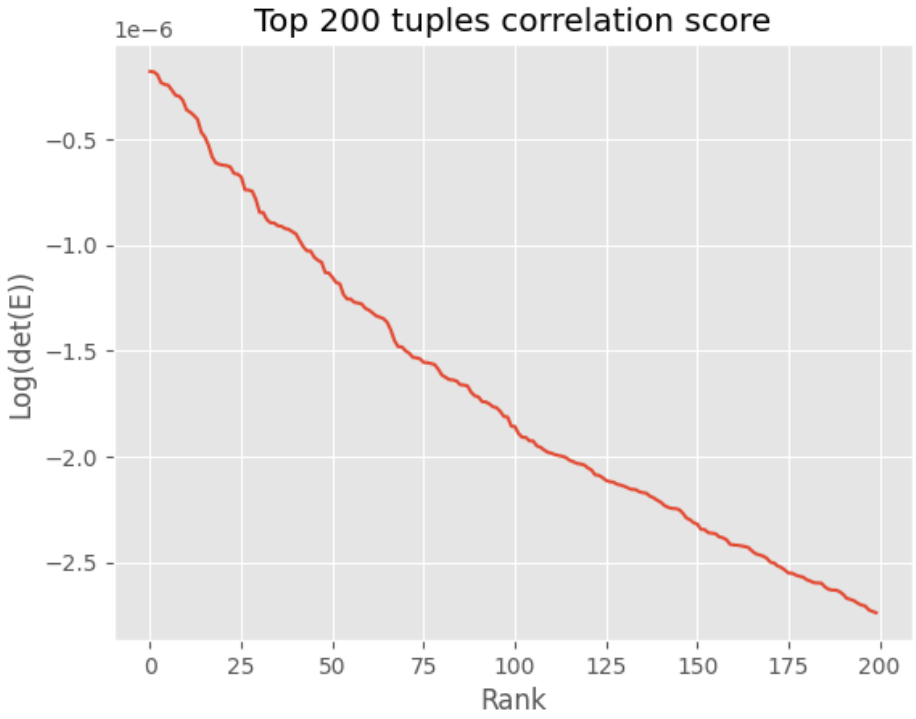

Figure 2.8: The MINE criterion log(det(E)) was used to select 80 accessions for use in a GWAS experiment at Wellbrook Farm, GA in 2022. The top 200 selected triples of accessions are ranked by det(E). From these top 200 triples 80 distinct accessions were selected.

2.7 Application of MINE to population and systems ecology

One of the fundamental aspects to the partnership between AMF and plants is the assembly of an AMF biome in the roots of plants. Part of this community assembly is driven by the competition between AMF in the roots. Understanding competition is a central problem for population ecology[61], but is also key to understanding the partnership between AMF and plants. The development of an understanding of competition led to the notion of an ecological niche.

The root space occupied by AMF and other fungi is part of their ecological niche. One of the earliest experimental studies of competition was by Gause and Witt[62] and involved examination of competition between fungal yeasts, a work that led to the development of the Competitive Exclusion Principle. This work still retains a high degree of relevance to understanding the symbiosis of AMF and land plants. AMF differ in their ability to capture N and P and to share this with their plant host. The N and P available to AMF are additional dimensions to the niche of each AMF species in addition to their root space. A final major dimension is the carbon shared by the plant host. The relationship of the AMF to the plant depends on an economic exchange between the plant and host, which can vary on a continuum from parasitism to mutualism[63].

A simple model has been proposed for how the plant mediates the competition between AMF and their relation to plant host[3]. The model is used to illustrate MINE. The density of the plant host population is denoted by H, and the density of the AMF taxa by S_1 and S_2. The population ecology model is:

$$\frac{dH}{dt} = rH(1 - \frac{H}{K + \gamma S_1}) - a_1 H S_1 - a_2 H S_2 \tag{2.39}$$

$$\frac{dS_1}{dt} = b_1 H S_1 - d_1 S_1 (1 + e_1 S_1 + c_1 S_2) \tag{2.40}$$

$$\frac{dS_2}{dt} = b_2 H S_2 - d_2 S_2 (1 + e_2 S_2 + c_2 S_1) \tag{2.41}$$

In this model of Neuhauser and Fargione[3] the plant host (H) has logistic growth at rate r to carrying capacity K in the absence of AMF. The AMF S_1 can have both positive and negative effects on the host – positive effects occur directly on host carrying capacity through γ , but negative effects happen through an increased death rate a_1. The other species S_2 is strictly parasitic on the plant through the death rate a_2 and competes under the Lotka-Volterra formalism of Gause and Witt[62] with the mutualist species S_1.

The interaction between plant and AMF S_1 is parasitic when γ is small, but can be mutualistic for intermediate values of γ. Species S_1 becomes parasitic again for high γ. This interaction can be thought of as reflecting the amount of phosphorous in the soil[49]. To make this explicit one further assumption to the model is added, namely that the interaction is linear in Phosphorous (P) applied to the field:

$$\gamma = fP \tag{2.42}$$

where the level U = P defines the experimental condition and f is a scaling constant. For low P, the benefit to the host γ is low, and parasitism occurs. If P is sufficiently high, then the host benefit γ is intermediate, and mutualism occurs. If P is extremely high, then the host γ can lead to parasitism again[3].

Growth of each AMF benefits by the plant through b_1 or b_2, but also there is a self-inhibition through e_1 and e_2 and a competitive interaction between AMF through c_1 and c_2.

The measurements in this system are the densities (or possibly hyphal extension) of AMF in roots over time, $S_1, 1, S_1, 2, \ldots S_{1,n}$ and $S_2, 1, S_2, 2, \ldots S_{2,n}$ and of the plant host, $H_1, H_2, \ldots, H_n$. Ensemble methods have been proposed to fit these kinds of models[64].

To implement MINE the problem begins by identifying the predictions of the model in a formulation very similar to that of the biological clock[1] in section VI. There are predictions about the plant host and AMF density:

$f_i^H, i = 1, \ldots, n, f_i^{S_1}, i = 1, \ldots, n, \quad and \quad f_i^{S_2} = i = 1, \ldots, n.$

These predictions are derived by solving the three ODEs above. The prediction components are then compared to the measured values of the densities under a normality assumption for errors between the predictions and measurements as in [36]:

$$P_d(\delta|\sigma_\delta^2) = \prod_{i=1}^{n} (\frac{1}{2\pi\sigma_\delta^2})^{1/2} e^{(-\frac{1}{2\sigma_\delta^2}\delta_i^2)} \tag{2.43}$$

$$P_e(\epsilon|\sigma_\epsilon^2) = \prod_{i=1}^{n} (\frac{1}{2\pi\sigma_\epsilon^2})^{1/2} e^{(-\frac{1}{2\sigma_\epsilon^2}\epsilon_i^2)} \tag{2.44}$$

$$P_f(\zeta|\sigma_\zeta^2) = \prod_{i=1}^{n} (\frac{1}{2\pi\sigma_\zeta^2})^{1/2} e^{(-\frac{1}{2\zeta_\zeta^2}\zeta_i^2)} \tag{2.45}$$

The errors are: $\delta_i = H_i - f_i^H$ and $\epsilon_i = S_{1,i} - f_i^{S_1}$ and $\zeta_i = S_{2,i} - f_i^{S_2}$.

The data in the model are $Y = (H_1, \ldots, H_n, S_1, \ldots, S_n)$ with parameters $\theta = (r, K, \gamma, a_1, a_2, b_1, b_2, c_1, c_2 d_1, d_2, e_1, e_2, \sigma_\delta^2, \sigma_\epsilon^2, \sigma_\zeta^2)$ and experimental condition U=(P). The measurements on host and symbionts are done independently so that the model specification is:

$$P(Y|\theta, U) = P_d(\delta|\sigma_\delta^2) P_e(\epsilon|\sigma_\epsilon^2) P_f(\zeta|\sigma_\zeta^2) \tag{2.46}$$

Considering the model specification as a function of the parameters yields the ensemble $Q(\theta|Y, U)$.

In this example, as in the first example of a mixture experiment, the initial densities of the plant, AMF Species 1, and AMF Species 2 are treated as known

and measured. The limitations of this competition experiment is that only one AMF species, namely Species 1, is measured over time in each of 8 years in an agricultural plot so that the number of time points n was 8. In this hypothetical example the values of the parameters were taken from Table 263 with the two further specifications that the plant carrying capacity K was 2, the benefit γ was allowed to play the role of the experimental condition U, and the error $\sigma_\epsilon^2 = 0.02$.

A Markov Chain Monte Carlo (MCMC) experiment is used to identify the ensemble $Q(\theta|Y, U)$ using the Metropolis-Hastings updating rule described in the first example using 40,000 equilibration moves and 40,000 accumulation moves. The chi-squared statistic,

$$\chi^2 = \sum_{i=1}^{n}(-\frac{1}{2\sigma_\epsilon^2}\epsilon_i^2) \tag{2.47}$$

only involves minimizing the errors in the predictions about AMF Species 1 since that is all that was measured in this hypothetical example. With the ensemble in hand, the correlation matrix E(U) between the predictions $f_i^{S_1}, i = 1, \ldots, 8$ was computed from the identified ensemble for several experimental conditions captured as U = phosphorous level is varied. In that the phosphorous level was linearly related to the benefit of AMF 1 to the plant host, in the MINE calculation the benefit γ was simply varied from 0 up to 10.0. For example, the 8 x 8 correlation matrix E(U) in [7] for a benefit $\gamma = 2.67$ was estimated from the ensemble:

Table 2.2: Estimated ensemble values.

1.0000	0.9467	0.9548	0.9525	0.9500	0.9487	0.9477	0.9472
0.9467	1.0000	0.9922	0.9928	0.9921	0.9909	0.9903	0.9898
0.9548	0.9922	1.0000	0.9985	0.9981	0.9977	0.9970	0.9967
0.9525	0.9928	0.9985	1.0000	0.9977	0.9992	0.9988	0.9987
0.9500	0.9921	0.9981	0.9970	1.0000	0.9998	0.9994	0.9993
0.9487	0.9909	0.9977	0.9993	0.9998	1.0000	0.9999	0.9998
0.9477	0.9903	0.9970	0.9988	0.9994	0.9999	1.0000	1.0000
0.9472	0.9967	0.9898	0.9987	0.9993	0.9998	1.0000	1.0000

This value of the benefit is a very interesting value because its value is right at a bifurcation point for the competition model, and at this point the qualitative behavior of the ODEs changes. Below a benefit of 2.67 for the host, only one competing AMF species survives, and at or above this benefit value 2.67 both AMF species stably coexist. In other words, when the benefit is sufficiently large, both the mutualist and pathogen AMF species are maintained in the plant roots.

The expectation for the determinant of this correlation matrix between the 8 prediction components det(E(U)) can then be used to calculate the MINE experiment (Figure 2.9) from the fitted ensemble. One of the challenges is that this correlation matrix E(U) is nearly singular and hence the det(E(U)) is a very small number. Particular care must be taken in evaluating the correlation matrix used in MINE by the correlation ellipsoid method.

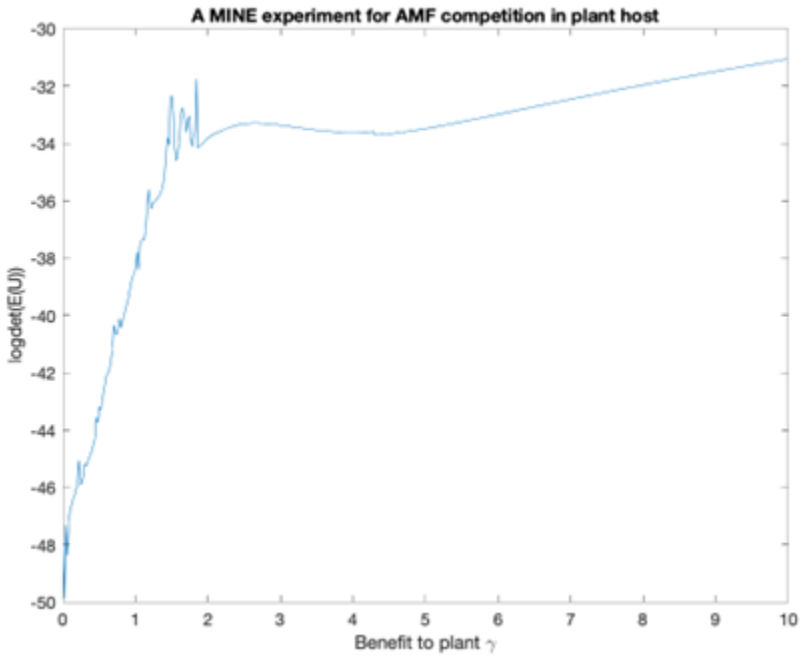

Figure 2.9: The MINE criterion based on the correlation ellipsoid allows a choice of phosphate (or equivalently plant benefit γ) in a simple Lotka-Volterra model of competition between two AMF species within a plant host[3].

The resulting MINE experiment for AMF competition in a plant host allows a choice of plant benefit or equivalently phosphorous level (Figure 2.9). There are particular phosphorous levels that yield more information in the next experiment about AMF competition in the plant. There appear to be two values, one at lower phosphate with the plant benefit varying between $\gamma = 1.5 - 1.84$ and one at a high phosphate value with the plant benefit being around $\gamma = 10.0$. The emphasis is on a range of phosphorous levels as carried out by Propster and Johnson[49]. It would be very interesting to know how this MINE experiment would change if the ensemble were identified under a very different phosphorous level.

2.8 Conclusion

Ensemble methods[6] address the developments in the analysis of large genetic studies made possible in Genomics, Transcriptomics, and other related omics and originated in statistical physics[21]. In most omics experiments to characterize genes, genetic networks, and genomes the number of parameters or effects (p) greatly exceeds the number of data points (n). In any network only a few of the species (RNAs or proteins) are measured, but the behavior of a network underlying a complex trait involves many genes and their products. In carrying our experiments on RIL or GWAS populations the problem is similar. Again there are many more SNP or gene effects on a trait than there are measurements on individuals. The same situation arises in the study of microbiomes in host species. R. A. Fisher and colleagues have built the main framework of experimental design in an agricultural setting with methods that require $n \gg p$1. How do we design current genetic experiments in this context when classical design requires $n > p$, but modern day genetics lives in the world of $n \ll p$?

The problem is to design and analyze these kinds of genetic experiments and field studies when $n \ll p$. We have developed a new model-guided discovery approach to experimental design called the Maximally Informative Next Experiment or MINE for designing such experiments[23]. This approach is distinct from the criteria of experimental design in the case of linear models[19]. MINE focuses on discovery while classic experimental design focuses on the precision of effects in the design. MINE is adaptive and uses a sequence of experiments or field studies to identify the parameters in the model, whether they be rate constants in a genetic network or SNP effects in a GWAS on a complex trait. The focus is not on the precision of the parameter estimates, but on the exploration of the parameter space for the purposes of discovery. The discovery may be new regulatory components of the model in a genetic network or the discovery of SNPs that underly a complex trait in a GWAS. This was demonstrated in the use of MINE with the study of clock in *Neurospora crassa*.

MINE is a discovery tool designed specifically for very large genetic data sets and is illustrated in a variety of problems within genetics here. MINE is built upon other ensemble methods[25] that have been developed for fitting models with $n \ll p$. These ensemble methods of model identification coupled with MINE complete a discovery cycle (Figure 2.7) for exploring problems in genetics. This discovery cycle has been called computing life[15]. MINE as a discovery tool completes the cycle in both analyzing and designing future costly omics experiments arising in genetics and allows an adaptive approach to solving problems in genetics.

Chapter 3

MINE: Maximally Informative Next Experiment – Towards a new experimental design and methodology

3.1 Introduction

Genome Wide Association Studies (GWAS) have become a standard approach to gain insights about genes that control a complex trait. As far as we know there is no literature addressing GWAS experimental design (genotypes selection) adaptively; in this work we propose a method to select the annual most informative genotypes in a sequence of GWAS experiments for discovering chromosomal regions related to a quantitative trait. The advantage of this approach is to deconstruct a GWAS into a smaller series of more tractable field experiments that may be more informative than one large classical design. This approach is called MINE: Maximally Informative Next Experiment, and maximizes the prediction uncertainty volume in a complex trait from linear and mixed linear models; it was originally presented for analysis of *Neurospora crassa* genetic networks[1][23], and it represents an alternative to classic experimental design[10].

The modeling part of GWAS also allows us to rethink classic methodology in experimental design. In fact, linear regression is the technique used to fit linear models[5] to GWAS data, where researchers generally have fewer phenotypic observations (rows) than SNPs or chromosomal regions (columns) in

the design matrix. The solution proposed here for this problem is ensemble methods for fitting [65]. Geneticists regularly take one SNP at a time to create a model (one model per SNP); therefore, a single model encapsulating available SNP interaction is not possible, and linkage disequilibrium(LD) is present implicitly due to the low number of bases separating SNPs. To overcome this problem a tool was created that puts together available SNPs forming wider chromosomal regions so that distinct regions are free of linkage disequilibrium. The result was one matrix accounting for all SNPs, and in order to identify a model ensemble methods are used and computed by Markov Chain Monte Carlo (MCMC). Specifically, the Metropolis algorithm[31] was used to fit both linear models with fixed effects and mixed linear models to data on complex traits from a GWAS. When we refer to linear models herein, it will be understood it is linear models with fixed effects throughout. To select the most significant chromosomal regions the Bayesian interval[66] method and Benjamini Hochberg criteria[67] are used for feature selection. A tool was also developed to extract the currently known genes within these chromosomal regions from the Phytozome database[68]. Finally, the last issue is the choice of columns (i.e., accessions) to include in the design matrix for succeeding years of a GWAS. This design question was addressed by a particular MCMC method called MINE [1] [23]

In order to illustrate this new design methodology for GWAS, an adaptive GWAS was implemented on a Bioenergy Association Panel (BAP) [2] for Sorghum bicolor on Wellbrook Farm, Watkinsville, GA, over three years. Roughly 80 accessions each year were arranged in a randomized complete block design with three blocks over three years to examine a variety of quantitative traits measuring plant health: log dry weight, tiller number, fungal disease burden, and height. There are over 10,000 genotypes worldwide that have been characterized in this tractable diploid genetic system for potential use in the GWAS [8].

3.2 Materials and Methods

3.2.1 Field experimental design

The location for the field experiment was Wellbrook Farm at Watkinsville, Georgia, USA; 81 accessions were planted for 3 years randomly in 3 blocks, each accession appearing 6 times in a block. Accessions were randomized within a block by plot (i.e., row). The blocks were uniform in soil composition.

Figure 3.1: Aerial photo of Sorghum plants at Wellbrook Farm. Courtesy of Dr. Peng Qi.

The accessions were taken from the Bioenergy Accession Panel (BAP)[2], which means these accessions have already been sequenced to determine their SNP genotype; this BAP collected is maintained by the USDA in Griffin, GA. Initially the seed were ordered through USDA and germinated in pots at UGA greenhouse; watering was done daily, and after a period of 2 weeks small sorghum seedlings were transplanted to Wellbrook Farm. Once in the soil at the farm, a weeding regime was established twice per week, and harvesting took place 3 months later. During harvesting disease and height were determined and recorded; canopies were chopped and put into bags to be taken to ovens for drying and weighing. Foliar fungal disease (and referred to as disease hereafter) was scored for Leaf Blight, Target Leaf Spot, Zonate Leaf Spot, Gray Leaf Spot, and Anthracnose. Disease and height data were put directly in the database; however, dry weight canopies spent one week in the oven and were weighed immediately after. The dry weight was recorded in grams, height in meters, and disease with a number from 1-10 representing its severity on the leaves of the plant. All software developed and used in this work can be found at the following link: https://github.com/JArnoldLab/MINE

3.2.2 Modeling

GWAS modeling has traditionally used linear regression [5]; however, the first challenge to be addressed is to increase the number of observations so they supersede the number of columns in the design matrix for classic methods. MCMC methods for implementing ensemble methods are a good alternative when the

number of observations is less than the number of parameters [65], and the Metropolis algorithm was used as a backbone for solving the optimization problem for finding the parameters fitting the GWAS data ; its description is the following:

Algorithm 1 Customized Metropolis algorithm

Set a stepwidth
Set N number of equilibration sweeps
Set K number of parameters in the β vector
Set P number of decorrelation sweeps
Set M number of β vectors to accumulate
Set a random initial β vector
for N **do**
  for K **do**
    Choose an element β_k from β randomly
    Propose an update $\beta_k = \beta_k + (stepwidth * U(-1, 1))$
    Accept the change with probability $min(1, \frac{Q(\beta', X)}{Q(\beta, X)})$
for M times **do**
  for P times **do**
    for K times **do**
      Choose an element β_k from β randomly
      Propose an update $\beta_k = \beta_k + (stepwidth * U(-1, 1))$
      Accept the change with probability $min(1, \frac{Q(\beta', X)}{Q(\beta, X)})$
  Store the β vector
return Accumulated β vectors

In order to use all the data in the fitting procedure, a design matrix was introduced (Regular GWAS methodology considers one SNP at a time, and consequently SNP interactions and carries LD implicitly). A tool was implemented that puts all available SNPs together to form chromosomal regions, a method known as the sum method [60]. The size of each chromosomal region was increased to cover an adjacent SNP until the size of the region was at least 50 kb. SNPs from different chromosomal regions are unlikely to be in linkage disequilibrium; this yields a design matrix that covers the entire genome if desired and sidesteps the issue of linkage disequilibrium [69].

The tool is designed to work individually with each chromosome, so if the user wanted to cover all the genome, an additional tool would be implemented

to join the small matrices coming from each chromosome into one large design matrix.

Linear Model

The Metropolis algorithm is based on a particular model, which is then used to setup a minimization problem characterizing model fit; the first model chosen was a linear model with fixed effects using the previously defined design matrix. The model is defined as follows:

$$Y = X\beta + \varepsilon \tag{3.1}$$

The Greek letter β represents the parameters describing the effects of particular chromosomal regions formed from SNPs (design matrix columns). In total there were 2748 regions in the Sorghum bicolor genome. The rows of the X matrix (design matrix) represent the genotype of each accession. The ϵ are the error in each observation on each accession. The minimization method chosen is that of least squares. In the Metropolis algorithm the objective function is called the Hamiltonian [7], and defined as:

$$H(\beta, X) = \frac{1}{2}(Y - X\beta)' V^{-1}(Y - X\beta) \tag{3.2}$$

This formula represents a distance metric of how far the model prediction is from the observed value of the quantitative trait. The V matrix is diagonal and fixed at the sample variance in the complex trait for each accession. The Metropolis Algorithm carries out a random walk in the parameter space to minimize the Hamiltonian. In order to determine if the proposed random step is good or not, the Metropolis algorithm utilizes a probability function. The Boltzmann distribution was chosen and is very well-known in statistical physics. Its definition is the following:

$$Q(\beta, X) = \frac{1}{\Omega(X)} \epsilon^{-H(\beta, X)} \tag{3.3}$$

The Metropolis algorithm implementation has two phases; the first phase is to equilibrate the system at a minimum in the Hamiltonian, which means that toward the end of the equilibration stage the parameters being estimated change very little. We say that they reached an equilibrium in the Monte Carlo experiment. The second accumulation phase comprises collecting the n almost best solutions. In our case 1000 sets of parameters were collected.

Mixed Linear Model

The mixed linear model is an elaboration of the linear model and a standard now for GWAS [59]. We included variance components into the effects of different chromosomal regions, and it is defined as follows:

$$Y = X\beta + Zu + \varepsilon \tag{3.4}$$

The first term (XB) represents the trait prediction, such as biomass; the second term (ZU) represents the variance components and heritability, and it is defined as:

$$V = \sum_{i=1}^{n} X_i' X_i \sigma^2_{j(i)} + \sigma^2 I \tag{3.5}$$

The index i runs over samples. Some samples share the same variance component. The index j(i) returns the variance component of the ith observation for the jth accession. The Hamiltonian will be represented by the negative natural logarithm of the following likelihood function:

$$L = \frac{\epsilon^{-\frac{1}{2}(Y-X\beta)'V^{-1}(Y-X\beta)}}{(2\pi)^{n/2}|V|^{1/2}} \tag{3.6}$$

$$H = \frac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta) + \frac{1}{2}n\ln(2\pi) + \frac{1}{2}\ln(|V|) \tag{3.7}$$

The Boltzmann distribution was also chosen for the mixed linear model. The resulting Hamiltonian has two more terms than the linear model with fixed effects.

3.2.3 Stepwidth adjuster feature in the Metropolis algorithm

In the Metropolis algorithm model parameters are sampled in random steps from the current position in the parameter space around a region defined by the local minimum in the Hamiltonian; however, the random step depends on a number selected by the user, the stepwidth; later, the random step is accepted or rejected. It is recommended that the rate of accepted steps is between 30 and 70 percent during the equilibration phase; this acceptance rate range is not always achieved, and may be an effect of being stuck in a small subregion during the Monte Carlo experiment. Consequently, the parameters collected in the

accumulation phase (MC sample) may include models that represent a bad fit mixed in with the almost best solutions.

To overcome this issue, a dynamic stepwidth adjuster was designed and implemented within the Metropolis equilibration phase; its mathematical definition is the following [70]:

$$S_n = f_n * S_{n-1} \tag{3.8}$$

$$f_n = \begin{cases} 1 & if\,(r_{min} \leq r_n \leq r_{max})\,|\,(S_{n-1} < S_{min})\,|\,(S_{n-1} > S_{max}) \\ f_{n-1} & if\,(r_{n-1} < r_{min}\,\&\,r_n < r_{min})\,|\,(r_{n-1} > r_{max}\,\&\,r_n > r_{max}) \\ \frac{1}{\sqrt{f_{n-1}}} & if\,(r_{n-1} < r_{min}\,\&\,r_n > r_{max})\,|\,(r_{n-1} > r_{max}\,\&\,r_n < r_{min}) \\ 2/3 & if\,(r_{min} \leq r_{n-1} \leq r_{max}\,\&\,r_n < r_{min})\,|\,(n = 1\,\&\,r_1 < r_{min}) \\ 3/2 & if\,(r_{min} \leq r_{n-1} \leq r_{max}\,\&\,r_n > r_{max})\,|\,(n = 1\,\&\,r_1 > r_{max}) \end{cases} \tag{3.9}$$

The values r_{min} and r_{max} are the low and high acceptance rate limits set by the user, and r_n is the current acceptance rate

$$S_{min} = E_s \bar{S} \tag{3.10}$$

$$S_{max} = (\frac{1}{E_s})\bar{S} \tag{3.11}$$

$$\bar{S}_n = \frac{1}{n}\sum_{n'=0}^{n-1} S_{n'} \tag{3.12}$$

$$E_s = 10^{-4} \ldots 10^{-6} \tag{3.13}$$

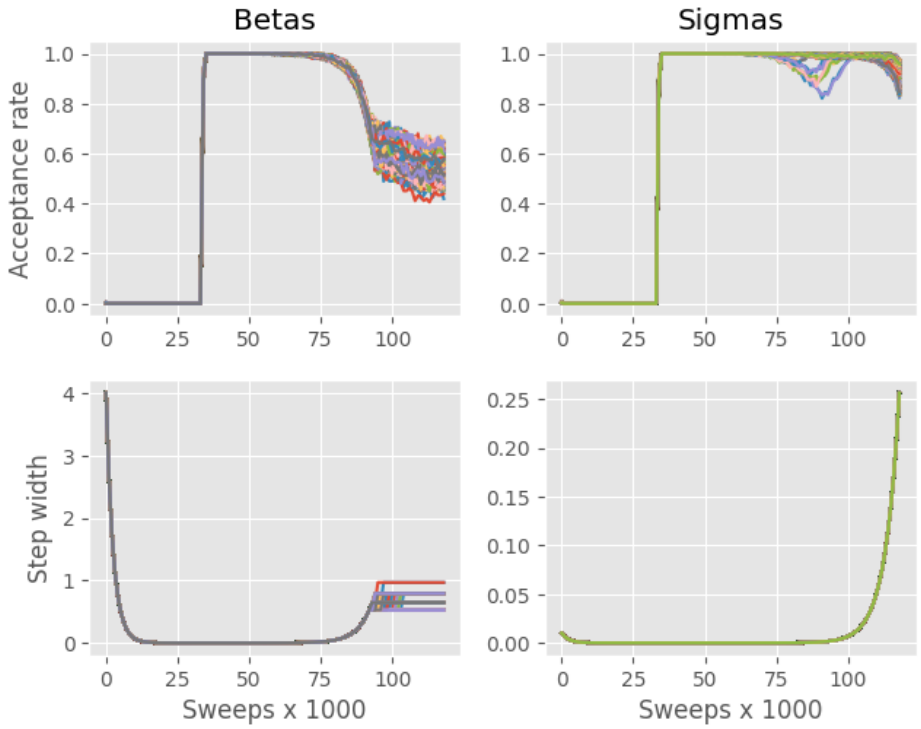

Figure 3.2: Dry weight acceptance rate and stepwidth for Beta and Sigma parameters using the stepwidth adjuster. The mixed linear model was computed with all the data available.

Reasonable results were obtained with the stepwith adjuster in our Metropolis algorithm implementation as shown in Figure 3.2. The stepwidth adjustor takes the acceptance rate to the range we desired during the equilibration phase. The step width adjuster stabilized for the chromosomal effects but increased for the variance components to achieve the desired acceptance rate range. Sometimes it was necessary to use an acceptance rate above 70 percent to achieve system equilibration. The dry weight target acceptance rate was set between 0.3 and 0.7 for the mixed linear model.

3.2.4 MINE

The MINE approach, a model-guided discovery tool, is an experimental design tool to obtain the genotypes that yield the most trait information and its relation to chromosomal regions by maximizing the model prediction uncertainty volume about a complex trait (see review [23]). This is done by two different criteria: covariance and Pearson correlation between predictions of the quantitative trait in the next GWAS experiment one year later.

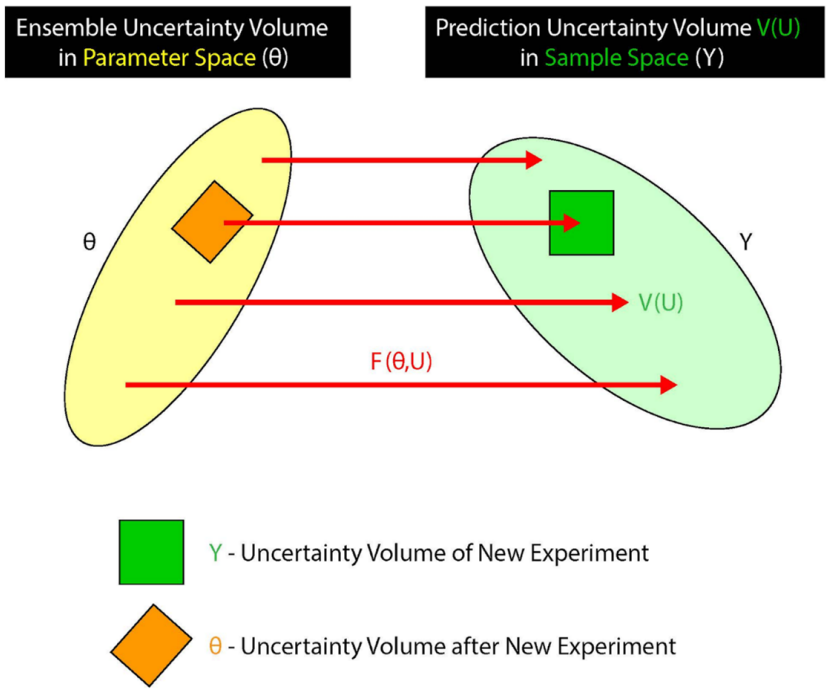

Figure 3.3: A visual explanation of the relation between parameter space and phenotype space (Y). If we maximize the volume (green square) of our phenotypic observations on the quantitative trait, then the choice of parameters (brown square) will be shrunk. If we set up various experiments, adding one more experiment each time, then the next parameter choice will be better and volume, tighter.

The covariance approach defines a volume (in the trait observation space) based on the uncertainty ellipsoid from the covariances of K observable predictions to reduce the volume in the parameter space. On the other hand, the correlation approach uses the Pearson correlation of K observable predictions for the ellipsoid. The MINE approach then uses the estimated parameters for the model by ensemble methods to calculate expected values in the next year.

$$E_{mc}[G_k(.,X)] = \sum_{n=1}^{N} G_k(\beta^{(n)}, X) \tag{3.14}$$

$$E_{mc}[G_k(.,X)G_j(.,X)] = \sum_{n=1}^{N} G_k(\beta^{(n)}, X)G_j(\beta^{(n)}, X) \tag{3.15}$$

where k = 1,2,3...K; j = 1,2,3...K. The constant K represents the number of genotypes to be chosen in a given year to obtain the maximum amount of information about the model parameters. The covariance and correlation criteria

between predictions in the next experiment are based on a matrix of covariances or correlations, respectively. Each element of the covariance matrix is defined as follows:

$$D_{kj}(X) = E_{mc}[G_k(.,X)G_j(.,X)] - E_{mc}[G_k(.,X)]E_{mc}[G_j(.,X)] \quad (3.16)$$

Each element of the correlation matrix can be obtained from the covariances.

$$C_{kj}(X) = \frac{D_{kj}(X)}{\sqrt{D_{kk}}\sqrt{D_{jj}}} \quad (3.17)$$

The matrix determinant represents the ellipsoid uncertainty volume, and since both matrices are square and positive semi-definite, then the determinant is computed in the following way from the eigenvalues of the covariance or correlation matrix:

$$Det(D) = \prod_{k=1}^{K} \lambda_k \quad (3.18)$$

$$Det(C) = \prod_{k=1}^{K} \chi_k \quad (3.19)$$

The λ and χ represent an eigenvalue from the covariance and correlation matrices, respectively calculated from the model predictions.

3.2.5 Optimization algorithms

Imagine there are 350 genotypes (accessions), but resources are only available to plant 80 genotypes per year; using the MINE procedure, it will be necessary to analyze each subset of 80 genotypes; therefore consideration will need to be given to analyzing 350 choose 80 subsets. This number alone is $2.59x10^{80}$, which is not feasible to examine each subset on the fastest computers. In order to have a result in reasonable time, 4 optimization algorithms were designed to have a result in a matter of hours.

Suboptimal algorithms

The objective of this algorithm is to reduce the size of each subset to be analyzed; the size should be a number that divides the total number of accessions to be planted. In the case here, every subset was chosen to be of size 3 because 81

accessions were planted (81 is divisible by 3). When all subsets are analyzed the top n are chosen until completing the number of accessions to plant. Its description is as follows:

Algorithm 2 Suboptimal algorithm

Set n number of elements in a tuple
Set p number of accessions to be selected
Generate all combinations of n elements from the accessions
Score all tuples
Sort the tuples based on the score
Select the top p individual accessions from the tuples
return List of top p accessions

Monte Carlo

This MINE computation algorithm intends to select a subset of size equal to the number of accessions to plant, and the rest of the accessions to go into a pool as candidates for later choices (but currently categorized as not for use). After n steps, one accession in the subset and one accession in the pool are swapped, and the MINE score is calculated to decide if the change is an improvement by the MINE criterion. The algorithm is described as follows:

Algorithm 3 Monte Carlo algorithm

Set p number of accessions to be selected
Set m number of swapping steps
Generate an initial random sample of p accessions
for m times **do**
    Swap one accession from the sample with one from the accessions pool
    Score the sample
    Accept or reject the change using Boltzmann probability and the score as a Hamiltonian
return Sample of p accessions

Suboptimal combination algorithm (Nc3+2)

This MINE computing algorithm combines two suboptimal search approaches; the length of subsets in both must be small. A suboptimal search is performed first. The top subset is taken, and from that point the search with the smaller subset length suboptimal search is performed successively until having the total

number of accessions to be planted. In our case, the first suboptimal search was length 3, and the second one length 2. This algorithm is the following:

Algorithm 4 Suboptimal combination algorithm

```
Set p number of accessions to be selected
Set n number of elements per tuple in the first suboptimal algorithm
Set m number of elements per tuple in the second suboptimal algorithm
Get the top tuple of length n from the first suboptimal algorithm
Remove the accessions in the tuple from the pool
Create a list initializing the accessions in the tuple
while number of accessions in the list < p do
    Run the second suboptimal algorithm for the accessions in the pool
    Get the top tuple of length m
    Add the top tuple to the accessions list
    Remove the accessions in the tuple from the pool
return List of p accessions
```

Greedy algorithm

This algorithm starts by taking as reference the top subset of the suboptimal search. It adds one accession at a time from the pool of other accessions, scores the subset, removes the previously added accession and goes on to the next one; after all accessions are passed, the highest score subset remains. The process is repeated until the remaining subset size equals to the number of accessions to be planted. Its description is the following:

Algorithm 5 Greedy algorithm

```
Set p number of accessions to be selected
Set n number of accessions per tuple in the suboptimal algorithm
Get the top tuple of length n from the suboptimal algorithm
Remove the accessions in the tuple from the pool
Create a list initializing the accessions in the tuple
while number of accessions in the list < p do
    for each accession in the pool do
        add the accession to the list
        Score the list
        Store the list and score
        Remove the accession from the list
    Sort the stored lists based on their score
    Select the top list
    Make the selected list the main list
return List of p accessions
```

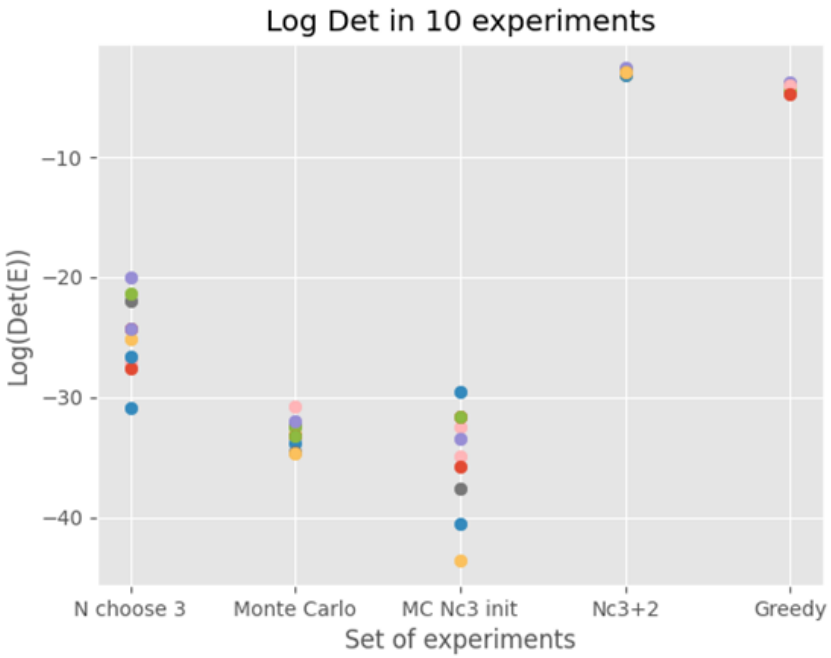

Figure 3.4: Optimization algorithms MINE score over 10 experiments. Suboptimal is labeled "N choose 3", Suboptimal combination is labeled "Nc3 + 2", MC Nc3 representing the Monte Carlo algorithm initialized with the suboptimal algorithm results.

To compare the performance of the 5 MINE computational algorithms over ten different MCMC simulations were computed using the same data, but different random seed. The results show that the greedy and suboptimal combination algorithm are the best (Figure 3.4).

3.2.6 Marker selection

It is the norm that GWAS selects the most significant chromosomal regions (markers) by means of a t-test applied to an individual marker modeling [5]. Since all markers here were encapsulated into a large design matrix and have only one model, a linear projection, Bayesian interval, and Benjamini-Hochberg criteria were used to filter out the most significant markers. The linear projection method is intended to remove noise due to the parameters not constrained by the observations on the quantitative trait. The procedure is as follows:

Algorithm 6 Linear projection algorithm

Create a diagonal matrix from the data variance V
From design matrix X, create a matrix: $X^T V^{-1} X$
Create a rotation matrix rot by extracting and sorting the eigenvectors from $X^T V^{-1} X$
Rotate the β vector $\beta^* = \beta^T rot^T$
Remove the values that wandered around $-\infty$ to $+\infty$ during estimation
Get the projected β by rotating back β^*: $\beta^p = \beta^* rot$

Once the linear projection method is applied, the Bayesian interval procedure is then applied. It consists of retaining those parameters outside of a 95 percent Bayesian confidence interval of about zero. The Bayesian Confidence Interval is computed from the ensemble.

Algorithm 7 Bayesian interval algorithm

Sort each β parameter sample
for each β parameter sample **do**
    if first $2.5\% < 0$ and last $2.5\% > 0$ **then**
        Remove β parameter

In parallel with the Bayesian interval Method [66], parameters were also filtered by the Benjamini-Hochberg criterion [67] using a false discovery rate threshold, in our case 0.05, and z-scores from the MC sample.

Algorithm 8 Benjamini-Hochberg algorithm

Get z-scores and p-values across all β parameters
Sort the p-values
Set a threshold α for false discovery rate
Set the Benjamini Hochberg critical value $BH = \frac{rank}{length} * \alpha$
if p-value < BH **then**
    Keep the β that corresponds to the p-value

We keep the markers that passed all filters.

3.2.7 Gene finder tool

Having the final chromosomal regions (markers) is only part of the results. A tool was implemented that allows the retrieval of the known genes through Biomart in Phytozome within each chromosomal region. This functionality can be expanded to sequences or other entities. This tool takes the filtered chromosomal regions and its limits in the genome, and connects to the Phytozome database [68] (maintained by the DOE) to retrieve the genes within each region.

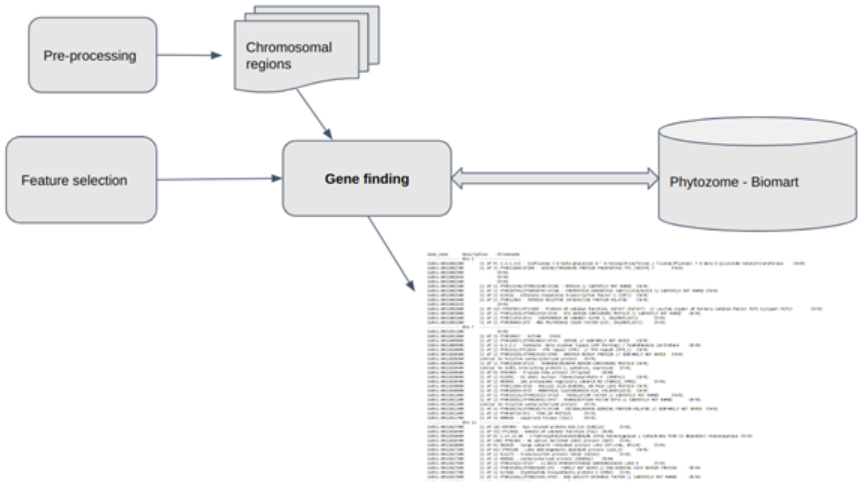

Figure 3.5: Gene finder tool diagram

3.3 Results

Modeling

A 3 year adaptive GWAS experiment using MINE was performed on Sorghum data. Our initial data (year 0) came from the Kresovich laboratory[2], and for

3 consecutive years selected Sorghum genotypes (accessions) were planted, creating an expanding database for dry weight, height and disease. Most omic studies will have prior data to initialize a MINE experiment (see, for example, [1]); however, if this were not the case, an alternative is to choose a diversity of accessions with different phenotypes to initialize the MINE procedure. MINE is a discovery tool, and such a choice in a diversity of accessions would tend to reflect its discovery purpose of MINE. In year 1 our accessions were selected from the study made by Kresovich. A total of 3 blocks of accessions were used for planting, and 81 genotypes were planted in each block using a randomized block design with 3 blocks having 6-9 replicates per row. The parameters estimation procedure (computed with the Metropolis algorithm) performance was reasonable (Figure 3.6), and a set of 1000 sets of parameters were collected in the ensemble (MC sample) for each trait in the accumulation phase. The equilibration phase appeared successful (Figure 3.6).

In the second year the MINE procedure was used to select the genotypes to plant for year 3. 81 genotypes were selected with MINE; 3 blocks were set up, each of them containing the 81 genotypes randomly arranged into rows with 9 replicates. Data collected from the previous year were merged with the Kresovich data (year 0) to run the parameters estimation and the MINE procedure for year 3.

For the third year, data from years 2, 1, and Kresovich were combined to run the parameter estimation by the ensemble method and the MINE procedure was computed for year 4. Again 3 blocks were used in the same location to plant the 81 new genotypes.

In order to obtain the most significant chromosomal regions and the genes within each region, a GWAS analysis was performed with the ensemble method using all of the data from Kresovich, year 1, year 2 and year 3.

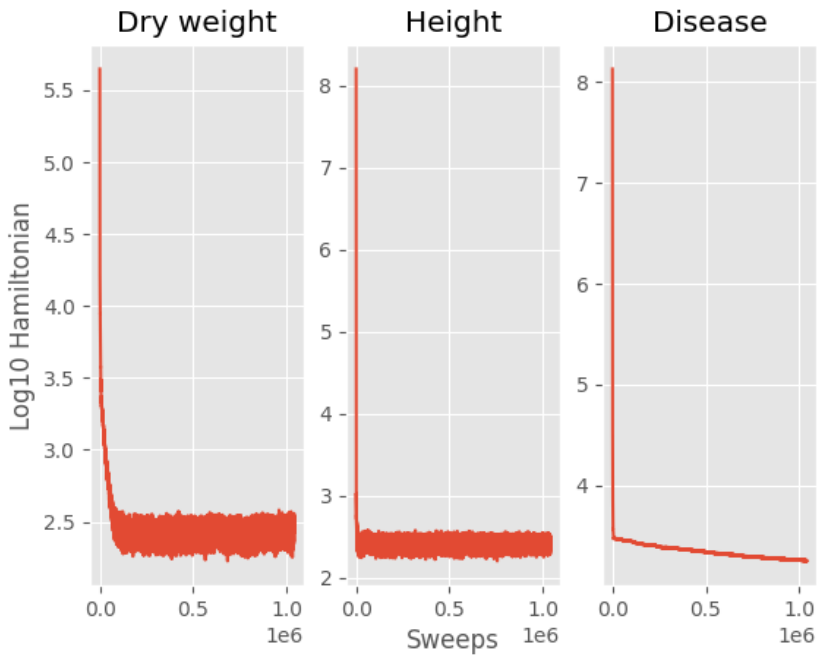

Figure 3.6: Dry weight, height, disease linear model Hamiltonian using Kresovich, year 1, year 2, year 3 data against sweep (a visit on average to each model parameter once).

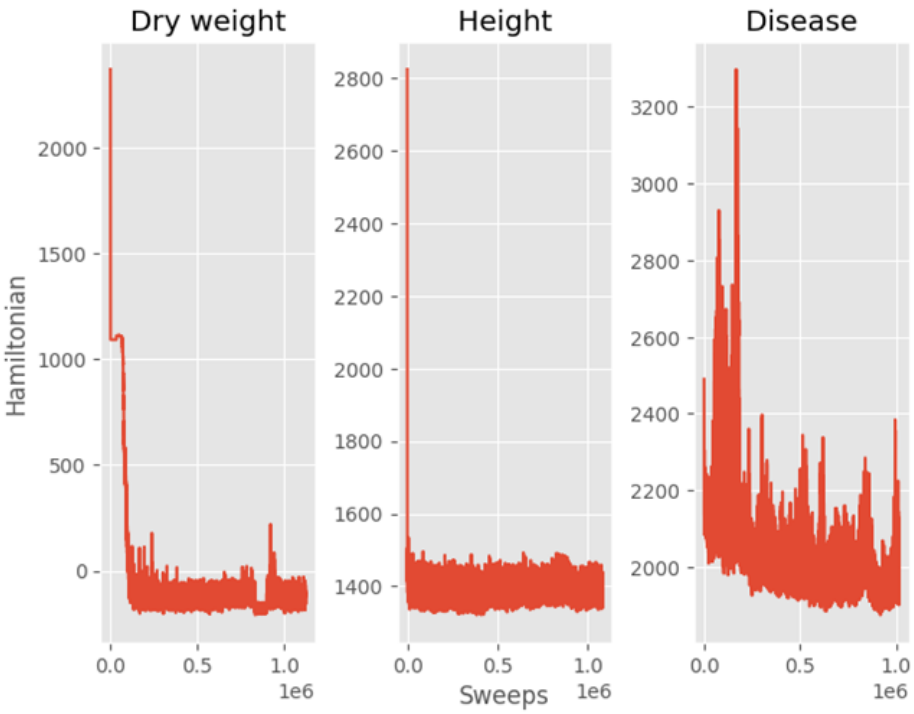

Figure 3.7: Dry weight, height, disease mixed linear model Hamiltonian using Kresovich, year 1, year 2, year 3 data against sweep (a visit on average to each model parameter once).

As a control on the Monte Carlo experiment, the Hamiltonians on both linear model (Figure 3.6) and mixed linear model (Figure 3.7) were computed (a visit on average to each model parameter once) to demonstrate equilibration. The number of collected parameter vectors was 1000 for the accumulation phase, and the number of decorrelation sweeps were 1000 between each member of the ensemble in the accumulation phase. That is why Figures 3.6 and 3.7 show an equilibrium during at least one million sweeps.

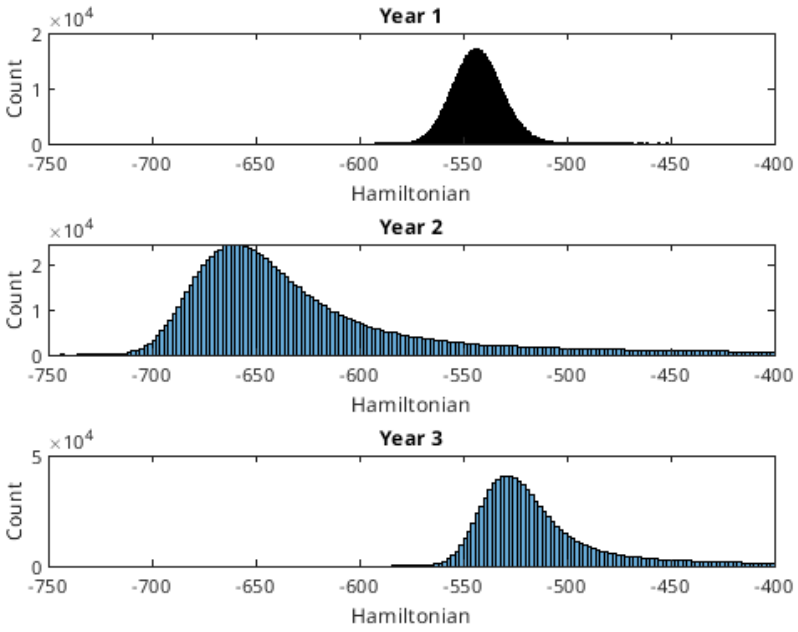

Figure 3.8: Ensembles separately fitted by year are overlapping with respect to their Hamiltonians. Hamiltonian histograms from ensembles of the mixed linear model for height were separately fitted in each year and computed.

An assumption of the MINE approach to GWAS is that there is no year effect on the complex trait, and as can be seen in Figure 3.8 the Hamiltonians for year 1, 2, and 3 are overlapping, suggesting no year effect. When there is a year effect, two approaches are available. One approach is to introduce a year effect into the mixed linear model. A second approach is to recognize that when there are small yearly effects, this environmental noise provides an additional filter for significant features in the genome. Researchers are only going to be interested in chromosomal regions or SNP effects that survive the yearly environmental effects of planting, such as those due to variation in rainfall.

MINE

The MINE procedure was run to select the most informative genotypes for the second year. The data used were from Kresovich and year 1. The data for the third year was from Kresovich, year 1, year 2.

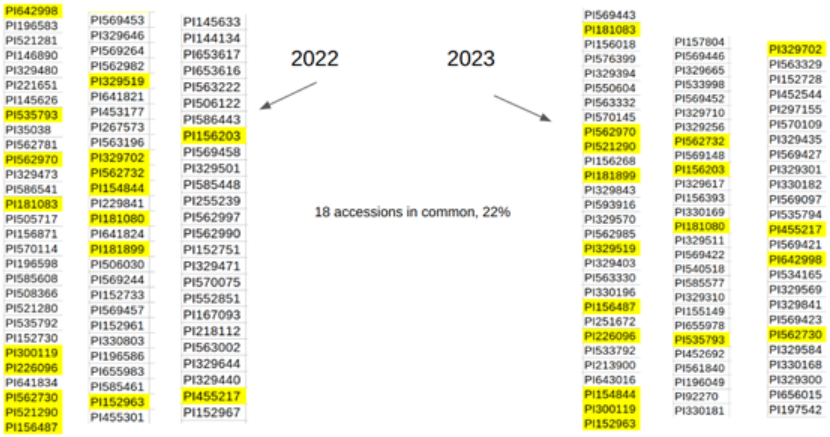

Figure 3.9: Accessions selected by the MINE procedure for planting in year 2 (2022) and year 3 (2023). The selected accessions are in yellow

The MINE approach kept 22 percent of the accessions from year 2 to year 3.

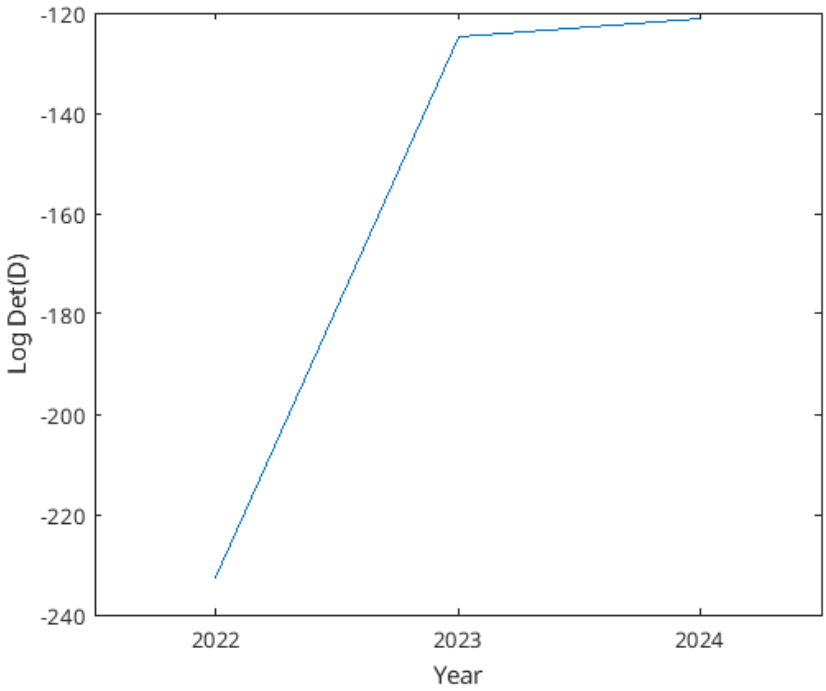

Figure 3.10: MINE score for covariance criterion in equation 3.18 increases over the 3 years.

The MINE criterion (D), used to select the most informative accessions, increased in the experiment over years (Figure 3.10), which shows that the MINE

procedure was improving on our knowledge about the relation of a complex trait and its relation to the effects on the complex trait.

Marker selection

Markers were selected only using the final cumulative data in year 3, and both linear and mixed linear models were utilized in feature selection. The results of the feature selection of significant chromosomal regions with each class of models for each trait are summarized in Venn Diagrams [4]. (Figures 3.11 - 3.13):

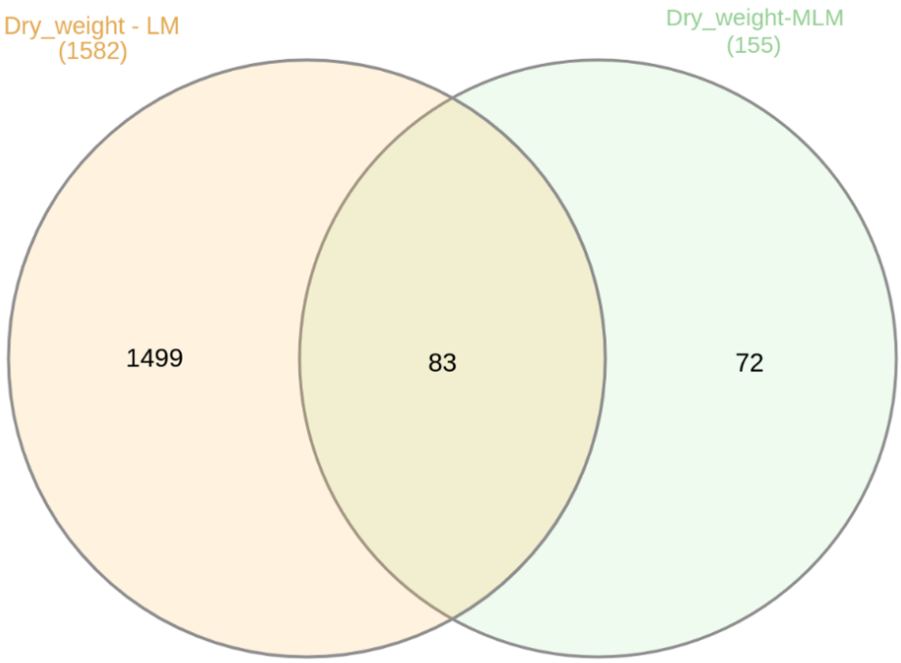

Figure 3.11: Significant markers (chromosomal regions) using all data available. LM are linear model results, and MLM mixed linear model results. Computed with [4].

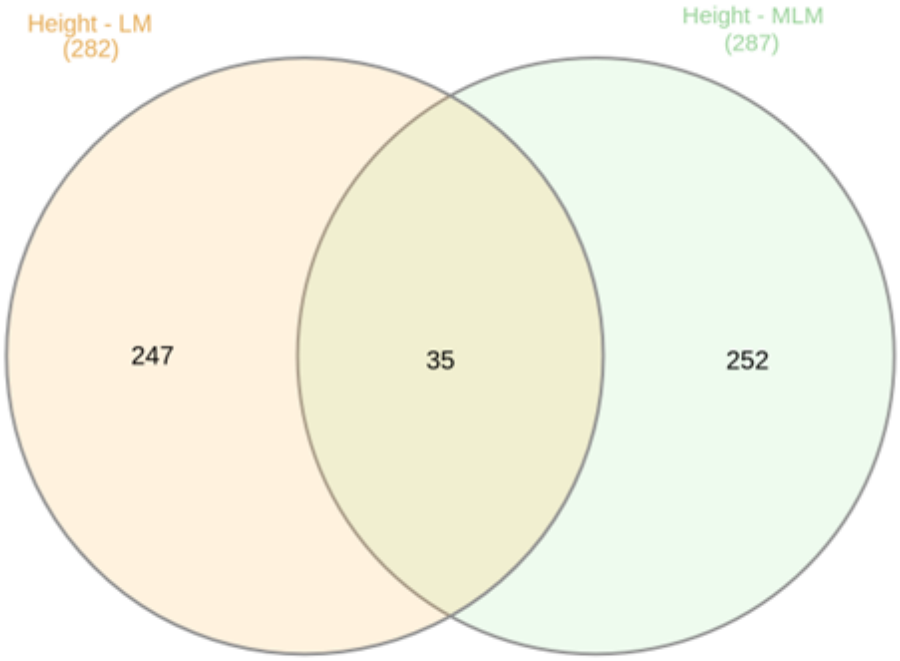

Figure 3.12: Significant markers (chromosomal regions) using all data available. LM are linear model results, and MLM mixed linear model results. Computed with [4].

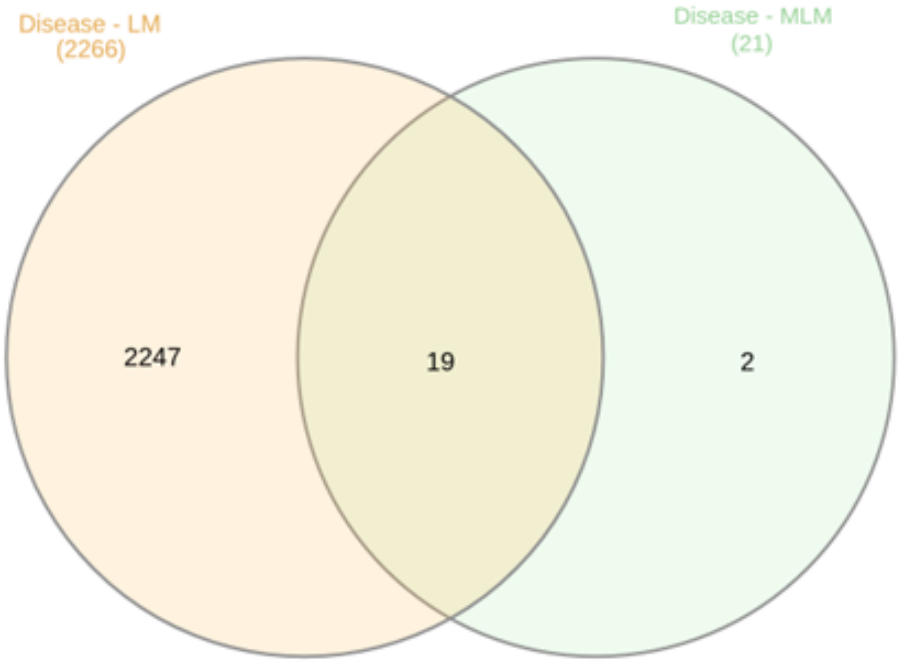

Figure 3.13: Significant markers (chromosomal regions) using all data available. LM are linear model results, and MLM mixed linear model results. Computed with [4].

Comparison of mixed linear model with fixed effects vs. linear model in GWAS analysis

As seen in Figures 3.11 , 3.12, and 3.13, the mixed linear model generally shows a higher filtering capacity for the effects of chromosomal region than the linear model. Only for height was the filtering of features comparable for the linear and mixed linear models. This means our approach of attributing a variance component to each accession as a model parameter and also considering the fixed variance across observations translates into fewer selected features.

Marker analysis

To understand what genes underly each complex trait, we retrieved all of the genes within each selected chromosomal region from the Phytozome database [68] using our "gene finding" tool previously described, an average of 12 genes

per chromosomal region were retrieved. The markers from the mixed linear model were chosen because the filters showed a higher capacity as shown in the previous figures. The genes identified arise from an adaptive GWAS performed over 3 years. Data were collected for 3 traits: dry weight, height, and disease; however, the construction of the design matrix selection (of genotypes) was guided by the dry weight observations. Then, an exhaustive search of the GWAS literature on each gene was carried out in sorghum, and all relevant sorghum GWAS papers were downloaded from Google Scholar. A script was created to take a gene found here in our GWAS and located in each paper. The results of this search are summarized in Table S1-S3. The script was run on gene lists found in our GWAS from three traits. Disease had 241 genes in the GWAS here associated with significant chromosomal regions. Dry weight had 1807 genes identified in 155 chromosomal regions, and height had 3603 genes identified in 287 chromosomal regions.

Dry weight genes

Any gene listed in this section was found to have a significant effect on dry weight in our GWAS analysis. Its heritability score was 0.998996. This heritability score obtained here is consistent with the ranges reported [71]. Gene Sobic.001G112500 has been found to be important for biomass previously and is the closest gene to a significant marker, explaining 16.2 percent of the variation and making it the major one [72]. Gene Sobic.004G044200 and Sobic.004G273900 were related to tannin and starch content. Sobic.004G044200 was found 1010 base pairs away from a selected marker, and Sobic.004G273900 was 33720 base pairs[73] from a marker. Gene Sobic.002G116000 was downregulated in mucilage-secreting aerial roots, and Sobic.010G120200 was determined to be a candidate by a previous GWAS and transcriptome analysis[74]. Yield per panicle was related to gene Sobic.005G064900, which was found in linkage disequilibrium with a significant SNP [75]. Gene Sobic.006G122200 was associated to biomass composition [76]. Tiller number was found to be associated with gene Sobic.007G151400 (and found here in its effect on dry weight) in a GWAS for forage yield [77], and gene Sobic.008G186400 in a GWAS for plant architecture and bioenergy [78]. Our GWAS approach on dry weight also recovered significant chromosomal regions containing genes from closely related traits such as height or disease, as supported by additional literature. So, genes below were pulled in our GWAS as affecting dry weight, but in other GWAS studies affected other traits, such as disease or height. Amino acid traits affecting grain quality were related to genes Sobic.001G241200 and Sobic.001G405500, which are 21770 and 4080 base pairs away from an impor-

tant marker[73]. Sobic.004G273600 was a direct hit for tannin content and Sobic.004G273800 was 28900 base pairs away from a significant marker [73]. Seed width was found to be related to gene Sobic.001G271500 (as well as dry weight here), its distance to a selected SNP being 22206 base pairs[79]; on the other hand, gene Sobic.007G093100 was related to seed perimeter and was located 37632 base pairs away from a significant SNP[79]. Gene Sobic.001G328500 was found near a significant QTL in a GWAS for grain color and tannin content[80]. For parasitic plant (Striga) resistance GWAS found genes Sobic.002G021700, Sobic.009G056400, and Sobic.010G032000 that were close to significant markers[81]. Nodal root length was associated with gene Sobic.002G188800 in a previous GWAS as well as Sobic.002G188600[82]. Genes Sobic.002G280400, Sobic.002G280600, Sobic.002G280700, Sobic.002G280300, and Sobic.002G280800 showed resistance to anthracnose, downy mildew, grain mold, and heat smut[83]. Sobic.002G416400 and Sobic.005G165632 were found to be a determinant for plant color[84]. Leaf senescence was determined to be related to gene Sobic.003G052200 in a GWAS[85]. Epi-cuticular wax genes were also discovered in our GWAS analysis. For example, Sobic.004G154200 was found 2880 base pairs away from a significant marker, and Sobic.004G154300 5771 base pairs away from another SNP [86]. A dwarf locus encodes a protein kinase, Sobic.006G067600, which was related to height [87]. Gene Sobic.006G067700 was found in 3 GWAS papers, and it was found in a dwarf locus [88] [89] [76]. An important marker representing 1 percent of the variance on a days to flowering GWAS was found in gene Sobic.006G120000 [87]. A circadian rhythm gene was identified in a GWAS for forage yield Sobic.010G045100 [77].

Disease genes

All genes listed in this section were identified as candidates in the GWAS analysis of Disease here and tied to previous GWAS results. Its heritability score was 0.648086 [90]. For disease genes, we found Sobic.004G002200 related to starch content, which is an important factor in plant fitness under abiotic stress[91]. Gene Sobic.004G002300 was associated with grain mold resistance in a GWAS performed by [92] locating it 0 base pairs away from a significant SNP. Resistance to parasitic plant (Striga) was also associated with gene Sobic.004G158901, where it was found in a significant association with a SNP [81]. A GWAS on seed morphology mentioned gene Sobic.004G340100 being 2091 base pairs away from a selected marker, as well as gene Sobic.007G093100 being 37632 base pairs away. Both genes were associated with the seed perimeter[79]. Genes Sobic.006G149650 and Sobic.006G149700 were found associated with plant color [84].

Height genes

All genes in this section were found to be associated with height in the GWAS analysis of this paper. Its heritability score was 0.273398 [71]. Gene Sobic.003G202000 is associated with plant height and was found 40100 base pairs away from a significant SNP [78]. Gene Sobic.007G161700 was found close to a dwarf locus, specifically 2000 base pairs away, and Sobic.007G161800 and Sobic.009G024600 were close to significant markers [93]. Another gene found close to a dwarf locus is Sobic.007G160400, 94100 base pairs away [94]. Height is also associated with gene Sobic.009G223500, which was found within two significant SNPs [87]. Similar to dry weight, the height GWAS here also picked up significant chromosomal regions containing genes cited by other traits in previous GWAS literature, such as biomass or disease. Gene Sobic.001G270200 was related to grain mold resistance, and was found 27000 base pairs away from a significant SNP. Gene Sobic.001G270301 was also related to grain mold resistance and was found 166000 base pairs away from another significant marker [95]. Anthracnose resistance was found related to gene Sobic.001G377200, which contains two significant markers [96]. Gene Sobic.001G405500, Sobic.002G113600, Sobic.003G033900, Sobic.004G156000, Sobic.006G187900, Sobic.010G080300 were found related to grain quality variation [73]. Gene Sobic.002G113900 was related to head smut resistance [83]. Gene Sobic.002G208200 was related to grain yield, and gene Sobic.010G216600 was related to grain number and weight [75]. Tiller number was associated to gene Sobic.002G253000, which was found 3700 base pairs away from a significant SNP [78]. Genes Sobic.003G052200, Sobic.004G299500, Sobic.004G299600, Sobic.004G299700, and Sobic.006G261100 were associated to leaf senescence [85]. Grain carotenoids were associated with gene Sobic.003G197500 and Sobic.007G156300, which were found 150000 and 40000 base pairs away from different significant SNPs [97]. Response to anthracnose was related to gene Sobic.003G203500 [98]. Concentrations of iron and zinc were studied and we found gene Sobic.003G350800 highly expressed in a GWAS [99]. Epi-cuticular wax was associated to genes Sobic.004G154200, Sobic.004G154300, Sobic.004G156000, Sobic.005G222000, and Sobic.010G065300, which were found 2880, 5771, 3821, 7607, 14138 base pairs away from different significant SNPs, respectively [86]. Gene Sobic.004G163700 and Sobic.007G004500 were found to be associated with parasitic plant striga resistance [81]. Grain color and tannin content were associated to gene Sobic.004G230000 [80]. Sobic.005G033801 and Sobic.006G248300 were related to resistance to sorghum aphid and were found within 200000 base pairs of different significant SNPs. Gene Sobic.010G091100 was found within 408000 base pairs away from a significant SNP [100]. Glume cover was found to be associated with genes So-

bic.006G095550 and Sobic.006G095400 [93]. Panicle exsertion was associated with gene Sobic.006G094600 and Sobic.006G094800 [93]. Biomass accumulation under cold stress was related to gene Sobic.007G033300, which was found within a significant SNP [101]. The number of nodes in aerial roots was associated with gene Sobic.007G155900, which contained a significant SNP [102]. Panicle length was found associated with gene Sobic.008G120200 [103]. Anthracnose response was associated with gene Sobic.009G162500 [104]. Biomass related traits were associated to several genes, plant maturity with Sobic.009G250500, Sobic.009G250600 and Sobic.009G250700; however, gene Sobic.009G250800 was the closest to a significant SNP, 55 base pairs away [72]. Biomass composition was found associated with gene Sobic.009G250600 [76]. Number of nodal roots had significant SNP found within gene Sobic.010G198000 in a GWAS for root system architecture [82].

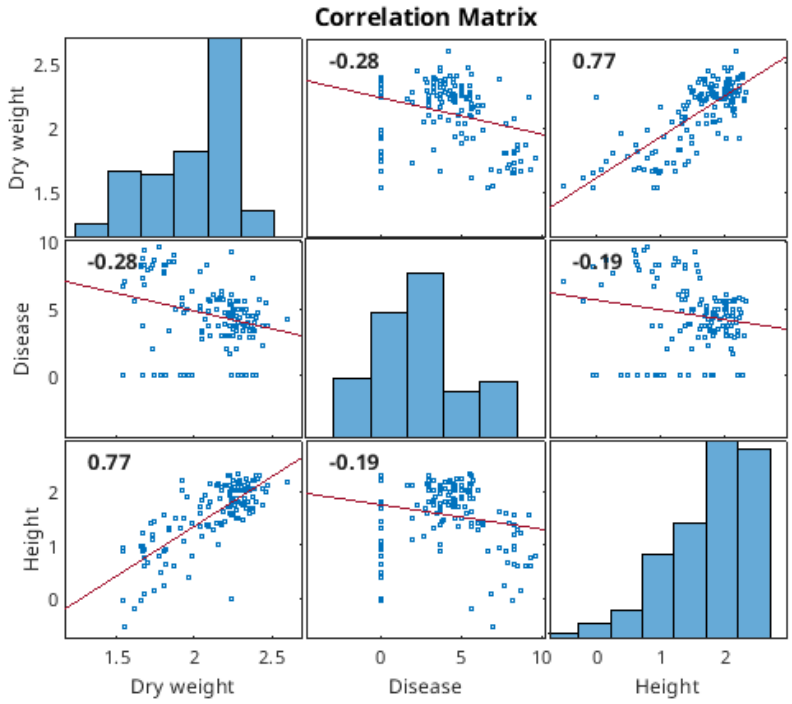

Figure 3.14: Dry weight, height and disease correlations and histograms

In Figure 3.14, there is a correlation displayed between dry weight, height, and disease; therefore, these correlations are consistent with some candidate genes appearing in the analysis of other traits.

Markers in the BAP original study vs. our study

The following figure shows the number of selected chromosomal regions in the BAP original study and our study, using the same tools presented in this work.

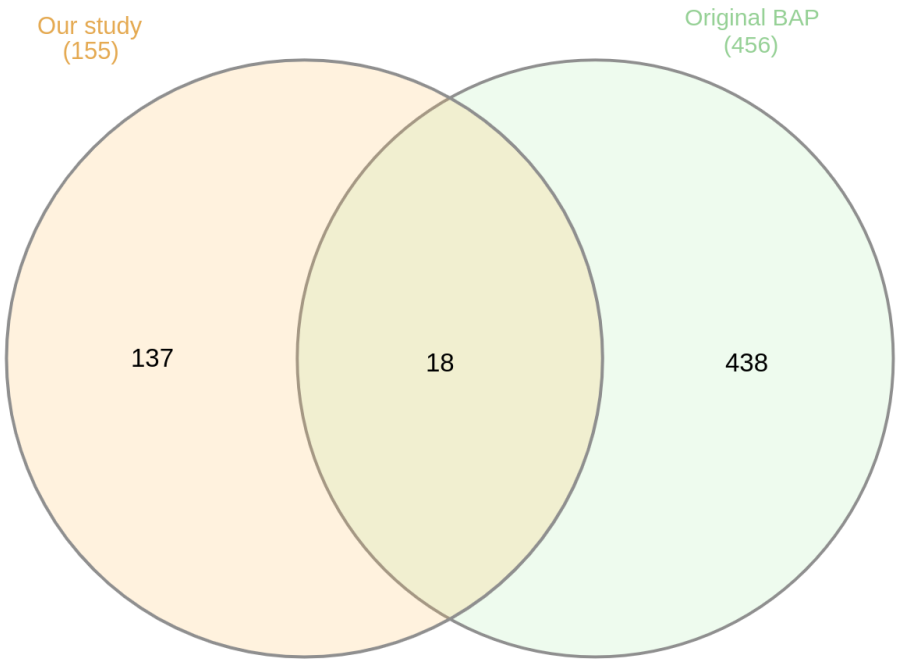

Figure 3.15: Selected chromosomal regions in the BAP original study [2] and Arnold lab study for log dry weight under the mixed linear model. Computed with [4].

The original BAP study had 244 different accessions in its analysis [2]. The total number of accessions in the BAP collection is $\sim$ 343. In the MINE procedure in the current GWAS, a total of 155 novel chromosomal regions were selected for follow up validation studies. The MINE procedure took a subset of these 343 accessions each year; therefore, the MINE procedure ended up adding more observations for some accessions and utilizing others in exploration, and that is what we see in the previous Venn diagram (Figure 3.15). We validated the chromosomal regions for both our and BAP data alone. There were 31 confirmed chromosomal regions contributing to dry weight in our study, resulting in 20 percent contribution; on the other hand, there were 65 confirmed chro-

mosomal regions contributing to dry weight in the BAP study, resulting in 15 percent contribution. The ultimate validation of these chromosomal regions are further experiments examining the effects of genes through knockouts and expression studies within the chromosomal region.

3.4 Discussion

In this work we have presented a new GWAS experimental design approach supported by non-traditional methodology for adaptive (in time) GWAS. The MINE procedure was proposed by Bernd Schuttler [1] and has been previously used to identify the dynamics of cellular processes[23] as well as to choose taxa in phylogenetic problems [105] [106], both sets of prior work focusing on experimental design. In this work we have applied it to GWAS, in order to discover what genotypes are the most informative for an adaptive sequence of experiments over consecutive years. The ensemble method including the MINE design tool were specifically designed for precision agriculture in which the number of effects (p) greatly exceeds the number of observations (n). Traditional approaches in experimental design cannot accommodate this situation [10]. The second novelty of MINE is its focus on discovery rather than the precision of effects of genes or chromosomal regions on the complex trait of interest. In the situation where there are potentially 10^5 effects but only 10^3 observations (plants) in the field experiment, the focus must be on discovery of the important effects rather than the precision of the estimates of effects. The results here shows us that MINE performs reasonably well (Figure 3.10); there are limitations and benefits to using ensemble methods including MINE. One of the main benefits is being able to perform an analysis having less observations than markers. Rather than utilizing a single SNP modeling procedure, all of the data are considered together. The approach taken here differs from methods using LD directly to hunt down QTLs by sidestepping the LD problem with the use of chromosomal regions [60]. The main limitation is accommodating changes in the model over time. If the effects are stable in time, then smaller yearly experiments can be carried out to identify the effects and satisfy resource constraints (see Figure 3.8). For example, our laboratory was only capable of planting 80 genotypes/accession per year, while still taking into account statistical design requirements, such as 3 randomized blocks containing 6 replicates of each genotype. The resulting field experiment still had 1440 plants to manage, and only 720 (3 replicates per block) were harvested. Results show that around 20 percent of the genotypes were kept from one year to the next one, achieving a good degree of diversity across the selected genotypes. Finally, we filtered out

the most significant chromosomal regions and examined all of the genes within each region; our literature review reveals that our GWAS was able to accurately identify markers containing genes related specifically to a particular trait in previous studies, such as height. Among the genes we found directly related to height, Sobic.007G161700 and Sobic.007G160400 were previously described by other GWAS to be associated to dwarfism, and genes Sobic.003G202000, Sobic.007G161800, Sobic.009G024600, and Sobic.009G223500 were found directly influencing height. Genes found directly related to dry weight were Sobic.001G112500, Sobic.004G044200, Sobic.004G273900, Sobic.004G044200, Sobic.004G273900, Sobic.002G116000, Sobic.010G120200, Sobic.005G064900, Sobic.006G122200, Sobic.007G151400, and Sobic.008G186400. For disease, the genes directly related were Sobic.004G002200, Sobic.004G002300, and Sobic.004G158901. We also showed that height, dry weight and disease traits have a degree of correlation (Figure 3.14); therefore, our GWAS on height picked genes previously associated to dry weight and disease; similarly the GWAS on dry weight picked genes related to height and disease, and the same observation was made for the GWAS on disease. Interestingly dry weight genes found on height GWAS were Sobic.002G208200, related to grain yield, Sobic.010G216600, related to grain number and weight, and Sobic.007G033300, related to biomass accumulation under cold stress. There were many disease genes found in dry weight and height GWAS, the most interesting being Sobic.002G021700, Sobic.009G056400, and Sobic.010G032000, related to parasitic plant Striga resistance; Sobic.002G280400, Sobic.002G280600, Sobic.002G280700, Sobic.002G280300, and Sobic.002G280800, related to fungal resistance; Sobic.003G203500, related to response to Anthracnose; Sobic.005G033801, and Sobic.006G248300, related to resistance to sorghum aphid. Height genes found in dry weight GWAS were Sobic.006G067600, and Sobic.006G067700, related to dwarf loci. In this work we used two different models, a linear model with fixed effects and a mixed linear model. Both of them utilized the same chromosomal regions as explanatory variables; however, for the mixed linear model we added two terms to the Hamiltonian that allowed us to free the accession variances and convert them into additional parameters to represent genetic variance components. The result was a better filter for important chromosomal regions affecting the complex trait (refer to Venn diagrams in Figure 3.11 - 3.13) The mixed linear model showed a higher filtering capacity, achieving a reduction of 90 percent in the final number of chromosomal regions for dry weight, and 99 percent for disease.

The use of MINE challenges the way experiments are done in precision agriculture on the home turf of the subject of experimental design [10]. The focus of the MINE approach is on discovery of relationships in large data sets rather

than the precision of effects in the experimental design. The parameter space is large and potentially could involves 10^5 effects where there are only on the order of 10^3 observations. This is a common problem in systems biology. The solution to this problem is not only using ensemble methods to address the p > n problem, but also having an adaptive approach that combines both intelligent data collection and the use of the model to guide future experiments. Model-guided discovery then leads to relationships with the underlying genes that can be tested in the context of classical experimental design, once the relationships between the complex trait and genes are found. The solution proposed here for GWAS is only one example of where ensemble methods and MINE are critical for the new omics experiments in genetics.

ACKNOWLEDGEMENTS. This work was supported by DOE DE-SC0021386

Chapter 4

Limitations of the Scope of Work in this thesis and future work

4.1 Random effects on the chromosomal regions

The mixed linear model described in this work considers the random effects by accessions on various plant health indicators, such as log dry weight, however, the random effects by chromosomal regions are not considered . A future work may involve generating mixed linear models that consider both types of random effects or only the one over the chromosomal regions. This change implies updating the implementation of the customized Metropolis algorithm by considering more variables to estimate as well as modifying the Hamiltonian, there is also a possibility that the Boltzmann probability function needs an update as well.

4.2 Large field experiment comparison

This work addresses an adaptive GWAS over 3 years, with a purpose of overcoming the lack of resources to process a large amount of plants without a decrease in accuracy. Part of a future work is to compare the adaptive GWAS approach (using the MINE) with a large GWAS experiment using all of the BAP panel in the same field. To do the comparison the data from the large experiment will have to be processed using the tools developed in this work. It will be of interest to know whether the MINE guided procedure outperforms the classical design using the entire BAP panel.

4.3 Projection method in mixed linear model

In this work a data projection was implemented to remove noise from the resulting set of parameters not well specified by the data on the complex trait. This method was only implemented for the linear model with fixed effects because the mixed linear model is nonlinear in form for the variance components in the Hamiltonian used to guide estimation of model parameters. Therefore, a future work would be to come up with a method to remove noise on parameters from the mixed linear model, or avoid high fluctuations during modeling.

4.4 GWAS on AMF colonization data

Arnold lab has an ongoing project that measures AMF colonization in Sorghum roots. Plants from this work are part of that project as well. A future work would examine the AMF colonization data in a GWAS using the tools developed in this work as well as the MINE approach.

4.5 Inclusion of field treatment in mixed linear model

This work didn't consider field treatment with phosphorus and nitrogen in the adaptive experiment over three years;. Future work might consider treatment as a new variable of the mixed linear model; this would imply to include additional parameters in the modeling phase, and update the customized Metropolis algorithm. The Bennetzen lab has a project involving field treatment; this data can be considered in the proposal.

4.6 Generation of synthetic data to validate this work

The lack of workforce to plant and harvest is a main constraint in this work. A way to further compare the approaches presented here is to generate synthetic data on the computer and run the tools to evaluate the procedures developed here. Future work might consider both generating synthetic data to evaluate the new methodologies developed here.

Bibliography

[1] Wubei Dong, Xiaojia Tang, Yihai Yu, Roger Nilsen, Rosemary Kim, James Griffith, Jonathan Arnold, and H. Bernd Schüttler. System biology of the clock in neurospora crassa. *PLoS ONE*, 3, 2008. ISSN 19326203. doi: 10.1371/journal.pone.0003105.

[2] Zachary W. Brenton, Elizabeth A. Cooper, Mathew T. Myers, Richard E. Boyles, Nadia Shakoor, Kelsey J. Zielinski, Bradley L. Rauh, William C. Bridges, Geoffrey P. Morris, and Stephen Kresovich. A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. *Genetics*, 204, 2016. ISSN 19432631. doi: 10.1534/genetics.115.183947.

[3] Claudia Neuhauser and Joseph E. Fargione. A mutualism–parasitism continuum model and its application to plant–mycorrhizae interactions. *Ecological Modelling*, 177(3):337–352, 2004. ISSN 0304-3800. doi: https://doi.org/10.1016/j.ecolmodel.2004.02.010. URL `https://www.sciencedirect.com/science/article/pii/S0304380004001450`.

[4] Heberle H., Meirelles G. V., da Silva F. R., and Telles G. P. Interactivenn: a web-based tool for the analysis of sets through venn diagrams. *BMC Bioinformatics*, 16, 2015. doi: 10.1186/s12859-015-0611-3.

[5] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies, 2021. ISSN 26628449.

[6] D. Battogtokh, D. K. Asch, M. E. Case, J. Arnold, and H. B. Schuttler. An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of neurospora crassa. *Proc Natl Acad Sci U S A*, 99(26):16904–9, 2002. ISSN 0027-8424 (Print). URL `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=`

Retrieve&db=PubMed&dopt=Citation&list_uids=12477937. Journal Article United States.

[7] David Landau and Kurt Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2021.

[8] Zhenbin Hu, Marcus O. Olatoye, Sandeep Marla, and Geoffrey P. Morris. An integrated genotyping-by-sequencing polymorphism map for over 10,000 sorghum genotypes. *The Plant Genome*, 12, 2019. ISSN 1940-3372. doi: 10.3835/plantgenome2018.06.0044.

[9] Ronald Aylmer Fisher. *The design of experiments*. Oliver and Boyd, Edinburgh, London„ 1935. 36004973 incl. tables. 23 cm.

[10] R.A. Fisher. *The design of experiments*. Oliver and Boyd, 5 edition, 1949.

[11] Peter William Meredith John. *Statistical design and analysis of experiments*. Macmillan, New York„ 1971. 75121672 [by] Peter W. M. John. 24 cm. Bibliography: p. 339-350.

[12] Jesse R. Lasky, Hari D. Upadhyaya, Punna Ramu, Santosh Deshpande, C. Tom Hash, Jason Bonnette, Thomas E. Juenger, Katie Hyma, Charlotte Acharya, Sharon E. Mitchell, Edward S. Buckler, Zachary Brenton, Stephen Kresovich, and Geoffrey P. Morris. Genome-environment associations in sorghum landraces predict adaptive traits. *Science Advances*, 1(6):e1400218, 2015. doi: 10.1126/sciadv.1400218. URL http://advances.sciencemag.org/content/1/6/e1400218.abstract.

[13] Gabriel Castrillo, Paulo José Pereira Lima Teixeira, Sur Herrera Paredes, Theresa F. Law, Laura De Lorenzo, Meghan E. Feltcher, Omri M. Finkel, Natalie W. Breakfield, Piotr Mieczkowski, Corbin D. Jones, Javier Paz-Ares, and Jeffery L. Dangl. Root microbiota drive direct integration of phosphate stress and immunity. *Nature*, 543, 2017. ISSN 14764687. doi: 10.1038/nature21417.

[14] Yanqing Chen, Jun Zhu, Pek Yee Lum, Xia Yang, Shirly Pinto, Douglas J. MacNeil, Chunsheng Zhang, John Lamb, Stephen Edwards, Solveig K. Sieberts, Amy Leonardson, Lawrence W. Castellini, Susanna Wang, Marie France Champy, Bin Zhang, Valur Emilsson, Sudheer Doss, Anatole Ghazalpour, Steve Horvath, Thomas A. Drake, Aldons J. Lusis, and Eric E. Schadt. Variations in dna elucidate molecular networks that cause disease. *Nature*, 452, 2008. ISSN 14764687. doi: 10.1038/nature06757.

[15] T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292(5518):929–934, 2001. ISSN 0036-8075. URL <GotoISI>://000168514900044. 429KB Times Cited:887 Cited References Count:29.

[16] Nancy Collins Johnson, Gail W.T. Wilson, Jacqueline A. Wilson, R. Michael Miller, and Matthew A. Bowker. Mycorrhizal phenotypes and the law of the minimum. *New Phytologist*, 205, 2015. ISSN 14698137. doi: 10.1111/nph.13172.

[17] Cheng Gao, Liliam Montoya, Ling Xu, Mary Madera, Joy Hollingsworth, Elizabeth Purdom, Robert B. Hutmacher, Jeffery A. Dahlberg, Devin Coleman-Derr, Peggy G. Lemaux, and John W. Taylor. Strong succession in arbuscular mycorrhizal fungal communities. *The ISME Journal*, 13(1):214–226, 2019. ISSN 1751-7370. doi: 10.1038/s41396-018-0264-0. URL https://doi.org/10.1038/s41396-018-0264-0.

[18] Nancy Collins Johnson, Jason D. Hoeksema, James D. Bever, V. Bala Chaudhary, Catherine Gehring, John Klironomos, Roger Koide, R. Michael Miller, John Moore, Peter Moutoglis, Mark Schwartz, Suzanne Simard, William Swenson, James Umbanhowar, Gail Wilson, and Catherine Zabinski. From lilliput to brobdingnag: Extending models of mycorrhizal function across scales. *BioScience*, 56(11):889–900, 2006. ISSN 0006-3568. doi: 10.1641/0006-3568(2006)56[889:FLTBEM]2.0.CO;2. URL https://doi.org/10.1641/0006-3568(2006)56[889:FLTBEM]2.0.CO;2.

[19] Amanda M. Bouffier, Jonathan Arnold, and H. Bernd Schüttler. A mine alternative to d-optimal designs for the linear model. *PLOS ONE*, 9(10):e110234, 2014. doi: 10.1371/journal.pone.0110234. URL https://doi.org/10.1371/journal.pone.0110234.

[20] Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U. Eliasen, Yunxuan Jiang, Sridharan Raghavan, Jenkai Miao, Joshua D. Arias, Sarah E. Graham, Ronen E. Mukamel, Cassandra N. Spracklen, Xianyong Yin, Shyh Huei Chen, Teresa Ferreira, Heather H. Highland, Yingjie Ji, Tugce Karaderi, Kuang Lin, Kreete Lüll, Deborah E. Malden, Carolina Medina-Gomez, Moara Machado, Amy Moore, Sina Rüeger, Xueling

Sim, Scott Vrieze, Tarunveer S. Ahluwalia, Masato Akiyama, Matthew A. Allison, Marcus Alvarez, Mette K. Andersen, Alireza Ani, Vivek Appadurai, Liubov Arbeeva, Seema Bhaskar, Lawrence F. Bielak, Sailalitha Bollepalli, Lori L. Bonnycastle, Jette Bork-Jensen, Jonathan P. Bradfield, Yuki Bradford, Peter S. Braund, Jennifer A. Brody, Kristoffer S. Burgdorf, Brian E. Cade, Hui Cai, Qiuyin Cai, Archie Campbell, Marisa Cañadas-Garre, Eulalia Catamo, Jin Fang Chai, Xiaoran Chai, Li Ching Chang, Yi Cheng Chang, Chien Hsiun Chen, Alessandra Chesi, Seung Hoan Choi, Ren Hua Chung, Massimiliano Cocca, Maria Pina Concas, Christian Couture, Gabriel Cuellar-Partida, Rebecca Danning, E. Warwick Daw, Frauke Degenhard, Graciela E. Delgado, Alessandro Delitala, Ayse Demirkan, Xuan Deng, Poornima Devineni, Alexander Dietl, Maria Dimitriou, Latchezar Dimitrov, Rajkumar Dorajoo, Arif B. Ekici, Jorgen E. Engmann, Zammy Fairhurst-Hunter, Aliki Eleni Farmaki, Jessica D. Faul, Juan Carlos Fernandez-Lopez, Lukas Forer, Margherita Francescatto, Sandra Freitag-Wolf, Christian Fuchsberger, Tessel E. Galesloot, Yan Gao, Zishan Gao, Frank Geller, Olga Giannakopoulou, Franco Giulianini, Anette P. Gjesing, Anuj Goel, Scott D. Gordon, Mathias Gorski, Jakob Grove, Xiuqing Guo, Stefan Gustafsson, Jeffrey Haessler, Thomas F. Hansen, Aki S. Havulinna, Simon J. Haworth, Jing He, Nancy Heard-Costa, Prashantha Hebbar, George Hindy, Yuk Lam A. Ho, Edith Hofer, Elizabeth Holliday, Katrin Horn, Whitney E. Hornsby, Jouke Jan Hottenga, Hongyan Huang, Jie Huang, Alicia Huerta-Chagoya, Jennifer E. Huffman, Yi Jen Hung, Shaofeng Huo, Mi Yeong Hwang, Hiroyuki Iha, Daisuke D. Ikeda, Masato Isono, Anne U. Jackson, Susanne Jäger, Iris E. Jansen, Ingegerd Johansson, Jost B. Jonas, Anna Jonsson, Torben Jørgensen, Ioanna Panagiota Kalafati, Masahiro Kanai, Stavroula Kanoni, Line L. Kårhus, Anuradhani Kasturiratne, Tomohiro Katsuya, Takahisa Kawaguchi, Rachel L. Kember, Katherine A. Kentistou, Han Na Kim, Young Jin Kim, Marcus E. Kleber, Maria J. Knol, Azra Kurbasic, Marie Lauzon, Phuong Le, Rodney Lea, Jong Young Lee, Hampton L. Leonard, Shengchao A. Li, Xiaohui Li, Xiaoyin Li, Jingjing Liang, Honghuang Lin, Shih Yi Lin, Jun Liu, Xueping Liu, Ken Sin Lo, Jirong Long, Laura Lores-Motta, Jian'an Luan, Valeriya Lyssenko, Leo Pekka Lyytikäinen, Anubha Mahajan, Vasiliki Mamakou, Massimo Mangino, Ani Manichaikul, Jonathan Marten, Manuel Mattheisen, Laven Mavarani, Aaron F. McDaid, Karina Meidtner, Tori L. Melendez, Josep M. Mercader, Yuri Milaneschi, Jason E. Miller, Iona Y. Millwood, Pashupati P. Mishra, Ruth E. Mitchell, Line T.

Møllehave, Anna Morgan, Soeren Mucha, Matthias Munz, Masahiro Nakatochi, Christopher P. Nelson, Maria Nethander, Chu Won Nho, Aneta A. Nielsen, Ilja M. Nolte, Suraj S. Nongmaithem, Raymond Noordam, Ioanna Ntalla, Teresa Nutile, Anita Pandit, Paraskevi Christofidou, Katri Pärna, Marc Pauper, Eva R.B. Petersen, Liselotte V. Petersen, Niina Pitkänen, Ozren Polašek, Alaitz Poveda, Michael H. Preuss, Saiju Pyarajan, Laura M. Raffield, Hiromi Rakugi, Julia Ramirez, Asif Rasheed, Dennis Raven, Nigel W. Rayner, Carlos Riveros, Rebecca Rohde, Daniela Ruggiero, Sanni E. Ruotsalainen, Kathleen A. Ryan, Maria Sabater-Lleal, Richa Saxena, Markus Scholz, Anoop Sendamarai, Botong Shen, Jingchunzi Shi, Jae Hun Shin, Carlo Sidore, Colleen M. Sitlani, Roderick C. Slieker, Roelof A.J. Smit, Albert V. Smith, Jennifer A. Smith, Laura J. Smyth, Lorraine Southam, Valgerdur Steinthorsdottir, Liang Sun, Fumihiko Takeuchi, Divya Sri Priyanka Tallapragada, Kent D. Taylor, Bamidele O. Tayo, Catherine Tcheandjieu, Natalie Terzikhan, Paola Tesolin, Alexander Teumer, Elizabeth Theusch, Deborah J. Thompson, Gudmar Thorleifsson, Paul R.H.J. Timmers, Stella Trompet, Constance Turman, Simona Vaccargiu, Sander W. van der Laan, Peter J. van der Most, Jan B. van Klinken, Jessica van Setten, Shefali S. Verma, Niek Verweij, Yogasudha Veturi, Carol A. Wang, Chaolong Wang, Lihua Wang, Zhe Wang, Helen R. Warren, Wen Bin Wei, Ananda R. Wickremasinghe, Matthias Wielscher, Kerri L. Wiggins, Bendik S. Winsvold, Andrew Wong, Yang Wu, Matthias Wuttke, Rui Xia, Tian Xie, Ken Yamamoto, Jingyun Yang, Jie Yao, Hannah Young, Noha A. Yousri, Lei Yu, Lingyao Zeng, Weihua Zhang, Xinyuan Zhang, Jing Hua Zhao, Wei Zhao, Wei Zhou, Martina E. Zimmermann, Magdalena Zoledziewska, Linda S. Adair, Hieab H.H. Adams, Carlos A. Aguilar-Salinas, Fahd Al-Mulla, Donna K. Arnett, Folkert W. Asselbergs, Bjørn Olav Åsvold, John Attia, Bernhard Banas, Stefania Bandinelli, David A. Bennett, Tobias Bergler, Dwaipayan Bharadwaj, Ginevra Biino, Hans Bisgaard, Eric Boerwinkle, Carsten A. Böger, Klaus Bønnelykke, Dorret I. Boomsma, Anders D. Børglum, Judith B. Borja, Claude Bouchard, Donald W. Bowden, Ivan Brandslund, Ben Brumpton, Julie E. Buring, Mark J. Caulfield, John C. Chambers, Giriraj R. Chandak, Stephen J. Chanock, Nish Chaturvedi, Yii Der Ida Chen, Zhengming Chen, Ching Yu Cheng, Ingrid E. Christophersen, Marina Ciullo, John W. Cole, Francis S. Collins, Richard S. Cooper, Miguel Cruz, Francesco Cucca, L. Adrienne Cupples, Michael J. Cutler, Scott M. Damrauer, Thomas M. Dantoft, Gert J. de Borst,

Lisette C.P.G.M. de Groot, Philip L. De Jager, Dominique P.V. de Kleijn, H. Janaka de Silva, George V. Dedoussis, Anneke I. den Hollander, Shufa Du, Douglas F. Easton, Petra J.M. Elders, A. Heather Eliassen, Patrick T. Ellinor, Sölve Elmståhl, Jeanette Erdmann, Michele K. Evans, Diane Fatkin, Bjarke Feenstra, Mary F. Feitosa, Luigi Ferrucci, Ian Ford, Myriam Fornage, Andre Franke, Paul W. Franks, Barry I. Freedman, Paolo Gasparini, Christian Gieger, Giorgia Girotto, Michael E. Goddard, Yvonne M. Golightly, Clicerio Gonzalez-Villalpando, Penny Gordon-Larsen, Harald Grallert, Struan F.A. Grant, Niels Grarup, Lyn Griffiths, Vilmundur Gudnason, Christopher Haiman, Hakon Hakonarson, Torben Hansen, Catharina A. Hartman, Andrew T. Hattersley, Caroline Hayward, Susan R. Heckbert, Chew Kiat Heng, Christian Hengstenberg, Alex W. Hewitt, Haretsugu Hishigaki, Carel B. Hoyng, Paul L. Huang, Wei Huang, Steven C. Hunt, Kristian Hveem, Elina Hyppönen, William G. Iacono, Sahoko Ichihara, M. Arfan Ikram, Carmen R. Isasi, Rebecca D. Jackson, Marjo Riitta Jarvelin, Zi Bing Jin, Karl Heinz Jöckel, Peter K. Joshi, Pekka Jousilahti, J. Wouter Jukema, Mika Kähönen, Yoichiro Kamatani, Kui Dong Kang, Jaakko Kaprio, Sharon L.R. Kardia, Fredrik Karpe, Norihiro Kato, Frank Kee, Thorsten Kessler, Amit V. Khera, Chiea Chuen Khor, Lambertus A.L.M. Kiemeney, Bong Jo Kim, Eung Kweon Kim, Hyung Lae Kim, Paulus Kirchhof, Mika Kivimaki, Woon Puay Koh, Heikki A. Koistinen, Genovefa D. Kolovou, Jaspal S. Kooner, Charles Kooperberg, Anna Köttgen, Peter Kovacs, Adriaan Kraaijeveld, Peter Kraft, Ronald M. Krauss, Meena Kumari, Zoltan Kutalik, Markku Laakso, Leslie A. Lange, Claudia Langenberg, Lenore J. Launer, Loic Le Marchand, Hyejin Lee, Nanette R. Lee, Terho Lehtimäki, Huaixing Li, Liming Li, Wolfgang Lieb, Xu Lin, Lars Lind, Allan Linneberg, Ching Ti Liu, Jianjun Liu, Markus Loeffler, Barry London, Steven A. Lubitz, Stephen J. Lye, David A. Mackey, Reedik Mägi, Patrik K.E. Magnusson, Gregory M. Marcus, Pedro Marques Vidal, Nicholas G. Martin, Winfried März, Fumihiko Matsuda, Robert W. McGarrah, Matt McGue, Amy Jayne McKnight, Sarah E. Medland, Dan Mellström, Andres Metspalu, Braxton D. Mitchell, Paul Mitchell, Dennis O. Mook-Kanamori, Andrew D. Morris, Lorelei A. Mucci, Patricia B. Munroe, Mike A. Nalls, Saman Nazarian, Amanda E. Nelson, Matt J. Neville, Christopher Newton-Cheh, Christopher S. Nielsen, Markus M. Nöthen, Claes Ohlsson, Albertine J. Oldehinkel, Lorena Orozco, Katja Pahkala, Päivi Pajukanta, Colin N.A. Palmer, Esteban J. Parra, Cristian Pattaro, Oluf Pedersen, Craig E. Pennell, Brenda W.J.H.

Penninx, Louis Perusse, Annette Peters, Patricia A. Peyser, David J. Porteous, Danielle Posthuma, Chris Power, Peter P. Pramstaller, Michael A. Province, Qibin Qi, Jia Qu, Daniel J. Rader, Olli T. Raitakari, Sarju Ralhan, Loukianos S. Rallidis, Dabeeru C. Rao, Susan Redline, Dermot F. Reilly, Alexander P. Reiner, Sang Youl Rhee, Paul M. Ridker, Michiel Rienstra, Samuli Ripatti, Marylyn D. Ritchie, Dan M. Roden, Frits R. Rosendaal, Jerome I. Rotter, Igor Rudan, Femke Rutters, Charumathi Sabanayagam, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Dharambir K. Sanghera, Naveed Sattar, Börge Schmidt, Helena Schmidt, Reinhold Schmidt, Matthias B. Schulze, Heribert Schunkert, Laura J. Scott, Rodney J. Scott, Peter Sever, Eric J. Shiroma, M. Benjamin Shoemaker, Xiao Ou Shu, Eleanor M. Simonsick, Mario Sims, Jai Rup Singh, Andrew B. Singleton, Moritz F. Sinner, J. Gustav Smith, Harold Snieder, Tim D. Spector, Meir J. Stampfer, Klaus J. Stark, David P. Strachan, Leen M. 't Hart, Yasuharu Tabara, Hua Tang, Jean Claude Tardif, Thangavel A. Thanaraj, Nicholas J. Timpson, Anke Tönjes, Angelo Tremblay, Tiinamaija Tuomi, Jaakko Tuomilehto, Maria Teresa Tusié-Luna, Andre G. Uitterlinden, Rob M. van Dam, Pim van der Harst, Nathalie Van der Velde, Cornelia M. van Duijn, Natasja M. van Schoor, Veronique Vitart, Uwe Völker, Peter Vollenweider, Henry Völzke, Niels H. Wacher-Rodarte, Mark Walker, Ya Xing Wang, Nicholas J. Wareham, Richard M. Watanabe, Hugh Watkins, David R. Weir, Thomas M. Werge, Elisabeth Widen, Lynne R. Wilkens, Gonneke Willemsen, Walter C. Willett, James F. Wilson, Tien Yin Wong, Jeong Taek Woo, Alan F. Wright, Jer Yuarn Wu, Huichun Xu, Chittaranjan S. Yajnik, Mitsuhiro Yokota, Jian Min Yuan, Eleftheria Zeggini, Babette S. Zemel, Wei Zheng, Xiaofeng Zhu, Joseph M. Zmuda, Alan B. Zonderman, John Anker Zwart, Gabriel Cuellar Partida, Yan Sun, Damien Croteau-Chonka, Judith M. Vonk, Loic Le Marchand, Daniel I. Chasman, Yoon Shin Cho, Iris M. Heid, Mark I. McCarthy, Maggie C.Y. Ng, Christopher J. O'Donnell, Fernando Rivadeneira, Unnur Thorsteinsdottir, Yan V. Sun, E. Shyong Tai, Michael Boehnke, Panos Deloukas, Anne E. Justice, Cecilia M. Lindgren, Ruth J.F. Loos, Karen L. Mohlke, Kari E. North, Kari Stefansson, Robin G. Walters, Thomas W. Winkler, Kristin L. Young, Po Ru Loh, Jian Yang, Tõnu Esko, Themistocles L. Assimes, Adam Auton, Goncalo R. Abecasis, Cristen J. Willer, Adam E. Locke, Sonja I. Berndt, Guillaume Lettre, Timothy M. Frayling, Yukinori Okada, Andrew R. Wood, Peter M. Visscher, and Joel N. Hirschhorn. A saturated map of common ge-

netic variants associated with human height. *Nature*, 610, 2022. ISSN 14764687. doi: 10.1038/s41586-022-05275-y.

[21] D. P. Landau, K. Binder, D. P. Landau, and K. Binder. Monte carlo simulations at the periphery of physics and beyond. *A Guide to Monte Carlo Simulations in Statistical Physics; Cambridge University Press: New York, NY, USA*, pages 13–22, 2014.

[22] E. A. Guggenheim. *Boltzmann's distribution law*. Series in physics. North-Holland Pub. Co.; Interscience Publishers, Amsterdam, New York„ 1955. 56058095 (Edward Armand), 19 cm.

[23] Reginald L. McGee and Gregery T. Buzzard. Maximally informative next experiments for nonlinear models. *Mathematical Biosciences*, 302, 2018. ISSN 18793134. doi: 10.1016/j.mbs.2018.04.007.

[24] John A. Cornell. *Experiments with mixtures : designs, models, and the analysis of mixture data*. Wiley series in probability and statistics. Wiley, New York, 3rd edition, 2002. ISBN 0471393673 (cloth alk. paper). URL `Contributorbiographicalinformationhttp://www.loc.gov/catdir/bios/wiley042/2001045457.htmlPublisherdescriptionhttp://www.loc.gov/catdir/description/wiley035/2001045457.htmlTableofcontentshttp://www.loc.gov/catdir/toc/wiley021/2001045457.html`. 2001045457 John A. Cornell. ill. ; 25 cm. "A Wiley-Interscience publication." Includes bibliographical references and index.

[25] Y. Yu, W. Dong, C. Altimus, X. Tang, J. Griffith, M. Morello, L. Dudek, J. Arnold, and H. B. Schuttler. A genetic network for the clock of neurospora crassa. *Proc Natl Acad Sci USA*, 104(8):2809–14, 2007. ISSN 0027-8424 (Print). URL `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17301235`. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States.

[26] A. Al-Omari, J. Griffith, C. Caranica, T. Taha, H. Schüttler, and J. Arnold. Discovering regulators in post-transcriptional control of the biological clock of neurospora crassa using variable topology ensemble methods on gpus. *IEEE Access*, 6:54582–54594, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2871876.

[27] Ahmad M. Al-Omari, James Griffith, Ashley Scruse, Robert W. Robinson, Heinz Bernd Schuttler, and Jonathan Arnold. Ensemble methods for identifying rna operons and regulons in the clock network of neurospora crassa. *IEEE Access*, 10, 2022. ISSN 21693536. doi: 10.1109/ACCESS.2022.3160481.

[28] R. A. Fisher and Edward John Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922. doi: 10.1098/rsta.1922.0009. URL https://doi.org/10.1098/rsta.1922.0009.

[29] Leonard J. Savage. *The foundations of statistics*. Dover Publications, New York„ 2d rev. edition, 1972. ISBN 0486623491. URL Publisherdescriptionhttp://www.loc.gov/catdir/description/dover032/79188245.htmlTableofcontentsonlyhttp://www.loc.gov/catdir/enhancements/fy1318/79188245-t.html. 79188245 [by] Leonard J. Savage. illus. 22 cm. Includes bibliographies.

[30] Anita J. Antoninka, Mark E. Ritchie, and Nancy C. Johnson. The hidden serengeti—mycorrhizal fungi respond to environmental gradients. *Pedobiologia*, 58(5):165–176, 2015. ISSN 0031-4056. doi: https://doi.org/10.1016/j.pedobi.2015.08.001. URL http://www.sciencedirect.com/science/article/pii/S0031405615300056.

[31] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 1953. ISSN 00219606. doi: 10.1063/1.1699114.

[32] George D. Mostow and Joseph H. Sampson. *Linear algebra*. International series in pure and applied mathematics. McGraw-Hill, New York„ 1969. 68055422 [by] George D. Mostow [and] Joseph H. Sampson. illus. 23 cm.

[33] Wyatt W. Anderson, Jonathan Arnold, Scott A. Sammons, and Darrell G. Yardley. Frequency-dependent viabilities of drosophila pseudoobscura karyotypes. *Heredity*, 56(1):7–17, 1986. ISSN 1365-2540. doi: 10.1038/hdy.1986.2. URL https://doi.org/10.1038/hdy.1986.2.

[34] John A. Cornell. Experiments with mixtures: A review. *Technometrics*, 15(3):437–455, 1973. ISSN 0040-1706. doi: 10.1080/00401706.1973.10489071. URL https://amstat.tandfonline.com/doi/abs/10.1080/00401706.1973.10489071.

[35] Adam B. Cobb, Gail W. T. Wilson, Carla L. Goad, Scott R. Bean, Rhett C. Kaufman, Thomas J. Herald, and Jeff D. Wilson. The role of arbuscular mycorrhizal fungi in grain production and nutrition of sorghum genotypes: Enhancing sustainability through plant-microbial partnership. *Agriculture, Ecosystems and Environment*, 233: 432–440, 2016. ISSN 0167-8809. doi: https://doi.org/10.1016/j.agee.2016.09.024. URL https://www.sciencedirect.com/science/article/pii/S0167880916304753.

[36] Stephanie J. Watts-Williams, Bryan D. Emmett, Veronique Levesque-Tremblay, Allyson M. MacLean, Xuepeng Sun, James W. Satterlee, Zhangjun Fei, and Maria J. Harrison. Diverse sorghum bicolor accessions show marked variation in growth and transcriptional responses to arbuscular mycorrhizal fungi. *Plant, Cell and Environment*, 42(5):1758–1774, 2019. ISSN 0140-7791. doi: https://doi.org/10.1111/pce.13509. URL https://doi.org/10.1111/pce.13509. https://doi.org/10.1111/pce.13509.

[37] Callie R. Chappell and Tadashi Fukami. Nectar yeasts: a natural microcosm for ecology. *Yeast*, 35(6):417–423, 2018. ISSN 0749-503X. doi: https://doi.org/10.1002/yea.3311. URL https://doi.org/10.1002/yea.3311. https://doi.org/10.1002/yea.3311.

[38] Yongjun Liu, Nancy Collins Johnson, Lin Mao, Guoxi Shi, Shengjing Jiang, Xiaojun Ma, Guozhen Du, Lizhe An, and Huyuan Feng. Phylogenetic structure of arbuscular mycorrhizal community shifts in response to increasing soil fertility. *Soil Biology and Biochemistry*, 89: 196–205, 2015. ISSN 0038-0717. doi: https://doi.org/10.1016/j.soilbio.2015.07.007. URL https://www.sciencedirect.com/science/article/pii/S0038071715002412.

[39] Shengjing Jiang, Yongjun Liu, Jiajia Luo, Mingsen Qin, Nancy Collins Johnson, Maarja Öpik, Martti Vasar, Yuxing Chai, Xiaolong Zhou, Lin Mao, Guozhen Du, Lizhe An, and Huyuan Feng. Dynamics of arbuscular mycorrhizal fungal community structure and functioning along a nitrogen enrichment gradient in an alpine meadow ecosystem. *New Phytologist*, 220(4):1222–1235, 2018. ISSN 0028-646X. doi: https://doi.org/

10.1111/nph.15112. URL https://doi.org/10.1111/nph.15112. https://doi.org/10.1111/nph.15112.

[40] Daniel Revillini, Catherine A. Gehring, and Nancy Collins Johnson. The role of locally adapted mycorrhizas and rhizobacteria in plant–soil feedback systems. *Functional Ecology*, 30(7):1086–1098, 2016. ISSN 0269-8463. doi: https://doi.org/10.1111/1365-2435.12668. URL https://doi.org/10.1111/1365-2435.12668. https://doi.org/10.1111/1365-2435.12668.

[41] Subhuti Dharmananda. *Studies of the circadian clock of Neurospora crassa: light-induced phase shifting*. University of California, 1980.

[42] T. P. McGONIGLE, M. H. MILLER, D. G. EVANS, G. L. FAIRCHILD, and J. A. SWAN. A new method which gives an objective measure of colonization of roots by vesicular—arbuscular mycorrhizal fungi. *New Phytologist*, 115, 1990. ISSN 14698137. doi: 10.1111/j.1469-8137.1990.tb00476.x.

[43] Pasquale De Vita, Luciano Avio, Cristiana Sbrana, Giovanni Laidò, Daniela Marone, Anna M. Mastrangelo, Luigi Cattivelli, and Manuela Giovannetti. Genetic markers associated to arbuscular mycorrhizal colonization in durum wheat. *Scientific Reports*, 8, 2018. ISSN 20452322. doi: 10.1038/s41598-018-29020-6.

[44] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997. ISSN 0036-8075. URL <GotoISI>://A1997YC32300053. Yc323 Times Cited:2537 Cited References Count:41.

[45] J. C. Dunlap. Molecular bases for circadian clocks. *Cell*, 96(2):271–90, 1999. ISSN 0092-8674 (Print). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9988221. GM 34985/GM/NIGMS NIH HHS/United States MH011186/MH/NIMH NIH HHS/United States MH44651/MH/NIMH NIH HHS/United States Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. Review United states.

[46] S. K. Crosthwaite, J. C. Dunlap, and J. J. Loros. Neurospora wc-1 and wc-2: transcription, photoresponses, and the origins of

circadian rhythmicity. *Science*, 276(5313):763–9, 1997. ISSN 0036-8075 (Print). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9115195. GM 34985/GM/NIGMS NIH HHS/United States MH01186/MH/NIMH NIH HHS/United States MH44651/MH/NIMH NIH HHS/United States Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. United states.

[47] C. R. McClung, B. A. Fox, and J. C. Dunlap. The neurospora clock gene frequency shares a sequence element with the drosophila clock gene period. *Nature*, 339(6225):558–62, 1989. ISSN 0028-0836 (Print). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2525233. Journal Article Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S. England.

[48] Kathleen K. Treseder and Michael F. Allen. Direct nitrogen and phosphorus limitation of arbuscular mycorrhizal fungi: a model and field test. *New Phytologist*, 155(3):507–515, 2002. ISSN 0028-646X. doi: https://doi.org/10.1046/j.1469-8137.2002.00470.x. URLhttps://doi.org/10.1046/j.1469-8137.2002.00470.x. https://doi.org/10.1046/j.1469-8137.2002.00470.x.

[49] Jeffrey Propster and Nancy Johnson. Uncoupling the effects of phosphorus and precipitation on arbuscular mycorrhizas in the serengeti. *Plant and Soil*, pages 21–34, 2015. doi: 10.1007/s11104-014-2369-1.

[50] Cline Jouffe, Gaspard Cretenet, Laura Symul, Eva Martin, Florian Atger, Felix Naef, and Frdric Gachon. The circadian clock coordinates ribosome biogenesis. *Plos Biology*, 11(1):e1001455, 2013. ISSN 1545-7885.

[51] A Al-Omari, J. Griffith, M Judge, T. R. Taha, J. Arnold, and H Schuttler. Discovering regulatory network topologies using ensemble methods on gpgpus with special reference to the biological clock of neurospora crassa. *IEEE Access*, 3:27–42, 2015.

[52] Rajanikanth Govindarajulu, Ashley N. Hostetler, Yuguo Xiao, Srinivasa R. Chaluvadi, Margarita Mauro-Herrera, Muriel L. Siddoway, Clinton Whipple, Jeffrey L. Bennetzen, Katrien M. Devos, Andrew N. Doust, and Jennifer S. Hawkins. Integration of high-density genetic mapping with transcriptome analysis uncovers numerous agro-

nomic qtl and reveals candidate genes for the control of tillering in sorghum. *G3 Genes|Genomes|Genetics*, 11(2), 2021. ISSN 2160-1836. doi: 10.1093/g3journal/jkab024. URL https://doi.org/10.1093/g3journal/jkab024.

[53] R. W. Doerge, Z. B. Zeng, and B. S. Weir. Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science*, 12(3):195–219, 1997. ISSN 08834237. URL http://www.jstor.org/stable/2246370.

[54] Ronald Aylmer Fisher. *The genetical theory of natural selection*. Dover books on science. Dover Publications, New York„ 2d rev. edition, 1958. 58013362 illus. 21 cm. Includes bibliography.

[55] R. A. Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. ISSN 0080-4568. doi: 10.1017/S0080456800012163. URL https://www.cambridge.org/core/article/xvthe-correlation-between-relatives-on-the-supposition-of-mendelian-inheritance/A60675052E0FB78C561F66C670BC75DE.

[56] Shufan Zhang, Yue Wu, Michael Skaro, Jia-Hwei Cheong, Amanda Bouffier-Landrum, Isaac Torrres, Yinping Guo, Lauren Stupp, Brooke Lincoln, Anna Prestel, et al. Computer vision models enable mixed linear modeling to predict arbuscular mycorrhizal fungal colonization using fungal morphology. *Scientific Reports*, 14(1):10866, 2024.

[57] Maurice G. Kendall, Alan Stuart, J. K. Ord, and Anthony O'Hagan. *Kendall's advanced theory of statistics*. Edward Arnold ; Halsted Press, London New York, 6th edition, 1994. ISBN 0340614307 (v. 1 Edward Arnold) 047023380X (v. 1 Halsted Press) 0340529229 (v. 2B Edward Arnold). 94188490 Alan Stuart, J. Keith Ord. ill. ; 26 cm. Vol. 2B: 1st ed., 1994; 2d ed., 2004. Includes bibliographical references and indexes. v. 1. Distribution theory – v. 2A. Classical inference and the linear model / Alan Stuart ... [et. al.] – v. 2B. Bayesian inference / Anthony O'Hagan.

[58] Anita Antoninka, Julie E. Wolf, Matthew Bowker, AimÉE T. Classen, and Nancy Collins Johnson. Linking above- and belowground responses to global change at community and ecosystem scales. *Global Change Biology*, 15(4):914–929, 2009. ISSN 1354-1013. doi: 10.1111/j.1365-2486.2008.01760.x. URL https://doi.org/10.1111/j.1365-2486.2008.01760.x.

[59] Zhiwu Zhang, Elhan Ersoz, Chao Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, Jianming Yu, Donna K. Arnett, Jose M. Ordovas, and Edward S. Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42, 2010. ISSN 10614036. doi: 10.1038/ng.546.

[60] Alberto Romagnoni, Simon Jégou, Kristel Van Steen, Gilles Wainrib, Jean Pierre Hugot, Laurent Peyrin-Biroulet, Mathias Chamaillard, Jean Frederick Colombel, Mario Cottone, Mauro D'Amato, Renata D'Incà, Jonas Halfvarson, Paul Henderson, Amir Karban, Nicholas A. Kennedy, Mohammed Azam Khan, Marc Lémann, Arie Levine, Dunecan Massey, Monica Milla, Sok Meng Evelyn Ng, Ioannis Oikonomou, Harald Peeters, Deborah D. Proctor, Jean Francois Rahier, Paul Rutgeerts, Frank Seibold, Laura Stronati, Kirstin M. Taylor, Leif Törkvist, Kullak Ublick, Johan Van Limbergen, Andre Van Gossum, Morten H. Vatn, Hu Zhang, Wei Zhang, Jane M. Andrews, Peter A. Bampton, Murray Barclay, Timothy H. Florin, Richard Gearry, Krupa Krishnaprasad, Ian C. Lawrance, Gillian Mahy, Grant W. Montgomery, Graham Radford-Smith, Rebecca L. Roberts, Lisa A. Simms, Katherine Hanigan, Anthony Croft, Leila Amininijad, Isabelle Cleynen, Olivier Dewit, Denis Franchimont, Michel Georges, Debby Laukens, Harald Peeters, Jean Francois Rahier, Paul Rutgeerts, Emilie Theatre, André Van Gossum, Severine Vermeire, Guy Aumais, Leonard Baidoo, Arthur M. Barrie, Karen Beck, Edmond Jean Bernard, David G. Binion, Alain Bitton, Steve R. Brant, Judy H. Cho, Albert Cohen, Kenneth Croitoru, Mark J. Daly, Lisa W. Datta, Colette Deslandres, Richard H. Duerr, Debra Dutridge, John Ferguson, Joann Fultz, Philippe Goyette, Gordon R. Greenberg, Talin Haritunians, Gilles Jobin, Seymour Katz, Raymond G. Lahaie, Dermot P. McGovern, Linda Nelson, Sok Meng Ng, Kaida Ning, Ioannis Oikonomou, Pierre Paré, Deborah D. Proctor, Miguel D. Regueiro, John D. Rioux, Elizabeth Ruggiero, L. Philip Schumm, Marc Schwartz, Regan Scott, Yashoda Sharma, Mark S. Silverberg, Denise Spears, A. Hillary Steinhart, Joanne M. Stempak, Jason M. Swoger, Constantina Tsagarelis, Wei Zhang, Clarence Zhang, Hongyu Zhao, Jan Aerts, Tariq Ahmad, Hazel Arbury, Anthony Attwood, Adam Auton, Stephen G. Ball, Anthony J. Balmforth, Chris Barnes, Jeffrey C. Barrett, Inês Barroso, Anne Barton, Amanda J. Bennett, Sanjeev Bhaskar, Katarzyna Blaszczyk, John Bowes, Oliver J. Brand, Peter S. Braund, Francesca Bredin, Gerome Breen, Morris J. Brown, Ian N. Bruce, Jaswinder Bull, Oliver S. Burren, John Burton, Jake Byrnes,

Sian Caesar, Niall Cardin, Chris M. Clee, Alison J. Coffey, John MC Connell, Donald F. Conrad, Jason D. Cooper, Anna F. Dominiczak, Kate Downes, Hazel E. Drummond, Darshna Dudakia, Andrew Dunham, Bernadette Ebbs, Diana Eccles, Sarah Edkins, Cathryn Edwards, Anna Elliot, Paul Emery, David M. Evans, Gareth Evans, Steve Eyre, Anne Farmer, I. Nicol Ferrier, Edward Flynn, Alistair Forbes, Liz Forty, Jayne A. Franklyn, Timothy M. Frayling, Rachel M. Freathy, Eleni Giannoulatou, Polly Gibbs, Paul Gilbert, Katherine Gordon-Smith, Emma Gray, Elaine Green, Chris J. Groves, Detelina Grozeva, Rhian Gwilliam, Anita Hall, Naomi Hammond, Matt Hardy, Pile Harrison, Neelam Hassanali, Husam Hebaishi, Sarah Hines, Anne Hinks, Graham A. Hitman, Lynne Hocking, Chris Holmes, Eleanor Howard, Philip Howard, Joanna M.M. Howson, Debbie Hughes, Sarah Hunt, John D. Isaacs, Mahim Jain, Derek P. Jewell, Toby Johnson, Jennifer D. Jolley, Ian R. Jones, Lisa A. Jones, George Kirov, Cordelia F. Langford, Hana Lango-Allen, G. Mark Lathrop, James Lee, Kate L. Lee, Charlie Lees, Kevin Lewis, Cecilia M. Lindgren, Meeta Maisuria-Armer, Julian Maller, John Mansfield, Jonathan L. Marchini, Paul Martin, Dunecan Co Massey, Wendy L. McArdle, Peter McGuffin, Kirsten E. McLay, Gil McVean, Alex Mentzer, Michael L. Mimmack, Ann E. Morgan, Andrew P. Morris, Craig Mowat, Patricia B. Munroe, Simon Myers, William Newman, Elaine R. Nimmo, Michael C. O'Donovan, Abiodun Onipinla, Nigel R. Ovington, Michael J. Owen, Kimmo Palin, Aarno Palotie, Kirstie Parnell, Richard Pearson, David Pernet, John Rb Perry, Anne Phillips, Vincent Plagnol, Natalie J. Prescott, Inga Prokopenko, Michael A. Quail, Suzanne Rafelt, Nigel W. Rayner, David M. Reid, Anthony Renwick, Susan M. Ring, Neil Robertson, Samuel Robson, Ellie Russell, David St Clair, Jennifer G. Sambrook, Jeremy D. Sanderson, Stephen J. Sawcer, Helen Schuilenburg, Carol E. Scott, Richard Scott, Sheila Seal, Sue Shaw-Hawkins, Beverley M. Shields, Matthew J. Simmonds, Debbie J. Smyth, Elilan Somaskantharajah, Katarina Spanova, Sophia Steer, Jonathan Stephens, Helen E. Stevens, Kathy Stirrups, Millicent A. Stone, David P. Strachan, Zhan Su, Deborah P.M. Symmons, John R. Thompson, Wendy Thomson, Martin D. Tobin, Mary E. Travers, Clare Turnbull, Damjan Vukcevic, Louise V. Wain, Mark Walker, Neil M. Walker, Chris Wallace, Margaret Warren-Perry, Nicholas A. Watkins, John Webster, Michael N. Weedon, Anthony G. Wilson, Matthew Woodburn, B. Paul Wordsworth, Chris Yau, Allan H. Young, Eleftheria Zeggini, Matthew A. Brown, Paul R. Burton, Mark J. Caulfield, Alastair

Compston, Martin Farrall, Stephen C.L. Gough, Alistair S. Hall, Andrew T. Hattersley, Adrian V.S. Hill, Christopher G. Mathew, Marcus Pembrey, Jack Satsangi, Michael R. Stratton, Jane Worthington, Matthew E. Hurles, Audrey Duncanson, Willem H. Ouwehand, Miles Parkes, Nazneen Rahman, John A. Todd, Nilesh J. Samani, Dominic P. Kwiatkowski, Mark I. McCarthy, Nick Craddock, Panos Deloukas, Peter Donnelly, Jenefer M. Blackwell, Elvira Bramon, Juan P. Casas, Aiden Corvin, Janusz Jankowski, Hugh S. Markus, Colin Na Palmer, Robert Plomin, Anna Rautanen, Richard C. Trembath, Ananth C. Viswanathan, Nicholas W. Wood, Chris C.A. Spencer, Gavin Band, Céline Bellenguez, Colin Freeman, Garrett Hellenthal, Eleni Giannoulatou, Matti Pirinen, Richard Pearson, Amy Strange, Hannah Blackburn, Suzannah J. Bumpstead, Serge Dronov, Matthew Gillman, Alagurevathi Jayakumar, Owen T. McCann, Jennifer Liddle, Simon C. Potter, Radhi Ravindrarajah, Michelle Ricketts, Matthew Waller, Paul Weston, Sara Widaa, and Pamela Whittaker. Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data. *Scientific Reports*, 9, 2019. ISSN 20452322. doi: 10.1038/s41598-019-46649-z.

[61] G. Evelyn Hutchinson. *An introduction to population ecology*. Yale University Press, New Haven, 1978. ISBN 0300021550. 77011005 (George Evelyn), G. Evelyn Hutchinson. ill. ; 28 cm. Includes bibliographical references and indexes.

[62] G. F. Gause and A. A. Witt. Behavior of mixed populations and the problem of natural selection. *The American Naturalist*, 69(725):596–609, 1935. ISSN 0003-0147. doi: 10.1086/280628. URL `https://doi.org/10.1086/280628`.

[63] Nancy Collins Johnson, Diane L. Rowland, Lea Corkidi, Louise M. Egerton-Warburton, and Edith B. Allen. Nitrogen enrichment alters mycorrhizal allocation at five mesic to semiarid grasslands. *Ecology*, 84(7):1895–1908, 2003. ISSN 0012-9658. doi: https://doi.org/10.1890/0012-9658(2003)084[1895:NEAMAA]2.0.CO;2. URL `https://doi.org/10.1890/0012-9658(2003)084[1895:NEAMAA]2.0.CO;2`. https://doi.org/10.1890/0012-9658(2003)084[1895:NEAMAA]2.0.CO;2.

[64] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P. H. Stumpf. Approximate bayesian computation scheme for parameter

inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009. doi: 10.1098/rsif.2008.0172. URL https://doi.org/10.1098/rsif.2008.0172.

[65] Alan E. Gelfand and Adrian F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 1990. ISSN 1537274X. doi: 10.1080/01621459.1990.10476213.

[66] E. T. Jaynes and Oscar Kempthorne. *Confidence Intervals vs Bayesian Intervals*. 1976. doi: 10.1007/978-94-010-1436-6_6.

[67] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57, 1995. ISSN 1369-7412. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[68] David M. Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, and Daniel S. Rokhsar. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40, 2012. ISSN 03051048. doi: 10.1093/nar/gkr944.

[69] Bruce S Weir. *Genetic data analysis. Methods for discrete population genetic data.* 1990.

[70] Jianliang Dai, Li Li, Sangkyu Kim, Beth Kimball, S. Michal Jazwinski, and Jonathan Arnold. Exact sample size needed to detect dependence in 2 × 2 × 2 tables. *Biometrics*, 63, 2007. doi: https://doi.org/10.1111/j.1541-0420.2007.00801.x.

[71] J. D. Liedtke, C. H. Hunt, B. George-Jaeggli, K. Laws, J. Watson, A. B. Potgieter, A. Cruickshank, and D. R. Jordan. High-throughput phenotyping of dynamic canopy traits associated with stay-green in grain sorghum. *Plant Phenomics*, 2020, 2020. doi: 10.34133/2020/4635153. URL https://spj.science.org/doi/abs/10.34133/2020/4635153.

[72] Ephrem Habyarimana, Paolo De Franceschi, Sezai Ercisli, Faheem Shehzad Baloch, and Michela Dall'Agata. Genome-wide association study for biomass related traits in a panel of sorghum bicolor and s. bicolor × s. halepense populations. *Frontiers in Plant Science*, 11, 2020. ISSN 1664462X. doi: 10.3389/fpls.2020.551305.

[73] Wilson Kimani, Li Min Zhang, Xiao Yuan Wu, Huai Qing Hao, and Hai Chun Jing. Genome-wide association study reveals that different pathways contribute to grain quality variation in sorghum (sorghum bicolor). *BMC Genomics*, 21, 2020. ISSN 14712164. doi: 10.1186/s12864-020-6538-8.

[74] Si Xu, Xiu Qing Li, Hong Guo, Xiao Yuan Wu, Ning Wang, Zhi Quan Liu, Huai Qing Hao, and Hai Chun Jing. Mucilage secretion by aerial roots in sorghum (sorghum bicolor): sugar profile, genetic diversity, gwas and transcriptomic analysis. *Plant Molecular Biology*, 112, 2023. ISSN 15735028. doi: 10.1007/s11103-023-01365-1.

[75] Richard E. Boyles, Elizabeth A. Cooper, Matthew T. Myers, Zachary Brenton, Bradley L. Rauh, Geoffrey P. Morris, and Stephen Kresovich. Genome-wide association studies of grain yield components in diverse sorghum germplasm. *The Plant Genome*, 9, 2016. ISSN 1940-3372. doi: 10.3835/plantgenome2015.09.0091.

[76] Neeraj Kumar, J. Lucas Boatwright, Richard E. Boyles, Zachary W. Brenton, and Stephen Kresovich. Identification of pleiotropic loci mediating structural and non-structural carbohydrate accumulation within the sorghum bioenergy association panel using high-throughput markers. *Frontiers in Plant Science*, 15, 2024. ISSN 1664462X. doi: 10.3389/fpls.2024.1356619.

[77] Lihua Wang, Yanlong Liu, Li Gao, Xiaocui Yang, Xu Zhang, Shaoping Xie, Meng Chen, Yi Hong Wang, Jieqin Li, and Yixin Shen. Identification of candidate forage yield genes in sorghum (sorghum bicolor l.) using integrated genome-wide association studies and rna-seq. *Frontiers in Plant Science*, 12, 2022. ISSN 1664462X. doi: 10.3389/fpls.2021.788433.

[78] Feng Luo, Zhongyou Pei, Xiongwei Zhao, Huifen Liu, Yiwei Jiang, and Shoujun Sun. Genome-wide association study for plant architecture and bioenergy traits in diverse sorghum and sudangrass germplasm. *Agronomy*, 10, 2020. ISSN 20734395. doi: 10.3390/agronomy10101602.

[79] Ezekiel Ahn, Jacob Botkin, Vishnutej Ellur, Yoonjung Lee, Kabita Poudel, Louis K. Prom, and Clint Magill. Genome-wide association study of seed morphology traits in senegalese sorghum cultivars. *Plants*, 12, 2023. ISSN 22237747. doi: 10.3390/plants12122344.

[80] Liyi Zhang, Jianxia Xu, Yanqing Ding, Ning Cao, Xu Gao, Zhou Feng, Kuiying Li, Bing Cheng, Lengbo Zhou, Mingjian Ren, Yuezhi Tao, and

Guihua Zou. Gwas of grain color and tannin content in chinese sorghum based on whole-genome sequencing. *Theoretical and Applied Genetics*, 136, 2023. ISSN 14322242. doi: 10.1007/s00122-023-04307-z.

[81] Jacinta Kavuluko, Magdaline Kibe, Irine Sugut, Willy Kibet, Joel Masanga, Sylvia Mutinda, Mark Wamalwa, Titus Magomere, Damaris Odeny, and Steven Runo. Gwas provides biological insights into mechanisms of the parasitic plant (striga) resistance in sorghum. *BMC Plant Biology*, 21, 2021. ISSN 14712229. doi: 10.1186/s12870-021-03155-7.

[82] Masarat Elias, Diriba Chere, Dagnachew Lule, Desalegn Serba, Alemu Tirfessa, Dandena Gelmesa, Tesfaye Tesso, Kassahun Bantte, and Temesgen M. Menamo. Multi-locus genome-wide association study reveal genomic regions underlying root system architecture traits in ethiopian sorghum germplasm. *Plant Genome*, 17, 2024. ISSN 19403372. doi: 10.1002/tpg2.20436.

[83] Ezekiel Ahn, Louis K. Prom, and Clint Magill. Multi-trait genome-wide association studies of sorghum bicolor regarding resistance to anthracnose, downy mildew, grain mold and head smut. *Pathogens*, 12, 2023. ISSN 20760817. doi: 10.3390/pathogens12060779.

[84] Lihua Wang, Wenmiao Tu, Peng Jin, Yanlong Liu, Junli Du, Jiacheng Zheng, Yi-Hong Wang, and Jieqin Li. Genome-wide association study of plant color in sorghum bicolor. *Frontiers in Plant Science*, 15, 2024. ISSN 1664-462X. doi: 10.3389/fpls.2024.1320844. URL `https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1320844`.

[85] Lidong Wang, Li Shang, Xiaoyuan Wu, Huaiqing Hao, and Hai Chun Jing. Genomic architecture of leaf senescence in sorghum (sorghum bicolor). *Theoretical and Applied Genetics*, 136, 2023. ISSN 14322242. doi: 10.1007/s00122-023-04315-z.

[86] Dinakaran Elango, Weiya Xue, and Surinder Chopra. Genome wide association mapping of epi-cuticular wax genes in sorghum bicolor. *Physiology and Molecular Biology of Plants*, 26, 2020. ISSN 09740430. doi: 10.1007/s12298-020-00848-5.

[87] Muluken Enyew, Tileye Feyissa, Anders S. Carlsson, Kassahun Tesfaye, Cecilia Hammenhag, Amare Seyoum, and Mulatu Geleta. Genome-wide analyses using multi-locus models revealed marker-trait associations

for major agronomic traits in sorghum bicolor. *Frontiers in Plant Science*, 13, 2022. ISSN 1664462X. doi: 10.3389/fpls.2022.999692.

[88] Chunming Bai, Chunyu Wang, Ping Wang, Zhenxing Zhu, Ling Cong, Dan Li, Yifei Liu, Wenjing Zheng, and Xiaochun Lu. Qtl mapping of agronomically important traits in sorghum (sorghum bicolor l.). *Euphytica*, 213, 2017. ISSN 15735060. doi: 10.1007/s10681-017-2075-1.

[89] Chenyong Miao, Yuhang Xu, Sanzhen Liu, Patrick S. Schnable, and James C. Schnable. Increased power and accuracy of causal locus identification in time series genome-wide association in sorghum1[open]. *Plant Physiology*, 183, 2020. ISSN 15322548. doi: 10.1104/pp.20.00277.

[90] Anthony Wenndt, Richard Boyles, Arlyn Ackerman, Sirjan Sapkota, Ace Repka, and Rebecca Nelson. Host determinants of fungal species composition and symptom manifestation in the sorghum grain mold disease complex. *Plant Disease*, 107(2):315–325, 2023. doi: 10.1094/PDIS-03-22-0675-RE. URL `https://doi.org/10.1094/PDIS-03-22-0675-RE`. PMID: 36800304.

[91] Sirjan Sapkota, J. Lucas Boatwright, Kathleen Jordan, Richard Boyles, and Stephen Kresovich. Identification of novel genomic associations and gene candidates for grain starch content in sorghum. *Genes*, 11, 2020. ISSN 20734425. doi: 10.3390/genes11121448.

[92] Louis K. Prom, Hugo E. Cuevas, Ezekiel Ahn, Thomas Isakeit, William L. Rooney, and Clint Magill. Genome-wide association study of grain mold resistance in sorghum association panel as affected by inoculation with alternaria alternata alone and alternaria alternata, fusarium thapsinum, and curvularia lunata combined. *European Journal of Plant Pathology*, 157, 2020. ISSN 15738469. doi: 10.1007/s10658-020-02036-3.

[93] Gezahegn Girma, Habte Nida, Amare Seyoum, Moges Mekonen, Amare Nega, Dagnachew Lule, Kebede Dessalegn, Alemnesh Bekele, Adane Gebreyohannes, Adedayo Adeyanju, Alemu Tirfessa, Getachew Ayana, Taye Taddese, Firew Mekbib, Ketema Belete, Tesfaye Tesso, Gebisa Ejeta, and Tesfaye Mengiste. A large-scale genome-wide association analyses of ethiopian sorghum landrace collection reveal loci associated with important traits. *Frontiers in Plant Science*, 10, 2019. ISSN 1664462X. doi: 10.3389/fpls.2019.00691.

[94] Juan S. Panelo, Yin Bao, Lie Tang, Patrick S. Schnable, and Maria G. Salas-Fernandez. Genetics of canopy architecture dynamics in

photoperiod-sensitive and photoperiod-insensitive sorghum. *Plant Phenome Journal*, 7, 2024. ISSN 25782703. doi: 10.1002/ppj2.20092.

[95] Habte Nida, Gezahegn Girma, Moges Mekonen, Alemu Tirfessa, Amare Seyoum, Tamirat Bejiga, Chemeda Birhanu, Kebede Dessalegn, Tsegau Senbetay, Getachew Ayana, Tesfaye Tesso, Gebisa Ejeta, and Tesfaye Mengiste. Genome-wide association analysis reveals seed protein loci as determinants of variations in grain mold resistance in sorghum. *Theoretical and Applied Genetics*, 134, 2021. ISSN 14322242. doi: 10.1007/s00122-020-03762-2.

[96] Hugo E. Cuevas, Louis K. Prom, Elizabeth A. Cooper, Joseph E. Knoll, and Xinzhi Ni. Genome-wide association mapping of anthracnose (colletotrichum sublineolum) resistance in the u.s. sorghum association panel. *The Plant Genome*, 11, 2018. ISSN 1940-3372. doi: 10.3835/plantgenome2017.11.0099.

[97] Clara Cruet-Burgos, Sarah Cox, Brian P. Ioerger, Ramasamy Perumal, Zhenbin Hu, Thomas J. Herald, Scott R. Bean, and Davina H. Rhodes. Advancing provitamin a biofortification in sorghum: Genome-wide association studies of grain carotenoids in global germplasm. *Plant Genome*, 13, 2020. ISSN 19403372. doi: 10.1002/tpg2.20013.

[98] Ezekiel Ahn, Louis K. Prom, Zhenbin Hu, Gary Odvody, and Clint Magill. Genome-wide association analysis for response of senegalese sorghum accessions to texas isolates of anthracnose. *Plant Genome*, 14, 2021. ISSN 19403372. doi: 10.1002/tpg2.20097.

[99] Niranjan Ravindra Thakur, Sunita Gorthy, Anilkumar Vemula, Damaris A. Odeny, Pradeep Ruperao, Pramod Ramchandra Sargar, Shivaji Pandurang Mehtre, Hirakant V. Kalpande, and Ephrem Habyarimana. Genome-wide association study and expression of candidate genes for fe and zn concentration in sorghum grains. *Scientific Reports*, 14, 2024. doi: https://doi.org/10.1038/s41598-024-63308-0.

[100] Somashekhar M. Punnuri, Addissu G. Ayele, Karen R. Harris-Shultz, Joseph E. Knoll, Alisa W. Coffin, Haile K. Tadesse, J. Scott Armstrong, Trahmad K. Wiggins, Hanxia Li, Scott Sattler, and Jason G. Wallace. Genome-wide association mapping of resistance to the sorghum aphid in sorghum bicolor. *Genomics*, 114, 2022. ISSN 10898646. doi: 10.1016/j.ygeno.2022.110408.

[101] Erica Agnew, Greg Ziegler, Scott Lee, César Lizárraga, Noah Fahlgren, Ivan Baxter, Todd C. Mockler, and Nadia Shakoor. Longitudinal genome-wide association study reveals early qtl that predict biomass accumulation under cold stress in sorghum. *Frontiers in Plant Science*, 15, 2024. ISSN 1664-462X. doi: 10.3389/fpls.2024.1278802. URL https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1278802.

[102] Emily S.A. Wolf, Saddie Vela, Jennifer Wilker, Alyssa Davis, Madalen Robert, Valentina Infante, Rafael E. Venado, Cătălin Voiniciuc, Jean Michel Ané, and Wilfred Vermerris. Identification of genetic and environmental factors influencing aerial root traits that support biological nitrogen fixation in sorghum. *G3: Genes, Genomes, Genetics*, 14, 2024. ISSN 21601836. doi: 10.1093/g3journal/jkad285.

[103] Lihua Wang, Hari D. Upadhyaya, Jian Zheng, Yanlong Liu, Shailesh Kumar Singh, C. L.L. Gowda, Rajendra Kumar, Yongqun Zhu, Yi Hong Wang, and Jieqin Li. Genome-wide association mapping identifies novel panicle morphology loci and candidate genes in sorghum. *Frontiers in Plant Science*, 12, 2021. ISSN 1664462X. doi: 10.3389/fpls.2021.743838.

[104] Ezekiel Ahn, Coumba Fall, Louis K. Prom, and Clint Magill. Genome-wide association study of senegalese sorghum seedlings responding to a texas isolate of colletotrichum sublineola. *Scientific Reports*, 12, 2022. ISSN 20452322. doi: 10.1038/s41598-022-16844-6.

[105] Jeffrey P. Townsend and Christoph Leuenberger. Taxon Sampling and the Optimal Rates of Evolution for Phylogenetic Inference. *Systematic Biology*, 60(3):358–365, 02 2011. ISSN 1063-5157. doi: 10.1093/sysbio/syq097. URL https://doi.org/10.1093/sysbio/syq097.

[106] Jeffrey P. Townsend and Francesc Lopez-Giraldez. Optimal Selection of Gene and Ingroup Taxon Sampling for Resolving Phylogenetic Relationships. *Systematic Biology*, 59(4):446–457, 05 2010. ISSN 1063-5157. doi: 10.1093/sysbio/syq025. URL https://doi.org/10.1093/sysbio/syq025.