

AN ANALYSIS OF GENETIC VARIATION  
IN COMPLEX TRAITS OF MAIZE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Jason Andrew Peiffer

August 2012

© 2012 Jason Andrew Peiffer

# AN ANALYSIS OF GENETIC VARIATION IN COMPLEX TRAITS OF MAIZE

Jason Andrew Peiffer, Ph. D.

Cornell University 2012

Maize (*Zea mays* L.) is a complex crop. Governed by the universal processes of evolution that dictate the differential reproduction of all life, maize germplasm has been gradually adapted to better suit societal needs through domestication and breeding. However, these modifications were largely accomplished with little knowledge of the genetic architecture or molecular mechanics of its traits. Investigating the reaches of the rhizosphere to the top of the tassel, the following studies analyze the natural variation of complex maize traits to better understand both their means and degree of inheritance.

First, the heritability and environmental specificity of maize-microbe interactions were estimated by pyrosequence profiling 16s rRNA gene amplicons from rhizosphere bacterial populations of diverse inbreds grown in multiple maize field environments. We found substantial variation in bacterial diversity was attributable to environment. Nonetheless, a small but significant proportion of variation was heritable. While kinship inferred from a simple additive model assuming contributions from all polymorphisms did not explain this heritable variation, its discovery is a step toward identifying those genes responsible for novel plant-microbe interactions in natural environments.

Second, maize stalk strength variation was analyzed to delineate the accuracy of genomic prediction in a low heritability trait. While few robust loci were associated with stalk strength, a

significant proportion of heritable variation was captured by kinship among the inbreds. This revealed the efficacy of genomic prediction and suggested the potential to accurately predict other low heritability phenotypes such as yield. These and similar efforts to facilitate the selection of genotyped seed with desirable qualities before planting will enhance breeding efficiency.

Finally, variation in the most classic and heritable of complex traits, maize height was partitioned to reveal its genetic architecture and pleiotropy with other traits such as flowering time and node counts. As anticipated height was highly polygenic and well captured by kinship; however, an interesting finding was the lacking concordance between mapped loci and those established through previous cloning efforts. Equally intriguing was the paucity of pleiotropic loci identified for height and flowering time. These findings reveal the potential for independent evolvability of these traits during maize breeding.

## BIOGRAPHICAL SKETCH

Jason Andrew Peiffer began his career as an orchard hand working high school summers for Bill Serfass at Prydenjoy Farms Ltd. of Allentown, Pennsylvania. It was during this time he came to appreciate the abundance of diverse crop species he had the privilege to help manage and harvest. Depending upon the season, numerous tree varieties of apples, peaches, pears, plums, apricots, nectarines, and cherries could be found blooming or fruiting across the orchard. Similarly, the smell of strawberries filled the orchard each June and blueberries, raspberries, currants, and various melon crops were ripe for the picking later each summer. Working with these numerous species each year led him to possess an affinity for understanding the intricacies of plant biology and development. His interest in the molecular genetics of plants was further spurred by biotechnological advancements in the development of transgenic crops that were popularly debated at the time.

In pursuit of this interest, Jason enrolled at the University of Delaware's College of Agriculture and Natural Resources and undertook courses in the major of plant science with minors in the fields of biochemical engineering and biochemistry. However, it was the encouragement and mentoring he obtained from four years of undergraduate research assistantship in the molecular genetics of *Arabidopsis thaliana* at the Delaware Biotechnology Institute with Dr. Blake Meyers, and Dr. Kan Nobuta, and later by assisting Dr. James Hawk with maize fieldwork at the College of Agriculture and Natural Resources that inspired him to pursue a plant genetics research career. After obtaining an honors bachelor's degree with distinction in plant science, he decided to step away from the lab bench, get back to working

outside, and expand his genetic, computational, and statistical skills in the fields of plant breeding, quantitative genetics, and high-throughput crop genomics.

He was admitted to the Department of Plant Breeding and Genetics at Cornell University's College of Agriculture and Life Sciences wherein he joined the USDA-ARS laboratory of Dr. Edward S. Buckler. During his tenure at Cornell, he worked primarily on dissecting the genetic architecture of the complex traits described within this dissertation. Chief empirical exploits included marker development and the fine mapping of a quantitative trait locus for height on the long arm of chromosome nine in two diverse near isogenic lines (NILs) for over 10,000 F<sub>2</sub> progeny. The engineering and development of a rapid method to phenotype rind penetrometer resistance or stalk strength, and the application of that method across three environments of a nested association mapping population (NAM) of approximately 5,000 F<sub>2</sub>:S<sub>5</sub> recombinant inbred lines (RILs). Principal analytical endeavors included development of an automated pipeline for partitioning the phenotypic variance and generating best linear unbiased predictors (BLUPs) from a multi-population, multi-environment field trial using Java and ASREML, coding a bootstrapping routine for joint-linkage quantitative trait loci (QTL) mapping in SAS, developing an R pipeline for ordination and multivariate statistical analysis of the maize and microbial data sets. In addition, he developed a query program for identifying association proximal to candidate genes from an SQL database with Java as well as a primer design program for constructing molecular markers to aid fine mapping efforts from the first generation maize HapMap. He also used the extensive code resources developed by others in the lab, most notably Drs. Peter J. Bradbury, Edward S. Buckler, and Zhiwu Zhang, as well as Terry Casstevens, for the analysis of genetic variation in approximately fifteen complex traits measured in the NAM panel across multiple field environments.

I gratefully dedicate this dissertation to my parents, Dennis and Lynne Peiffer, who provided me with both the opportunities that accompany a college education and the loving support needed during my further pursuit of knowledge.

## ACKNOWLEDGMENTS

I am much indebted to the many individuals who have assumed the roles of advisor, mentor, colleague, and friend in helping to make both this dissertation and the considerable empirical and analytical efforts essential to its construction a possibility over the past five years. I am especially grateful for my advisor **Dr. Edward S. Buckler's** scientific and professional insights throughout my graduate experience at Cornell University. I remain inspired by the efficiency of the extensive infrastructure that has been developed within our lab. The manner in which our laboratory collects phenotypic and genotypic data, manages that data, and statistically summarizes it remains the envy of many other genetic research groups, and fosters an exceptionally diverse learning environment. None of this would be possible without Ed's ambitious and clever guidance and the skill, creativity, and competence of past and present members in the Buckler laboratory and the Institute for Genomic Diversity.

I extend great thanks to the empiricists who provided all the required technical assistance and shared their extensive field experience in collection of the phenotypic data required for all of these analyses. I would especially like to acknowledge our field manager **Nick K. Lepak** for his tireless help in ensuring proper seed preparation and storage, planting, phenotyping, and harvesting. Thanks are also due to the numerous collaborators from Cornell and other institutions that through similar perspiration and persistence provided the phenotypic data necessary to dissect maize microbial, height, and stalk strength genetic architectures with the exceptional precision which was obtained. These include the laboratories of **Dr. Margaret E. Smith** at Cornell University and **Dr. Stephen P. Moose** at University of Illinois for their assistance in planting multiple field environments for our analysis of maize-microbe interactions. For assistance in dissecting the genetic architecture of plant height, I thank the labs of **Dr.**

**Candice A.C. Gardner** and **Dr. Mark J. Millard** at Iowa State University, **Dr. Sherry Flint-Garcia** and **Dr. Michael D. McMullen** at the University of Missouri, **Dr. James B. Holland** at North Carolina State University, and **Dr. John F. Doebley** at the University of Wisconsin. Assistance in the planting and management of plots used in measuring stalk strength was also provided by **Dr. Sherry Flint-Garcia** of University of Missouri and **Dr. Natalia De Leon** of the University of Wisconsin. Sincere gratitude for insights in the construction of the phenotyping apparatus constructed to measure stalk strength or rind penetrometer resistance are due to **Dr. Sherry Flint-Garcia** for providing earlier versions of the apparatus, **Dennis J. Peiffer** for guidance on the fabrication and mechanical aspects which improved the device, and **Dallas E. Kroon** for sharing his programming experiences and providing suggestions which aided in collection of the phenotypic data. Thanks are also due to my colleague **Sara J. Larsson**, and **Drs. Cinta Romay, Feng Tian, Nengyi Zhang, Denise E. Costich, and Elhan Ersoz**, as well as the numerous undergraduates, most notably **Sarah S. Asman**, who over the years have helped to make seed packing, planting, phenotyping, and harvesting less daunting.

For quality preparation and production of the pyrosequence data required in dissecting the genetic architecture of the maize-microbe interactions, thanks are due to **Drs. Ayme Spor, Ruth E. Ley, Jeffrey Dangel**, and those laboratories at the U.S. Department of Energy who performed the requisite sequencing. Without the impressive throughput and magnitude of sequence data developed through construction of HapMap Generation 1, 2, and more recently the Genotyping-by-Sequencing pipeline the extensive analysis of the genetic architecture of maize height and pleiotropically related complex traits would not have been possible. For their efforts in these endeavors, I would like to express my gratitude to **Robert J. Elshire, Drs. Sharon E. Mitchell, Dr. Michael A. Gore**, and the members of the Institute for Genomic Diversity. For

assistance and camaraderie during the trials of molecular marker generation for our fine mapping efforts, I thank **Dr. Moira J. Sheehan**.

Given the massive amounts of genotypic and phenotypic data generated by both aforementioned empirical groups, many thanks are in order for the bioinformatics and statistical genetics groups with whom I have had the privilege to work extensively. **Dr. Peter J. Bradbury** provided his admirable experience and numerous programs aiding our mapping of genotypes to phenotypes. Without his assistance the following analyses would never have been possible. **Dr. James B. Holland** provided exceptional insights into the use and theory behind fitting the expansive models and the experimental designs needed to control for spatial variation and the effects of environment across our numerous field trials. Also thanks are due to the entire bioinformatics team that made Genotyping-By-Sequencing the success it has become. This includes **Drs. Jer-Ming Chia, Jeff C. Glaubitz, Qi Sun, and James Harriman**, as well as many others who have helped along the way. Useful discussions regarding statistics with **Dr. Zhiwu Zhang** and **Dr. Alex E. Lipka** have also been helpful in further understanding the underlying mechanics and analysis of genetic data.

Last but not least, I extend appreciation to my committee **Dr. John C. Schimenti** for Genomics and **Dr. Thomas P. Brutnell** for Plant Molecular Biology. It was a pleasure working with **Dr. John C. Schimenti** as a teaching assistant for his BioGD 4000 Genomics course. Although my research interests have slightly strayed from the application of methods in plant molecular genetics to exploring population and quantitative genetics, biometry, and bioinformatics, I am very grateful for the committee's continued input at our annual meetings.

## TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	i
DEDICATION	iii
ACKNOWLEDGMENTS	iv
TABLE OF CONTENTS	vii
LIST OF TABLES	viii
LIST OF FIGURES	ix
<b>CHAPTER 1 PARTITIONING THE GENETIC VARIATION IN COMPLEX TRAITS OF PLANTS</b>	<b>1</b>
EVOLUTIONARY DYNAMICS OF PLANT POPULATIONS	1
GENETIC ARCHITECTURE OF COMPLEX TRAITS IN PLANTS	11
GENOMIC PREDICTION OF COMPLEX TRAITS IN PLANTS	19
REFERENCES	23
<b>CHAPTER 2 BIODIVERSITY AND HERITABILITY OF THE MAIZE RHIZOSPHERE MICROBIOME UNDER FIELD CONDITIONS</b>	<b>31</b>
ABSTRACT	31
INTRODUCTION	32
RESULTS	36
DISCUSSION	55
MATERIALS AND METHODS	59
REFERENCES	65
<b>CHAPTER 3 MAPPING AND PREDICTION OF MAIZE STALK STRENGTH IN A NESTED ASSOCIATION MAPPING POPULATION</b>	<b>67</b>
ABSTRACT	67
INTRODUCTION	68
MATERIALS AND METHODS	71
RESULTS	77
DISCUSSION	91
REFERENCES	96
<b>CHAPTER 4 MAPPING THE GENETIC ARCHITECTURE OF MAIZE HEIGHT AND CORRELATED COMPLEX TRAITS</b>	<b>99</b>
ABSTRACT	99
INTRODUCTION	101
MATERIALS AND METHODS	105
RESULTS	110
DISCUSSION	122
REFERENCES	128

## LIST OF TABLES

<b>Table 2.1</b> Maize rhizosphere and bulk soil sample sequence characteristics and proportion of Green Genes classifiable sequences obtained by pyrosequencing	38
<b>Table 2.2</b> Soil characteristic of surveyed field environments based on ICP-AES, modified Mehlich buffer testing, and organic matter by loss on ignition	54
<b>Table 3.1</b> Heritability of stalk strength, flowering time, and ear height in the maize nested association mapping panel	80
<b>Table 4.1</b> Heritability of height in the maize nested association mapping panel	111
<b>Table 4.2</b> Heritability of height in the Ames maize inbred diversity panel	116
<b>Table 4.3</b> Co-localization of joint-linkage-assisted GWAS and published maize height candidate genes	121

## LIST OF FIGURES

<b>Figure 2.1</b> Proportion of 16S rRNA classifiable pyrosequence reads obtained from maize rhizosphere or bulk soil samples	39
<b>Figure 2.2</b> Rarefaction curves of the deepest microbiome extractions within study remained nearly linear revealing extensive diversity yet to be surveyed.	41
<b>Figure 2.3</b> Rarefaction curves of alpha diversity revealed significant variation between the rhizosphere and bulk soil microbiome extractions.	43
<b>Figure 2.4</b> Substantial maize inbred by environment interactions were observed for rhizosphere microbiome profiles.	44
<b>Figure 2.5</b> Unweighted UniFrac beta diversity revealed substantial differentiation between microbiome extractions.	47
<b>Figure 2.6</b> Weighted UniFrac beta diversity largely corroborated results of the unweighted UniFrac metric.	48
<b>Figure 2.7</b> Preston's log normal models of abundance data reveal substantial proportions of veiled OTU as compared to observed OTU richness.	50
<b>Figure 2.8</b> Whittaker's rank-abundance curves were plotted for the median abundance of each OTU across microbiome extractions.	51
<b>Figure 2.9</b> Chloroplast enrichment profile across microbiome samples in each sample type, environment, and maize inbred.	53
<b>Figure 3.1</b> Apparatus constructed for measuring rind penetrometer resistance	73
<b>Figure 3.2</b> Distribution of estimated breeding values for stalk strength	78
<b>Figure 3.3</b> Variation for stalk strength in the maize NAM panel	81
<b>Figure 3.4</b> Correlation of stalk strength in the maize NAM panel and diversity panel	81
<b>Figure 3.5</b> Heat map of QTL for strength, flowering time, and ear height chromosomes one to five	84
<b>Figure 3.6</b> Heat map of QTL for strength, flowering time, and ear height chromosomes six to ten	85

<b>Figure 3.7</b> Genomic prediction of stalk strength by ridge regression BLUP in the NAM and the diversity panel	91
<b>Figure 4.1</b> Partitioning phenotypic variation of the maize NAM population	111
<b>Figure 4.2</b> Modular clustering of maize traits	114
<b>Figure 4.3</b> Partitioning phenotypic variation of Ames inbred diversity panel	115
<b>Figure 4.4</b> Distribution of heritability height variation in NAM by QTL selection	117
<b>Figure 4.5</b> Allele effect sizes of most significant height QTL by NAM family	118
<b>Figure 4.6</b> Pleiotropy of QTL capturing height variation in NAM	119
<b>Figure 4.7</b> Genomic prediction of plant height by ridge regression BLUP in the NAM panel and the Ames inbred diversity panel	122

**CHAPTER 1**  
**PARTITIONING GENETIC VARIATION**  
**IN COMPLEX TRAITS**

**EVOLUTIONARY DYNAMICS OF PLANT POPULATIONS**

*Population and Quantitative Genetics*

Studying the present phenotypic and allelic composition of a population, estimating its past structure, and predicting the population's future trajectory given defined selection pressures and other processes has occupied a central role in evolutionary biology for over a century. However, until the recent advancement of high-throughput genotyping, allelic information was not available to empiricists on a genomic scale (Hudson, 2008). Instead, for the first half of the 20<sup>th</sup> century the role of population genetics was largely constrained to theory, most of which could only be empirically validated in simple Mendelian traits such as pigment mutations. These foundational studies debunked Lamarckian evolution, supported the particulate model of inheritance, elucidated concepts such as Hardy-Weinberg equilibrium, Wright's adaptive landscapes, natural selection, genetic drift, mutation, and gene flow, and ultimately led to development of the modern evolutionary synthesis (Dobzhansky, 1937; Fisher, 1911; Fisher, 1918a; Ford, 1931; Haldane, 1924; Mayr, 1942; Wright, 1931; Wright, 1932).

Nonetheless, with an inability to genotype and track allele frequencies over time, population genetics was of somewhat limited practical application in more than a few case examples detailing theoretical principles of interest. Instead, quantitative genetic analyses tracking phenotypic variation based on pedigree remained the primary toolset for the assessment of complex continuous traits. This was especially true of plant science where, despite

substantial improvements since the domestication of cultivated plants approximately 13,000 years ago, the allelic basis for these improvements was not understood and could only be inferred by the observation of heritable phenotypes (Allard, 1999).

Throughout and following development of the modern evolutionary synthesis, the complex nature of most quantitative traits, especially those of agronomic consequence, led to considerable improvements in applied statistical methods. R.A. Fisher continued to improve upon modeling approaches and bested the least squares method of moments in defining parameter estimates through the introduction of his method of maximum likelihood (Aldrich, 1997; Fisher, 1918b; Fisher, 1922a; Fisher, 1922b). C.R. Henderson's mixed linear modeling approaches unified Fisher's earlier fixed and random modeling methods into a single design framework (Henderson et al., 1959; McLean et al., 1991). Efforts to incorporate pedigree among individuals as well as the complex spatial and temporal nature of environmental effects in plant breeding trials led to the development of numerous kernel-based approaches to better model these relationships and estimate heritability (Piepho and Williams, 2010). Further developments in quantitative genetics, such as R. Lande's G matrix expanded these univariate linear modeling procedures into multivariate approximations of phenotypic coevolution (Lande and Arnold, 1983; Stepan et al., 2002). However, nearly all these vector space and kernel methods remained focused on linear multivariate normal approximations of heritable phenotypes with only theoretical regard to segregation and recombination of their true allelic underpinnings.

During the last quarter of the 20<sup>th</sup> century, allozymes provided the first key insights into the dynamics of population genetics without being constrained to visible mutations in a developmental or morphological Mendelian trait (C.W. Stuber, 1980; Hamrick and Godt, 1989). The biochemical nature of these enzyme assays provided markers with substantially more

penetrance than previous traits that remained environmentally dependant, and gave geneticists the ability to observe genetic diversity in a codominant manner (C.W. Stuber, 1980; Hamrick and Godt, 1989). However, they were also very low throughput, highly labor intensive, and could under some circumstances lack complete penetrance as a result of environmentally-induced posttranslational modifications. To overcome these shortcomings and directly assay polymorphisms prior to translation, numerous DNA based markers were developed throughout the last decades of the 20<sup>th</sup> century. The most recognized of these included restriction fragment length polymorphisms (RFLPs) and microsatellites (SSRs), which provided researchers with codominant markers at low throughput (Gupta et al., 1999; Spritz, 1981; Wyman and White, 1980). Random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphisms (AFLPs) were also developed and provided dominant markers that could achieve considerably higher throughput (Botstein et al., 1980; Gupta et al., 1999; Vos et al., 1995).

The utility of genetic information provided by these marker types varied substantially. Microstatelites captured selectively neutral allele series at a single locus and allowed researchers to infer natural population structure (Gupta et al., 1999; Spritz, 1981). However, to cover more of the genome and provide denser bi-parental linkage maps, RAPDs and AFLPs were of higher value (Botstein et al., 1980; Gupta et al., 1999; Vos et al., 1995). By semi-automating these techniques, tens to hundreds of markers could be scored across a genome. From the forward genetic perspective, these various methods provided some of the first serious opportunities to precisely track the inheritance of alleles directly at the DNA level. This genotyping of alleles allowed researchers to construct higher density linkage maps and facilitated cloning of genomic regions controlling traits such as fruitcase architecture in maize (*teosinte-glume-architecture1* (Dorweiler et al., 1993)), apical dominance in maize *teosinte-branched1* (Doebley J, 1995), fruit

weight in tomato (*fruit-weight2.2* (Frary et al., 2000)), and flowering time in maize (*vegetative-to-generative1* (Salvi et al., 2002)). Similarly, ecologists, conservationists, and evolutionary biologists could now begin to track and infer the true diversity and allelic dynamics of natural plant populations. It was becoming possible to make refined estimates of genetic distance and to determine those regions of the genome under recent or ancestral selection, those explaining the population's structure, and genomic regions simply neutrally evolving.

In parallel with marker methods, DNA sequencing technologies also exponentially improved over the last quarter of the 20<sup>th</sup> century. From the 24 base pairs first scored by “chemical sequencing” to the chain-termination methods of Sanger sequencing, and eventually to fluorescently labeled capillary Sanger sequencing of up to 1,000 base pair lengths (Pettersson et al., 2009). These methods provided local estimates of genetic diversity and were often used to infer a population's genomic diversity. Furthermore, the sequencing methods facilitated our understanding of the diversity and evolutionary history of candidate genes and allowed for the characterization of gene families. Beyond sequencing these candidate regions, expressed sequence tags (Adams et al., 1991) provided us one of the first methods to characterize the exome and to exponentially increase the amount of sequence data within the public *Genbank* repository (Benson et al., 2008). This increase in sequence information began the push for basic local alignment search tools (BLAST) and other homology searches for conserved sequences across species and substantially improved and simplified efforts in candidate gene analysis (Altschul et al., 1990).

Through both BAC-by-BAC (Bacteria Artificial Chromosome) and shotgun Sanger sequencing, drafts of the first plant genomes, *Arabidopsis thaliana* (Initiative, 2000) and rice (*Oryza sativa*) (Goff et al., 2002; Yu et al., 2002), were published within the first two years of

the 21<sup>st</sup> century. These and similar sequencing efforts initiated in other plant species were monumental technological and biological achievements. They provided a foundational scaffold on which future surveys of genetic diversity across entire populations could be anchored (Gore et al., 2009; Schneeberger et al., 2011). The allelic basis of the phenomena theorized in population genetics for the past century is progressively becoming empirically obtainable information at the genomic scale. The molecular groundwork is now in place to begin fusing the population genetic view of allele frequencies with the linear regression approaches of phenotypic relationships in quantitative genetics.

Now, over a decade into the 21<sup>st</sup> century, the earlier molecular marker and sequencing technologies have been superseded by array and high-throughput short-read sequencing based approaches. For many of their former applications earlier molecular marker technologies are nearing or have reached obsolescence. Beginning with Illumina's (formerly Solexa/Lynx Therapeutics) cleavable reversible terminator-based sequencing by synthesis technology, massively parallel signature sequencing became obtainable by 2000 (Brenner et al., 2000). By 2004, this technology permitted sequencing of over three million base pairs per run. That same year, Roche (formerly 454 Life Science Technologies) commercialized a pyrosequencing technology enabling the sequencing of 300-600 million base pairs per run (Nyrén, 2007). By 2006, Illumina technology was boasting over 1 billion base pairs per run, a number that has since risen to as much as 600 billion base pairs per run. Using capillary Sanger sequencing, these platforms, and numerous other competing high-throughput sequencing platforms, such as Applied Biosystem's polony-sequencing based SOLiD platform (Shendure et al., 2005), hundreds of plant, animal, and microbial genomes have now been sequenced to near completion. These have included both additional ecotypes of *Arabidopsis thaliana* (Schneeberger et al.,

2011) and species of rice (Goff et al., 2002; Yu et al., 2002), as well as more crop plant genomes such as maize (Schnable et al., 2009), sorghum (Paterson et al., 2009), soybean (Schmutz et al., 2010), potato (Consortium, 2011), cucumber (Huang et al., 2009), grape (Velasco et al., 2007), apple (Velasco et al., 2010), strawberry (Shulaev et al., 2011), and cacao (Argout et al., 2011).

In addition to whole genome sequencing through the use of these short-read resequencing technologies, HapMaps detailing genome wide single nucleotide polymorphisms (SNPs) for several species have also been constructed through a combination of both deep coverage sequencing of single genotypes and subsequent array based approaches or skim sequencing through highly-multiplexed Genotyping-By-Sequencing methods (Baird et al., 2008; Elshire et al., 2011; Gore et al., 2009). With all this information becoming inexpensive and rapidly accessible, potential approaches to optimize experimental design and the analyses required to assess population and quantitative genetics parameters during experimentation are rapidly changing. As biologists, we must rethink our current analytical framework for understanding these complex systems and predicting their future evolution.

### *Modern Conceptualizations of Evolutionary Theory*

Our understanding of any phenomenon is influenced by the abstractions we use to conceptualize its fundamentals and the way we seek and incorporate new found information as knowledge. Until recently, most biological sciences were empirically reductionist, contrasting treatment and control while ignoring the multivariate nature of biological reality. This approach required a myriad of biologists, many of whom were rather independently accruing inherently conditional facts explaining cause and effect under often artificial circumstances of study. While these approaches unveiled aspects of the overwhelmingly large and complex network of life,

they were often limited in their ability to unite disparate biological disciplines. The size and complexity of biology as well as the difficulty of accruing information in a systematic high-throughput manner have arguably retarded the evolution of our understanding by intellectually isolating fields of biology to their own network neighborhoods. In contrast, the evolution of physics and other mathematical sciences has been less restricted as researchers overcame impositions of structure and embraced an intellectual panmixis. This flow of ideas was fostered considerably by the relative ease in acquisition of information, a uniformity in mathematical language, and the analytically reductionist (in contrast to empirically reductionist) nature of physics in its quest for increasingly universal explanations (Hoppensteadt, 1995).

Given the rapid deluge of sequence data afforded to the field of genomics by technological advances in high-throughput molecular biology, several limitations once imposed on biologists have been lifted. The network of life has not become any smaller or less complex; however, we are rapidly gaining the information needed to expand our knowledge in all biological fields of inquiry, and in doing so, bridge the gaps existing among them. As we shift focus from data acquisition to data analysis, the concepts and methods in population and evolutionary genetics conceived by theoretical biologists throughout the 20<sup>th</sup> century are now applied to massive empirical data sets with exceptional computational speed (Conrad et al., 2006; Coop et al., 2008; Voight et al., 2006) . More researchers familiar with the concepts underpinning these abstractions in population genetics and evolutionary theory are needed in research and to train future generations of biologists to conceptualize the rules of evolution from a mathematical perspective at the onset of their education. Although the complexities of biology ostensibly warrant focus on empirical detail and exceptions to the rule, we frequently neglect to adequately address what the rules are and to convey the sense of universality in the process of

evolution as an explanation of biological organization. Idealized mathematical abstractions such as the multivariate breeder's equation may fail to precisely describe concrete phenomena, nonetheless these equations enhance our understanding and provide the conceptual framework on which to hang our exceptions. As we gain more information, perhaps the serialized nature of studies in physics and other mathematical disciplines will replace the disconnected modules found in current biological study.

The compatibility of the sciences is becoming ever apparent as we champion multidisciplinary studies and observe the benefit of their relationships. The universality and utility of many fundamental laws and simplifying equations first drafted in physics have long been recognized by the analysts of other scientific fields (de Vladar and Barton, 2011). Arguably, the strongest parallel drawn in the fields of quantitative and population genetics is that to statistical thermodynamics. One of the first individuals to observe the similarities of these fields was Fisher in his book, *The Genetical Theory of Natural Selection* (Fisher, 1930). While working with theoretical physicist E.T. Jaynes, founder of the principle of maximum entropy, Fisher noted the uncanny resemblance of his fundamental theorem of natural selection and the principles of thermodynamics, yet stated "while it is possible that both [thermodynamics and evolution] may ultimately be absorbed by some more general principle, for the present we should note that the laws as they stand present profound differences." (Fisher, 1930) Without C.E. Shannon's later advances in the field of information theory and its own unique applications of statistical thermodynamic principles to cryptography and computer science, Fisher conceived but failed to accurately develop the analogy between the disciplines. He equated the randomness of total genotypic fitness to thermal entropy as opposed to the randomness of marginal allele fitness (de Vladar and Barton, 2011; Fisher, 1930). Motoo Kimura's later work on diffusion models and

their application to the neutral theory of evolution served to develop solutions to several of the problems which arose in explaining population genetic phenomenon (Kimura, 1983). However, analogies between population genetics and statistical thermodynamics beyond his groundbreaking work have been scant until recently. During the past five years over two hundred articles have been published in statistical physics and thermodynamics related journals addressing evolutionary theoretic approaches to biological phenomena. Unfortunately, few physicists have consulted biologists before diving into the development of novel methods (de Vladar and Barton, 2011). As a result, few of these publications have any useful application for biologists or breeders.

Sadi Carnot's second law of thermodynamics states that in the absence of a change in mass-energy any isolated physical system will increase in entropy until reaching thermal equilibrium (Carnot, 1824). In genetics a population's allelic dynamics abide by this fundamental physical law (de Vladar and Barton, 2011). Any isolated population not gaining genetic variation by mutation or migration, losing variation to selection, and sizable enough to overcome drift, will approach Hardy-Weinberg equilibrium. Nonetheless, no natural population exists in isolation, and few populations even approximate equilibrium. Identifying the factors responsible for perturbing equilibrium remains an intriguing endeavor and continues to expand our understanding of biology's most unifying theory, evolution. Perhaps more than any other profession, breeders are acquainted with the existing breadth of natural genetic variation, its natural population structure, and the application of methods to select upon it. Any breeder will readily acknowledge artificial selection requires an input of work. As Darwin first observed, this input of work is analogous to the solar and other energy inputs driving the complex ecology responsible for natural selection (Darwin, 1868) and the population structure it imposes in

concert with mutation, migration, and drift. Although the mechanism is not direct and acts through the breeder and the environment, or what is known in statistical thermodynamics as Maxwell's demon (Andrade, 2004), energy inputs maintain a population's disequilibrium and dictate the balance between the imperfect inheritance of adaptive information (order) and the persistence of genetic diversity or allelic entropy (randomness). Further exploration of the analogy reveals the energy inputs to Maxwell's demon can only decrease a system's entropy if the demon's actions are thermodynamically irreversible or the information enabling their reversibility is lost. This has been likened to the irreversibility of a population's evolution once a path on Wright's adaptive landscape is assumed by selection (de Vladar and Barton, 2011).

At first, the aforementioned similarities may appear to be a trivial rehashing of parallel phenomena discovered independently in two disciplines. However, such analogies improve our understanding of evolution, and provide biologists with enhanced analytical tools to direct crossing decisions that improve our choice of evolutionary trajectory and enhance breeding efficiency. These gains in understanding accrued through synthesis of disciplines are not isolated. Numerous others are emerging such as the realization that Bayes theorem of statistics, the quasi-species equation of molecular evolution, Lotka - Volterra equation of ecology, replicator-mutator equation of game theory, Price equation of population genetics, and breeder's equation of plant and animal breeding are all part of a unified framework of evolutionary dynamics (Harper, 2010; Page and Nowak, 2002). Under many situations these difference and differential equations may be shown to be limits or mathematical equivalents of one another under their respective discrete or continuous generational assumptions (Page and Nowak, 2002). Seeking to condense the numerous theories that have developed in biology may help us to stop re-inventing methods and simplify needless complexities in the already complex discipline of

biology. Specifically, further fusion of analytical methods in the fields of ecology/evolutionary biology, quantitative/population genetics, and plant/animal breeding are needed. Fisher's introduction and expansion of least squares approaches as ANOVA in biology has served its purpose by increasing the quantitative nature of biological analyses (Fisher, 1922a; Fisher, 1922b). Similar bridges connecting seemingly disparate disciplines such as statistical thermodynamics and information geometry to biology may be necessary due to the vast differences in application, nomenclature, and language among biology and these more traditionally mathematical disciplines. Nonetheless, we should seek to better unify the analogous concepts existing within biology itself.

## **GENETIC ARCHITECTURE OF COMPLEX TRAITS IN PLANTS**

### *Methods and Pitfalls in Defining Genetic Architectures*

Genetic architecture refers to the underlying basis of a trait with respect to a given population and environment. Plants have exceptional experimental flexibility to characterize genetic architecture due in part to researchers' ability to replicate genotypes across unique environments through selfing genotypes to pure breeding lines, clonal propagation techniques, and more recently the use of doubled haploid technologies (Xu et al., 2007). Furthermore, researchers have the ability to control crosses and select progeny, allowing for the manipulation of allele frequencies as well as patterns of linkage disequilibrium both within and across mapping populations. Despite recombination suppression, population structure, and other practical constraints naturally imposed by the biology of a plant population, the potential to design large populations and experiments tailored to capture the desired information detailing genetic architecture on a genomic scale is much greater and of higher direct value in crop plants than many other organisms of study.

Depending upon the population of interest and experimental design, studies of genetic architecture seek to capture a wide assortment of information. This includes such attributes as the polygenicity or number of genetic effects impacting heritable variation in a trait, as well as the size distribution, allele frequency, genomic location, and founder genotype of these genetic effects or quantitative trait loci (QTL). The pleiotropy or number of traits affected by a given QTL and the consequent genome-wide modularity of traits, the proportion of heritable variation captured by dominant and epistatic QTL, the genotypic and phenotypic plasticity and environmental canalization of QTL, and numerous other phenomena are also important aspects of genetic architecture. No population or single experimental design can optimally address all of these genetic parameters of interest simultaneously. However, the information afforded by all of these studies aids in determining the evolvability of the population and is of great utility in better directing breeding efforts.

From Sturtevant and Morgan's expansion and refinement of Mendel's second law of independent assortment in 1913 (Sturtevant, 1913) to the end of 20<sup>th</sup> century, many research experiments mapping genotypes to phenotypes were performed in segregating F<sub>2</sub> bi-parental linkage mapping populations. These populations provide a very coarse-grained view of the functional genetic diversity segregating within a severely constricted pool of total genetic diversity. Having only two potential allelic states at a 1:1 ratio at every locus of the genome and the recombination of a single effective meiotic generation affords considerable statistical power to identify cumulatively large effects on a chromosomal arm in one parental haplotype versus the other and to assess the degree of dominance present between these segments. However, the level of resolution afforded by F<sub>2</sub> bi-parental linkage mapping provided minimal information to aid molecular inferences.

More recent modifications of the F<sub>2</sub> bi-parental linkage population design have provided the opportunity to focus on specific aspects of genetic architecture. At the expense of an ability to assess heterosis and dominance, the development of recombinant inbred lines (RILs) in which F<sub>2</sub> individuals are selfed to near fixation now immortalize the lines and increase the number of environments in which a given genotype can be evaluated (Broman, 2005). This process also effectively adds an additional meiotic generation to further recombine and more finely resolve these loci. Advanced inter-mated linkage mapping populations, such as the B73 x Mo17 (IBM) population in maize (Lee et al., 2002), further increase recombination and thus resolution, while maintaining the 1:1 ratio of each parental allele. Moreover, complex mapping designs such as the North Carolina Designs I, II, and III allow the estimation of non-additive genetic variation (Robinson et al., 1954).

Until recently, the primary means to further resolve these coarse-grained linkage mapping approaches in plants have been conceptually simple, yet highly labor intensive. These include the development of introgression libraries of near isogenic lines (NILs) and the implementation of fine mapping efforts to resolve large effect loci through many generations of recurrent backcrossing and marker-assisted selection. These methods were responsible for resolving some of the first QTL, including *teosinte-glume-architecture1* (Dorweiler et al., 1993), *teosinte-branched1* (Doebley J, 1995), *fruit weight2.2* (Frary et al., 2000), *vegetative-to-generative1* (Salvi et al., 2002) and *vernalization1* (Yan et al., 2003). Once Sanger sequencing methods and sufficient sequence data were available, candidate gene association mapping across diverse inbred panels became a reverse genetics approach to rapidly refining linkage mapping approaches. Although some successes were achieved (Harjes et al., 2008), this method was limited to

understanding the allelic variation present in candidate genes already implicated in a phenotype of interest.

With the advent of genotyping arrays and high-throughput sequencing technologies, genome wide association mapping studies (GWAS) across diverse inbred lines took center stage as a powerful approach to rapidly resolve natural allelic diversity (Atwell et al., 2010; Klein et al., 2005). However, in most plant populations this approach remains confounded by population structure, limited in instance of low allele frequency, and creates additional issues such as synthetic associations (Wray et al., 2011). While mixed linear modeling approaches are prepared to statistically control for population structure (Zhang et al., 2010), no linear regression can maintain sufficient power to dissect highly collinear parameters such as a flowering time locus confounded with population structure. To empirically address these issues and strengthen analyses, the concept of nested association mapping (NAM) in populations such as the maize NAM panel was developed (McMullen et al., 2009).

In the maize NAM panel, 25 inbreds were crossed to the common reference parent B73. Two hundred progeny from each of the F<sub>2</sub> bi-parental linkage mapping population were selfed for five generation to develop 25 immortalized families in a panel of 5,000 RILs (McMullen et al., 2009). This panel of families allowed mapping at the level of recent recombination through joint-linkage mapping within families as well as ancestral recombination by mapping in a GWAS across families. Moreover, it facilitated the control of background genetic variance identified in linkage mapping during GWAS, reduced those false discovery problems associated with population structure, ensured minor allele frequencies within families were optimal, and that across families these frequencies were maintained at a level providing adequate statistical power (Buckler et al., 2009; Kump et al., 2011; Poland et al., 2011; Tian et al., 2011).

While these methods have greatly increased the accuracy, efficiency, and proportion of relevant information obtainable by researchers in their elucidation of genetic architecture, any fixed population fails to address contextual dependencies and merely represents one instance of an infinite number of potential populations. For many traits mapped within a fixed population, the bulk of genetic variation appears to be additive; however, statistically determined additivity at a locus does not sufficiently reflect physiological/biochemical additivity. Through the properties of emergence, the numerous pairwise and higher epistatic interactions existing in molecular networks may act in a statistically additive manner depending upon the allelic composition (frequency and LD structure) of the population in which the effects were mapped. This is similar to the issues associated with synthetic associations (Wray et al., 2011) and lends truth to the cliché “correlation does not imply causation.” Given the millions of polymorphisms already known to segregate within plant species such as maize, no plant population or study will ever possess enough degrees of freedom to map all loci in a perfectly orthogonal over-determined system. Furthermore, by performing the single marker tests that are commonly employed in GWAS, the Beavis effect (Beavis, 1994) or similar winner’s curse (Thaler, 1988) ensures polymorphism effects are grossly inflated in polygenic traits, as a result of their capture of variance that should be attributed to environmental differences or the genetic variance of other polymorphisms.

With these realizations, additional methods are needed to accurately address the uncertainty we possess when mapping the genetic architecture of a trait and further dissecting molecular networks for basic research. From an applied plant breeding perspective, future efforts to discern the utility of more coarsely-grained haplotypes will be necessary. Similarly, efforts to define the optimal level of resolution at which to map these haplotype effects and

methods to best predict the recombination rates between them are needed. Using this information detailing genetic architecture, simulated random walks through Wright's adaptive landscapes may assist future breeding decisions with respect to which genotypes are selected, which crosses are made, and in which environments these occur. These informed management decisions may offer breeders the potential to adapt a population to a desired environment in a manner which supersedes the efficiency of natural phenotypic selection itself. Natural selection acts on phenotypic variance without knowledge of allelic covariance structures. With genome wide sequencing of entire populations we may now select on both phenotypic variance as it has been partitioned across the genome, and the covariance of these effects, thus ensuring the stacking of desirable haplotypes in phase and facilitating a more rapid approach toward a fitness maximum.

#### *Environmental Adaptation, Ecology, and the Extended Phenotype of Plants*

Natural selection is the only adaptive evolutionary process. It is the only means by which environmental information is encoded into a plant's genome (Frank, 2009). Despite cytosine methylation and select instances of epigenetic inheritance which may be seen as inheriting information in a Lamarckian sense at the level of individuals (Bird, 2007), most information encoding is performed at the level of populations through the fixation of desirable mutations under directional selection. This process acts in a Bayesian manner (Harper, 2010) whereby the prior probability distribution may be seen as an allele's initial frequency within a population. This probability or allele frequency is updated by the ratio of the conditional and marginal probabilities. This may be viewed as the ratio of the mean number of plant progeny surviving in the next generation given they possess the allele to the mean number of plant

progeny surviving in the next generation given they possess any allele at the locus. The final posterior probability is equal to the allele frequency in the next generation, at which time it may serve as the prior for the next round of Bayesian updating or natural selection.

The adaptive environmental information accrued by natural selection determines which alleles approach fixation as a result of local ecology. While abiotic factors play a substantial role in shaping what information is encoded in the genome as a result of direct competition for resources among plants within the same population, further complications such as frequency dependent selection also arise, depending on other traits of the population itself. These are further complicated by interactions with other populations in the community. The extended phenotype is a phenomenon by which phenotypic attributes resulting from selection in one population influence the selective pressures of other populations in the community, which in turn may impact the initial population. This feedback loop is apparent in numerous symbiotic relationships; however, perhaps better known than any other relationship is that between microbes and plant species.

Plants have evolved numerous symbiotic relationships with various microbial species, including the 70-90% of terrestrial plants which barter carbon exudates for nitrogen with strains of arbuscular mycorrhizae (Shannon and Kendrick, 1982), the mutualistic organogenic behavior of rhizobia and leguminous plant species (Patriarca et al., 2004), and the countless parasitic interactions between plants and countless fungal and bacterial species. However, most of these relationships were characterized by a readily apparent and highly influential plant phenotype. Using targeted high-throughput sequencing technologies, most notably Roche's pyrosequencing approach, it is now possible to profile entire microbiomes, thereby characterizing novel microbial diversity that may have a less obvious influence on the growth and development of a

plant species (Leveau, 2007). Similarly, it is feasible to see which microbial species are strongly influenced by the presence of the plant. Given recent advances in multiplexing of high-throughput sequencing runs, these assessments may be done across numerous samples. This allows researchers to begin dissecting at the genetic level the manner in which plants regulate these complex microbial networks and reveals just how those alleles governing plant development impact the environment beyond their direct influences on the plant. Furthermore, plant genotypes adapted to unique environments, such as crops bred before and after the Green Revolution, may be compared for influences on microbial community structure.

#### *Constraining Evolvability, Pleiotropy, Modularity, and the Cost of Complexity*

The evolvability of a plant population refers to its ability to generate adaptive heritable phenotypic variation and thus allow the process of natural selection to change the population's survival and reproductive fitness. In breeding environments this simply refers to the rate at which the desired breeding gains can be made in a given population. The ultimate source of genetic variation is mutation; however, most mutations are considered to be nearly neutral or deleterious. According to Fisher's geometric model (Fisher, 1930) this is especially true as a population approaches a fitness maximum. Therefore for adequate evolvability, the genetic architecture of a population must be composed such that advantageous mutations may be easily recombined and selected away from deleterious ones. This may explain the relative dearth of open reading frames around the centromeres and other low recombinogenic regions of the genome. Without recombination, the Hill-Robertson effect (Hill, 1966) will reduce the overall fitness of the population as the more rapidly occurring deleterious mutations are linked with advantageous

alleles and natural selection cannot independently act upon them. In a similar phenomenon, the Bulmer effect (Bulmer, 1973) reduces evolvability by ensuring negative linkage disequilibrium develops between beneficial alleles in the selected individuals. This limits the population's ability to stack desirable alleles within the same haplotype and reduces the total genetic variance of a population thus limiting future gains.

In addition to dictating the evolvability of a population through its interplay with mutation rates, the genetic architecture of a population also impacts evolvability through the relative degree of pleiotropy among the traits that are under selection. A universally pleiotropic genetic architecture severely reduces evolvability, as selection on every trait will impact that of other traits in what has been called “the cost of complexity” (Wagner et al., 2008). Given these constraints on evolution, many trait combinations may not be possible. For instance, if a plant's maturation rate and height are pleiotropic at most loci regulating the variance in these traits, the evolvability of a phenotype which breaks covariation of these traits may be constrained. However, if the genetic architecture is organized in a modular manner whereby functional modules influencing similar traits are pleiotropic but, these modules are independent of one another, then the effective reduction in evolvability as a function of complexity is much less, and traits are free to evolve in a largely independent manner (Wagner et al., 2008).

## **GENOMIC PREDICTION OF COMPLEX TRAITS IN PLANTS**

### *Methods and Pitfalls in Predicting Phenotypes*

Genomic prediction is a relatively new analytical approach (Meuwissen et al., 2001) which seeks to predict the genomic estimated breeding value (GEBV) of plant genotypes based on all scored polymorphisms within the population. Given its drastic improvement of predictive

accuracy relative to past marker-assisted modeling approaches, it has achieved widespread acceptance within the field of plant breeding and quantitative genetics. Using this suite of analytical tools, breeders may now take advantage of the information supplied by high-throughput sequencing and predict the value of seed before expending the resources necessary to evaluate it in field trials. This allows selection from a substantially larger initial pool of potential genotypes and thus will increase selection gains per year (Jannink et al., 2010).

To achieve these gains, a training population is genotyped and phenotyped to develop a predictive model for the phenotypes of interest. Subsequently, related seeds are all genotyped through seed chipping procedures and the aforementioned model is used to predict their GEBVs. The accuracy of genomic prediction is influenced by numerous factors: the genetic architecture of the trait; size of the population under study; the density of polymorphisms genotyped and their imputation accuracy; the bias of polymorphisms genotyped with respect to true population structure; the genetic distance between the training and breeding populations, as well as the modeling method employed in prediction.

Those previous analyses seeking to characterize the genetic architecture of a trait, within a population and environment relied upon various model selection criteria, such as Bayesian information criterion, to discern significant QTL within the genome which could then be fitted in a multiple regression for prediction (Bogdan et al., 2004). This approach overcame the modeling issues attributable to an under-determined system. In under-determined systems the number of parameters ( $p$ ) or polymorphisms under an additive model greatly exceeds that of the number of observations ( $n$ ) or plants. This ill-posed  $p > n$  situation results in too many unconstrained solutions (too few degrees of freedom) and ultimately an infinite number of potential allelic effect estimates when partitioning variation across the genome (Jannink et al., 2010).

Similar to genetic architecture studies, in genomic prediction a process known as regularization is imposed on estimates of allelic effects. This merely means additional information is provided to the modeling framework to either penalize the complexity of the model and impose some degree of selection among all possible parameters as was performed in genetic architecture analyses, or to restrict the norm or absolute size of the allelic effect estimates. The most commonly employed genomic prediction methods in a frequentist framework, or a framework which seeks to find the most probable conclusion upon theoretically infinite repetition of experimentation, have been ridge regression (Yang et al., 2010), least absolute shrinkage and selection operator (LASSO), and the combination of these, known as elastic net regularization (Ogutu et al., 2012).

While most frequentist models addressing ill-posed problems have analogous models in a Bayesian framework, numerous other hierarchical Bayesian model fitting frameworks such as Bayes A, B,  $C_\pi$  and  $D_\pi$  also exist that better accommodate heterogeneous variances (Gianola et al., 2009; Meuwissen et al., 2001). With the exception of Bayes A which, like ridge regression, assumes all polymorphisms affect the trait of interest (Meuwissen et al., 2001), these models allow for prior specification (Bayes B) or data based inference (Bayes  $C_\pi$  and  $D_\pi$ ) of the probability that a polymorphism has an effect depending upon the polygenicity of the trait. They also allow specification of prior variance attributed to the polymorphism as drawn from a scaled inverse chi-squared distribution determined for each polymorphism (Bayes A and B), assuming either equally scaled inverse chi-squared distribution for all polymorphisms (Bayes  $C_\pi$ ), or the scaling parameters for the prior distribution are drawn from hyperprior distributions themselves (Bayes  $D_\pi$ ) (Habier et al., 2011). These parametric methods have also been complemented by the non-parametric Reproducing Kernel Hilbert Space (RKHS) (Gianola and van Kaam, 2008).

In most empirical situations, complex polygenic traits are often modeled with comparable accuracy by all the aforementioned methods. However, in simpler more Mendelian traits, Bayes B with a properly selected prior for the reduced number of QTL will often outperform other model building frameworks (Meuwissen et al., 2009) in a manner similar to those model selection frameworks often used to dissect genetic architecture.

### *Future Prospects*

Given the current popularity of genomic prediction methods and realizations that many modeling methods are comparable for complex traits, the next steps in genomic prediction are discerning optimal training populations, predicting coarse-grained haplotypic effects rather than genotypic differences, and leveraging knowledge of the covariance of these haplotype effects to improve selection, stack QTL and overcome the Bulmer effect (Bulmer, 1973). These approaches must essentially combine and optimize the analytical frameworks currently used to understand genetic architecture or what has been called “back-end” mathematics with the “front-end” mathematics (Sherwin, 2010) of genomic prediction. The end goal will be a coarse-grained characterization of the genome that allows us to surmount the contextual dependencies (Cooper et al., 2009) that confound fine-grained characterizations such as GWAS while also overcoming the lack of knowledge detailing genetic architecture when implementing current genomic prediction approaches. In a simulation framework this may allow for improved breeding decisions regarding which crosses should be made and how to best improve the evolvability of the population.

## REFERENCES

- Adams M., Kelley J., Gocayne J., Dubnick M., Polymeropoulos M., Xiao H., Merril C., Wu A., Olde B., Moreno R., et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252:1651-1656.
- Aldrich J. (1997) R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science* 12:162-176.
- Allard R.W. (1999) History Of Plant Population Genetics. *Annual Review of Genetics* 33:1-27.
- Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215:403-410.
- Andrade E. (2004) On Maxwell's Demons and the Origin of Evolutionary Variations: An Internalist Perspective. *Acta Biotheoretica* 52:17-40.
- Argout X., Salse J., Aury J.-M., Guiltinan M.J., Droc G., Gouzy J., Allegre M., Chaparro C., Legavre T., Maximova S.N., Abrouk M., Murat F., Fouet O., Poulain J., Ruiz M., Roguet Y., Rodier-Goud M., Barbosa-Neto J.F., Sabot F., Kudrna D., Ammiraju J.S.S., Schuster S.C., Carlson J.E., Sallet E., Schiex T., Dievart A., Kramer M., Gelley L., Shi Z., Berard A., Viot C., Boccara M., Risterucci A.M., Guignon V., Sabau X., Axtell M.J., Ma Z., Zhang Y., Brown S., Bourge M., Golser W., Song X., Clement D., Rivallan R., Tahi M., Akaza J.M., Pitollat B., Gramacho K., D'Hont A., Brunel D., Infante D., Kebe I., Costet P., Wing R., McCombie W.R., Guiderdoni E., Quetier F., Panaud O., Wincker P., Bocs S., Lanaud C. (2011) The genome of *Theobroma cacao*. *Nat Genet* 43:101-108.
- Atwell S., Huang Y.S., Vilhjalmsson B.J., Willems G., Horton M., Li Y., Meng D., Platt A., Tarone A.M., Hu T.T., Jiang R., Mulyati N.W., Zhang X., Amer M.A., Baxter I., Brachi B., Chory J., Dean C., Debieu M., de Meaux J., Ecker J.R., Faure N., Kniskern J.M., Jones J.D.G., Michael T., Nemri A., Roux F., Salt D.E., Tang C., Todesco M., Traw M.B., Weigel D., Marjoram P., Borevitz J.O., Bergelson J., Nordborg M. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627-631.
- Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U., Cresko W.A., Johnson E.A. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3:3376.
- Beavis W.D. (1994) The power and deceit of QTL experiments: Lessons from comparative QTL studies. D. B. Wilkinson (ed.) 49th Ann Corn Sorghum Res Conf. Am Seed Trade Assoc.,:250-266.
- Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. (2008) GenBank. *Nucleic Acids Research* 36:D25-D30.
- Bird A. (2007) Perceptions of epigenetics. *Nature* 447:396-398.
- Bogdan M., Ghosh J.K., Doerge R.W. (2004) Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. *Genetics* 167:989-999.
- Botstein D., White R.L., Skolnick M., Davis R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet.* 32:314-331.
- Brenner S., Johnson M., Bridgham J., Golda G., Lloyd D.H., Johnson D., Luo S., McCurdy S., Foy M., Ewan M., Roth R., George D., Eletr S., Albrecht G., Vermaas E., Williams S.R., Moon K., Burcham T., Pallas M., DuBridge R.B., Kirchner J., Fearon K., Mao J.-i., Corcoran K. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotech* 18:630-634.

- Broman K.W. (2005) The Genomes of Recombinant Inbred Lines. *Genetics* 169:1133-1146.
- Buckler E.S., Holland J.B., Bradbury P.J., Acharya C.B., Brown P.J., Browne C., Ersoz E., Flint-Garcia S., Garcia A., Glaubitz J.C., Goodman M.M., Harjes C., Guill K., Kroon D.E., Larsson S., Lepak N.K., Li H., Mitchell S.E., Pressoir G., Peiffer J.A., Rosas M.O., Rocheford T.R., Romay M.C., Romero S., Salvo S., Villeda H.S., Sofia da Silva H., Sun Q., Tian F., Upadyayula N., Ware D., Yates H., Yu J., Zhang Z., Kresovich S., McMullen M.D. (2009) The Genetic Architecture of Maize Flowering Time. *Science* 325:714-718. DOI: 10.1126/science.1174276.
- Bulmer M.G. (1973) The maintenance of the genetic variability of polygenic characters by heterozygous advantage. *Genetical Research*. 22.
- C.W. Stuber R.H.M., M.M. Goodman, H.E. Schaffer, B.S. Weir. (1980) Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea Mays* L.). *Genetics* 95:225.
- Carnot C. (1824) *Reflections on the Motive Power of the Fire*, Paris.
- Conrad D.F., Jakobsson M., Coop G., Wen X., Wall J.D., Rosenberg N.A., Pritchard J.K. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38:1251-1260.
- Consortium T.P.G.S. (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475:189-195.
- Coop G., Wen X., Ober C., Pritchard J.K., Przeworski M. (2008) High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. *Science* 319:1395-1398.
- Cooper M., van Eeuwijk F.A., Hammer G.L., Podlich D.W., Messina C. (2009) Modeling QTL for complex traits: detection and context for plant breeding. *Current Opinion in Plant Biology* 12:231-240.
- Darwin C. (1868) *The variation of animals and plants under domestication*, London.
- de Vladar H.P., Barton N.H. (2011) The contribution of statistical physics to evolutionary biology. *Trends in ecology & evolution (Personal edition)* 26:424-432.
- Dobzhansky T., G. Sturtevant, A. H. . (1937) *Genetics and the Origin of Species* Columbia University Press.
- Doebley J S.A., Gustus C. (1995) Teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141:333-346.
- Dorweiler J., Stec A., Kermicle J., Doebley J. (1993) Teosinte glume architecture 1: A Genetic Locus Controlling a Key Step in Maize Evolution. *Science* 262:233-235.
- Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6:e19379.
- Fisher R.A. (1911) *Mendelism and biometry*. Unpublished.
- Fisher R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*.
- Fisher R.A. (1922a) On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*.
- Fisher R.A. (1922b) The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. Roy.*
- Fisher R.A. (1930) *The genetical theory of natural selection* Oxford, England: Clarendon Press.
- Ford E.B. (1931) *Mendelism and evolution* Methuen, London.

- Frank S.A. (2009) Natural selection maximizes Fisher information. *Journal of Evolutionary Biology* 22:231-244.
- Frary A., Nesbitt T.C., Frary A., Grandillo S., Knaap E.v.d., Cong B., Liu J., Meller J., Elber R., Alpert K.B., Tanksley S.D. (2000) fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size. *Science* 289:85-88.
- Gianola D., van Kaam J.B.C.H.M. (2008) Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics* 178:2289-2303.
- Gianola D., de los Campos G., Hill W.G., Manfredi E., Fernando R. (2009) Additive Genetic Variability and the Bayesian Alphabet. *Genetics* 183:347-363.
- Goff S.A., Ricke D., Lan T.-H., Presting G., Wang R., Dunn M., Glazebrook J., Sessions A., Oeller P., Varma H., Hadley D., Hutchison D., Martin C., Katagiri F., Lange B.M., Moughamer T., Xia Y., Budworth P., Zhong J., Miguel T., Paszkowski U., Zhang S., Colbert M., Sun W.-l., Chen L., Cooper B., Park S., Wood T.C., Mao L., Quail P., Wing R., Dean R., Yu Y., Zharkikh A., Shen R., Sahasrabudhe S., Thomas A., Cannings R., Gutin A., Pruss D., Reid J., Tavtigian S., Mitchell J., Eldredge G., Scholl T., Miller R.M., Bhatnagar S., Adey N., Rubano T., Tusneem N., Robinson R., Feldhaus J., Macalma T., Oliphant A., Briggs S. (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296:92-100.
- Gore M.A., Chia J.-M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S. (2009) A First-Generation Haplotype Map of Maize. *Science* 326:1115-1117.
- Gupta P.K., Varshney R.K., Sharma P.C., Ramesh B. (1999) Molecular markers and their applications in wheat breeding. *Plant Breeding* 118:369-390.
- Habier D., Fernando R., Kizilkaya K., Garrick D. (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186.
- Haldane J.B.S. (1924) A Mathematical Theory of Natural Selection. *Biological Reviews* 1:158-163.
- Hamrick J.L., Godt M.J.W. (1989) Allozyme Diversity in Cultivated Crops. *Crop Sci.* 37:26-30.
- Harjes C.E., Rocheford T.R., Bai L., Brutnell T.P., Kandianis C.B., Sowinski S.G., Stapleton A.E., Vallabhaneni R., Williams M., Wurtzel E.T., Yan J., Buckler E.S. (2008) Natural Genetic Variation in Lycopene Epsilon Cyclase Tapped for Maize Biofortification. *Science* 319:330-333.
- Harper M. (2010) The Replicator Equation as an Inference Dynamic. *Dynamical Systems* 3.
- Henderson C.R., Kempthorne O., Searle S.R., Krosigk C.M.v. (1959) The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics* 15:192-218.
- Hill W.G., and A. Robertson. (1966) The effect of linkage on limits to artificial selection. *Genetical Research.* 8:269-294.
- Hoppensteadt F. (1995) Getting Started in Mathematical Biology. *Notices of the American Mathematical Society* 42.
- Huang S., Li R., Zhang Z., Li L., Gu X., Fan W., Lucas W.J., Wang X., Xie B., Ni P., Ren Y., Zhu H., Li J., Lin K., Jin W., Fei Z., Li G., Staub J., Kilian A., van der Vossen E.A.G., Wu Y., Guo J., He J., Jia Z., Ren Y., Tian G., Lu Y., Ruan J., Qian W., Wang M., Huang Q., Li B., Xuan Z., Cao J., Asan, Wu Z., Zhang J., Cai Q., Bai Y., Zhao B., Han Y., Li Y., Li X., Wang S., Shi Q., Liu S., Cho W.K., Kim J.-Y., Xu Y., Heller-Uszynska K.,

- Miao H., Cheng Z., Zhang S., Wu J., Yang Y., Kang H., Li M., Liang H., Ren X., Shi Z., Wen M., Jian M., Yang H., Zhang G., Yang Z., Chen R., Liu S., Li J., Ma L., Liu H., Zhou Y., Zhao J., Fang X., Li G., Fang L., Li Y., Liu D., Zheng H., Zhang Y., Qin N., Li Z., Yang G., Yang S., Bolund L., Kristiansen K., Zheng H., Li S., Zhang X., Yang H., Wang J., Sun R., Zhang B., Jiang S., Wang J., Du Y., Li S. (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41:1275-1281.
- Hudson M.E. (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources* 8:3-17.
- Initiative T.A.G. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796-815.
- Jannink J.-L., Lorenz A.J., Iwata H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* 9:166-177. DOI: 10.1093/bfpg/elq001.
- Kimura M. (1983) *The neutral theory of molecular evolution* Cambridge.
- Klein R.J., Zeiss C., Chew E.Y., Tsai J.-Y., Sackler R.S., Haynes C., Henning A.K., SanGiovanni J.P., Mane S.M., Mayne S.T., Bracken M.B., Ferris F.L., Ott J., Barnstable C., Hoh J. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* 308:385-389.
- Kump K.L., Bradbury P.J., Wisser R.J., Buckler E.S., Belcher A.R., Oropeza-Rosas M.A., Zwonitzer J.C., Kresovich S., McMullen M.D., Ware D., Balint-Kurti P.J., Holland J.B. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 43:163-168.
- Lande R., Arnold S.J. (1983) The Measurement of Selection on Correlated Characters. *Evolution* 37:1210-1226.
- Lee M., Sharopova N., Beavis W.D., Grant D., Katt M., Blair D., Hallauer A. (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. *Plant Molecular Biology* 48:453-461.
- Leveau J. (2007) The magic and menace of metagenomics: prospects for the study of plant growth-promoting rhizobacteria. *European Journal of Plant Pathology* 119:279-300.
- Mayr E. (1942) *Systematics and the Origin of Species from the Viewpoint of a Zoologist* Harvard University Press, Boston.
- McLean R.A., Sanders W.L., Stroup W.W. (1991) A Unified Approach to Mixed Linear Models. *The American Statistician* 45:54-64.
- McMullen M.D., Kresovich S., Villeda H.S., Bradbury P., Li H., Sun Q., Flint-Garcia S., Thornsberry J., Acharya C., Bottoms C., Brown P., Browne C., Eller M., Guill K., Harjes C., Kroon D., Lepak N., Mitchell S.E., Peterson B., Pressoir G., Romero S., Rosas M.O., Salvo S., Yates H., Hanson M., Jones E., Smith S., Glaubitz J.C., Goodman M., Ware D., Holland J.B., Buckler E.S. (2009) Genetic Properties of the Maize Nested Association Mapping Population. *Science* 325:737-740.
- Meuwissen T., Solberg T., Shepherd R., Woolliams J. (2009) A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genetics Selection Evolution* 41:2.
- Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157:1819-1829.
- Nyrén P. (2007) The History of Pyrosequencing. *Methods in Molecular Biology* 373:1-13.
- Ogutu J., Schulz-Streeck T., Piepho H.-P. (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC*

Proceedings 6:S10.

- Page K.M., Nowak M.A. (2002) Unifying Evolutionary Dynamics. *Journal of Theoretical Biology* 219:93-98.
- Paterson A.H., Bowers J.E., Bruggmann R., Dubchak I., Grimwood J., Gundlach H., Haberer G., Hellsten U., Mitros T., Poliakov A., Schmutz J., Spannagl M., Tang H., Wang X., Wicker T., Bharti A.K., Chapman J., Feltus F.A., Gowik U., Grigoriev I.V., Lyons E., Maher C.A., Martis M., Narechania A., Ollillar R.P., Penning B.W., Salamov A.A., Wang Y., Zhang L., Carpita N.C., Freeling M., Gingle A.R., Hash C.T., Keller B., Klein P., Kresovich S., McCann M.C., Ming R., Peterson D.G., Mehboob ur R., Ware D., Westhoff P., Mayer K.F.X., Messing J., Rokhsar D.S. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* 457:551-556.
- Patriarca E.J., Tatè R., Ferraioli S., Iaccarino M. (2004) Organogenesis of Legume Root Nodules, *International Review of Cytology*, Academic Press. pp. 201-262.
- Pettersson E., Lundeberg J., Ahmadian A. (2009) Generations of sequencing technologies. *Genomics* 93:105-111.
- Piepho H.P., Williams E.R. (2010) Linear variance models for plant breeding trials. *Plant Breeding* 129:1-8.
- Poland J.A., Bradbury P.J., Buckler E.S., Nelson R.J. (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences*.
- Robinson H.F., Mann T.J., Comstock R.E. (1954) An analysis of quantitative variability in *Nicotiana tabacum*. *Heredity* 8:365-376.
- Salvi S., Tuberosa R., Chiapparino E., Maccaferri M., Veillet S., van Beuningen L., Isaac P., Edwards K., Phillips R.L. (2002) Toward positional cloning of *Vgt1*, a QTL controlling the transition from the vegetative to the reproductive phase in maize. *Plant Molecular Biology* 48:601-613.
- Schmutz J., Cannon S.B., Schlueter J., Ma J., Mitros T., Nelson W., Hyten D.L., Song Q., Thelen J.J., Cheng J., Xu D., Hellsten U., May G.D., Yu Y., Sakurai T., Umezawa T., Bhattacharyya M.K., Sandhu D., Valliyodan B., Lindquist E., Peto M., Grant D., Shu S., Goodstein D., Barry K., Futrell-Griggs M., Abernathy B., Du J., Tian Z., Zhu L., Gill N., Joshi T., Libault M., Sethuraman A., Zhang X.-C., Shinozaki K., Nguyen H.T., Wing R.A., Cregan P., Specht J., Grimwood J., Rokhsar D., Stacey G., Shoemaker R.C., Jackson S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178-183.
- Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., Liang C., Zhang J., Fulton L., Graves T.A., Minx P., Reily A.D., Courtney L., Kruchowski S.S., Tomlinson C., Strong C., Delehaunty K., Fronick C., Courtney B., Rock S.M., Belter E., Du F., Kim K., Abbott R.M., Cotton M., Levy A., Marchetto P., Ochoa K., Jackson S.M., Gillam B., Chen W., Yan L., Higginbotham J., Cardenas M., Waligorski J., Applebaum E., Phelps L., Falcone J., Kanchi K., Thane T., Scimone A., Thane N., Henke J., Wang T., Ruppert J., Shah N., Rotter K., Hodges J., Ingenthron E., Cordes M., Kohlberg S., Sgro J., Delgado B., Mead K., Chinwalla A., Leonard S., Crouse K., Collura K., Kudrna D., Currie J., He R., Angelova A., Rajasekar S., Mueller T., Lomeli R., Scara G., Ko A., Delaney K., Wissotski M., Lopez G., Campos D., Braidotti M., Ashley E., Golser W., Kim H., Lee S., Lin J., Dujmic Z., Kim W., Talag J., Zuccolo A., Fan C., Sebastian A., Kramer M., Spiegel L., Nascimento L., Zutavern T., Miller B., Ambroise C., Muller S.,

- Spooner W., Narechania A., Ren L., Wei S., Kumari S., Faga B., Levy M.J., McMahan L., Van Buren P., Vaughn M.W., *et al.* (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326:1112-1115.
- Schneeberger K., Ossowski S., Ott F., Klein J.D., Wang X., Lanz C., Smith L.M., Cao J., Fitz J., Warthmann N., Henz S.R., Huson D.H., Weigel D. (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proceedings of the National Academy of Sciences* 108:10249-10254.
- Shannon M.B., Kendrick B. (1982) Vesicular-Arbuscular Mycorrhizae of Southern Ontario Ferns and Fern-Allies. *Mycologia* 74:769-776.
- Shendure J., Porreca G.J., Reppas N.B., Lin X., McCutcheon J.P., Rosenbaum A.M., Wang M.D., Zhang K., Mitra R.D., Church G.M. (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* 309:1728-1732.
- Sherwin W.B. (2010) Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography. *Entropy* 12:1765-1798.
- Shulaev V., Sargent D.J., Crowhurst R.N., Mockler T.C., Folkerts O., Delcher A.L., Jaiswal P., Mockaitis K., Liston A., Mane S.P., Burns P., Davis T.M., Slovin J.P., Bassil N., Hellens R.P., Evans C., Harkins T., Kodira C., Desany B., Crasta O.R., Jensen R.V., Allan A.C., Michael T.P., Setubal J.C., Celton J.-M., Rees D.J.G., Williams K.P., Holt S.H., Rojas J.J.R., Chatterjee M., Liu B., Silva H., Meisel L., Adato A., Filichkin S.A., Troggio M., Viola R., Ashman T.-L., Wang H., Dharmawardhana P., Elser J., Raja R., Priest H.D., Bryant D.W., Fox S.E., Givan S.A., Wilhelm L.J., Naithani S., Christoffels A., Salama D.Y., Carter J., Girona E.L., Zdepski A., Wang W., Kerstetter R.A., Schwab W., Korban S.S., Davik J., Monfort A., Denoyes-Rothan B., Arus P., Mittler R., Flinn B., Aharoni A., Bennetzen J.L., Salzberg S.L., Dickerman A.W., Velasco R., Borodovsky M., Veilleux R.E., Folta K.M. (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* 43:109-116.
- Spritz R.A. (1981) Duplication/deletion polymorphism 5'- to the human  $\beta$  globin gene. *Nucleic Acids Research* 9:5037-5048.
- Steppan S.J., Phillips P.C., Houle D. (2002) Comparative quantitative genetics: evolution of the G matrix. *Trends in ecology & evolution (Personal edition)* 17:320-327.
- Sturtevant A.H. (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* 14:43-59.
- Thaler R.H. (1988) Anomalies: The Winner's Curse. *The Journal of Economic Perspectives* 2:191-202.
- Tian F., Bradbury P.J., Brown P.J., Hung H., Sun Q., Flint-Garcia S., Rocheford T.R., McMullen M.D., Holland J.B., Buckler E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159-162.
- Velasco R., Zharkikh A., Troggio M., Cartwright D.A., Cestaro A., Pruss D., Pindo M., FitzGerald L.M., Vezzulli S., Reid J., Malacarne G., Iliev D., Coppola G., Wardell B., Micheletti D., Macalma T., Facci M., Mitchell J.T., Perazzolli M., Eldredge G., Gatto P., Oyzerski R., Moretto M., Gutin N., Stefanini M., Chen Y., Segala C., Davenport C., Demattè L., Mraz A., Battilana J., Stormo K., Costa F., Tao Q., Si-Ammour A., Harkins T., Lackey A., Perbost C., Taillon B., Stella A., Solovyev V., Fawcett J.A., Sterck L., Vandepoele K., Grando S.M., Toppo S., Moser C., Lanchbury J., Bogden R., Skolnick M., Sgaramella V., Bhatnagar S.K., Fontana P., Gutin A., Van de Peer Y., Salamini F., Viola R. (2007) A High Quality Draft Consensus Sequence of the Genome of a

- Heterozygous Grapevine Variety. PLoS ONE 2:e1326.
- Velasco R., Zharkikh A., Affourtit J., Dhingra A., Cestaro A., Kalyanaraman A., Fontana P., Bhatnagar S.K., Troggio M., Pruss D., Salvi S., Pindo M., Baldi P., Castelletti S., Cavaiuolo M., Coppola G., Costa F., Cova V., Dal Ri A., Goremykin V., Komjanc M., Longhi S., Magnago P., Malacarne G., Malnoy M., Micheletti D., Moretto M., Perazzolli M., Si-Ammour A., Vezzulli S., Zini E., Eldredge G., Fitzgerald L.M., Gutin N., Lanchbury J., Macalma T., Mitchell J.T., Reid J., Wardell B., Kodira C., Chen Z., Desany B., Niazi F., Palmer M., Koepke T., Jiwan D., Schaeffer S., Krishnan V., Wu C., Chu V.T., King S.T., Vick J., Tao Q., Mraz A., Stormo A., Stormo K., Bogden R., Ederle D., Stella A., Vecchietti A., Kater M.M., Masiero S., Lasserre P., Lespinasse Y., Allan A.C., Bus V., Chagne D., Crowhurst R.N., Gleave A.P., Lavezzo E., Fawcett J.A., Proost S., Rouze P., Sterck L., Toppo S., Lazzari B., Hellens R.P., Durel C.-E., Gutin A., Bumgarner R.E., Gardiner S.E., Skolnick M., Egholm M., Van de Peer Y., Salamini F., Viola R. (2010) The genome of the domesticated apple (*Malus [times] domestica* Borkh.). *Nat Genet* 42:833-839.
- Voight B.F., Kudravalli S., Wen X., Pritchard J.K. (2006) A Map of Recent Positive Selection in the Human Genome. *PLoS Biol* 4:e72.
- Vos P., Hogers R., Bleeker M., Reijans M., Lee T.v.d., Hornes M., Friters A., Pot J., Paleman J., Kuiper M., Zabeau M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23:4407-4414.
- Wagner G.P., Kenney-Hunt J.P., Pavlicev M., Peck J.R., Waxman D., Cheverud J.M. (2008) Pleiotropic scaling of gene effects and the 'cost of complexity'. *Nature* 452:470-472.
- Wray N.R., Purcell S.M., Visscher P.M. (2011) Synthetic Associations Created by Rare Variants Do Not Explain Most GWAS Results. *PLoS Biol* 9:e1000579.
- Wright S. (1931) Evolution in Mendelian Populations. *Genetics* 16:97-159.
- Wright S. (1932) The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. *Proceedings of the Sixth Annual Congress of Genetics* 1:356-366.
- Wyman A.R., White R. (1980) A highly polymorphic locus in human DNA. *Proceedings of the National Academy of Sciences* 77:6754-6758.
- Xu L., Najeeb U., Tang G.X., Gu H.H., Zhang G.Q., He Y., Zhou W.J. (2007) Haploid and Doubled Haploid Technology, in: M. D. Surinder Kumar Gupta and J. C. Kader (Eds.), *Advances in Botanical Research*, Academic Press. pp. 181-216.
- Yan L., Loukoianov A., Tranquilli G., Helguera M., Fahima T., Dubcovsky J. (2003) Positional cloning of the wheat vernalization gene VRN1. *Proceedings of the National Academy of Sciences* 100:6263-6268.
- Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E., Visscher P.M. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42:565-569.
- Yu J., Hu S., Wang J., Wong G.K.-S., Li S., Liu B., Deng Y., Dai L., Zhou Y., Zhang X., Cao M., Liu J., Sun J., Tang J., Chen Y., Huang X., Lin W., Ye C., Tong W., Cong L., Geng J., Han Y., Li L., Li W., Hu G., Huang X., Li W., Li J., Liu Z., Li L., Liu J., Qi Q., Liu J., Li L., Li T., Wang X., Lu H., Wu T., Zhu M., Ni P., Han H., Dong W., Ren X., Feng X., Cui P., Li X., Wang H., Xu X., Zhai W., Xu Z., Zhang J., He S., Zhang J., Xu J., Zhang K., Zheng X., Dong J., Zeng W., Tao L., Ye J., Tan J., Ren X., Chen X., He J., Liu D., Tian W., Tian C., Xia H., Bao Q., Li G., Gao H., Cao T., Wang J., Zhao W., Li P., Chen

- W., Wang X., Zhang Y., Hu J., Wang J., Liu S., Yang J., Zhang G., Xiong Y., Li Z., Mao L., Zhou C., Zhu Z., Chen R., Hao B., Zheng W., Chen S., Guo W., Li G., Liu S., Tao M., Wang J., Zhu L., Yuan L., Yang H. (2002) A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. indica). *Science* 296:79-92.
- Zhang Z., Ersoz E., Lai C.-Q., Todhunter R.J., Tiwari H.K., Gore M.A., Bradbury P.J., Yu J., Arnett D.K., Ordovas J.M., Buckler E.S. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42:355-360.

**CHAPTER 2**  
**BIODIVERSITY AND HERITABILITY OF THE MAIZE RHIZOSPHERE**  
**MICROBIOME UNDER FIELD CONDITIONS**

**ABSTRACT**

Understanding interactions between plants and their environment is vital to the genetic improvement of crops and facilitates the development of more sustainable agronomic practices. While considerable efforts in agroecology have focused above ground, we are increasingly realizing the importance and influence of the complex biological networks inhabiting our soils. The rhizosphere is a critical interface supporting exchange of resources between crops and their soil environment. Many of these interactions affect plant fitness and may be influenced by biodiversity existing in rhizosphere microbial communities; however, more exploration is needed to distinguish the environmental and host genotype specificity of the rhizosphere microbiome. By pyrosequencing a hypervariable region of the bacterial 16S ribosomal RNA gene, we estimated taxonomic richness and relative abundances of Operational Taxonomic Units (OTUs) approximating bacterial community structure present in bulk soil and the rhizospheres of 27 modern maize inbreds of exceptional genetic diversity. Replicated plots of the inbreds were planted in five fields each with unique soil properties and management conditions. We discerned variation in bacterial richness and relative abundances between bulk soil and the maize rhizosphere as well as between fields. The maize inbred rhizospheres possessed a small but significant proportion of heritable variation in bacterial diversity across fields and substantially more variation between maize inbred rhizospheres within each field. However, in both instances the observed variation could not be explained by total genetic relatedness among inbreds. Given the bountiful biodiversity residing in soils, further studies possessing enhanced sequencing or

focusing on functional guilds of bacteria and surveying more maize diversity grown under replicated field conditions are needed. These studies will profile the microbiome and may identify plant-microbe interactions at the level of polymorphisms by genome wide association.

### **AUTHOR SUMMARY**

Deeper insights of microbial ecology in agricultural environments are needed to gain a more holistic perspective of these complex ecosystems. The maize rhizosphere, despite its fundamental importance as an interface for plant-microbe interactions, has lacked research attention because of the technical difficulties involved in studying it. By pyrosequencing a region of the bacterial 16S rRNA gene, it is now possible to analyze the microbial diversity present in the soil and in the rhizosphere environment. In five agroecosystems, we sampled the rhizospheres of twenty-seven modern maize inbreds and determined the taxonomic richness and relative abundances of bacteria within and between bulk soil and rhizosphere microbiomes. Most of the variation in bacterial diversity existed between fields. However, the maize rhizosphere was considerably differentiated from bulk soil and less taxonomically diverse in all surveyed fields. A significant, but small, proportion of variation in bacterial diversity was identified between maize inbred rhizospheres. Nonetheless, the estimates of relatedness among maize inbreds based on total genetic diversity were not found to be associated with the differentiation of total rhizosphere bacterial diversity.

### **INTRODUCTION**

Exposing the dynamic interactions between plant populations and their natural environment is critical to unraveling the underlying mechanics driving an ecosystem. In agricultural environments, knowledge of these interactions also facilitates crop improvement by directing the breeding of locally adapted germplasm and facilitating the development of more

sustainable agronomic practices. Modern technological advances in both precision agriculture (John V, 2000) and phenomics (Houle et al., 2010) continue to increase the soil, meteorological, and biotic factors we may effectively query for their potential impacts on crop growth and development in a high throughput manner. Although modern advances have greatly improved upon both the accuracy and precision of their measurement, the effects of soil and meteorological factors on crop plants has long been focal points of crop modeling efforts. In contrast, only those macroscopic biotic interactions and plant-microbe symbioses with readily perceivable effects on crop growth have been objects of intense study under natural field conditions.

Well established plant-microbe symbioses include such detrimental pathogenic interactions as those observed between the filamentous fungus *Fusarium* and *Poaceae*, the water mold *Phytophthora* and *Solanaceae*, and the bacteria *Erwinia* and *Rosaceae* (Agrios, 2005). Several beneficial mutualistic symbioses such as the bartering of nitrogen for carbon among arbuscular mycorrhiza fungi and 70-90% of terrestrial plants (Smith, 2008) and between rhizobia diazotrophs and leguminous plants (Fox et al., 2007) are also well known. These canonical plant-microbe symbioses are of outstanding agricultural and ecological significance and exert substantial influence on crop stress tolerance (Rodriguez et al., 2008), quality, and yield (Fox et al., 2007), as well as on an environment's biogeochemical cycle (Falkowski et al., 2008). However, culturable microbes such as these represent less than 1 % of existing soil microbial diversity (Torsvik and Øvreås, 2002), with much of the remaining biodiversity housed in bacterial clades. After accounting for known factors such as genetic diversity between crop varieties, established environmental effects, and recognized symbiotic relationships, substantial variation present in crop growth, development, and nutrient use efficiency still remains to be

explained in many field environments. The vast tracts of microbial diversity recalcitrant to culturing and study by traditional techniques in microbiology may be responsible for this variation. As such, a deeper understanding of this biodiversity and its relationship with crops remains crucial to characterizing the agricultural ecosystem in a more holistic manner and will illuminate the interactions that must be exploited to both improve the breeding of future crops and facilitate the development of superior management practices.

Pyrosequencing and related high throughput sequencing technologies now offer an unprecedented window to perceive the underlying mechanics of the microbial world (Wooley et al., 2010). They are greatly expanding the current repertoire of plant-microbe symbiotic relationships and will soon further detail the many digraphs of ecosystems and established biogeochemical cycles. Through the use of targeted 16s rRNA gene sequencing, we now regularly infer the richness, relative abundance distributions, and phylogeny of Operational Taxonomic Units (OTUs) composing the bacterial populations of an entire natural microbial community or microbiome without the limitations and biases associated with culturing microbes in a laboratory (Wooley et al., 2010). Moreover, estimates of microbiome structure may be directly compared both across and within unique field environments and host genotypes at various spatial and temporal scales to infer elements of specificity and interaction. Although sequence based approaches possess their own inherent amplification and sequencing biases that must be properly addressed (Kim and Bae, 2011; Unterseher et al., 2011), these efforts will continue to unravel the complex mechanics of microbial ecology in a much more quantitative manner than previously possible.

The maize rhizosphere is an important interface in which a deeper understanding of plant-microbe symbiotic interactions is greatly needed. Maize is one of the most economically

significant crops in the world, possesses exceptional phenotypic and molecular diversity (Gore et al., 2009), and is substantially influenced by environmentally conditional genetic variation (Buckler et al., 2009). Also, given its prevalent growth under monoculture, maize may be viewed as an ecosystem engineer strongly responsible for shaping the agricultural environment for cohabitating species. One of a plant's greatest areas of influence is the rhizosphere, where roots may expend up to 21% (Marschner, 1995) of their fixed carbon in exudates such as sugars, organic acids, aromatics, and enzymes. These compounds interact with soil qualities such as pH, water potential, texture, and nutrient availability, as well as existing microbial populations to promote growth and development of the plant. Given these critical interactions and the extensive genetic diversity present in maize as well as previous indications of gross scale genotype specificity of microbes under both greenhouse (Bouffaud et al., 2012) and field conditions (Aira et al., 2010), it is likely rhizosphere microbial communities are influenced by maize host genotypes and their differing root exudation and secretion profiles under field conditions. Furthermore, with sufficient generations of co evolution it is plausible that a significant proportion of inter-specific adaptations occurring in maize and its microbiome may be mutualistic in nature. It is also reasonable to infer these relationships may be contingent upon other microbes, soil qualities, meteorology, and other natural attributes present within an agricultural field environment.

To begin exploring these many hypotheses, we performed targeted high-throughput pyrosequencing of the bacterial 16S rRNA gene. We assessed the taxonomic richness and relative abundance distributions of OTUs in the maize rhizosphere and bulk soil microbiomes derived from five agroecosystems at their median flowering time. Each of these environments possessed unique soil and management conditions as well as replicated plots of twenty-seven

modern maize inbreds. We approximated a balanced randomized complete block design with respect to maize inbreds, field environments, and sample preparation factors such as primer set, PCR amplification batch, and pyrosequencing run. After selecting a primer set with desirable qualities, the effects of field environment, sample type (bulk soil or rhizosphere), and host maize inbred on bacterial alpha and beta diversity as well as the abundances of common OTUs were then inferred by permutation-based analysis of variance and partial canonical principal coordinate analysis.

## RESULTS

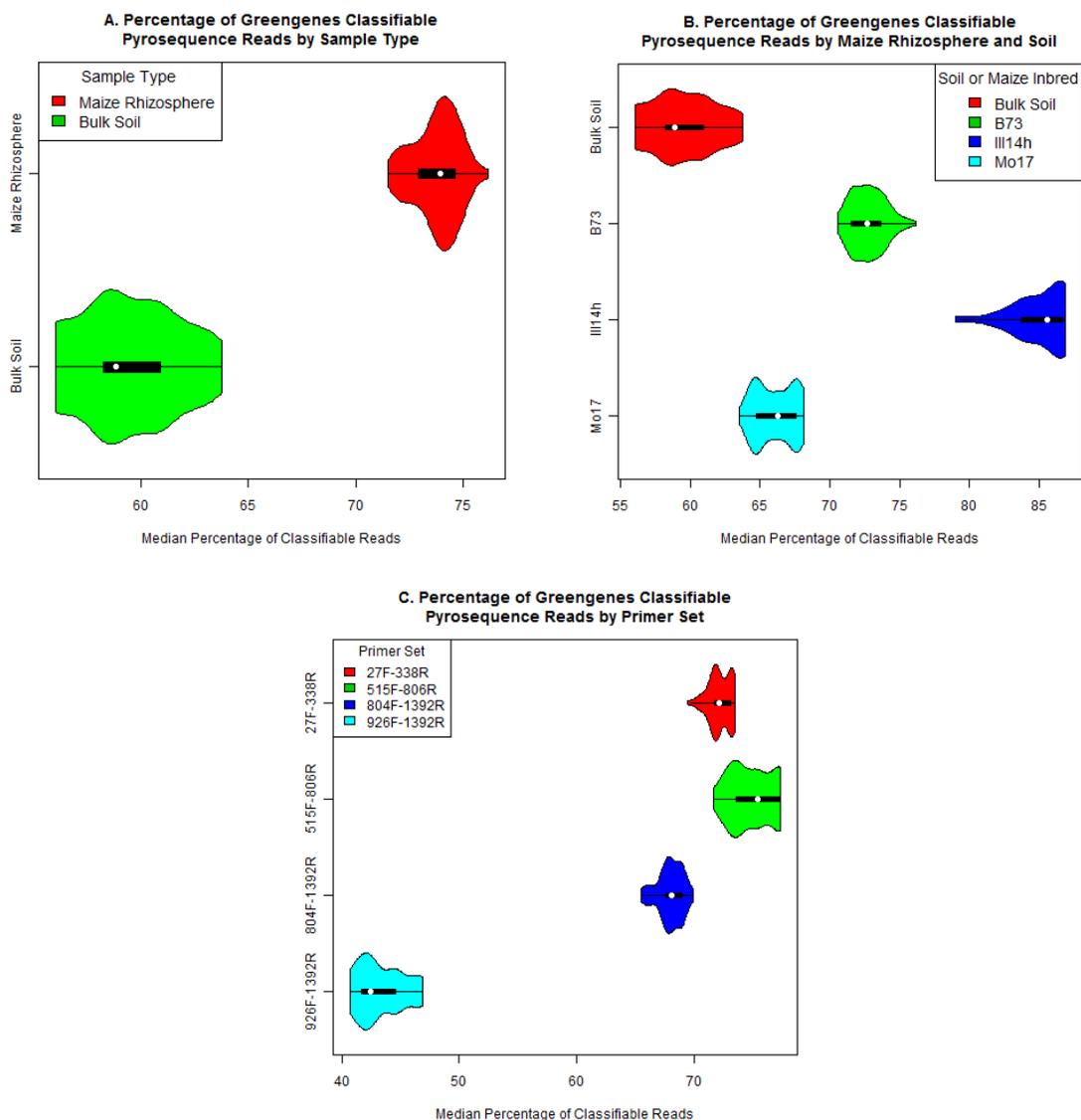
### *Variation in Proportion of Taxonomically Classifiable Diversity*

In a pilot study to select a desirable primer set for discrimination of the microbiome DNA extractions, a subset of the maize rhizosphere and bulk soil samples all collected at median flowering time from a field near Columbia, MO (Table 2.1) were prepared and analyzed. Significant variation in the percentage of total pyrosequence reads that were taxonomically classifiable by the Greengenes 16S rRNA gene database was observed between sample types (bulk soil or rhizosphere), maize inbreds (B73, Mo17, Ill14h), and primer sets (27F-338R, 515F-806R, 804F-1392R, and 926F-1392R). A significantly higher proportion ( $n \geq 17$ ;  $P < 2.00E-04$ ; Figure 2.1A) of the pyrosequence reads were classifiable in microbiome extractions of the maize rhizospheres (73.7%; 95% bootstrapped confidence interval ( $CI_{\text{Bootstrap}} = (71.5\%, 75.6\%)$ )) than those classifiable in bulk soil (58.6%;  $CI_{\text{Bootstrap}} = (54.9\%, 63.7\%)$ ). Similarly, significant variation in the proportion of classifiable pyrosequence reads was observed between maize inbreds ( $n \geq 11$ ;  $P < 2.00E-04$ ; Figure 2.1B). The most classifiable proportion of total reads was observed in the sweet corn inbred Ill14h (85.6%;  $CI_{\text{Bootstrap}} = (79.0\%, 86.8\%)$ ). Significant variation in the proportion of classifiable pyrosequence reads was also observed among the four primer sets (27F-338R, 515F-806R, 804F-1392R, and 926F-1392R) amplifying distinct

hypervariable regions of the 16S bacterial rRNA gene (V1-V2, V3-V4, V5-V7, and V6-V7) respectively ( $n \geq 14$ ,  $P < 2.00E-04$ ; Figure 2.1C). Primer set 515F-806R (V3-V4) amplified a significantly larger proportion of classifiable reads ( $n \geq 14$ ,  $P < 1.20E-03$ ) than the other primer sets tested (75.4%;  $CI_{\text{Bootstrap}} = (71.6\%, 77.3\%)$ ). It also remained reasonably consistent in the proportion of classifiable reads obtained within each of the surveyed microbiome extractions and did not possess a significantly larger variance than the other primer sets. For these reasons, 515F-806R was selected for characterization of the bacterial community profiles in the full set of maize inbred rhizosphere and bulk soil microbiome extractions across all five of the surveyed field environments (Urbana, IL; Columbia, MO; Aurora, NY; Ithaca, NY; Lansing, NY).

**Table 2.1** Pyrosequence abundance means with respect to primer set, sample type, and maize inbred revealed significant variation within each sample; however, no significant interactions were observed between maize inbred and sample type with primer set.

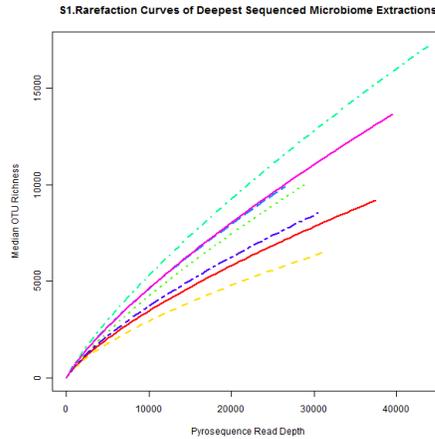
<i>Primer</i>	<i>Maize Inbred</i>	<i>Sample Type</i>	<i>Total Pyrosequence Reads</i>	<i>Pyrosequence Reads (Less Singletons)</i>	<i>Greengenes Classifiable Reads</i>	<i>Proportion Classifiable</i>
27F-338R		Bulk Soil	1093.5 (±608.2)	773.5 (±426.2)	605 (±350.7)	0.77 (±0.06)
27F-338R	B73	Rhizosphere	728.8 (±490)	580.3 (±376.5)	509.8 (±317.3)	0.89 (±0.03)
27F-338R	Ill14h	Rhizosphere	526 (±130.3)	468.7 (±121)	448.3 (±119.3)	0.96 (±0.02)
27F-338R	Mo17	Rhizosphere	502.8 (±73.1)	382.8 (±59.7)	326.3 (±50.5)	0.85 (±0.01)
515F-806R		Bulk Soil	11202.8 (±3695.8)	8307.3 (±2872.6)	7061.8 (±2538.3)	0.85 (±0.03)
515F-806R	B73	Rhizosphere	10355.5 (±1470.7)	8643 (±1343.5)	7674.3 (±1142.1)	0.89 (±0.01)
515F-806R	Ill14h	Rhizosphere	11138 (±1494.1)	9492 (±787.2)	8955.7 (±573.3)	0.94 (±0.03)
515F-806R	Mo17	Rhizosphere	12192 (±3476.4)	9563.5 (±2980.3)	8356.3 (±2750.7)	0.87 (±0.01)
804F-1392R		Bulk Soil	23001.8 (±14331.1)	18525.5 (±11063.6)	15125.8 (±9042.3)	0.82 (±0.03)
804F-1392R	B73	Rhizosphere	28840.5 (±8579.9)	26182.5 (±8339.4)	22085 (±7037.6)	0.85 (±0.02)
804F-1392R	Ill14h	Rhizosphere	17484 (±3547.5)	15773.7 (±2729.4)	14561.7 (±2209.1)	0.93 (±0.03)
804F-1392R	Mo17	Rhizosphere	27369 (±11531.6)	23944.8 (±9872.3)	19571.8 (±8296.6)	0.81 (±0.02)
926F-1392R		Bulk Soil	4509 (±109)	2861.5 (±198.2)	1968.8 (±170)	0.69 (±0.02)
926F-1392R	B73	Rhizosphere	5064 (±502.2)	4028.8 (±360.9)	2128.5 (±687.2)	0.53 (±0.13)
926F-1392R	Ill14h	Rhizosphere	4115 (±528.9)	3661.5 (±649.8)	3357 (±767.9)	0.92 (±0.05)
926F-1392R	Mo17	Rhizosphere	5505.8 (±854.3)	4133.5 (±886.6)	2292.3 (±482.8)	0.56 (±0.08)



**Figure 2.1** The median percentage of 16s ribosomal RNA gene pyrosequence reads that were taxonomically classifiable by the GreenGenes database varied by sample types, maize inbred rhizospheres, and primer sets in the pilot study. (A) The microbial diversity sequenced within bulk soil was significantly less classifiable than that observed within the maize rhizosphere ( $n \geq 16$ ,  $P < 2.00E-04$ ). The proportion of classifiable reads was also significantly less disperse across bulk soil extractions. (B) Maize inbreds possessed significant differences in the proportion of classifiable reads observed in their rhizosphere microbiomes ( $n \geq 11$ ,  $P < 2.00E-04$ ). (C) With significant differences observed between 926F-1392R (V6-V7) and the remaining primer sets, differential amplification of classifiable reads was also apparent ( $n \geq 14$ ,  $P < 1.20E-03$ ). Primer set 515F-806R was selected for further experimentation across all five fields.

### *Variation in Alpha Diversity*

Alpha diversity is an inventory of the species richness (Jurasinski et al., 2009) observed in an environment without regard to its taxonomic similarity or the proportion of species common to other environments. It is well known the richness of unique species observed is a function of the sampling efforts taken to identify them, and that this relationship is not linear when sampled to a depth sufficient for exhaustive characterization. In highly multiplexed microbial sequence profiling studies these realizations are critical (Wooley et al., 2010). It is not possible to ensure a negligible coefficient of variation in pyrosequence read depths or sampling efforts between microbiome extractions and therefore impossible to directly compare OTU abundances. To address this issue, rarefaction was performed using QIIME (Caporaso et al., 2010b). Rarefaction curves detailing observed OTU richness as a function of pyrosequencing efforts revealed none of the microbiome extractions possessed sufficient read depth to approach exhaustion of the OTU richness present in any extraction. All rarefaction curves were nearly linear revealing much of the expansive microbial diversity present within the rhizosphere and bulk soil remained unsampled. Even the most deeply sampled extractions within the study, possessing pyrosequence read depths in the range of 25-50,000 sequence reads, were never limited in the rate at which they revealed novel OTUs (Figure 2.2).

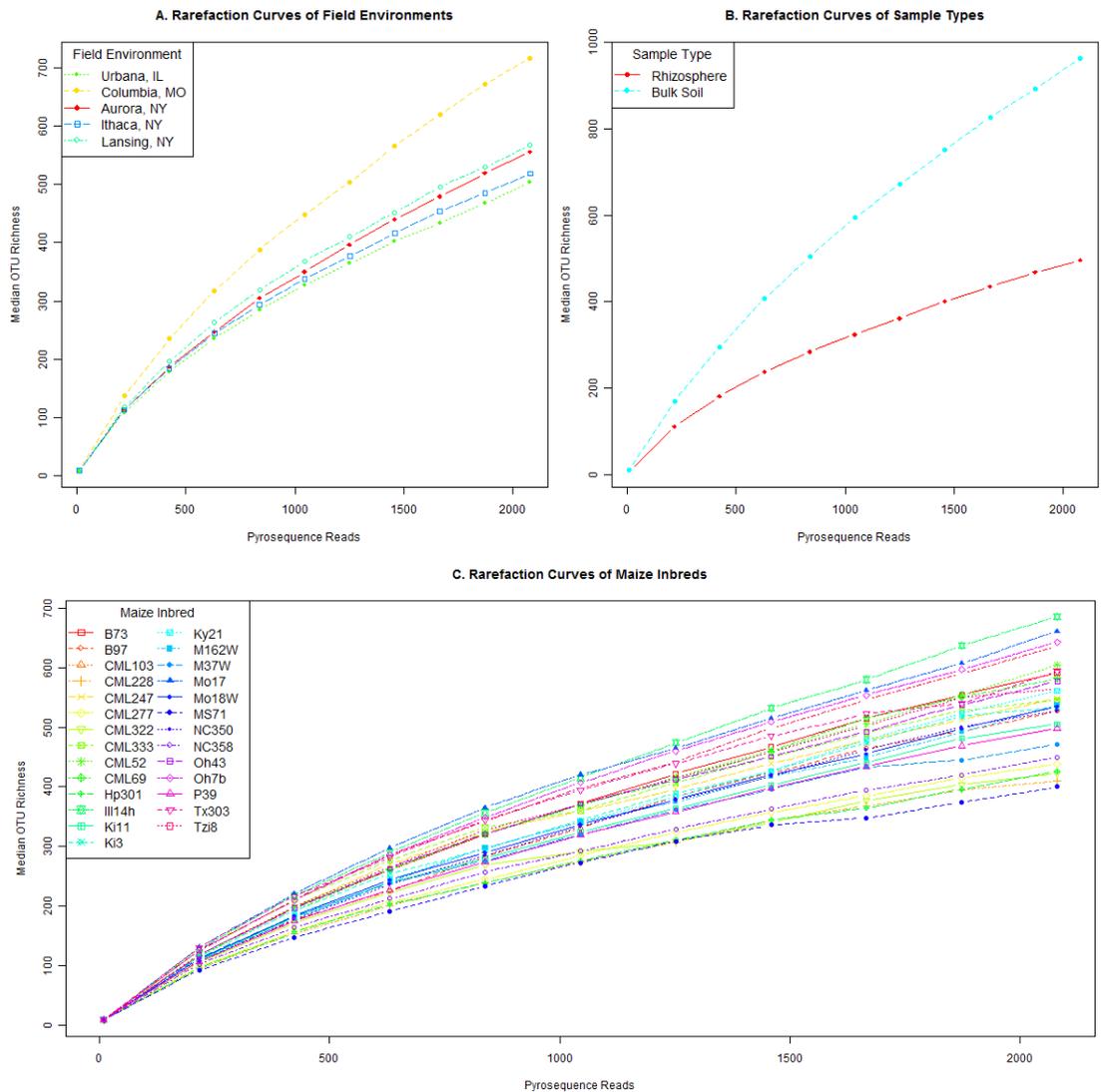


**Figure 2.2** Rarefaction curves of the seven deepest microbiome extractions within this study all remained nearly linear revealing little sign of any plateau or decline in available novel OTUs at even the highest sampled Greengenes classifiable pyrosequencing read depths (44,370 reads).

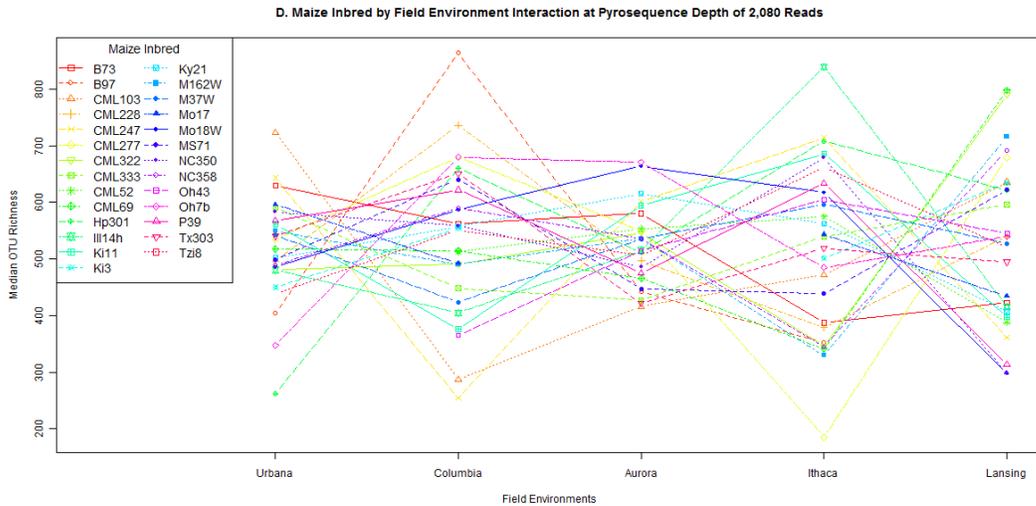
During rarefaction, equidistant pyrosequence read depths were selected ranging from 10 to 2,080 reads. Rarefied values from each sequence depth of every extraction were then bootstrapped one hundred times to maintain a balanced design with respect to the levels of the factors of interest. Median values for each environment, sample type, and maize inbred were then plotted as a function of sampling depth and the distributions of observed OTU richness at 2,080 pyrosequence reads were compared by permutation testing. Significant variation was explained by field environments (20.0%;  $P < 2.00E-04$ ;  $CI_{\text{Bootstrap}} = (19.8\%, 20.4\%)$ ; Figure 2.3A), sample types (32.3%;  $P < 2.00E-04$ ;  $CI_{\text{Bootstrap}} = (32.0\%, 32.6\%)$ ; Figure 2.3B), and maize inbreds (19.1%;  $P < 2.00E-04$ ;  $CI_{\text{Bootstrap}} = (18.9\%, 19.2\%)$ ; Figure 2.3C) after controlling for both the pyrosequencing run, PCR amplification batch, and the remaining factors. The most OTU rich field was located near Columbia, MO ( $n \geq 258$ ,  $P < 8.00E-04$ ). However, the remaining field environments located near Urbana, IL and the three fields sampled in New York did not significantly differ in OTU richness at current power. Similarly, the organically managed field in Ithaca, NY was not found to significantly differ in OTU richness from the conventionally managed field environments of the data set.

Further tests of interaction terms for sample type within each field environment revealed small but significant interactions in OTU richness (6.1%;  $P < 2.00E-04$ ;  $CI_{\text{Bootstrap}} = (5.8\%, 6.2\%)$ ) after accounting for both the main effect of sample type and field environment as well as pyrosequencing run and PCR amplification batch. Nonetheless, in every comparison the rhizosphere was identified as a more selective environment, less rich in bacterial diversity than bulk soil ( $n \geq 21$ ;  $P < 2.00E-03$ ). Partitioning variation in alpha diversity between rhizosphere within each environment also captured a significant proportion of total variation in OTU richness (48.7%;  $P < 2.00E-04$ ;  $CI_{\text{Bootstrap}} = (48.5\%, 49.1\%)$ ) after accounting for the main effect of maize inbred field environment and the remaining model factors. In further confirmation of the substantial maize inbred by field environment interaction, no strong correlation among the OTU richness estimates of maize inbreds between field environments were observed at a read depth of 2,080 reads after adjusting for field environment, maize inbred, pyrosequencing run, and PCR amplification batch effects (Figure 2.3D).

Rarefaction and analogous bootstrapped permutation tests to those performed to discern variation between field environments, sample types, and maize inbreds for observed OTU richness were also executed for the Chao-1 estimator of total OTU richness and the whole tree phylogenetic diversity estimator that considers the phylogenetic relatedness among observed species. However, these test results remained comparable to those obtained for observed OTU richness. This revealed the abundances of rare OTU used to infer Chao-1 estimates of total OTU richness did not substantially differ between fields, sample types, or maize inbreds. Similarly, comparable results across the estimates of whole tree phylogenetic diversity suggested the degree of taxonomic relatedness among OTU was not substantially different between the various levels of these factors.



**Figure 2.3** Rarefaction curves of alpha diversity revealed significant variation between microbiome extractions. **(A)** Variation in OTU richness by field environment was significant ( $P < 2.00E-04$ ); however, the OTU enriched field near Columbia, MO exhibited the only significant pairwise difference ( $n \geq 258$ ,  $P < 2.00E-03$ ). **(B)** Variation in OTU richness by sample type revealed the relative enrichment of bulk soil compared to that observed in the maize rhizosphere ( $P < 2.00E-04$ ). **(C)** Variation in OTU richness by maize inbred was significant across all field environments ( $P < 2.00E-04$ ).



**Figure 2.4** Substantial maize inbred by environment interactions were also observed ( $P < 2.00E-04$ ) for OTU richness and rank among maize inbreds within each field was not well maintained.

#### *Variation in Beta Diversity*

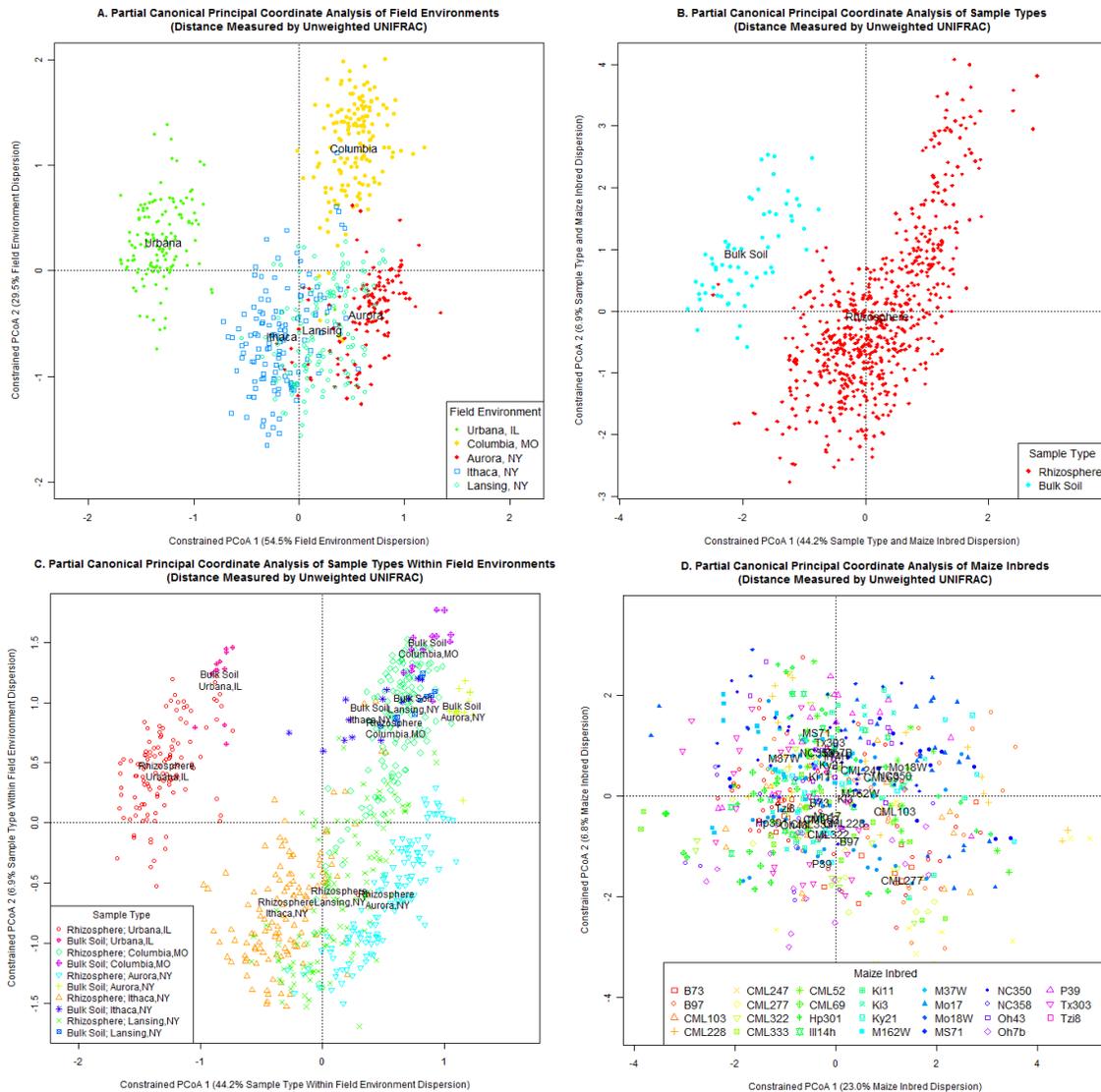
Although the resemblance of whole tree phylogenetic diversity and observed OTU richness suggested no substantial differences in the magnitude of OTU relatedness was present among field environments, sample types, and maize inbreds, estimates of beta diversity enable differentiation based on proportions of shared OTU. Unweighted and weighted UniFrac distance metrics, are two commonly employed measures of beta diversity that facilitate this contrast (Lozupone et al., 2007). While both distance metrics consider the phylogenetic relatedness of microbiome extractions, unweighted distances reflects the presence or absence of bacterial lineages whereas weighted distances detect abundance differences among these lineages. Using partial canonical principal coordinate analysis (Anderson and Willis, 2003), the dispersion in beta diversity as measured by unweighted and weighted UniFrac was partitioned into that attributable to fields, sample types, and maize inbreds (Figure 2.4, 2.5). While field environments, sample type, and maize inbred could explain slightly more of the total variation of the data set when measured in weighted than unweighted UniFrac distances, these differences

were not as readily captured by the two largest principal coordinates upon ordination.

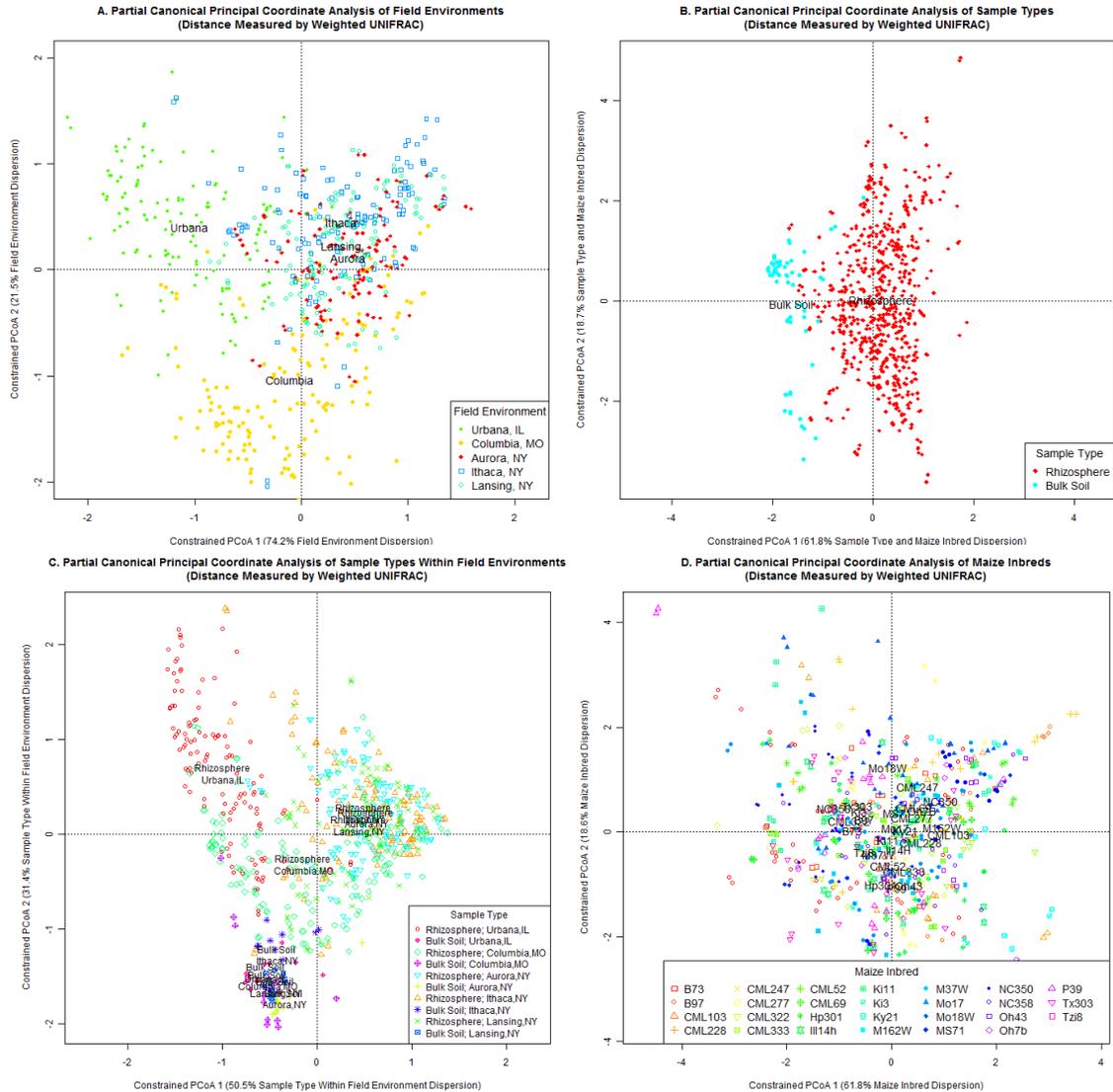
Field environments were found to explain the most variation in beta diversity by both unweighted (13.6% ;  $P < 5.00E-03$ ;  $CI_{\text{Bootstrap}} = (12.7\%, 14.4\%)$ ; Figure 2.5A) and weighted UniFrac distance metrics (18.3%;  $P < 5.00E-03$ ;  $CI_{\text{Bootstrap}} = (14.6\%, 18.5\%)$ ; Figure 2.6A) after conditioning on sample types, maize inbreds, amplification batch and sequence run. Comparing distances between centroids revealed field environments surveyed in New York were more similar to each other than they were to fields surveyed in the other states ( $n \geq 300$ ,  $P < 2.00E-04$ ). No significant differences were observed between the organically managed field located in Ithaca, NY and the conventionally managed field environments. Similarly, no significant differences in within field dispersion among microbiome extractions were noted.

Maize rhizosphere and bulk soil were also found to explain a substantial proportion of the beta diversity across the field environments surveyed for unweighted (29.6%;  $P < 2.00E-04$ ;  $CI_{\text{Bootstrap}} = (24.9\%, 31.2\%)$ ; Figure 2.5B) and weighted (46.7%;  $P < 5.00E-03$ ;  $CI_{\text{Bootstrap}} = (44.5\%, 48.8\%)$ ; Figure 2.6B) UniFrac distance metrics after conditioning on field environment, amplification batch and sequencing run. While significant, the proportion of variation captured by sample type within field environment remained small in both unweighted (3.7%;  $P < 5.00E-02$ ;  $CI_{\text{Bootstrap}} = (0.8\%, 4.7\%)$ ; Figure 2.5C) and weighted UniFrac measures (1.6%;  $P < 5.00E-02$ ;  $CI_{\text{Bootstrap}} = (1.0\%, 1.9\%)$ ; Figure 2.6C). In all environments, microbiome extractions collected from the maize rhizosphere were more disperse in beta diversity than those collected from bulk soil extractions. To discern if the variation among rhizosphere samples had a heritable component, the proportion of beta diversity between maize inbreds across and within each of the field environments was discerned. Comparable estimates of beta diversity were determined between maize inbred rhizospheres across the field environments for both unweighted (5.0%;  $P$

$< 5.00E-02$ ;  $CI_{\text{Bootstrap}} = (4.8\%, 5.6\%)$ ; Figure 2.5D) and weighted UniFrac (7.7%;  $P < 5.00E-02$ ;  $CI_{\text{Bootstrap}} = (7.1\%, 15.4\%)$ ; Figure 2.6D). However, the proportion of heritable beta diversity captured across all the fields was substantially less than that captured by the maize inbreds randomized and replicated within each field environment for both unweighted (17.9%;  $P < 5.00E-03$ ;  $CI_{\text{Bootstrap}} = (14.3\%, 19.9\%)$ ) and weighted (25.3%;  $P < 5.00E-03$ ;  $CI_{\text{Bootstrap}} = (21.8\%, 27.7\%)$ ) UniFrac distance measures.



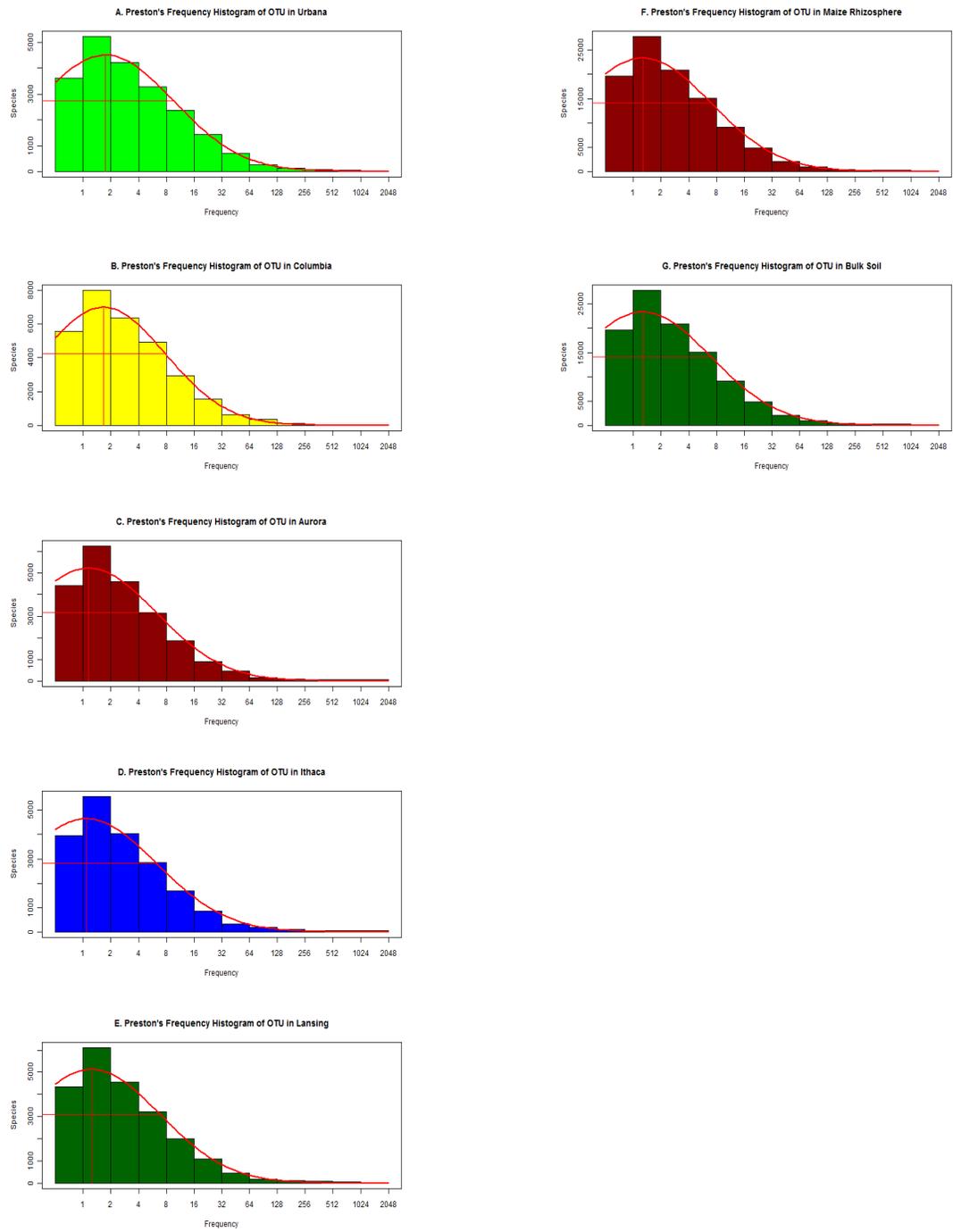
**Figure 2.5** Beta diversity revealed substantial differentiation between microbiome extractions derived from unique field environments, sample types, and maize inbreds. **(A)** Variation in unweighted UniFrac dispersion by field environment was significant ( $P < 5.00E-03$ ); however, New York environments tended to cluster more tightly. **(B)** Variation in beta diversity revealed large distinction between bulk soil and the maize rhizosphere ( $P < 5.00E-03$ ). **(C)** Variation in beta diversity by bulk soil and maize rhizosphere was substantial within all field environments ( $P < 5.00E-02$ ). Nonetheless, after accounting for main effect differences between soil and rhizosphere field specific difference in this measure were minimal. **(D)** A small but significant proportion of the variation in unweighted UniFrac distances was identified between maize inbreds ( $P < 5.00E-02$ ). However, it pailed in comparison to that observed between maize inbreds within each field environment ( $P < 5.00E-03$ ).



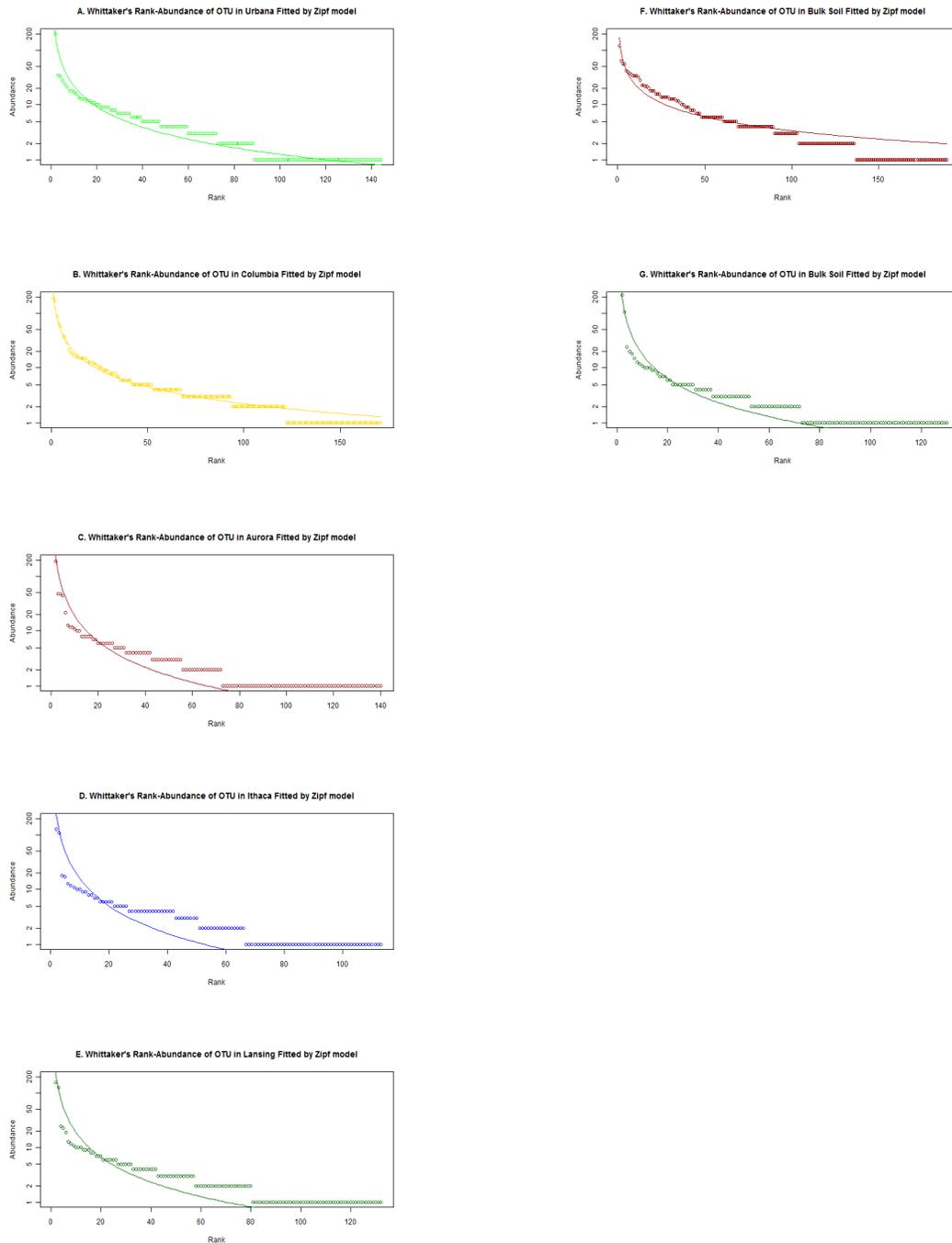
**Figure 2.6** Beta diversity as measured by weighted UniFrac revealed substantial differentiation between microbiome extractions derived from unique field environments, sample types, and maize inbreds. **(A)** Variation in weighted UniFrac dispersion by field environment was significant ( $P < 2.00E-04$ ); however, New York environments tended to cluster more tightly. **(B)** Variation in beta diversity revealed large distinction between bulk soil and the maize rhizosphere ( $P < 2.00E-04$ ). **(C)** Variation in beta diversity by bulk soil and maize rhizosphere was substantial within all field environments ( $P < 2.00E-04$ ). **(D)** A small but significant proportion of the variation in weighted UniFrac distances was identified between maize inbreds ( $P < 2.00E-04$ ). However, it paled in comparison to that observed between maize inbreds within each field environment ( $P < 2.00E-04$ ).

### *Variation in Common OTU Abundances*

Both Preston's frequency histogram (Figure 2.7) and Whittaker's rank abundance plots (Figure 2.8) detailing the relative abundance of OTU diversity were constructed for each field environment, and sample type from abundances rarefied to a depth of 2,080 pyrosequence reads. However, no substantial differences were noted between these distributions. In agreement with both the rarefaction curves and estimates of Chao1 total diversity, Preston's lognormal abundance histogram for all microbiome extractions were truncated near their mean and suggested substantial proportions of OTU diversity remained veiled and was not yet observed in all the environments and sample types. Zipf models characterizing an OTU's frequency as inversely proportional to its rank was found to most reasonably approximate (by Bayesian Information Criterion) the observed rank abundance distributions in all environments when compared to models constructed from both lognormal and geometric series. Nonetheless, no significant differences in distributions were noted between fields, and sample types. Most OTU diversity was found at sequence read depths at the level of singletons and doubletons.



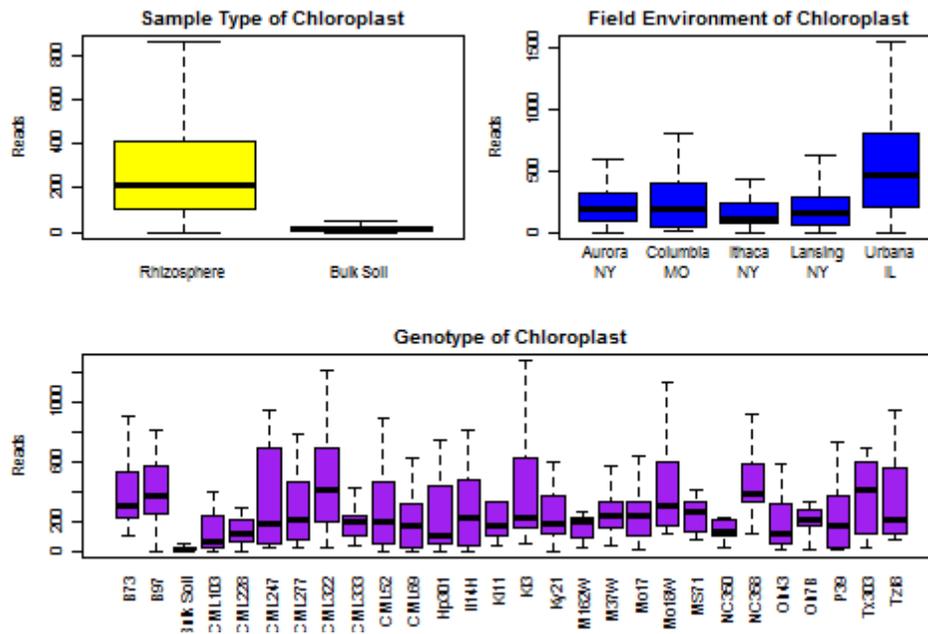
**Figure 2.7** Preston's lognormal models of abundance data reveal substantial proportions of veiled OTU as compared to the observed OTU richness. Also, little differentiation in abundance levels was noted between field environments or sample types.



**Figure 2.8** Whittaker's Rank Abundance Curves were plotted for the median abundance of each OTU across microbiome extractions for samples within each field environment as well as bulk soil and maize rhizosphere samples. Geometric, lognormal, and Zipf-Mandelbrot models were tested; however, the best fit model was obtained from a Zipf distribution as determined by BIC.

Given the exceptional levels of OTU richness, and limited sequencing depth, only the most abundant OTU and those of the highest taxonomic ranks could be quantified with a level of precision sufficient to compare them on an individual basis. While significant variation existed between the surveyed environments and maize inbred rhizospheres, the most marked contrast in the abundance of microbial taxa was observed between the maize rhizosphere and bulk soil samples. The primer set chosen to characterize the full data set was selected due to its enrichment of classifiable sequences as well as reduced amplification of chloroplast related sequences.

Nonetheless, chloroplast sequences still remained the most significantly enriched OTU in the rhizosphere microbiome extractions as compare to bulk soil ( $n \geq 120$ ,  $P < 4.00E-04$ ; Figure 2.9). Bacterial taxa with confirmed enrichment in the rhizosphere microbiome relative to bulk soil included *Burkholderia* ( $n \geq 120$ ,  $P < 2.00E-04$ ), *Oceanospirillales* ( $n \geq 120$ ,  $P < 2.00E-04$ ), and *Sphingobacteriales* ( $n \geq 120$ ,  $P < 2.00E-04$ ). In contrast, other phyla such as *Acidobacteria* ( $n \geq 120$ ,  $P < 2.00E-04$ ), *Chloroflexi* ( $n \geq 120$ ,  $P < 2.00E-04$ ), *Planctomycetes* ( $n \geq 120$ ,  $P < 2.00E-04$ ), and *Verucomicrobia* ( $n \geq 120$ ,  $P < 2.00E-04$ ) were found at higher concentrations within bulk soil as compared the rhizosphere samples. While variation in common OTU abundances was identified between maize inbreds, no outlying maize inbreds or trends with respect to relatedness were observed.



**Figure 2.9** Although the selected primer set was chosen not to enrich for maize chloroplast sequences, those OTU corresponding to chloroplasts remained the most significant contrast observed when comparing enrichment in the rhizosphere relative to bulk soil ( $P < 5.00E-95$ ).

### *Associations between Soil Bacterial Diversity and Physical properties*

Soil properties of fifteen randomly selected bulk soil samples were measured across each of the five field environments. These revealed substantial variation in the relative abundances of nitrate, phosphorous, potassium, and several other nutrients and minerals. Significant pairwise differences in all characteristics except moisture content were observed between field environments after Bonferroni multiple test correction (Table 2.2). These characteristics were each correlated with estimates of alpha and beta diversity as well as the rarefied abundances of the top 100 most common OTUs; however, no significant trends were observed across the bulk soil microbiome extractions. Similarly, after standardization to a common range all soil nutrients were used to construct a correlation matrix detailing soil relatedness among the environments. This matrix was correlated with estimates of beta diversity by both the Mantel and Procrustean superimposition tests; however, no trends were observed.

**Table 2.2** Soil property means of 15 measurements ( $\pm$  std error).

<i>Property</i>	<b>Aurora, NY</b>	<b>Lansing, NY</b>	<b>Ithaca, NY</b>	<b>Columbia, MO</b>	<b>Urbana, IL</b>
<i>Moisture</i>	2.08 ( $\pm 0.15$ )	2.13 ( $\pm 0.18$ )	1.93 ( $\pm 0.27$ )	1.92 ( $\pm 0.25$ )	2.16 ( $\pm 0.41$ )
<i>pH</i>	7.65 ( $\pm 0.24$ )	6.64 ( $\pm 0.13$ )	6.14 ( $\pm 0.26$ )	6.5 ( $\pm 0.23$ )	5.44 ( $\pm 0.55$ )
<i>Organic Matter</i>	2.57 ( $\pm 0.19$ )	3.34 ( $\pm 0.32$ )	3.89 ( $\pm 0.37$ )	3.09 ( $\pm 0.45$ )	3.89 ( $\pm 0.25$ )
<i>NO<sub>3</sub></i>	69.6 ( $\pm 32.02$ )	84.6 ( $\pm 29.2$ )	40.13 ( $\pm 9.37$ )	36.69 ( $\pm 25.74$ )	26.73 ( $\pm 18.08$ )
<i>P</i>	15.93 ( $\pm 7.22$ )	52.13 ( $\pm 14.64$ )	21.46 ( $\pm 6.88$ )	9.67 ( $\pm 3.82$ )	4.47 ( $\pm 8.84$ )
<i>K</i>	120.67 ( $\pm 52.26$ )	435.33 ( $\pm 162.33$ )	294 ( $\pm 111.21$ )	129.68 ( $\pm 22.24$ )	151.17 ( $\pm 36.44$ )
<i>Mg</i>	658.67 ( $\pm 52.59$ )	411 ( $\pm 59.76$ )	257 ( $\pm 49.24$ )	229.3 ( $\pm 41.12$ )	412.92 ( $\pm 95.31$ )
<i>Ca</i>	5087.33 ( $\pm 1418.34$ )	4236 ( $\pm 1336.23$ )	3139.33 ( $\pm 584.94$ )	2313.99 ( $\pm 295.54$ )	2138.75 ( $\pm 150.81$ )
<i>Fe</i>	1.2 ( $\pm 0.45$ )	1.77 ( $\pm 1.19$ )	4.27 ( $\pm 1.22$ )	1.61 ( $\pm 0.83$ )	3.174 ( $\pm 1.01$ )
<i>Al</i>	7.8 ( $\pm 2.45$ )	16.06 ( $\pm 5.24$ )	30.6 ( $\pm 11.11$ )	7.43 ( $\pm 3.57$ )	26.74 ( $\pm 5.54$ )
<i>Mn</i>	16.46 ( $\pm 5.94$ )	21.6 ( $\pm 16.06$ )	17.53 ( $\pm 3.31$ )	33.02 ( $\pm 5.86$ )	36.78 ( $\pm 5.54$ )
<i>Zn</i>	0.45 ( $\pm 0.15$ )	4.2 ( $\pm 1.24$ )	0.48 ( $\pm 0.15$ )	0.62 ( $\pm 0.52$ )	0.82 ( $\pm 0.45$ )

### *Associations between Maize Rhizosphere Bacterial Diversity and Kinship*

To determine if the diversification history of maize and its flow of total genetic diversity could explain the beta diversity between maize inbred rhizospheres a relationship matrix between all twenty-seven maize inbreds was constructed from the over 1.4 million polymorphisms composing the First Generation Maize Hapmap (Gore et al., 2009). Genetic relatedness or kinship among the lines was calculated by percent identity by state (Gore et al., 2009). Estimates of both alpha and beta diversity were tested for correlation with total genetic relatedness among the lines. Similarly, rarefied abundances of common OTUs were also compared to estimates of total genetic relatedness among the maize inbred rhizospheres. Despite the significant heritability noted between maize inbreds, the simple additive model constructed from the total genetic diversity captured by the twenty-seven maize inbreds was not significantly correlated with the rhizosphere OTU richness or beta diversity as measured by both weighted and

unweighted UniFrac distance metrics in Mantel or Procrustean superimposition tests. Likewise, no significant correlations with rarefied abundances of abundant OTU were observed.

## DISCUSSION

The fundamental goal of many ecological studies is to appreciate how the behavior of a biological system characterized at one level naturally influences that of other levels within the hierarchy, and to determine if any robust feedback loops exist between these levels. As technological advances continue to provide us with an increased ability to accurately and precisely resolve additional levels of the hierarchy, we must venture to describe and relate these phenomena to the rest of the system in a comprehensive manner. In this study, we performed pyrosequencing of a hypervariable region of the 16s rRNA gene to characterize the microbial community structure within bulk soil and the rhizospheres of twenty-seven genetically diverse modern maize inbreds across five unique field environments. Significant variation in the taxonomic richness, bacterial composition, and the relative abundances of several common bacterial taxa were observed between field environments, bulk soil and the rhizosphere, as well as between maize inbred rhizospheres.

The most substantial variation in bacterial community composition was observed between the five field environments surveyed. These field environments enabled us to much more robustly test the differences observed between bulk soil and the maize rhizosphere as well as between maize inbred rhizospheres. However, this number of environments does not provide sufficient information to delineate robust factors of causation for environmental differences. Characterizations of soil profiles within each surveyed field environment did not reveal significant similarities to microbial beta diversity as measured by Weighted or Unweighted

UniFrac distance. Similarly, the organic management of Ithaca as compared to conventional management performed in the remaining locations did not have a discernible effect. The only observed trend in bacterial composition across the five fields surveyed was clustering by each field's geographic proximity. All fields surveyed in New York had substantially lower beta diversity to each other than those surveyed in the remaining locations. One may infer commonality of climate plays a significant role in shaping the similarity in bacterial community profiles of the proximal field environments in New York; but, further testing of additional environments is required. Furthermore, UniFrac distance metrics of the total microbial community structure adequately represent differences in the taxonomic profile of the environments; however, they may not well capture the functional differences between the microbiomes present and other means to ascribe functional distances in the microbiomes differentiating environments are necessary. Improved precision in measuring bacterial community composition may only be attained by sequencing to much higher depths across more environments to characterize OTU abundance differences at lower taxonomic ranks and allow analyses of more than just the most common bacterial taxa present. This will enable future associations of this diversity with the meteorological or soil characteristics of unique environments.

The selective reduction in bacterial diversity observed in the maize rhizosphere as compared to bulk soil has been recognized in several studies. It is well established the rhizosphere is both metabolically busier, and a more competitive environment than bulk soil. While this increased competition for resources leads to a reduction in the total bacterial diversity present, it also enriches for several bacterial taxa which form loose symbiotic relationships with the rhizosphere in order to attain carbon resources from the plant root exudates, secretions,

mucilages, mucigels and lysates. Next to the variation noted between maize field environments, the contrast between bulk soil and rhizosphere were the next largest and most significant within the study. The observed reduction in microbial alpha diversity and extensive beta diversity between soil and rhizosphere was observed in all surveyed in environments. Furthermore, several common OTU such as those of *Burkholderiales*, *Oceanospirillales*, and *Sphingobacteriales* classes were found significantly enriched in the maize rhizospheres and other phyla such as *Acidobacteria*, *Chloroflexi*, *Planctomycetes*, and *Veruccomicrobia*, were depleted in the rhizosphere when compared to bulk soil. While most bacterial taxa of lower rank were not sequenced at a high enough read depth to enable powerful comparisons, several interesting relationships such as the enrichment of the aromatic carbon-degrading *Sphingobium herbicidium* and other carbon seeking taxa were observed enriched within the maize rhizosphere.

Characterization of host genotype-specific symbioses remains a primary interest in microbial ecology, and harbors some of the most promise in terms of applicability to crop improvement. Should robust feedback loops between microbial community structure and plant genotypes or phenotypes be discerned, they may provide avenues to breed improved crop varieties which augment or diminish these symbiotic relationships for crop improvement. Unfortunately, this has proven to be one of the most difficult and least robust associations discovered in this and several previous studies. In concurrence with past studies relating the diversification history of maize with rhizosphere bacterial profiles based on other distance metrics of microbial beta diversity and maize genetic distance, no significant associations between the total genetic diversity of maize and the total microbial diversity existing within its rhizosphere could be discerned (Bouffaud et al., 2012). Significant maize genotype by environment interaction was observed to explain a substantial portion of the variation in both

alpha and beta diversity between rhizospheres. Nonetheless, a small but significant fraction of variation in both alpha and beta diversity were characterized between maize inbred rhizospheres across all of the surveyed environments. This suggests that while total genetic variation as measured by identity by state among maize inbreds does not well explain total microbial variation as measured by Weighted and Unweighted UniFrac, some portion of microbial variation is explained by differences between maize inbreds. The question remains what segregating alleles are responsible for this variation, what phenotypic differences do they encode between inbred rhizospheres, and what are the precise differences in microbial diversity accountable for the heritability noted in total microbial diversity.

This study has surveyed more intra-species plant diversity for relationships with its rhizosphere microbial community profile than any prior study using 16s rRNA sequencing to date. Nonetheless, substantially more maize diversity and a deeper or more focused sequencing effort of the existing rhizosphere microbial diversity are necessary to characterize the symbioses that exist under natural environmental conditions. Surveying maize landraces and a larger pool of diversity capturing the allelic variation that existed prior to breeding for adaptation to the heavily fertilized field environments of modern industrial agriculture may reveal additional functional alleles and additional symbiotic relationships that were not captured within this analysis. Without prior biological insights of the molecular mechanics responsible for governing novel symbioses, hundreds to thousands of deeply sequenced microbial community profiles derived from maize rhizospheres replicated across multiple field environments are likely necessary to robustly infer the desired associations. While current advances in high throughput sequencing ensure this will soon be a feasible endeavor, it remains a current limitation to those studies seeking to discern the genetic basis of plant-microbial symbiosis under field conditions.

## MATERIALS AND METHODS

### *Maize Germplasm, Microbiome Sample Collection, and Soil Sample Analysis*

Twenty-seven diverse maize inbreds, all founder genotypes of the Nested Association Mapping panel (NAM), were selected to maximize genetic dissimilarity using previously established genotypic data (Yu et al., 2008). Seed for each of these inbreds were attained from a uniform stand grown at Muskgrave Research Station in Aurora, NY in 2009. In 2010, these lines were hand planted in a randomized complete block design in five field environments located in three states (University of Illinois - Crop Sciences Research and Education Center near Champaign-Urbana, IL (Drummer silty-clay loam soil); University of Missouri – South Farm near Columbia, MO (Mexico silt loam soil); Cornell University - Muskgrave Research Station near Aurora, NY (Honeoye silt loam soil); Cornell University - Ketola Organic Research Farm near Ithaca, NY (Erie Channery silt loam soil); Willet Dairy near Lansing, NY (Lyons silt loam soil)). Conventional cultural practices were employed including ammonium nitrate-based fertilization, weed, and pest control in all locations except Ketola Research Farm wherein an organic management regime was implemented including manure-based fertilization and no pesticide or chemical weed control. The rhizosphere microbiomes of all maize inbred plots as well as bulk soil samples were collected at their mean anthesis, approximately twelve weeks after planting. Plants were carefully removed from the soil using a drain spade. Avoiding border effects potentially attributable to increased nutrient availability in the end plant of a plot, the roots of three random plants were sampled from the middle of each plot composed of between twelve and twenty-five plants (varying by environment). An approximately 5cm long root segment of 0.5-3mm in diameter was collected near the base of the plant along with any adherent soil particles. Bulk soil samples across each of the fields were also taken mid-range between

maize plots. In preparation for total genomic DNA extraction, all samples were chilled on ice immediately following collection. All soil analyses were subsequently performed by the Cornell University Nutrient Analysis Laboratory using their standard operational procedures for the identification of extractable phosphorous and nitrate by the Morgan test method, as well as potassium, calcium, magnesium, iron, manganese, zinc, and aluminum by an inductively coupled plasma atomic emission spectrometer (ICP-AES). Buffer pH was discerned by the Modified Mehlich buffer test and organic matter by loss on ignition.

#### *DNA Extraction, PCR Amplification, Quantification, Pooling, and Pyrosequencing*

After all samples were thoroughly homogenized using a bead beater at maximum speed for approximately two minutes, total genomic DNA was isolated from the maize root tip associated soil and about 0.25 g of bulk soil using the PowerSoil High Throughput DNA Isolation Kit (Mo Bio Laboratories Inc., NY). The remaining DNA extraction steps were all performed following the standard operating procedures given by the manufacturer. Following DNA extraction, total 16S rRNA genes were amplified from each sample using primer set 515F-806R to amplify the V3-V4 region of the 16S subunit. The PCR primers were constructed as follow: forward primer = 454 Titanium Lib-I Primer A/5-base barcode/forward 16S primer and reverse primer = 454 Titanium Lib-I Primer B/reverse 16S primer. All PCR reactions were carried out in triplicate 50  $\mu$ L reactions with 5  $\mu$ L of Easy-A 10X buffer, 0.25  $\mu$ L Easy-A Taq, 1  $\mu$ L of 10 $\mu$ M forward and reverse primers, 7  $\mu$ L MgCl<sub>2</sub>, 1  $\mu$ L of dNTP and about 50 ng template DNA. Thermal cycling consisted of initial denaturation at 95°C for 2 min, followed by 30 cycles of denaturation at 95°C for 30 s, annealing at 53°C for 20 s, and elongation at 72 °C for 60 s. Negative control samples were treated similarly with the exclusion of template DNA; these negative controls failed to produce visible PCR products. Following PCR, DNA amplicons were

purified with Ampure magnetic purification beads (Agencourt) and quantified using the Quant-iT Picogreen dsDNA Assay Kit (Invitrogen). Amplicons were then combined in equimolar ratios into a single tube with a final concentration of 12.5 ng/ $\mu$ L. Pyrosequencing was then performed using Roche Titanium chemistry at the Department of Energy Joint Genome Institute.

### *16S rRNA Gene Pyrosequence Analysis and Taxonomic Classification*

Sequences were analyzed using the QIIME software package (Quantitative Insights into Microbial Ecology) using default parameters for each step (Caporaso et al., 2010c). Sequences were removed if their lengths were greater than 200 nucleotides, they contained ambiguous bases, primer mismatches, homopolymer runs in excess of six bases or error in barcodes. More than 3.8 million quality-filtered reads were obtained for the samples, an average of 8,315 reads per sample (min = 2,225, max = 22,346). Similar sequences were clustered into OTUs using UCLUST, using a minimum pairwise identity of 97% (Edgar, 2010). Each cluster was represented by its most abundant sequence. Representative OTUs sequences were then aligned to the Greengenes database (DeSantis et al., 2006) using the PyNAST algorithm (minimum percent identity was set at 80%) (Caporaso et al., 2010a). The lanemask PH was used to screen out the hypervariable region and a phylogenetic tree was built using FastTree (Price et al., 2009). Taxonomy was subsequently assigned to each representative OTUs using the Greengenes database classifier with a minimum support threshold of 80% (DeSantis et al., 2006).

### *Constrained Ordination and Statistical Inference*

Following the taxonomic classification of 16S rRNA gene pyrosequence reads into their representative OTUs by UCLUST, calculations of the percentage of classifiable reads were performed using custom R scripts and executed using R version 2.13.2. The median proportion

of Greengenes classifiable reads obtained from each combination of primer sets, maize inbreds, and bulk soil from the pilot experiment to validate primer sets (Table. 2.1) was calculated from 100 bootstrap resamplings of the microbiome extractions stratified by primer sets, maize inbreds, and bulk soil to maintain balance among these factors. Given the lack of normality noted in the distributions of many populations tested, the R library “ImPerm” v1.1.2 (Wheeler) was employed to perform the non-parametric tests used in discerning variation in the proportion of classifiable reads between each primer set, sample type, and maize inbred. Reported variances explained by each factor reflect the proportion of variance explained by that factor after accounting for the remaining factors and are calculated from the marginal sums of squares. The 95% confidence interval for variance explained was derived from the resulting distribution of variance estimates after fitting multiple regression models to each of the 100 bootstrap resamplings of the data. A minimum of 5,000 permutations of the data were used to construct null distribution for each of the bootstrap resamplings of the raw data in calculating significance. Significances for all pairwise comparisons among the primer sets, soil, and maize inbreds in the pilot experiment were adjusted for multiple comparisons using Bonferroni correction.

Rarefaction was performed using QIIME to discern levels of OTU richness, Chao-1 diversity, and Whole Tree Phylogenetic diversity with respect to sequence depth (Caporaso et al., 2010b). Following rarefaction, median abundances for each microbiome extraction were calculated at a level of 2,080 pyrosequence reads. Given an inability to accurately extrapolate OTU abundances beyond a microbiome extraction’s maximum read depth, 2,080 reads was selected as a balance between removing microbiome extractions which did not possess this minimum and seeking to attain as many reads and thus sensitivity as possible in the included microbiome extractions. To address the unbalanced design resulting from removing extractions

not possessing this minimum read depth, the microbiome extractions were bootstrapped for 100 resamplings stratified by field environment, soil, and maize inbred. Permutation based regression analyses were performed in a similar manner to that implemented in discerning variation in classifiable reads for partitioning variation in alpha diversity among extractions.

To calculate Beta diversity, unweighted and weighted UniFrac distance metrics were calculated and used to construct distance matrices using QIIME (Caporaso et al., 2010b). Subsequently, the entries composing these matrices were bootstrapped for 100 resamplings stratified by field environment, bulk soil, and maize inbred. The R package “vegan” v2.0.2 (Oksanen) was used in calculation of partial constrained principal coordinate analyses. The proportion of the total inertia explained by each factor was calculated after conditioning on amplification batch, pyrosequencing run, and the remaining factors and constraining variation to the factor of interest. The 95% confidence intervals for this variation explained were derived from the bootstrap resamplings. Significances of factors within the model were calculated using vegan’s implementation of permutation testing “permutest” for constrained analysis of principal coordinates with 5,000 permutations. Comparisons of between field environment centroid distances were performed using lmPerm’s permutation testing of the between centroid distances across the 100 bootstrap resamples of the microbiome extractions. Comparisons of the levels of within factor multivariate dispersion were performed using vegan’s implementation of PERMDISP2 (Anderson et al., 2006).

Construction of frequency histograms and comparisons of model fits were performed using “prestonfit” and “radfit” routines in the R package vegan from OTU tables rarefied to a common depth of 2,080 reads. All comparisons of relative abundance of individual OTU as well as comparisons among soil characteristics were all performed by permutation testing using the

Imperm package. Reported significance values are all adjusted by Bonferroni correction. Normalization of the soil characteristics data and construction of the correlation matrix was performed using routines in the R base package. Estimations of the relatedness matrix among maize lines were performed using percent identity by state (Hardy and Vekemans, 2002) as well as genotype data from the first Generation Maize Hapmap (Gore et al., 2009). Soil characteristic and maize kinship matrices were bootstrapped for 100 resamples stratified by field environment and maize inbred and performed using vegan's Mantel test (Legendre, 1998).

## REFERENCES

- Agrios G.N. (2005) Plant Pathology, 5th Edition.
- Aira M., Gómez-Brandón M., Lazcano C., Bååth E., Domínguez J. (2010) Plant genotype strongly modifies the structure and growth of maize rhizosphere microbial communities. *Soil Biology and Biochemistry* 42:2276-2281.
- Anderson M.J., Willis T.J. (2003) Canonical Analysis of Principal Coordinates. *Ecology* 84:511-525.
- Anderson M.J., Ellingsen K.E., McArdle B.H. (2006) Multivariate dispersion as a measure of beta diversity. *Ecology Letters* 9:683-693.
- Bouffaud M.-L., Kyselková M., Gouesnard B., Grundmann G., Muller D., Moënne-Loccoz Y. (2012) Is diversification history of maize influencing selection of soil bacteria by roots? *Molecular Ecology* 21:195-206.
- Buckler E.S., Holland J.B., Bradbury P.J., Acharya C.B., Brown P.J., Browne C., Ersoz E., Flint-Garcia S., Garcia A., Glaubitz J.C., Goodman M.M., Harjes C., Guill K., Kroon D.E., Larsson S., Lepak N.K., Li H., Mitchell S.E., Pressoir G., Peiffer J.A., Rosas M.O., Rocheford T.R., Romay M.C., Romero S., Salvo S., Villeda H.S., Sofia da Silva H., Sun Q., Tian F., Upadyayula N., Ware D., Yates H., Yu J., Zhang Z., Kresovich S., McMullen M.D. (2009) The Genetic Architecture of Maize Flowering Time. *Science* 325:714-718.
- Caporaso J.G., Bittinger K., Bushman F.D., DeSantis T.Z., Andersen G.L., Knight R. (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26:266-7.
- Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E.K., Fierer N., Pena A.G., Goodrich J.K., Gordon J.I., Huttley G.A., Kelley S.T., Knights D., Koenig J.E., Ley R.E., Lozupone C.A., McDonald D., Muegge B.D., Pirrung M., Reeder J., Sevinsky J.R., Turnbaugh P.J., Walters W.A., Widmann J., Yatsunenko T., Zaneveld J., Knight R. (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nat Meth* 7:335-336.
- Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E.K., Fierer N., Pena A.G., Goodrich J.K., Gordon J.I., Huttley G.A., Kelley S.T., Knights D., Koenig J.E., Ley R.E., Lozupone C.A., McDonald D., Muegge B.D., Pirrung M., Reeder J., Sevinsky J.R., Turnbaugh P.J., Walters W.A., Widmann J., Yatsunenko T., Zaneveld J., Knight R. (2010c) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335-6.
- DeSantis T.Z., Hugenholtz P., Larsen N., Rojas M., Brodie E.L., Keller K., Huber T., Dalevi D., Hu P., Andersen G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72:5069-72.
- Edgar R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460-2461.
- Falkowski P.G., Fenchel T., Delong E.F. (2008) The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science* 320:1034-1039.
- Fox J.E., Gullledge J., Engelhaupt E., Burow M.E., McLachlan J.A. (2007) Pesticides reduce symbiotic efficiency of nitrogen-fixing rhizobia and host plants. *Proceedings of the National Academy of Sciences* 104:10282-10287.
- Gore M.A., Chia J.-M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S. (2009) A First-Generation Haplotype Map of Maize. *Science* 326:1115-1117.
- Hardy O.J., Vekemans X. (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes* 2:618-620.
- Houle D., Govindaraju D.R., Omholt S. (2010) Phenomics: the next challenge. *Nat Rev Genet* 11:855-866.
- John V S. (2000) Implementing Precision Agriculture in the 21st Century. *Journal of Agricultural*

- Engineering Research 76:267-275.
- Jurasinski G., Retzer V., Beierkuhnlein C. (2009) Inventory, differentiation, and proportional diversity: a consistent terminology for quantifying species diversity. *Oecologia* 159:15-26.
- Kim K.-H., Bae J.-W. (2011) Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Applied and Environmental Microbiology* 77:7663-7668.
- Legendre P.L.a.L. (1998) *Numerical Ecology*.
- Lozupone C.A., Hamady M., Kelley S.T., Knight R. (2007) Quantitative and Qualitative  $\beta$  Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Microbiology* 73:1576-1585.
- Marschner H. (1995) *Mineral Nutrition of Higher Plants* (Academic Press, London, ed. 2).
- Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. (2011) vegan: Community Ecology Package. R package version 2.0-2. <http://CRAN.R-project.org/package=vegan>.
- Price M.N., Dehal P.S., Arkin A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26:1641-50.
- Rodriguez R.J., Henson J., Van Volkenburgh E., Hoy M., Wright L., Beckwith F., Kim Y.-O., Redman R.S. (2008) Stress tolerance in plants via habitat-adapted symbiosis. *ISME J* 2:404-416.
- Smith S.E.R., D. J. . (2008) Mycorrhizal Symbiosis:145–187.
- Torsvik V., Øvreås L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology* 5:240-245.
- Unterseher M., Jumpponen A.R.I., Öpik M., Tedersoo L., Moora M., Dormann C.F., Schnittler M. (2011) Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. *Molecular Ecology* 20:275-285.
- Wheeler, R.E. (2010) lmPerm: Permutation tests for linear models. R package version 1.1-2. <http://CRAN.R-project.org/package=lmPerm>.
- Wooley J.C., Godzik A., Friedberg I. (2010) A Primer on Metagenomics. *PLoS Comput Biol* 6:10667.
- Yu J., Holland J.B., McMullen M.D., Buckler E.S. (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539-51.

**CHAPTER 3**  
**MAPPING AND PREDICTION OF MAIZE STALK STRENGTH IN A NESTED**  
**ASSOCIATION MAPPING PANEL**

**ABSTRACT**

Due to the influence of stalk strength on lodging and therefore yield, evaluation of strength is critical in any maize (*Zea mays* L.) breeding program. By measuring rind penetrometer resistance (RPR) or the force required to pierce a stalk rind with a spike, previous studies have effectively proxied this complex trait. Through partial mechanization of RPR, we examined variation in stalk strength between and within 25 families of recombinant inbred lines (RILs) constructed from crosses of diverse Non-Stiff-Stalk inbreds to the common Stiff-Stalk, B73. These families are known as the maize Nested Association Mapping (NAM) panel. A B73 x Mo17 intermated RIL family (IBM) and a diversity panel of nearly 300 inbreds were also evaluated for RPR. We measured inbreds of these maize families across three environments. Breeding values were estimated and QTL were identified by joint-linkage mapping. We also performed a joint-linkage-assisted GWAS and genomic prediction by ridge regression best linear unbiased prediction (RRBLUP). Despite nomenclature, alleles of Stiff-Stalk and Non-Stiff-Stalk heterotic groups segregating at near equal frequency in all RIL families did not stimulate high heritability for RPR. Only 8 of 26 families possessed  $H^2_{\text{line}} > 0.20$ . The diversity panel was more heritable at ~0.46; but, a portion of this variation was attributable to differential maturation at scoring. Over 12 joint-linkage QTL and ~141 GWAS associations were identified for RPR. No GWAS associations possessed tight linkage disequilibrium with known genes involved in phenylpropanoid and cellulose synthesis or vegetative phase change. Prediction by RRBLUP revealed ~84% of heritable variation in RPR was captured by a genomic relationship matrix

constructed from ~1.2 million polymorphisms in the NAM panel. This indicates utility in the application of genomic prediction methods to stalk strength improvement during maize breeding.

## INTRODUCTION

As a result of its influence on lodging and stover composition, maize stalk strength requires consideration when breeding to maximize grain yield or enhance silage quality. It is especially important in fields plagued by European corn borer, *Ostrinia nubilalis* H. (Papst et al., 2004), or Southwestern corn borer, *Diatraea grandiosella* D. (Gibson et al., 2010), and impacts colonization of stalk rotting fungal pathogens such as *Gibberella zeae* (Enrico Pè et al., 1993) and *Diplodia zeae* (Chambers, 1987). In addition to these stressors, high winds and soils with poorly managed nitrogen to phosphorous ratios (Arnold et al., 1974) also reveal the role of stalk strength in ensuring higher returns at harvest.

Dissection of stalk strength into its constituent traits suggests the structural composition of the rind, and not the pith or total stalk girth, is the main determinant of strength (Berzonsky et al., 1986; P. J. Loesch et al., 1962; Zuber et al., 1980). Previous study of maize rinds from populations divergently selected for stalk strength have revealed several potential means for enhancement (Berzonsky et al., 1986). From anatomical analyses, increases in vascular bundles, rind-parenchyma inter-lumen thickness, and percent hypodermal cell wall area are known to correlate with superior strength (Berzonsky et al., 1986). Vegetative phase change was also observed to occur earlier in varieties with stronger stalks (Abedon et al., 1999). In addition, transcriptional and compositional analyses have revealed the influence of cellulose and lignin on maize stalk strength (Bosch et al., 2011; Jung and Buxton, 1994).

Given the numerous mechanisms mediating stalk strength and the continuous variation observed for the trait in diverse maize populations, several studies have been performed to quantitatively dissect its genetic architecture (Flint-Garcia et al., 2003a; Flint-Garcia et al., 2003b; HerediaDiaz et al., 1996; Hu et al., 2012; Lee et al., 1996). In the most extensive previous quantitative study of stalk strength, composite interval mapping of quantitative trait loci (QTL) controlling stalk strength was performed in four bi-parental maize families (Flint-Garcia et al., 2003b). Construction of three of the families sought to maximize genetic variation for stalk strength by using parents divergently selected for high (MoSCSSS-High) and low (MoSCSSS-Low, MoSQB-Low) strength (Flint-Garcia et al., 2003b). Since many metrics for stalk strength and lodging such as stand counts are environmentally dependant and not easily reproducible, strength was scored by rind penetrometer resistance (RPR) (Flint-Garcia et al., 2003a; Flint-Garcia et al., 2003b; Zuber and Grogan, 1961; Zuber et al., 1980). This refers to the force required to pierce a stalk rind with a spike fixed to a digital force gauge (Sibale et al., 1992; Zuber and Grogan, 1961). Using RPR to pierce stalks mid-internode below the primary ear, QTL controlling stalk strength were identified in all four families (Flint-Garcia et al., 2003b). Primary ear height was genetically correlated with RPR; however, most QTL remained significant after accounting for its variation (Flint-Garcia et al., 2003a).

This previous RPR mapping analysis and earlier studies laid a foundation for evaluating stalk strength. However, whole genome sequencing of B73 (Schnable et al., 2009) and construction of a HapMap detailing segregation of ~1.2 million single nucleotide polymorphisms (SNPs) (Gore et al., 2009) now afford higher mapping resolution. For several quantitative traits, putatively causal alleles have been identified at the gene level in joint-linkage-assisted genome wide association studies (GWAS) (Brown et al., 2011; Kump et al., 2011; Poland et al., 2011;

Tian et al., 2011). Furthermore, genomic prediction methods such as ridge regression best linear unbiased prediction (RRBLUP) promise to increase efficiency in breeding complex traits. RRBLUP employs all genotyped polymorphisms in construction of a genomic relationship matrix and builds a model from past genotypic and phenotypic data. This model may subsequently be used to predict breeding values of genotyped seed prior to field testing, and thus enables selection of seed with more predicted promise (Jannink et al., 2010; Meuwissen et al., 2001).

The relevance of stalk strength in the harvestability of grain and digestibility of silage has not diminished since earlier studies. Furthermore, new applications in cellulosic ethanol production and biopolymer synthesis have caused a surge of interest in stalk strength related traits such as cell wall composition and biosynthesis (Bosch et al., 2011). While discoveries were made by molecular methods (Bosch et al., 2011), there remains interest in quantitatively resolving the genetic architecture of natural variation in stalk strength to further define these pathways. Leveraging our advanced knowledge of the maize genome as well as new mapping and prediction methods will enable us to understand the functional allelic diversity of stalk strength and breed maize varieties tailored to suit our enduring traditional needs as well as more contemporary applications.

In this study, we measured the stalk strength of 200 recombinant inbred lines (RILs) from each of 25 bi-parental families recombined and fixed for diverse Non-Stiff-Stalk alleles and alleles of the Stiff-Stalk, B73. This mapping resource is known as the maize Nested Association Mapping panel (NAM) (McMullen et al., 2009). A B73 x Mo17 intermated family (IBM) of 200 RILs (Lee et al., 2002) and a 282 inbred diversity panel (Flint-Garcia et al., 2005) were also scored. We took stalk strength measures within the NAM and IBM families in three

environments. In two of these environments, strength was also scored across the diversity panel. All strength measures were determined by RPR. A partially-mechanized RPR measurement method was developed to increase repeatability of the approach (Flint-Garcia et al., 2003b). We performed joint-linkage QTL mapping (Buckler et al., 2009) and joint-linkage-assisted GWAS (Tian et al., 2011) to resolve the genetic architecture of RPR. Genomic prediction by ridge regression best linear unbiased prediction (RRBLUP) was also performed to assess prediction accuracy of genomic estimated breeding values for RPR (Endelman, 2011).

## **MATERIALS AND METHODS**

### *Plant materials and environments*

The maize nested association mapping (NAM) panel developed by the Genetic Architecture of Maize and Teosinte Project consortium was constructed as previously detailed (McMullen et al., 2009). Briefly, the NAM panel was created by selection of 25 diverse Non-Stiff-Stalk maize inbreds crossed to a common reference Stiff-Stalk inbred, B73. Following generation of the F<sub>1</sub> hybrids, 200 progeny from each of the 25 bi-parental crosses were selfed for five generations. This produced a mapping panel of 5,000 recombinant inbred lines (RILs). In addition to the NAM panel, 200 RILs of the B73 x Mo17 intermated RIL family (IBM) (Lee et al., 2002) and a diverse inbred panel of 282 lines (Flint-Garcia et al., 2005) were also analyzed.

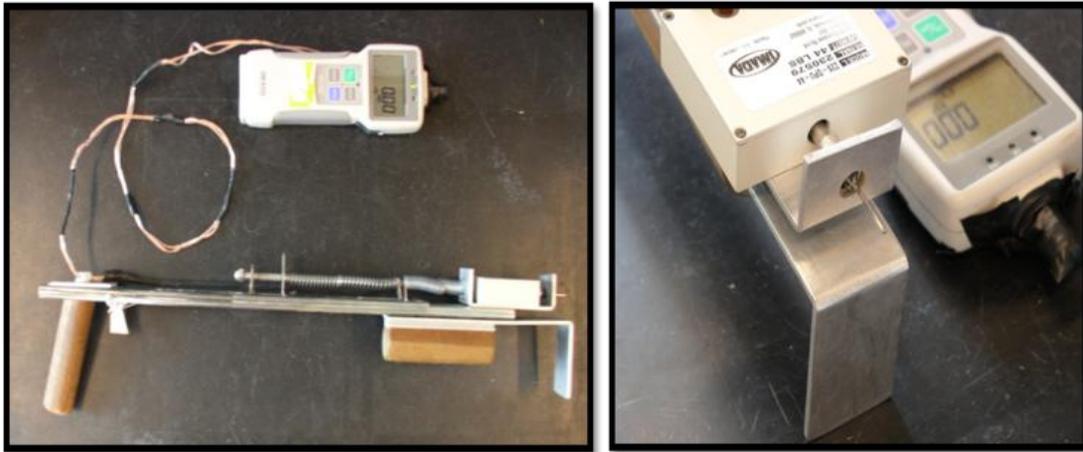
All inbreds were grown at Muskgrave Research Station in Aurora, NY (silt-loam soil) in the summer of 2008 and Rollins Bottoms Research Station in Columbia, MO (silt-loam soil) in the summer of 2009. The NAM and IBM families were also grown in Madison, WI at the Arlington Agricultural Research Station (silt-loam soil) in the summer of 2009. Within the NY and MO environments, one plot was grown for each inbred. However, B73 and other parental

founder of the NAM and IBM families were included in each sub-block of 22 plots to adjust for within environment block effects. Plots were composed of 12 plants in NY and 25 within MO. These plots were randomized within each family in both environments. The 282 inbreds of the diversity panel were also grown in both environments and B73 checks were planted to aid field correction. In WI fields, RILs of the NAM panel that matured early enough to permit flowering in the colder climate, were blocked in 10 maturity zone based on previous flowering data (Buckler et al., 2009). B73 checks were included in these blocks for field correction as well. All three environments were cultivated in a conventional manner with respect to fertilization, weed, and pest control.

#### *Phenotyping stalk strength and related traits*

To calculate stalk strength, a modified Accuforce Cadet digital force gauge (Ametek, Largo, FL) was assembled with a spike and used to manually pierce stalks as previously described (Flint-Garcia et al., 2003b) for all rind penetrometer resistance (RPR) measures collected in Aurora, NY 2008. Subsequent measures were collected using a mechanized Z2S-DPU digital force gauge reader (Imada, Northbrook, IL) to enhance the ease of data collection (Figure 3.1). In this mechanized apparatus, a spike was fixed to the digital force gauge. The gauge was then fastened to a track and driven by the release of a compressed spring. A trigger, cocking mechanism, and handle were fabricated to increase the ease and repeatability of phenotyping RPR by ensuring a more uniform acceleration of the spike when driven into a maize rind. A custom Java program was also developed to ensure all RPR measures from the gauge were stored to a text file for later analysis. Furthermore, audible commands were encoded into

the program to facilitate identification of which plots were phenotyped and which remained to be measured while collecting field data.



**Figure 3.1 Mechanization of stalk strength measure by rind penetrometer**

To assess stalk strength a rind penetrometer was fabricated. This device measured the kilograms of force required to pierce a stalk rind with a spike. To increase uniformity of measure the spike was driven into the rind by a compressed spring with a triggered release. Phenotypic values for each measure were automatically digitally recorded in a field book for later analysis.

All measures of RPR were collected near the middle of the stalk internode located immediately below the primary ear. Measures were taken from three randomly selected plants of each plot resulting in the collection of over 40,000 measurements across the NAM and IBM families as well as the maize diversity panel in three unique field environments. To avoid edge effects attributable to differential nutrient availability and light capture, the border plants of each plot were not measured. Phenotypic data for days to anthesis (DTA) and primary ear height (EHT) were all acquired from the same germplasm panels (Buckler et al., 2009; McMullen et al., 2009).

### *Genotyping families and diversity panels*

A total of 1,106 markers were scored on an Illumina Golden Gate Assay across the NAM RILs and IBM family to facilitate joint-linkage QTL mapping as previously described (McMullen et al., 2009). After QTL mapping, ~ 1.2 million polymorphisms reported in the maize HapMap (Gore et al., 2009) were projected onto the NAM and IBM RILs based on their respective parental lineage and the B73 genome as previously reported for a joint-linkage-assisted GWAS (Tian et al., 2011). These maize HapMap polymorphisms were further used in construction of genomic relatedness matrices between the RILs for genomic prediction (Endelman, 2011). Approximately 437,650 polymorphisms from the 282 inbred diversity panel were scored by means of Genotyping-By-Sequencing (GBS) directly on the RILs and inbred lines (Elshire et al., 2011). These GBS-characterized polymorphisms were imputed using TASSEL (Bradbury et al., 2007) and used for the mixed-model GWAS of the maize diversity panel as well as genomic prediction efforts (Endelman, 2011).

### *Statistical Analysis*

To partition phenotypic variation into genetic and environmental variance components, statistical analyses of RPR and correlated traits were performed using ASReml v3.00 (Gilmour et al., 1995) in coordination with custom Java code for backward selection of significant model terms. This code is available upon request. Best linear unbiased predictors (BLUPs) for RPR, height and flowering time of each of the inbred genotypes were calculated. Blocking effects were modeled as random independent effects when deemed significant by likelihood ratio testing with a critical value of  $\alpha = 0.05$ . A first order autoregressive by first order autoregressive (AR1 x AR1) error correlation structure was fitted for range and row within each of the fields as deemed

significant. Independence of residuals was assumed between fields and no heterogeneous measurement error variance or nugget variance for measures was fitted. Following the partitioning of phenotypic variation and calculation of BLUPs, measures of RPR, height, and flowering time were used to calculate phenotypic and genetic trait covariance and correlation matrices in R v2.12.0 (R, 2011).

To further partition genetic variance beyond the genotypic level, the SAS v9.2 statistics package (SAS, 2002-2004) was implemented. SAS PROC GLMSelect was executed to regress BLUPs for each of the traits against the 1,106 markers nested within the 19 most heritable NAM and IBM families in a joint-linkage QTL model. A model term was fitted for each family and family nested marker selection was performed by stepwise regression as previously described (Buckler et al., 2009). Covariates of DTA and EHT were also fitted as described. Significance of model entry and exit were set to  $p < 5e-4$  based on 1,000 null permutations of RPR. This stepwise model building routine was bootstrapped, re-sampling 20% of the RILs within each NAM family for 100 sampling iterations to calculate a resample model inclusion probability (RMIP) (Valdar et al., 2009) and improve identification of robust QTL. A RMIP greater than 10 out of 100 sampling iterations was attained from less than 5% of the selected markers at the given model entry and exit criteria in null permutation testing of RPR. Code for NAM and IBM family-stratified bootstrapping of joint-linkage QTL mapping is available on request.

After constructing a joint-linkage QTL model from the full data set, additional models were built for each chromosome. The model terms fitted for each family and the QTL identified as residing on a given chromosome were dropped and residual variance from the model was attributed to the missing genetic variance of the dropped chromosome in later GWAS analyses as previously described (Tian et al., 2011). This multi-stage analysis was performed with the

assumption of linkage equilibrium of QTL effects residing on independent chromosomes in the NAM and IBM families.

To perform a joint-linkage-assisted GWAS, residuals calculated from each of the 10 joint-linkage QTL models constructed from the full data set of RILs were regressed against polymorphisms of their respective chromosome in a re-sampled forward regression framework (Kump et al., 2011; Poland et al., 2011; Tian et al., 2011) using polymorphisms projected from the maize HapMap (Gore et al., 2009; Tian et al., 2011) and custom Java code (Tian et al., 2011). The threshold for model entry was set to  $p < 5e-8$  based on the results of null permutation testing. Re-sampling with replacement was performed for 100 sampling iterations to attain an estimate of its RMIP (Valdar et al., 2009) and assess the robustness of the observed associations.

Using Genome Association and Prediction Integrated Tool (GAPIT) (Lipka et al., in review) in R v2.12.0 (R, 2011), sequential single polymorphism mixed-model GWAS for RPR was also performed across the 282 inbred diversity panel. This approach accounted for the natural population structure and the false associates it creates (Yu et al., 2006) and allowed for potential identification of significant associations across the GBS characterized polymorphisms. Given, the covariation of RPR and flowering time measures across the diversity panel, these regressions were performed both before and after accounting for RPR covariation with DTA and EHT. The significance of associations was determined by measures of false discovery rate (Benjamini, 1995).

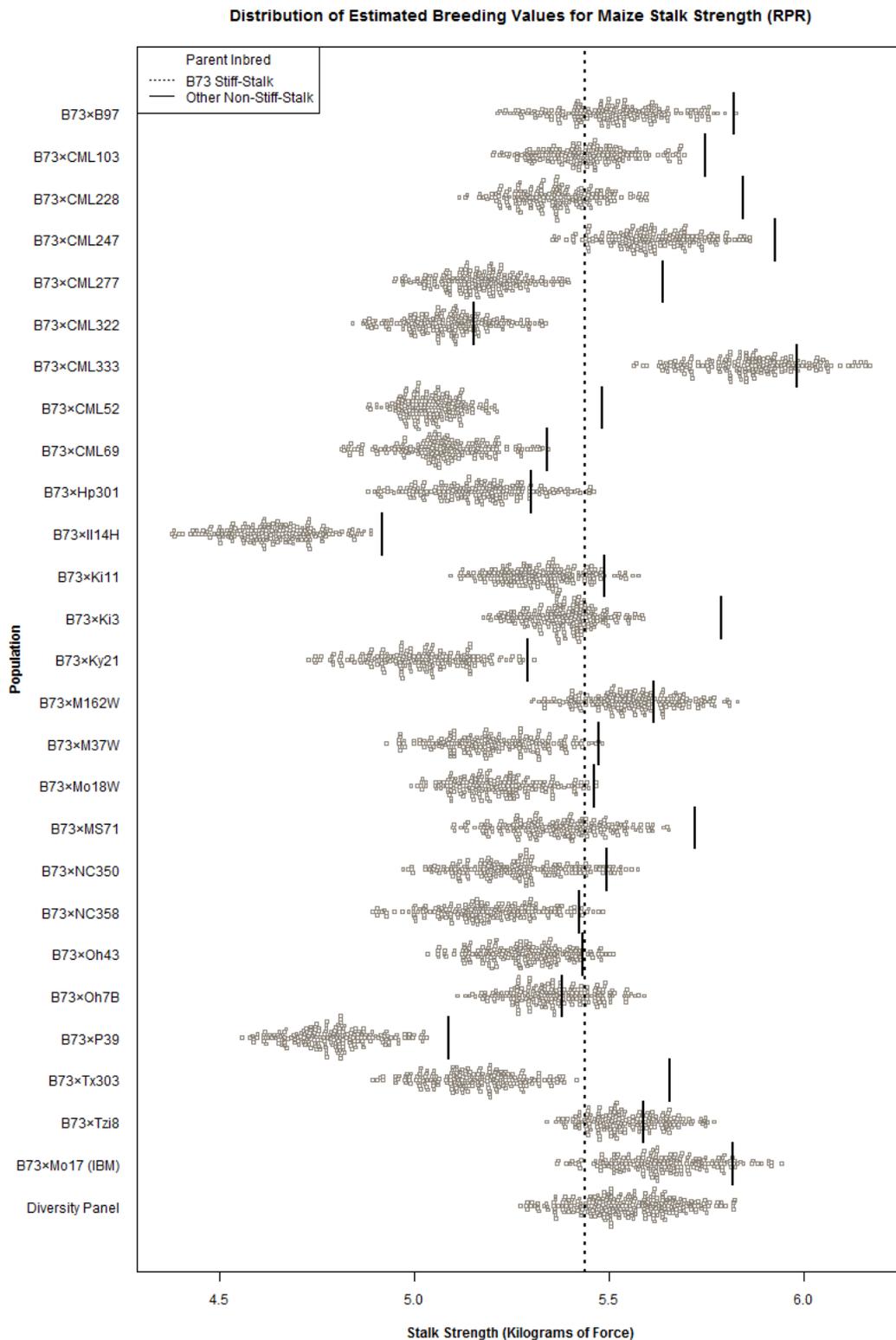
As an additional modeling method to partition genetic variation across the entire maize genome simultaneously, we implemented genomic prediction using the package rrBLUP (Endelman, 2011) in R v2.12.0 (R, 2011). Genomic relatedness matrices were constructed for the NAM and IBM families using the ~1.2 million polymorphisms of the maize HapMap (Gore et

al., 2009). Estimates of genomic relatedness among the 282 inbred diversity panel were calculated using the ~437,650 GBS polymorphisms. BLUPs for each panel were regressed against its genomic relatedness matrix to assess the ability of all polymorphisms to simultaneously predict the genomic estimated breeding value (GEBV) of each genotype. The prediction accuracy was determined by regressing actual breeding values against their respective GEBVs. The robustness of this relationship was tested to determine degree of shrinkage in the coefficient of determination upon fivefold family stratified cross-validation of the NAM families and random fivefold cross-validation of the NAM families, IBM family, and the diversity panel.

## **RESULTS**

### *Variation in stalk strength and related traits*

Estimated breeding values of the NAM and IBM families as well as the inbred diversity panel calculated for stalk strength, ear height, and flowering time possessed significant variation. Most NAM and IBM families exposed substantial transgressive segregation in all the surveyed traits (Figure 3.2). Estimated breeding values averaged across the 3 field environments revealed 95% of RPR stalk strength measures in the NAM and IBM families fell between 4.65 kilograms of force (KgF) and 5.87 KgF. In the diversity panel, 95% of RPR measures ranged from 5.30 KgF to 5.84 KgF. The weakest stalks among the NAM and IBM family parents were the Non-Stiff-Stalk sweet corn inbreds, Ill14h and P39. The sole Stiff-Stalk inbred of the NAM panel, B73, was ~5.44 KgF and stronger than less than one third of the inbreds in the diversity panel. Furthermore, B73 was only stronger than 9 of the 26 Non-Stiff-Stalk inbreds of the NAM families, and weaker than Mo17 of the IBM family.



**Figure 3.2 Distribution of estimated breeding values for maize stalk strength**  
 Substantial transgressive segregation of maize stalk strength in NAM and IBM families. Despite inclusion in the Non-Stiff-Stalk heterotic group, many Non-Stiff-Stalk inbreds (solid line segments) were stronger than the Stiff-Stalk B73 (dashed line). Rank was largely maintained after accounting for DTA covariation.

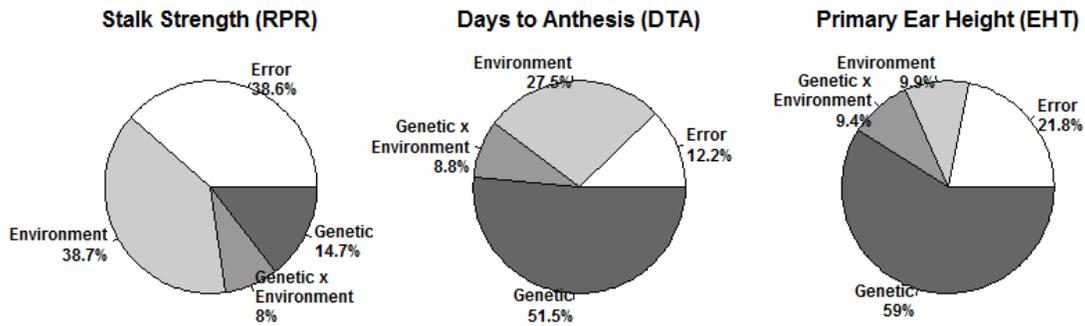
Low estimates of heritability were calculated for RPR measures of stalk strength in most of the surveyed germplasm. This occurred both across the heritable families of the NAM panel which possessed a broad sense line mean heritability of 0.21 and within the NAM families where the mean broad sense heritability was only 0.17. Of the 25 NAM families, 19 possessed heritable phenotypic variation in RPR above 0.05. At a broad sense line mean heritability of 0.34, a higher proportion of RPR variation was heritable in the IBM family than any of the NAM families studied. The 282 inbred diversity panel was even more heritable at a broad sense line mean heritability of 0.46. However, after accounting for covariation of maturational differences at scoring, as measured by DTA, this heritability estimate dropped to 0.38. A similar, less severe drop was noted after accounting for covariation of primary ear height 0.41 in the diversity panel. In contrast, accounting for flowering time or ear height in the NAM and IBM families did not reveal as sizeable a drop in broad sense line mean heritability across or within families (Table 3.1).

In addition to the heritable proportion of RPR variation, ~39% of stalk strength variation was captured between the 3 field environments when considering all families and panels of this study (Figure 3.3). This was much greater than that observed for both DTA and EHT. Nonetheless, measures of environmental variation were confounded with manual and mechanized RPR phenotyping methods. Measures taken in Aurora, NY in 2008 were performed with a manual RPR apparatus; while those stalk strength measures collected in Columbia, MO and Madison, WI were taken the following year with the mechanized RPR apparatus. Estimates of the proportion of environmentally conditional genetic variation in RPR while substantial, ~8%, and capturing approximately half as much stalk strength variation as that captured by genetic variation, ~15%, were also confounded by the phenotyping method employed.

Correlations among estimated breeding values for all inbreds under study calculated for the manual RPR method and mechanized approach were  $r = 0.33$ . The correlations observed among estimated breeding values between both field environments in which the mechanized RPR measures were taken were higher, at  $r = 0.41$ .

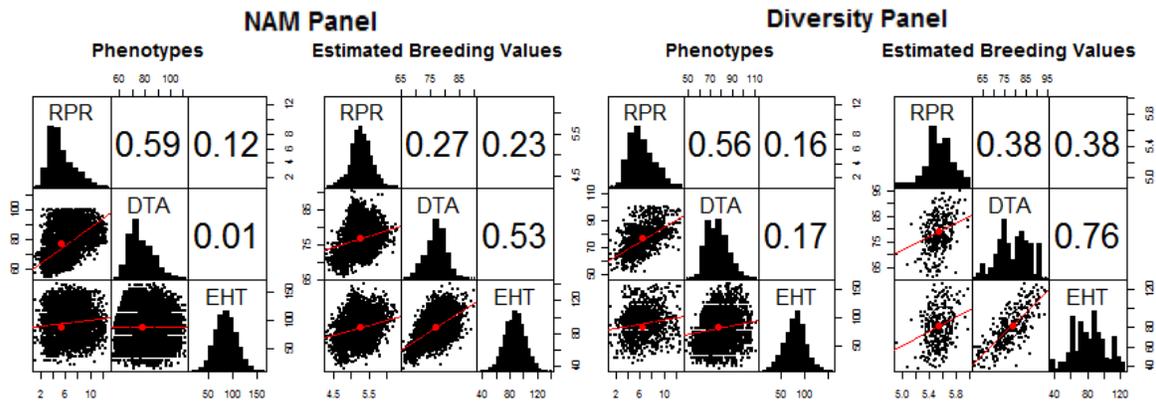
**Table 3.1 Heritability estimates of NAM and IBM families as well as Diversity Panel**

Trait $H^2_{line}$	n	RPR	RPR (DTA cov)	RPR (EHT cov)	DTA	EHT
NAM panel	37,548	0.21±0.02	0.20±0.02	0.21±0.02	0.94±0.01	0.93±0.01
B73 x B97	1,763	0.20±0.01	0.20±0.01	0.20±0.01	0.85±0.01	0.94±0.01
B73 x CML103	1,804	0.07±0.01	0.06±0.01	0.06±0.02	0.85±0.01	0.95±0.01
B73 x CML228	1,211	0.07±0.02	0.08±0.02	0.07±0.01	0.94±0.01	0.93±0.01
B73 x CML247	1,313	0.28±0.01	0.29±0.02	0.26±0.01	0.93±0.01	0.93±0.01
B73 x CML277	1,311	0.16±0.02	0.16±0.02	0.15±0.01	0.94±0.01	0.93±0.01
B73 x CML322	1,530	0.03±0.02	0.04±0.02	0.02±0.01	0.92±0.01	0.92±0.01
B73 x CML333	1,554	0.33±0.01	0.31±0.01	0.30±0.02	0.94±0.01	0.93±0.01
B73 x CML52	1,105	0.03±0.02	0.04±0.02	0.03±0.02	0.95±0.01	0.92±0.01
B73 x CML69	1,378	0.19±0.02	0.20±0.01	0.18±0.02	0.89±0.01	0.93±0.01
B73 x Hp301	1,815	0.11±0.02	0.12±0.02	0.10±0.01	0.90±0.01	0.95±0.01
B73 x Il14H	1,587	0.05±0.02	0.05±0.01	0.04±0.02	0.91±0.01	0.93±0.01
B73 x Ki11	1,352	0.12±0.03	0.11±0.02	0.12±0.03	0.94±0.01	0.94±0.01
B73 x Ki3	997	0.04±0.02	0.03±0.03	0.03±0.02	0.93±0.01	0.92±0.01
B73 x Ky21	1,611	0.17±0.01	0.18±0.01	0.17±0.02	0.84±0.01	0.93±0.01
B73 x M162W	1,556	0.15±0.02	0.14±0.01	0.14±0.02	0.91±0.01	0.92±0.01
B73 x M37W	1,641	0.15±0.03	0.14±0.03	0.13±0.01	0.90±0.01	0.91±0.01
B73 x Mo18W	1,365	0.08±0.02	0.08±0.02	0.07±0.02	0.93±0.01	0.93±0.01
B73 x MS71	1,684	0.08±0.02	0.09±0.01	0.08±0.02	0.78±0.01	0.90±0.01
B73 x NC350	1,517	0.28±0.02	0.24±0.02	0.26±0.02	0.92±0.01	0.93±0.01
B73 x NC358	1,685	0.24±0.01	0.25±0.01	0.23±0.01	0.84±0.01	0.92±0.01
B73 x Oh43	1,715	0.24±0.01	0.21±0.01	0.21±0.02	0.81±0.01	0.93±0.01
B73 x Oh7B	1,548	0.03±0.02	0.03±0.02	0.03±0.02	0.90±0.01	0.95±0.01
B73 x P39	1,473	0.02±0.02	0.04±0.02	0.01±0.01	0.95±0.01	0.93±0.01
B73 x Tx303	1,505	0.03±0.02	0.02±0.03	0.02±0.02	0.92±0.01	0.95±0.01
B73 x Tzi8	1,526	0.28±0.02	0.24±0.02	0.25±0.02	0.93±0.01	0.95±0.01
B73 x Mo17 (IBM)	1,735	0.34±0.01	0.30±0.01	0.31±0.01	0.92±0.01	0.96±0.01
Diversity panel	1,401	0.46±0.02	0.38±0.02	0.41±0.01	0.96±0.01	0.97±0.01



**Figure 3.3 Phenotypic variation of stalk strength, flowering time, and ear height across NAM panel, IBM family, and Diversity panel**

Nearly 15% of the total variation in stalk strength was captured between maize inbreds and families. This proportion remained smaller than that of flowering time and primary ear height. Despite this relative reduction in genetic variation of stalk strength, the proportion of environmentally conditional genetic variation was similar across all traits.



**Figure 3.4 Trait correlations between stalk strength, flowering time, and ear height**

Positive correlations between stalk strength and flowering time were reduced between estimated breeding values when compared to their phenotypic values across NAM families as well as within the diversity panel. In contrast, positive correlations between stalk strength and ear height were increased among estimated breeding values relative to their phenotypic value.

Correlations between stalk strength as measured by RPR, and flowering time (DTA), were positive,  $r = 0.59$ , at the phenotypic level across all NAM families (Figure 3.4) and within many of the families. However, these correlations were diminished among estimated breeding values,  $r = 0.27$  (Figure 3.4). This reduction among RPR and flowering time correlation between estimated breeding values was less severe across the inbred diversity panel where correlations among both phenotypes,  $r = 0.56$ , and estimated breeding value,  $r = 0.38$ , remained substantial.

Reduction in the RPR correlations with flowering time among breeding values relative to their phenotypic values appeared attributable to the positive correlations of the traits among environments. However, testing of additional environments is needed.

In contrast to the reduction in correlation between RPR and flowering time measures in estimated breeding values relative to phenotypic values, correlation between RPR and primary ear height increased among estimated breeding values,  $r = 0.23$ , relative to their respective phenotypic values,  $r = 0.12$  across all of the NAM families (Figure 3.4) as well as within most of them. This increase was even more pronounced in the inbred diversity panel, escalating from  $r = 0.16$  across phenotypic values to  $r = 0.38$  across estimated breeding values. The increased genetic correlation for these traits among estimated breeding values could not be attributed to negative correlations among surveyed environmental or environmentally conditional genetic factors explaining the covariation in stalk strength and primary ear height.

#### *Joint-linkage mapping of stalk strength QTL*

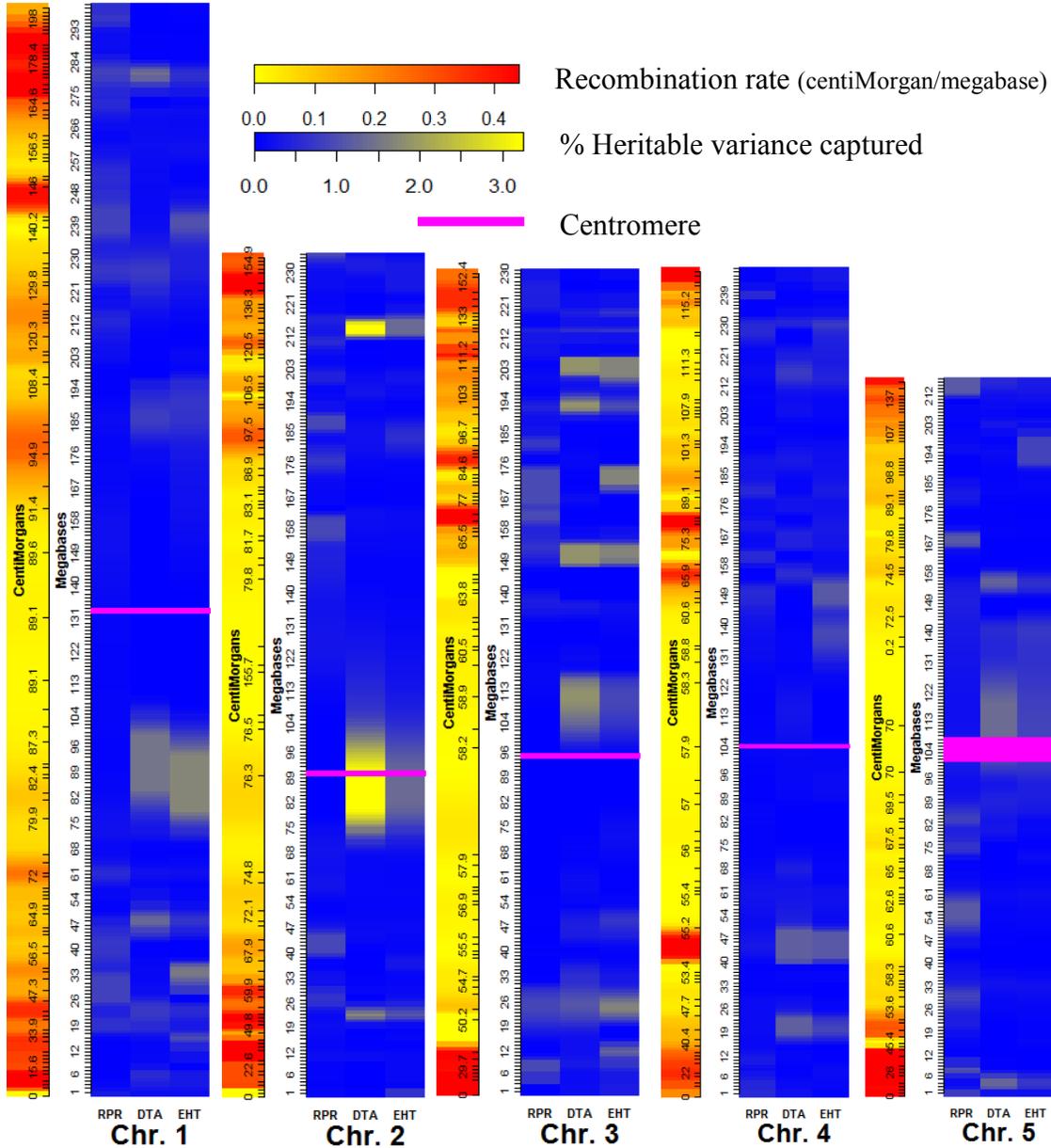
Using bootstrapped joint-linkage mapping, QTL capturing variation in RPR were detected across the heritable NAM and IBM families on all maize chromosomes (Figures 3.5 - 3.6). A 39 nested QTL model captured  $89 \pm 1\%$  of the heritable variation in stalk strength and dropped to  $81 \pm 2\%$  upon fivefold random cross-validation of the NAM and IBM families. Approximately 70 clusters of joint-linkage markers possessing a RMIP greater than or equal to 10 of 100 model builds were identified. The 12 most robust QTL all possessed RMIP greater than or equal to 20 and remained robust ( $\text{RMIP} > 15$  within 2.1 cM of marker association) after accounting for the covariance of flowering time (DTA) and primary ear height (EHT). At a RMIP of 61, the most robust QTL marker association was located on chromosome seven at

~105.2 cM as determined by the composite NAM linkage map (McMullen et al., 2009). Separated by ~1.7 cM, two neighboring markers on chromosome eight at ~97.4 cM possessed a combined RMIP of 62 and were only jointly selected in 1 of 100 model builds (RMIP= 35+28-1). Expanding this interval to ~6.2 cM achieved a combined RMIP of 94 of the 100 model builds. A strong positive correlation of median allele effect estimates across NAM and IBM families for all 15 pair wise comparisons of the six markers contained within this interval was also observed,  $r > 0.78$ . Similar significant clusters of joint-linkage markers were observed across the maize genome.

Partitioning genetic variation in RPR revealed no family nested QTL captured over 2.7% of stalk strength variation. Allele effect sizes of the joint-linkage mapped QTL were uniformly small across families. While 95% of estimated breeding values in the NAM and IBM families spanned a range of 1.22 KgF, the median significant (T-test allele effect within family,  $p < 1e-4$ ) negative and positive effects across 100 model builds were comparable and ~0.07 KgF from the family mean. The distributions of these positive and negative effect sizes possessed a median absolute deviation of ~0.01 KgF. No significant correlation was observed between median effect size of the significant alleles within a family and the broad sense line mean heritability of the family. The median number of significantly classified alleles within a family across the 100 model builds ranged from 1 to 21 QTL with an RMIP over 10.

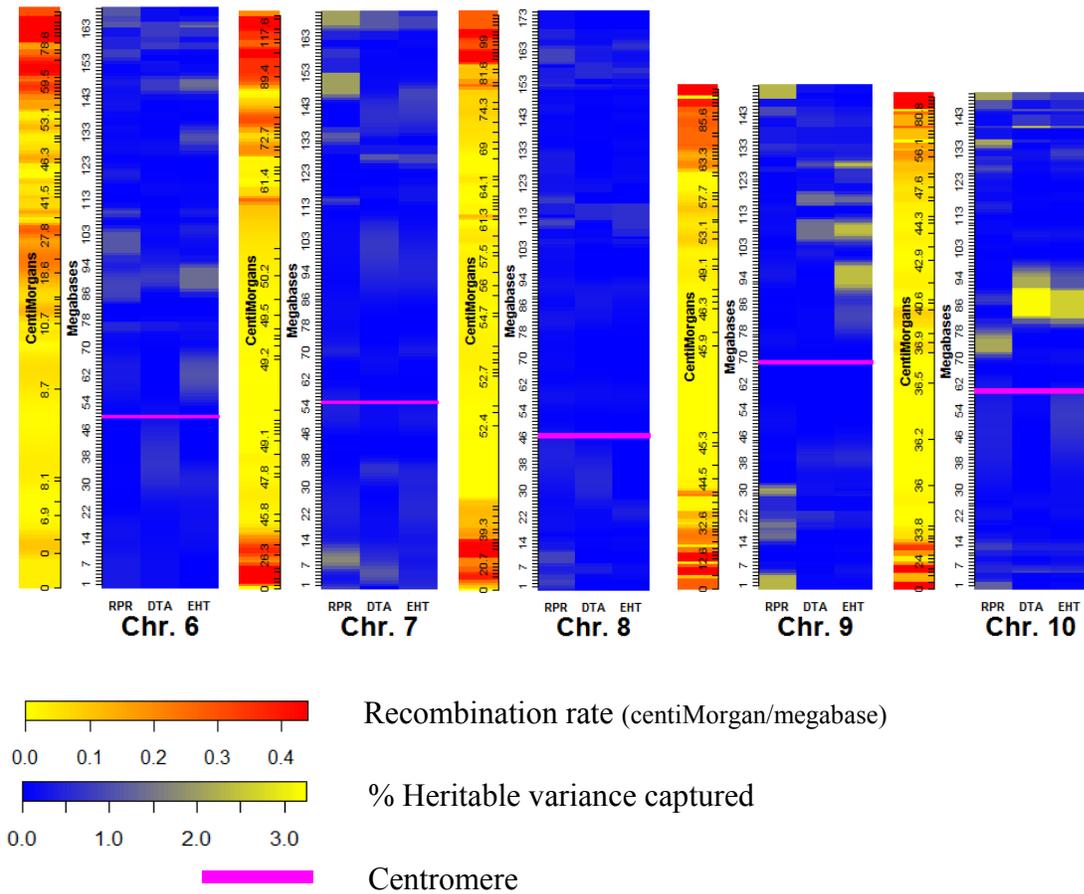
A median of  $7 \pm 2$  significant QTL were mapped across the NAM and IBM families. The number of significant QTL segregating in a family was highly correlated with that family's heritability estimate,  $r = 0.89$ . Similarly, the median families in which a QTL possessed a significant effect was  $9 \pm 3$  and was highly correlated with its RMIP,  $r = 0.81$ .

## Distribution of genetic variation in NAM and IBM families



**Figure 3.5. Partitioning genetic variation in stalk strength, flowering time, and ear height across the maize genome by joint-linkage mapping (Chromosomes 1-5)**  
 Using bootstrapped joint-linkage mapping of 1,106 markers nested within the 25 NAM families and the IBM family (ticks on centimorgan axis indicate all markers), QTL were mapped on every chromosome for all traits in both high and low recombinagenic regions of the genome. No obvious correlations were apparent among allele series of these QTL.

## Distribution of genetic variation in NAM and IBM families



**Figure 3.6. Partitioning genetic variation in stalk strength, flowering time, and ear height across the maize genome by joint-linkage mapping (Chromosomes 6-10)**

The common Stiff-Stalk, B73, and all 26 Non-Stiff-Stalk parental inbreds possessed both positive and negative allele effects for stalk strength across the 12 most robust QTL mapped by joint-linkage. Despite significant correlations among estimated breeding values for RPR and both DTA and EHT, no significant correlation of allele effect estimates across NAM and IBM families were observed at the resolution of these 12 independent stalk strength loci. Similarly, no significant correlation of RPR allele effects in any joint-linkage mapped QTL for DTA or EHT were characterized. In contrast, weak correlations between RPR and both DTA,  $r = 0.21$ , and EHT,  $r = 0.19$ , were observed across the NAM and IBM family means.

The resolution afforded by joint-linkage QTL mapping does not permit gene level characterization of stalk strength associations. In most cases, the significance of robust QTL instead persisted for an interval of approximately one to five cM before dropping below a RMIP of less than 10 of 100 sample iterations. Within these intervals, few known genes involved in phenylpropanoid or cellulose synthesis pathways were identified. No robust QTL were identified near the *brown midrib* mutants involved in lignin biosynthesis of the phenylpropanoid pathway. However, a putative 4-coumarate-CoA ligase-like gene (AF466202.2\_FG012) potentially involved in the phenylpropanoid pathway was located near a linkage marker on chromosome ten at ~69.2 cM with a RMIP of 18. Furthermore, a Caffeoyl-CoA O-methyltransferase (GRMZM2G077486) of the phenylpropanoid pathway was flanked by two linkage markers on chromosome ten at ~38.6 and 40.1 cM possessing RMIPs for stalk strength of 20 and 11 respectively.

Of the 12 characterized cellulose synthases in the maize genome, the only synthase whose nearest linkage marker possessed a RMIP over 10 was Cellulose Synthase-9 (GRMZM2G018241). This gene and marker are located on chromosome two at ~82.5 cM and possessed a RMIP of 24. Uncharacterized annotations (GRMZM2G157729, GRMZM2G110145) with predicted cellulose synthase activity by homology and known transcriptional evidence were also identified on chromosome nine near a linkage marker at ~42.8 cM and chromosome ten next to a linkage marker at ~38.6 cM. Both these linkage markers possessed a RMIP of 20, with the latter also flanked by a marker possessing a RMIP of 10.

Cloned loci previously identified for vegetative phase transition and other stalk strength related traits, such as the mutants of *brittle stalk2*, *glossy1-15*, and *teopod1*, 2 were not identified near QTL possessing RMIPs for stalk strength over 10 of 100 sampling iterations. Similarly, co-

localization of QTL mapped in the prior multi-family RPR study was not substantial. While a few overlapping QTL were identifiable, the 12 most robust QTL of this analysis did not significantly match those characterized in previous studies.

#### *GWAS of stalk strength associations*

In order to further resolve the joint-linkage mapped stalk strength QTL, 141 significant associations were identified by joint-linkage-assisted GWAS across the NAM and IBM families. QTL identified during joint-linkage mapping were used to account for background genetic variation in GWAS. The most robust of the GWAS associations co-localized with estimated joint-linkage QTL effects. However, many significant effects were found dispersed across the maize genome. No stalk strength associated polymorphisms were shared with the joint-linkage assisted GWAS of DTA (277) or EHT (304) in the NAM and IBM families. Approximately 5% (15) of the DTA associations were located within 1cM or 1Mb of a RPR association; whereas, ~10% (29) of the EHT associations were located within 1cM or 1Mb of an RPR association. Nearly one third (43) of the associated polymorphisms were identified within a known or hypothesized gene. However, no significant associations were located in genes known to be involved in the phenylpropanoid or cellulose synthesis pathways. Furthermore, no significant associations possessing a RMIP greater than 3 were identified within 100 kilobases of genes known to be involved in these pathways. The same was true of previously cloned loci and known genes implicated in vegetative phase transition.

The effect sizes of GWAS stalk strength associations across the NAM and IBM families were uniformly small and similar in size and distribution to the significant alleles nested within each family during joint-linkage QTL mapping. No RPR effects were greater than 0.05 KgF from the population mean. A median of ~0.02 KgF and median absolute deviation of ~0.006 KgF was observed for both the positive and negative GWAS effect estimates.

The most robust association identified across the NAM and IBM families was observed in every sampling of joint-linkage-assisted GWAS and therefore possessed a RMIP of 100. This polymorphism is located on chromosome three at 176,660,475 bp and was flanked by linkage markers that possessed RMIPs of 10 and 13 during joint linkage mapping. The nearest annotation is 5,139 bp downstream and encodes a transferase (GRMZM2G165192) responsible for transferring acyl groups other than amino-acyl. No annotations within a one cM interval surrounding the association were obvious candidates for stalk strength. The second and third most robust associations were both identified on chromosome eight at 163,943,201 bp and 8,415,595 bp. These possessed RMIP of 94 and 73, respectively. Both were also located near regions of the genome neighboring significant markers identified in joint-linkage analysis. The former is located in an interval wherein two linkage markers spaced ~1.7 cM possessed a combined RMIP of 62. The latter neighbors a linkage marker with a RMIP of 14 in joint-linkage mapping. In both instances no obvious candidates for stalk strength related pathways or developmental processes were apparent. The nearest respective annotations were a glycosyl-transferase (GRMZM2G002023) 4,605 bp downstream and a known protein with O-glycosyl hydrolyzing activity (AC234160.1\_FG003) 3,349 bp downstream.

In addition to joint-linkage-assisted GWAS across the NAM and IBM families, sequential single marker GWAS was also performed across the panel of 282 diverse inbreds. This enabled assessment of more diversity and examination if alleles common in natural maize diversity can be associated with stalk strength. To account for the inherent relatedness that was not reduced by the recent recombination of genetic diversity as it was within the bi-parental families of NAM and IBM, a mixed model framework was implemented (Yu et al., 2008). Using this method to query ~437,650 polymorphisms genotype by sequencing, no significant GWAS

associations were identified for RPR within the diversity panel. This lack of significant associations persisted after accounting for covariation of DTA and EHT. Despite a lack of significance as determined by false discovery rates, the strongest associations uncovered within the panel were reviewed. However, none were identified in tight linkage disequilibrium with obvious candidate genes for stalk strength.

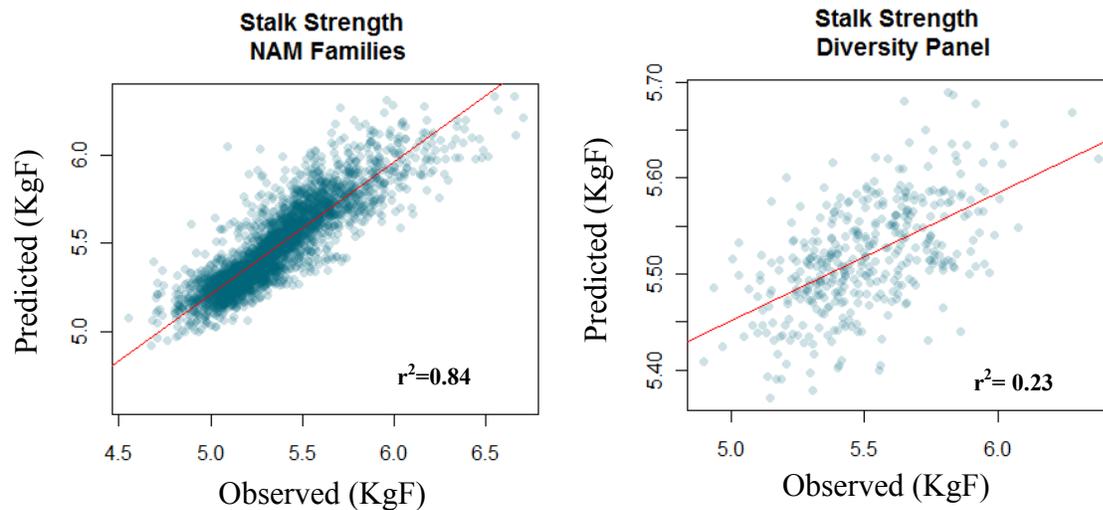
### *Genomic prediction of stalk strength*

Given the apparent polygenicity of RPR within the germplasm under study, genomic prediction was performed to determine the ability of all genotyped diversity to simultaneously capture the heritable variation in maize stalk strength as measured by RPR. A genomic relatedness matrix was constructed from ~1.2 million maize HapMap polymorphisms and was fitted in a RRBLUP framework (Endelman, 2011). This relatedness matrix was found to capture ~86% of the heritable variation across the 5,000 RILs of the NAM panel (Figure 3.7). A similar matrix constructed for the 200 RILs of the IBM family was found to capture ~55% of the heritable variation in RPR. Comparable levels of heritable variation to that captured within IBM were captured in each of the bi-parental NAM families possessing approximately the same number of individuals. The ability of total genotyped diversity to capture heritable differences among the 282 inbred diversity panel was assessed in addition to these bi-parental families. However, this was accomplished with a more limited set of 437,650 polymorphisms genotyped using GBS (Figure 3.7). In this panel the genomic relatedness matrix captured ~62% of heritable variation across the inbreds.

Fivefold cross-validation was performed to determine the stalk strength prediction accuracy of all genotyped diversity within and between families and within the diversity panel.

The shrinkage of heritable variation captured during prediction as compared to directly fitting all individuals was minimal in NAM. Estimates of variation captured after fivefold cross validation remained at  $85\pm 1\%$ . This reduction was only slightly less than that of random cross-validation without stratification by family, wherein  $84\pm 1\%$  of the heritable variation in stalk strength was captured in prediction. All prediction accuracies across the NAM panel greatly exceeded those observed within its families or within the IBM family,  $11\pm 1\%$ . This was also true in the 282 inbred diversity panel after fivefold cross-validation,  $23\pm 1\%$  (Figure 3.7).

The RPR effect estimates calculated by RRBLUP for alleles segregating in one family to predict heritable RPR variation in another family was also calculated for all pair wise comparisons among families. This revealed limited predictive accuracy between families. The variation in predictive accuracy existing among these comparisons was insufficient to assess if relatedness between Non-Stiff-Stalk inbreds of each family was correlated with the prediction accuracy of RPR between families.



**Figure 3.7 Genomic prediction of stalk strength in NAM and Diversity panel**  
 After fivefold cross-validation, approximately 84% of the variation in stalk strength was captured using ~1.6 million SNPs from the maize HapMap.

## DISCUSSION

Maize stalk strength is a highly complex polygenic trait that may be strongly influenced by numerous elements of a plant's phenotype as well as its interaction with the environment. Given the importance of stalk strength in determining harvestability and its interplay with compositional elements that influence silage digestibility, further characterization of the genetic architecture underlying heritable variation in strength are needed. The development of a means to accurately predict stalk strength from genotypic data is critical to further advancing maize breeding and crop improvement efforts. New applications of biomass and stover to cellulosic ethanol production and the synthesis of biopolymers have further augmented the importance of understanding this traits and how it influences related biochemical pathways that are critical to these products (Bosch et al., 2011).

In this study, we characterized the genetic architecture and genomic prediction accuracy of RPR as a proxy for maize stalk strength. RPR measurements were taken across the NAM and IBM families as well as a 282 inbred diversity panel. Measurements were collected in three field environments for NAM and IBM families. Two of these environments were also scored for the inbred diversity panel. All surveyed bi-parental families were composed of alleles of a Non-Stiff-Stalk inbred and the Stiff-Stalk B73 segregating at approximately equal frequency. Despite the use of inbred parents from these heterotic groups, substantial transgressive segregation for RPR was observed in most families scored. This result was not comparable to previous QTL mapping analyses performed in families constructed from parents divergently selected for stalk strength (Flint-Garcia et al., 2003b). Instead, it suggests repulsion phase QTL, non-additive gene actions, or numerous small effect polymorphisms that were not pyramided and fixed by directional phenotypic selection existed within most inbred parents used to construct the NAM and IBM families.

While B73 is one of the best known inbreds sourced from G.F. Sprague's Iowa Stiff-Stalk Synthetic (Hallauer, 2000), its stalk was weaker than two thirds of the Non-Stiff-Stalk inbreds used in construction of the NAM panel and remained so after accounting for covariation of flowering time and ear height. This suggests the nomenclature used in the initial designation of these heterotic groups no longer relates to current stalk strength as measured by RPR in the environments of this study and is further validated by weak relative strength rankings of several other inbreds classified in the Stiff-Stalk subpopulation of the maize diversity panel (Flint-Garcia et al., 2005; Pritchard et al., 2000).

Partitioning variation in RPR revealed the proportion attributable to genetic diversity within all NAM families and the 282 inbred diversity panels was diminished compared to

previous maize studies (Flint-Garcia et al., 2003b). Given the similarity in phenotyping method, this reduced heritability may be attributable to the reduced number or size of segregating functional loci within the families under study, an increase in repulsion phase linkages of these loci, or environmental differences. Of the genetic variation for stalk strength that was observed, a substantial proportion was captured by variation between NAM families and was not attributed to variation within them. With a shared B73 founder, all polymorphisms differing between families also segregated within them. Levels of heritable variation in a family may therefore be attributed to linkage patterns or interactions among extant polymorphisms within the family.

From the heritable variation that existed within NAM and IBM families, joint-linkage QTL mapping efforts revealed twelve highly robust QTL possessing a RMIP of over 20 across the NAM and IBM families. However, approximately 70 clusters of significant joint-linkage marker associations were identified and confirmed stalk strength as a highly polygenic complex trait. Both positive and negative effects relative to the common parent, B73, were observed revealing repulsion phase loci in the parental founders of these families. All QTL effect sizes were small. None captured greater than 2.7% of the heritable variation in RPR. When compared to allele series for DTA and EHT no significant correlations were apparent and no mutually pleiotropic QTL were identifiable. Although some covariation was observed between these traits and stalk strength at the level of genotypes, most was attributable to correlation of the traits among family means. This suggested pleiotropy may exist among small effects not well-defined by joint-linkage mapping.

In further resolution of the joint-linkage QTL mapping results and to better capture RPR variation existing across the NAM and IBM families, joint-linkage-assisted GWAS was performed. This analysis revealed the segregation of 141 significant associations across the

NAM and IBM families. However, no GWAS associations for stalk strength were shared with DTA or EHT associations mapped in joint-linkage-assisted GWAS and few were identified in close proximity. These were unable to explain the weak correlation between stalk strength and DTA or EHT measures. With the exception of the most robust polymorphisms, many RPR associations for GWAS did not co-localize with significant QTL mapped in each family. This is likely due to their capture of the substantial proportion of heritable stalk strength variation existing between NAM families that was not identified at the level of the family nested QTL. Although overlap of significant GWAS associations with past QTL studies was identified, the low mapping resolution of previous studies makes co-localization of QTL identified in this study a highly probable event for all but the least complex of traits (Flint-Garcia et al., 2003b).

No GWAS associations identified within the NAM and IBM families or the top associations identified within mixed model GWAS efforts in the 282 inbred diversity panel were characterized in close genomic proximity or tight linkage disequilibrium with established genes influencing phenylpropanoid genesis, cellulose synthesis or vegetative phase transition. This lack of significant co-localization with genes involved in pathways known to influence constituent traits of stalk strength may be due to a lack of segregation of these loci in natural variation as a result of their highly negative fitness effects. Nonetheless, given the low heritability of the trait it remains likely that numerous anatomical and compositional factors influence RPR and may complicate the mapping of causal loci. Future mapping efforts should seek to further decompose stalk strength into its constituent traits to reduce complexities limiting the power of genetic analyses.

In addition to mapping associations capturing heritable variation in RPR, genomic prediction was performed by RRBLUP. This method revealed substantial heritable variation in

RPR, ~84%, was captured using all polymorphisms simultaneously in construction and fitting of a genomic relatedness matrix. This was comparably achieved by model selection and construction of a nested QTL model, ~81%. However, the stalk strength variance captured by the nested QTL model may be attributed to the composition of the NAM and IBM panels and our ability to leverage an understanding of its structure and knowledge of the extended haplotypes which exist within each of its bi-parental families. Predicting germplasm outside of the bi-parental families of the NAM panel and building models from GWAS associations mapped in more diverse germplasm panels will not achieve the same degree of predictive accuracy as that attained in the NAM and IBM families. For this reason, RRBLUP and related genomic prediction methods remain critical to increase the rate of breeding enhancement beyond bi-parental crosses.

Stalk strength and its constituent traits will continue to play a key role in defining maize ideotypes. These analyses represent a comprehensive dissection of the genetic architecture of stalk strength. However, more studies detailing the genetic architecture responsible for those anatomical, compositional, and phase transition traits underlying the mechanics of stalk strength are needed. Future breeding efforts in low heritability traits as complex as stalk strength must develop methods incorporating an existing but inherently limited understanding of genetic architecture and an ability to accurately predict breeding values into selection frameworks to optimize crop improvement.

## REFERENCES

- Abedon B.G., Darrah L.L., Tracy W.F. (1999) Developmental changes associated with divergent selection for rind penetrometer resistance in the MoSCSSS maize synthetic. *Crop Sci.* 39:108-114.
- Arnold J.M., Josephson L.M., Parks W.L., Kincer H.C. (1974) Influence of Nitrogen, Phosphorus, and Potassium Applications on Stalk Quality Characteristics and Yield of Corn. *Agron. J.* 66:605-608.
- Benjamini Y., and Hochberg, Y. . (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57:289-300.
- Berzonsky W.A., Hawk J.A., Pizzolato T.D. (1986) Anatomical Characteristics of Three Inbred Lines and Two Maize Synthetics Recurrently selected for High and Low Stalk Crushing Strength. *Crop Sci.* 26:482-488.
- Bosch M., Mayer C.-D., Cookson A., Donnison I.S. (2011) Identification of genes involved in cell wall biogenesis in grasses by differential gene expression profiling of elongating and non-elongating maize internodes. *Journal of Experimental Botany.*
- Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y., Buckler E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635.
- Brown P.J., Upadaya N., Mahone G.S., Tian F., Bradbury P.J., Myles S., Holland J.B., Flint-Garcia S., McMullen M.D., Buckler E.S., Rocheford T.R. (2011) Distinct Genetic Architectures for Male and Female Inflorescence Traits of Maize. *PLoS Genet* 7:e1002383.
- Buckler E.S., Holland J.B., Bradbury P.J., Acharya C.B., Brown P.J., Browne C., Ersoz E., Flint-Garcia S., Garcia A., Glaubitz J.C., Goodman M.M., Harjes C., Guill K., Kroon D.E., Larsson S., Lepak N.K., Li H., Mitchell S.E., Pressoir G., Peiffer J.A., Rosas M.O., Rocheford T.R., Romay M.C., Romero S., Salvo S., Villeda H.S., Sofia da Silva H., Sun Q., Tian F., Upadaya N., Ware D., Yates H., Yu J., Zhang Z., Kresovich S., McMullen M.D. (2009) The Genetic Architecture of Maize Flowering Time. *Science* 325:714-718.
- Chambers K.R. (1987) Stalk Rot of Maize: Host-pathogen Interaction. *Journal of Phytopathology* 118:103-108.
- Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6:e19379.
- Endelman J.B. (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Gen.* 4:250-255.
- Enrico Pè M., Gianfranceschi L., Taramino G., Tarchini R., Angelini P., Dani M., Binelli G. (1993) Mapping quantitative trait loci (QTLs) for resistance to *Gibberella zeae* infection in maize. *Molecular and General Genetics MGG* 241:11-16.
- Flint-Garcia S.A., McMullen M.D., Darrah L.L. (2003a) Genetic Relationship of Stalk Strength and Ear Height in Maize. *Crop Sci.* 43:23-31.
- Flint-Garcia S.A., Jampatong C., Darrah L.L., McMullen M.D. (2003b) Quantitative Trait Locus Analysis of Stalk Strength in Four Maize Populations. *Crop Sci.* 43:13-22.
- Flint-Garcia S.A., Thuillet A.-C., Yu J., Pressoir G., Romero S.M., Mitchell S.E., Doebley J., Kresovich S., Goodman M.M., Buckler E.S. (2005) Maize association population: a high-

- resolution platform for quantitative trait locus dissection. *The Plant Journal* 44:1054-1064.
- Gibson B.K., Parker C.D., Musser F.R. (2010) Corn Stalk Penetration Resistance as a Predictor of Southwestern Corn Borer (Lepidoptera: Crambidae) Survival. *Midsouth Entomologist* 3:7-17.
- Gilmour A.R., Thompson R., Cullis B.R. (1995) Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 51:1440-1450.
- Gore M.A., Chia J.-M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S. (2009) A First-Generation Haplotype Map of Maize. *Science* 326:1115-1117.
- Hallauer A.R. (2000) *Biographical Memoirs V.78* The National Academies Press.
- HerediaDiaz O., Alsirt A., Darrah L.L., Coe E.H. (1996) Allelic frequency changes in the MoSCSSS maize synthetic in response to bi-directional recurrent selection for rind penetrometer resistance. *Maydica* 41:65-76.
- Hu H., Meng Y., Wang H., Liu H., Chen S. (2012) Identifying quantitative trait loci and determining closely related stalk traits for rind penetrometer resistance in a high-oil maize population. *TAG Theoretical and Applied Genetics* 124:1439-1447.
- Jannink J.-L., Lorenz A.J., Iwata H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* 9:166-177.
- Jung H.-J.G., Buxtono D.R. (1994) Forage quality variation among maize inbreds: Relationships of cell-wall composition and in-vitro degradability for stem internodes. *Journal of the Science of Food and Agriculture* 66:313-322.
- Kump K.L., Bradbury P.J., Wisser R.J., Buckler E.S., Belcher A.R., Oropeza-Rosas M.A., Zwonitzer J.C., Kresovich S., McMullen M.D., Ware D., Balint-Kurti P.J., Holland J.B. (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 43:163-168.
- Lee E.A., Darrah L.L., Coe E.H. (1996) Dosage effects on morphological and quantitative traits in maize aneuploids. *Genome* 39:898-908.
- Lee M., Sharopova N., Beavis W.D., Grant D., Katt M., Blair D., Hallauer A. (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. *Plant Molecular Biology* 48:453-461.
- McMullen M.D., Kresovich S., Villeda H.S., Bradbury P., Li H., Sun Q., Flint-Garcia S., Thornsberry J., Acharya C., Bottoms C., Brown P., Browne C., Eller M., Guill K., Harjes C., Kroon D., Lepak N., Mitchell S.E., Peterson B., Pressoir G., Romero S., Rosas M.O., Salvo S., Yates H., Hanson M., Jones E., Smith S., Glaubitz J.C., Goodman M., Ware D., Holland J.B., Buckler E.S. (2009) Genetic Properties of the Maize Nested Association Mapping Population. *Science* 325:737-740.
- Meuwissen T.H.E., Hayes B.J., Goddard M.E. (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157:1819-1829.
- P. J. Loesch J., Calvert O.H., Zuber M.S. (1962) Interrelations of Diplodia Stalk Rot and Two Morphological Traits Associated with Lodging of Corn. *Crop Sci.* 2:469-472.
- Papst C., Bohn M., Utz H.F., Melchinger A.E., Klein D., Eder J. (2004) QTL mapping for European corn borer resistance (*Ostrinia nubilalis* Hb.), agronomic and forage quality traits of testcross progenies in early-maturing European maize (*Zea mays* L.) germplasm. *TAG Theoretical and Applied Genetics* 108:1545-1554.

- Poland J.A., Bradbury P.J., Buckler E.S., Nelson R.J. (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences*.
- Pritchard J.K., Stephens M., Donnelly P. (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155:945-959.
- R D.C.T. (2011) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.
- SAS. (2002-2004) SAS 9.1.3, Cary, NC: SAS Institute Inc.
- Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S., Liang C., Zhang J., Fulton L., Graves T.A., Minx P., Reily A.D., Courtney L., Kruchowski S.S., Tomlinson C., Strong C., Delehaunty K., Fronick C., Courtney B., Rock S.M., Belter E., Du F., Kim K., Abbott R.M., Cotton M., Levy A., Marchetto P., Ochoa K., Jackson S.M., Gillam B., Chen W., Yan L., Higginbotham J., Cardenas M., Waligorski J., Applebaum E., Phelps L., Falcone J., Kanchi K., Thane T., Scimone A., Thane N., Henke J., Wang T., Ruppert J., Shah N., Rotter K., Hodges J., Ingenthron E., Cordes M., Kohlberg S., Sgro J., Delgado B., Mead K., Chinwalla A., Leonard S., Crouse K., Collura K., Kudrna D., Currie J., He R., Angelova A., Rajasekar S., Mueller T., Lomeli R., Scara G., Ko A., Delaney K., Wissotski M., Lopez G., Campos D., Braidotti M., Ashley E., Golser W., Kim H., Lee S., Lin J., Dujmic Z., Kim W., Talag J., Zuccolo A., Fan C., Sebastian A., Kramer M., Spiegel L., Nascimento L., Zutavern T., Miller B., Ambroise C., Muller S., Spooner W., Narechania A., Ren L., Wei S., Kumari S., Faga B., Levy M.J., McMahan L., Van Buren P., Vaughn M.W. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112-5.
- Sibale E.M., Darrah L.L., Zuber M.S. (1992) Comparison of two rind penetrometers for measurement of stalk strength in maize. *Maydica* 37:111-114.
- Tian F., Bradbury P.J., Brown P.J., Hung H., Sun Q., Flint-Garcia S., Rocheford T.R., McMullen M.D., Holland J.B., Buckler E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159-162.
- Valdar W., Holmes C.C., Mott R., Flint J. (2009) Mapping in Structured Populations by Resample Model Averaging. *Genetics* 182:1263-1277.
- Yu J., Holland J.B., McMullen M.D., Buckler E.S. (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539-51.
- Yu J., Pressoir G., Briggs W.H., Vroh Bi I., Yamasaki M., Doebley J.F., McMullen M.D., Gaut B.S., Nielsen D.M., Holland J.B., Kresovich S., Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203-208.
- Zuber M.S., Grogan C.O. (1961) A new technique for measuring stalk strength in corn. *Crop Sci.* 1:378-380.
- Zuber M.S., Colbert T.R., Darrah L.L. (1980) Effect of recurrent selection for crushing strength on several stalk components in maize. *Crop Sci.* 20:711-717.

**CHAPTER 4**  
**MAPPING THE GENETIC ARCHITECTURE OF MAIZE HEIGHT AND**  
**CORRELATED COMPLEX TRAITS**

**ABSTRACT**

Height is one of the most heritable and genetically complex maize traits. Its variation may be readily captured by pedigree, yet most molecular associations elude discovery and maize height proves nearly irreducibly complex in natural populations. To endeavor dissection of this complexity in constructed populations, we measured plant height and the correlated traits of ear height, node counts, and flowering time in a nested association mapping panel (NAM) of 5,000 recombinant inbred lines (RILs) across ten environments and a diverse panel of over 2,700 maize inbreds in three environments. These panels were genotyped for over 25 million and 680,000 single nucleotide polymorphisms (SNPs), respectively. Joint-linkage and genome wide association mapping (GWAS) found the largest effect captured about 2% of heritable height variation. Comparable effect sizes at the locus were confirmed by positional cloning in two near isogenic lines (NILs). However, most mapped natural height variation did not co localize with published mutations. Allele series estimates also revealed minimal co localization of loci capturing height variation with those explaining variation in flowering time despite the apparent genotypic correlation of these traits. About 75% of heritable height variation was captured by a genomic relationship matrix constructed from all genotyped SNPs. A comparable fraction of variation was captured by significant joint-linkage identified associations in the NAM panel. Given genotyping and phenotyping advances, a resolved empirical understanding of the architecture and evolution of complex traits such as maize height is becoming feasible on a genomic scale and will facilitate future crop improvement.

## **AUTHOR SUMMARY**

Despite its considerable heritability and the simplicity of its measurement, plant height is a decidedly complex trait controlled by numerous genetic factors. Nonetheless, high-throughput sequencing advances now afford unprecedented opportunities to evaluate the segregation of these factors on a genomic scale and to identify those responsible for heritable variation in complex traits such as height. In this study, we measured two large genetically diverse maize populations to identify regions of the genome associated with natural variation in plant height, ear height, node counts, and flowering time. In a crop averaging about 180 centimeters in height, the largest genetic effects for maize height remained less than about five centimeters. The presence and relative size of two such effects were validated by conformational fine mapping efforts. Published loci known to regulate height as identified by mutant screens do not explain a substantial fraction of the natural heritable height variation identified in this study. Similarly, loci identified as regulating height and flowering time in the surveyed diversity did not co-localize despite correlation of these traits at the level of genotypes. Given the number of genetic factors evidently regulating height variation, a genomic relationship matrix was formed to assess if all measured genetic diversity could simultaneously capture heritable height variation in maize. Using this approach, about three quarters of the heritable height variation was captured by all genetic factors. This revealed an ability to accurately predict height given knowledge of the genetic factors segregating within a population of closely related maize inbreds.

## INTRODUCTION

Adaptations in height are essential to plant fitness and agricultural performance. They are intrinsic to the evolutionary history, standing diversity, and genetic architecture of a population and impact the velocity of its phenotypic evolution. The height distribution of plant populations evolving in competitive environments is in part a product of natural selection imposed by the effects of light interception, weed competition, seed dispersal, carbon and nutrient capture (Lin et al., 1995) on individual and inclusive fitness. In maize and other domesticated crops, breeding efforts facilitating agricultural industrialization select adaptations maximizing breeding value as a function of yield gain under monoculture. Height adaptations buffering yield variation to any environmental instability that persists in fertile homogeneously planted fields are also desirable factors to be selected. Many of these phenomena are well illustrated in the yield gains and height reductions of rice and wheat during the ‘Green Revolution’ (Khush, 2001). During selection for industrial agriculture, height adaptations increase harvest uniformity, favorably partition carbon and nutrient resources between grain and non-grain biomass, and consequently enhance fertilizer, pesticide, and water-use efficiency (Khush, 2001). In many plants, especially grasses (*Poaceae*) such as maize, wheat, and rice, apical growth is terminated at reproductive maturity (Lin et al., 1995). This may further establish genetic correlations among height and maturation, and increases the biological complexity and potential number of selective forces concurrently impacting the evolution of plant height in a population.

Given plant height’s ease of measure, numerous loci have been successfully cloned aiding in elucidation of the gibberellin, brassinosteroid, auxin, and other biosynthetic, regulatory, and developmental networks. Our current understanding of the genetics of plant height and associated traits is primarily derived from mutant screens and nucleotide sequence homology as

opposed to the dissection of natural standing genetic variation (Salas Fernandez et al., 2009). This is especially true of the rice, *semi-dwarf1*, and wheat, *Reduced height1*, mutants popularized by the ‘Green Revolution’, and later identified as influencing an oxidase responsible for gibberellin biosynthesis (Sasaki et al., 2002) and a transcription factor modulating gibberellin signaling (Peng et al., 1999), respectively. In these instances, as well as in the maize mutant, *dwarf-8* (Winkler and Freeling, 1994), orthologs in *Arabidopsis thaliana* such as the *Gibberellin Insensitive* gene have been identified (Peng et al., 1999).

Previous studies reveal varied consensus among the allelic diversity identified in mutant screens and that naturally segregating in the standing diversity of plant populations and the correspondence appears to be largely trait and population dependant (Atwell et al., 2010; Tian et al., 2011). Furthermore, the applicability of the common disease common variant hypothesis (Buckler et al., 2009; Lander, 1996; Risch and Merikangas, 1996), or allelic heterogeneity of a single molecular function, has not been well established in many of these traits or populations. It remains unclear if genetic correlations among these traits occur through genetic variation captured by population structure, pleiotropy of single segregating alleles, or linkage disequilibrium among alleles with otherwise independent modes of action. It is also not well established if genetic correlations among these traits result from interactions at the molecular, biochemical, or physiological level (Wagner and Zhang, 2011). Empirical answers to these questions of causality and modularity provide a deeper understanding of adaptive landscapes and facilitate the optimal selection and recombination of heritable phenotypic diversity during crop improvement (Messina et al., 2011).

Maize breeders have selected and recombined the allelic variation underlying phenotypes for at least 7,000 years (Hamblin et al., 2007; Piperno et al., 2009); however, the past decade has

seen phenomenal advances in genotyping technologies and increasingly afforded breeders molecular markers to serve as potential proxies for heritable phenotypic diversity and facilitate the further dissection of its segregation. Marker density approaching whole genome coverage across genetically diverse plant populations is in the near future (Chia, 2012; Gore et al., 2009; Huang et al., 2010) . Nonetheless, our ability to associate a molecular marker and causal allele remains limited by effective population size, ancestral history, a trait's genetic architecture, genome size, and the analytical methods available to make the connection. Methods to explain heritable phenotypic variation and predict phenotypes using the recent deluge of genotypic data are in rapid development and have progressed from widespread use of sequential single marker analyses to multiple and multivariate regression methods including variable selection and regularized regression on massive marker sets (Logsdon et al., 2010; Tian et al., 2011; Wisser et al., 2011). Genome wide association studies (GWAS) pioneered in human genetics have been performed in *Arabidopsis* (Brachi et al.), maize (Tian et al., 2011) and several other species in field environments. In contrast to human GWAS, studies in plants are afforded several statistical advantages including experimental control of minor allele frequencies (MAFs) and population structure through controlled pollination, and replication of inbreds across environments. Many of these desirable properties were employed in the study of maize through crossing of a nested association mapping panel (NAM) with 5,000 recombinant inbred lines (RILs) (McMullen et al., 2009; Tian et al.; Tian et al., 2011).

Maize is a classic genetic model with exceptional phenotypic and molecular diversity; it is also one of the most economically significant crops and has been intensely bred over the last century. The diversity and predominately out-crossing mating system of maize lends to rapid decay in linkage disequilibrium (LD) (Remington et al., 2001) and average nucleotide diversity

estimates indicate over 30 million segregating polymorphisms may exist in a modestly sized population (Gore et al., 2009). This level of genetic diversity and decay of linkage disequilibrium afford high mapping resolution. However, GWAS is limited to the linkage disequilibrium of causal and genotyped polymorphisms and thus requires a high depth of sequence coverage. To address this limitation, construction of a 2<sup>nd</sup> generation maize HapMap has identified over 55 million segregating single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) in the genomes of 103 diverse maize inbreds, 27 of which were parents of the NAM panel and IBM family (Chia, 2012; Lee et al., 2002; McMullen et al., 2009). These densely genotyped SNPs segregating across NAM parents were imputed in their progeny (RILs) based on 1,106 markers genotyped directly on the RILs. This facilitated a finer-scale view of genetic diversity and improved our ability to identify significant genotype to phenotype associations in the NAM panel (Tian et al., 2011).

In this study, we employed the maize NAM panel and IBM family in joint-linkage mapping of quantitative trait loci (QTL) and a GWAS of polymorphisms explaining heritable variation in total height and genetically correlated phenotypes such as primary ear height, node counts, flowering time, and traits derived from these measures. A panel of 2,711 diverse maize inbreds collected from around the world (Ames) was also genotyped and phenotyped to further characterize the genetic architecture of height and flowering time. Fine mapping of two near isogenic lines (NILs) possessing two diverse introgressions in the same genetic background was performed to validate the NAM panel and IBM family mapping results on the long arm of chromosome nine wherein the largest variance was captured by joint-linkage mapping. The NAM and AMES panels were grown across ten and three temperate field environments, respectively. Fine mapping efforts were undertaken in three environments as well. By

accounting for confounding environmental effects, this multi-environment analysis allowed us to better partition heritable height variation across the maize genome and to refine the precision of our allele effect estimates and pleiotropy.

## **MATERIALS AND METHODS**

### *Germplasm and data collection*

The germplasm, crossing, and genotyping of the NAM panel have been previously described (Buckler et al., 2009; McMullen et al., 2009). Traits were scored in 10 environments: Aurora, New York; Columbia, Missouri; Urbana, Illinois; and Clayton, North Carolina in the summer of 2006 and 2007 and again in the summer of 2008 and 2009 in Aurora, New York and Columbia, Missouri, respectively. Days to silk, days to tassel, and the anthesis-silking interval were scored as previously described (Buckler et al., 2009). Plant height was measured as the distance from the soil line of the plant to the base of the flag leaf, ear height, as the distance from the soil line to the node of the primary ear. The ratio of ear to plant height was calculated as their quotient. Node measures below the primary ear were taken as the nodes between the node of the top brace root and the node of the primary ear. Node measures above the primary ear were taken as the nodes from the primary ear to the node of the flag leaf and were added to those measured below the ear to attain the total node count. Genotyping-By-Sequencing (Elshire et al., 2011) of RILs and deep sequencing of NAM parents through the first and second generation maize HapMaps (Chia, 2012; Gore et al., 2009) allowed testing of approximately 55 million SNPs across the NAM panel (Tian et al., 2011).

Germplasm composing the Ames inbred panel was requested from the USDA-ARS North Central Regional Plant Introduction Station (NCRPIS) located in Ames, IA. This germplasm resource consists of 2,711 diverse inbred lines collected from populations located around the world. Measures of total plant height, primary ear height, days to silks, and days to anthesis on

the AMES panel were phenotyped across three field environments: Aurora, New York, Columbia, Missouri, Clayton, NC in the same manner as was performed across the maize NAM panel. All Ames panel inbreds were also scored using Genotyping-By-Sequencing as previously described (Elshire et al., 2011; Romay, 2012).

Two near isogenic lines acquired from Syngenta AG consisting of chromosome nine introgressions of the tropical inbreds CML277 (58Mb) and CML333 (69Mb) in a Stiff-Stalk B73 background were grown across six unique field environments: Aurora, New York in the summer of 2008 as an F<sub>2</sub>. Aurora, New York ; Columbia, Missouri; Madison, WI; and Clayton, North Carolina in the summer of 2009 and again in the summer of 2010 in Aurora, New York and Columbia, Missouri, to select and fix recombinants. A panel of 200 recombinant lines that were selfed to fixation were selected and Genotyping-By-Sequencing as previously described (Elshire et al., 2011; Romay, 2012).

#### *Analysis of phenotypic variance*

Genetic, environment, and environmentally conditional genetic variance components and best linear unbiased predictors (BLUPs) for all phenotypes of each NAM RIL and AMES inbred across environments were calculated using ASREML version 3.0. Custom Java code was used to perform likelihood ratio testing ( $\alpha = 0.05$ ) and model selection in a backwards elimination framework using ASREML v. 3.00 (Gilmour et al., 1995). The Java code is available upon request. Blocking effects were modeled as random independent effects when deemed significant by likelihood ratio testing with a critical value of  $\alpha = 0.05$ . A first order autoregressive by first order autoregressive (AR1 x AR1) error correlation structure was fitted for range and row within each of the fields as deemed significant. Independence of residuals was assumed between fields and no heterogeneous measurement error variance or nugget variance for

measures was fitted. Following the partitioning of phenotypic variation and calculation of BLUPs, trait measures were used to calculate phenotypic, genetic, and environmental trait covariance and correlation matrices in R v2.12.0 (R, 2011).

Heritability on a plot and line mean basis were calculated as previously described (Hung et al., 2012) using ASREML. Clustering of traits to determine modularity was performed using Pearson's correlation coefficients and modulated modularity clustering (MMC) (Stone and Ayroles, 2009) of phenotypes and best linear unbiased predictors (BLUPs) for both environment and genotype. Testing for correlation among predicted evolutionary trajectory based on the multivariate breeders equation were performed between NAM families using the Random Skewer's method with 100,000 random skewers simulated per contrast (Cheverud and Marroig, 2007). Percent identity by state between the unshared parents of each NAM family was calculated as the proportion of shared alleles in the 1<sup>st</sup> generation HapMap (Gore et al., 2009) as previously described (Loiselle et al., 1995).

#### *Joint-Linkage Analysis of QTL within the NAM panel and IBM family*

To partition genetic variance beyond the genotypic level, the SAS v9.2 statistics package (SAS, 2002-2004) was implemented. SAS PROC GLMSelect was executed to regress BLUPs for each of the traits against the 1,106 markers nested within NAM and IBM families in a joint-linkage QTL model. A model term was fitted for each family and family nested marker selection was performed by stepwise regression as previously described (Buckler et al., 2009). The significance of model entry and exit were set to  $p < 5e-4$  based on 1,000 null permutations of plant height data. This stepwise model building routine was bootstrapped, re-sampling 20% of the RILs within each NAM family for 100 sampling iterations to calculate a resample model inclusion probability (RMIP) (Valdar et al., 2009) and improve identification of robust QTL. A

RMIP greater than 10 out of 100 sampling iterations was attained from less than 5% of the selected markers at the given model entry and exit criteria in null permutation testing of maize height. Code for NAM and IBM family-stratified bootstrapping of joint-linkage QTL mapping is available on request. To identify the best fit model built during resampling, family nested QTL in each of the fixed effect models built during the bootstrapping routine were refitted to the full data set and the model with the minimal Bayesian information criterion (BIC) was selected for each trait using the base library in R v2.12. This model was used to assess pleiotropy with related traits based on significant correlation of allele effects as previously described (Tian et al., 2011).

#### *Joint-Linkage Assisted GWAS Analysis within the NAM panel and IBM family*

To further dissect the joint linkage mapped QTL, we conducted a GWAS with approximately 5,000 RILs containing 1.6 million snps imputed as previously described (Tian et al., 2011) from NAM panel founders maize HapMapv1, as well as 25 million SNPs imputed from maize HapMapv2. After removing all QTL from a single linkage group, the BIC determined optimal family nested QTL model was fitted to BLUPs for each phenotype and residual variance attributed to QTLs of the missing linkage group as well as genetic variance not previously accounted for by the QTL model was determined. This procedure was repeated for all ten of the linkage groups in the maize genome. Using these estimates of residual variance and TASSEL v3.0 we performed a stepwise regression procedure as previously described (Bradbury et al., 2007; Tian et al., 2011) with a genome wide p-value significance threshold of  $5e-4$  based on null permutation testing of maize height. The procedure was repeated for a total of 100 stratified resamples. The fraction of resample model builds in which a SNP was included

revealed its RMIP (Valdar et al., 2009). A RMIP greater than 3 out of 100 sampling iterations was attained from less than 5% of the selected markers in null permutation testing of height.

#### *GWAS Analysis across the Ames Inbred Diversity Panel*

Using Genome Association and Prediction Integrated Tool (GAPIT) (Lipka et al., in review) in R v2.12.0 (R, 2011), sequential single polymorphism mixed-model GWAS for plant height was performed across the Ames inbred diversity panel. This approach accounted for the natural population structure and the false associates it creates (Yu et al., 2006) and allowed for potential identification of significant associations across the GBS characterized polymorphisms. The significance of these associations was determined by measures of false discovery rate at an ( $\alpha=0.05$ ) (Benjamini, 1995).

#### *Positional Cloning of Near Isogenic Lines Validating Joint-Linkage Associations*

Two near isogenic lines (NILs) possessing introgressions for CML277 and CML333 recurrently backcrossed into a B73 background were positionally cloned using R v2.12.0 (R, 2011). Imputation of markers across the regions of interest was performed using TASSEL v3.0. Sequential single marker testing of the 200 fixed lines genotyped by GBS was run and the regions controlling plant height were further refined and used to validate the NAM family mapping results within the CML277xB73 and CML333xB73 families.

#### *Analysis of Genomic Prediction Accuracy by RRBLUP in NAM and AMES panels*

To partition genetic variation across the entire maize genome simultaneously, we implemented genomic prediction using ridge regression best linear unbiased prediction as implemented by the package rrBLUP (Endelman, 2011) in R v2.12.0 (R, 2011). Genomic relatedness matrices were constructed for the NAM and IBM families using the ~1.2 million

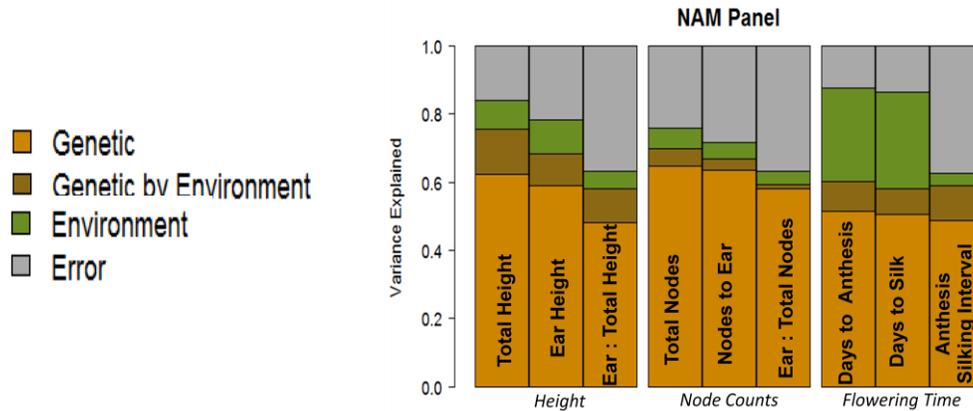
polymorphisms of the maize HapMap(Gore et al., 2009). Estimates of genomic relatedness among the Ames inbred diversity panel were calculated using the ~437,650 GBS polymorphisms. BLUPs for each panel were regressed against its genomic relatedness matrix to assess the ability of all polymorphisms to simultaneously predict the genomic estimated breeding value (GEBV) of each genotype. The prediction accuracy was determined by regressing actual breeding values against their respective GEBVs. The robustness of this relationship was tested to determine degree of shrinkage in the coefficient of determination upon fivefold cross-validation of the NAM and IBM families as well as the Ames inbred diversity panel.

## RESULTS

### *Partitioning height, node count, and flowering time variation in NAM panel*

Measures of total plant height and genetically correlated traits such as primary ear height, days to anthesis, days to silk, nodes to primary ear, and total node counts were harvested from the RILs of the maize NAM panel surveyed across ten temperate field environments. Variation in these phenotypes and traits derived from them such as the anthesis silking interval, and the ratios of both primary ear height: total height as well as primary ear node: total node count were partitioned into genetic, environmental, and environmentally conditional genetic variance components (Figure 4.1). Despite comparable photoperiods among the surveyed field environments, the proportion of variation captured by environmental differences for both days to silk and days to anthesis were substantially greater than that observed for either height or node counts measures. Normalization of temperature difference between environments through the conversion of days to anthesis and days to silk to growing degree days failed to produce a substantial reduction in the proportion of phenotypic variance captured between environments.

In contrast to environmental effects on phenotypic variation, the fraction of phenotypic variance captured by environmentally conditional genetic variation was found to be more substantial in height related traits than that observed in flowering time or node count measures.



**Figure 4.1 Partitioning Phenotypic Variation of NAM Panel**

The proportion of genetic, environmental, and environmentally conditional genetic variation captured for each surveyed phenotype was found to differ substantially.

**Table 4.1 Heritability of Height and Related Traits in NAM Panel**

All surveyed traits were found to be substantially heritable. Measures of flowering time were the most heritable across the NAM panel. These were followed by plant height and node count phenotypes.

Class	Trait	$H^2_{plot}$	$H^2_{line}$
Height	Total Height	.56	.92
Height	Ear Height	.59	.90
Height	Ear : Total Height	.50	.76
Node Counts	Total Nodes	.58	.82
Node Counts	Nodes to Ear	.55	.80
Node Counts	Ear: Total Nodes	.53	.79
Flowering Time	Days to Anthesis	.71	.94
Flowering Time	Days to Silk	.70	.94
Flowering Time	Anthesis-Silking Interval	.33	.78

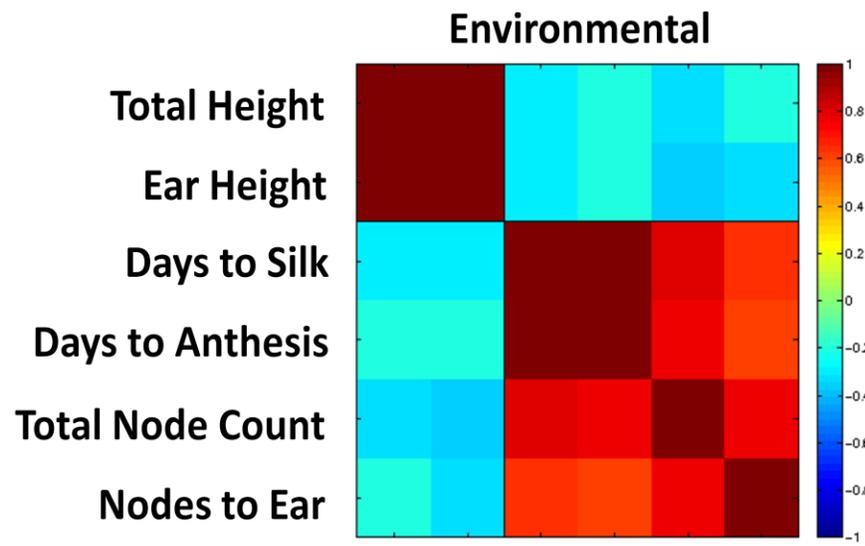
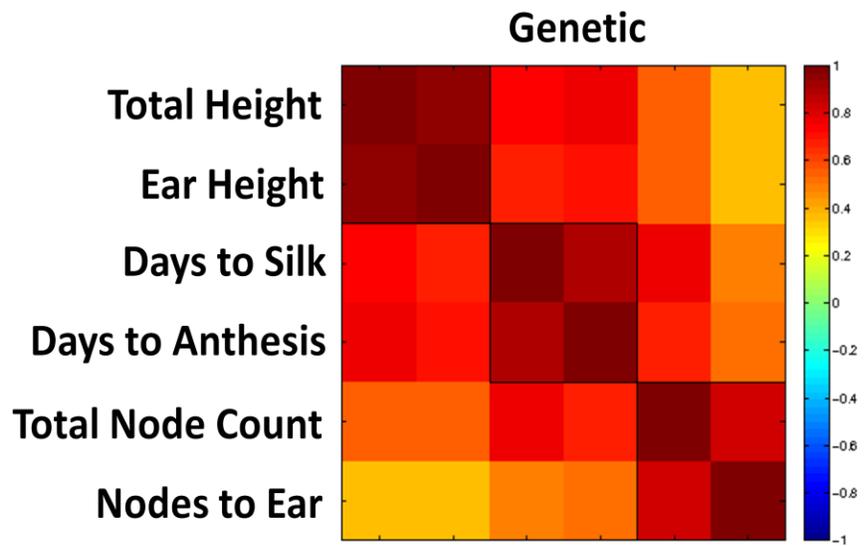
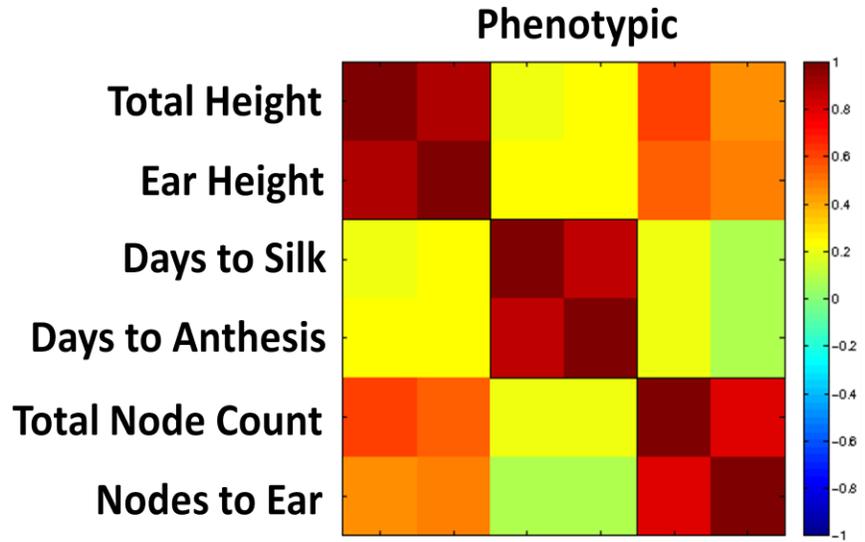
Despite the larger proportion of total phenotypic variation captured by genetic diversity in height and node count traits as compared to the flowering time traits, estimates of broad sense heritability on a line and plot mean basis indicated a higher proportion of phenotypic variation for both days to silk and days to anthesis was heritable after accounting for the environmental variance (Table 4.1).

Across the surveyed traits, the proportion of heritable variation that was attributable to between NAM family differences was highest in days to anthesis and days to silk at 68% and 64% respectively. Relative to these measures of variation, the proportion of heritable variance in total node counts and nodes to primary ear between NAM families was slightly reduced at 63% and 60% respectively. Substantially lower proportions of heritable variation were captured between the NAM families for both primary ear height measures at 46% and total height measures at 29%. To address if more heritable variation in height within a NAM family was correlated with more variation in another trait within the same family we compared heritabilities across families. Upon review of the variance in the variation of these traits within each NAM family, several weak positive correlations were observed between total height and flowering time ( $r_{\text{anthesis}} = 0.26$ ,  $r_{\text{silk}} = 0.28$ ) heritabilities as well as height and node counts heritabilities ( $r_{\text{total}} = 0.40$ ,  $r_{\text{ear}} = 0.35$ ) across the 25 NAM families. However, variation in height to primary ear among the 25 NAM families was by far the most significant trait correlated with total height heritability ( $r_{\text{ear}} = 0.67$ ).

Beyond measures of variation, to assess genetic covariation among the traits of the NAM panel we constructed and clustered (Stone and Ayroles, 2009) phenotypic, genetic, and environmental trait covariance matrices across plots, RILs, and field environments, respectively. Although correlations between traits were observed at the phenotype and environment level,

many of these significantly differed from their respective genetic correlations which were often significantly increased in strength. Little significant phenotypic correlation was noted between height and flowering time measurements across the NAM panel; however, a marginal correlation existed between height and node count measures ( $r = .43$ ). At the level of RILs, the phenotypic relationship between nodes and height measures was maintained ( $r = .39$ ). Nonetheless, a significant increase in correlation was observed between height and flowering time measures. Correlation among the traits across environments revealed a weakly negative correlation between height related traits and both flowering and node count traits; but, correlations remained strong between flowering time and node count related traits. Over all levels of observation, clustering of these matrices (Stone and Ayroles, 2009) revealed maturational and morphological traits retained substantial modularity. While this level of independence was most reduced at the level of genetic variation, it remained readily apparent.

In addition to clustering the covariation of traits at the level of plots, RILs, and environments across the entire NAM panel, construction of covariance matrices detailing relationships among the same six traits at the levels of RILs within each NAM family facilitated a multi-trait comparison of genetic architectures. Similarities in predicted trait responses to 10,000 random selection gradients were calculated between NAM families by application of the random-skewers method (Cheverud and Marroig, 2007) to the breeder's equation. Given the B73 reference-based design of NAM, variation in predicted responses to selection between NAM families may be attributed to the genetic architecture of each non-reference parent and how it complements B73 in of RIL progeny. Nonetheless, comparisons of identity by state (Loiselle et al., 1995) among the NAM parents revealed no significant correlation with similarity in their respective NAM family's predicted response to selection ( $r = -0.06$ ).

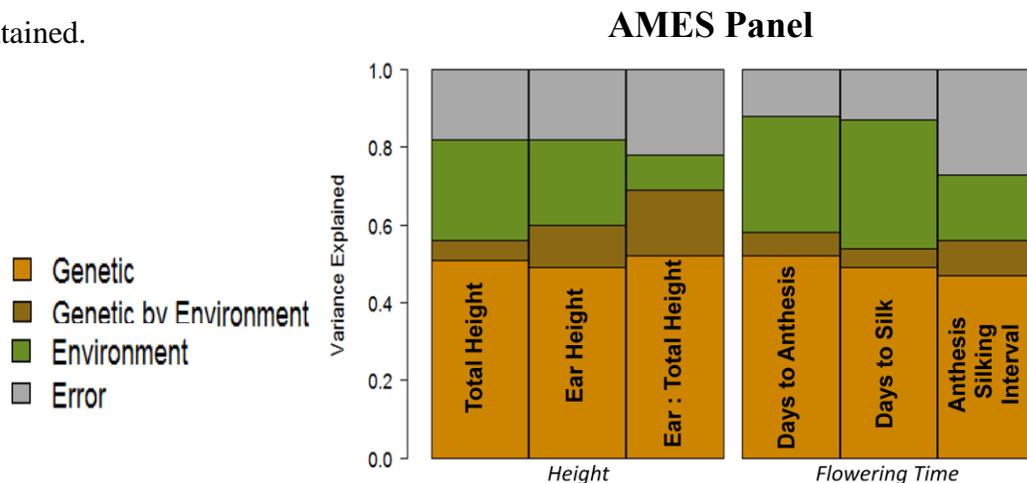


### Figure 4.2 Modular Clustering of Traits

Height, flowering time, and node count related traits exhibited substantial modularity at the phenotypic level (across plots). Their apparent independence was reduced at the genetic level (across RILs). This reduction was particularly notable between flowering time and height related traits. Environmental correlations among height related traits and both flowering and node counts were weakly negative (across field environments).

#### *Partitioning phenotypic variation in AMES panel*

Phenotypic variation in the AMES panel was partitioned in a manner analogous to that performed for the NAM panel (Figure 4.3). Despite the slightly larger proportion of the variation that was attributed to differences between surveyed field environments for all traits, the proportions of genetic, environmental, and environmentally conditional genetic variation captured for each trait was comparable to that identified in the NAM panel. Similarly, substantially less of the variation in derived traits such as anthesis silking interval was explainable when compared to directly measured phenotypes. In contrast to the NAM panel, less variation was noted in the proportions of heritable variance observed between the six surveyed traits. Upon comparing plot and line mean broad sense heritabilities, all measures of heritable variation were slightly reduced. Nonetheless, rank among the traits in both panels was largely maintained.



### Figure 4.3 Partitioning Phenotypic Variation of AMES Panel

The proportion of genetic, environmental, and environmentally conditional genetic variation captured for each surveyed phenotype was found to largely parallel that observed in the NAM panel.

**Table 4.2 Heritability of Height and Related Traits in AMES Panel**

All surveyed traits were found to be substantially heritable. Measures of flowering time were the most heritable across the AMES panel. These were followed by plant height and node count phenotypes.

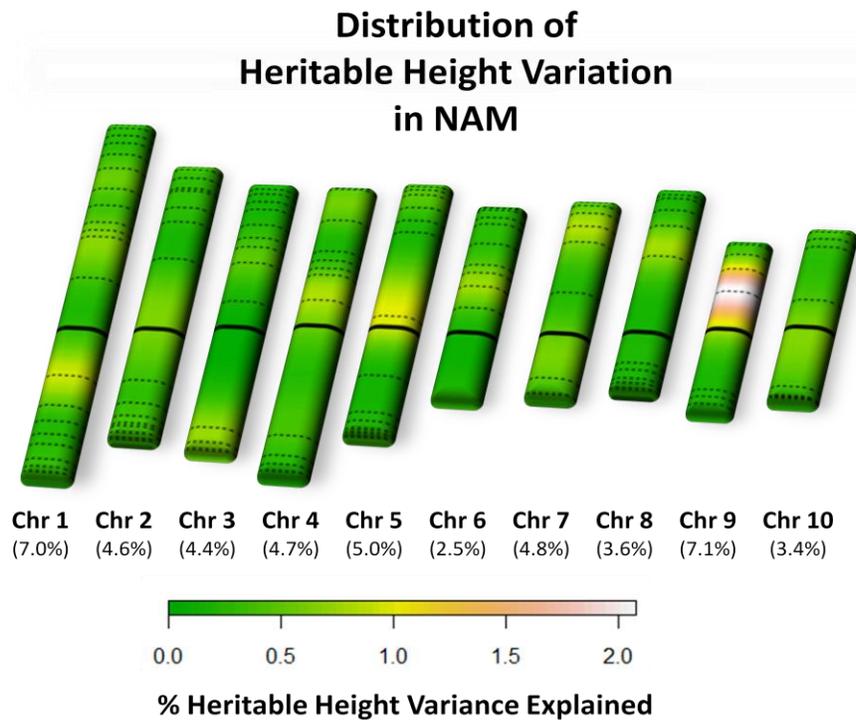
Class	Trait	$H^2_{plot}$	$H^2_{line}$
Height	Total Height	.53	.89
Height	Ear Height	.49	.88
Height	Ear : Total Height	.47	.86
Flowering Time	Days to Anthesis	.66	.91
Flowering Time	Days to Silk	.68	.91
Flowering Time	Anthesis-Silking Interval	.41	.83

*Joint-Linkage Mapping: Partitioning heritable variation within NAM families*

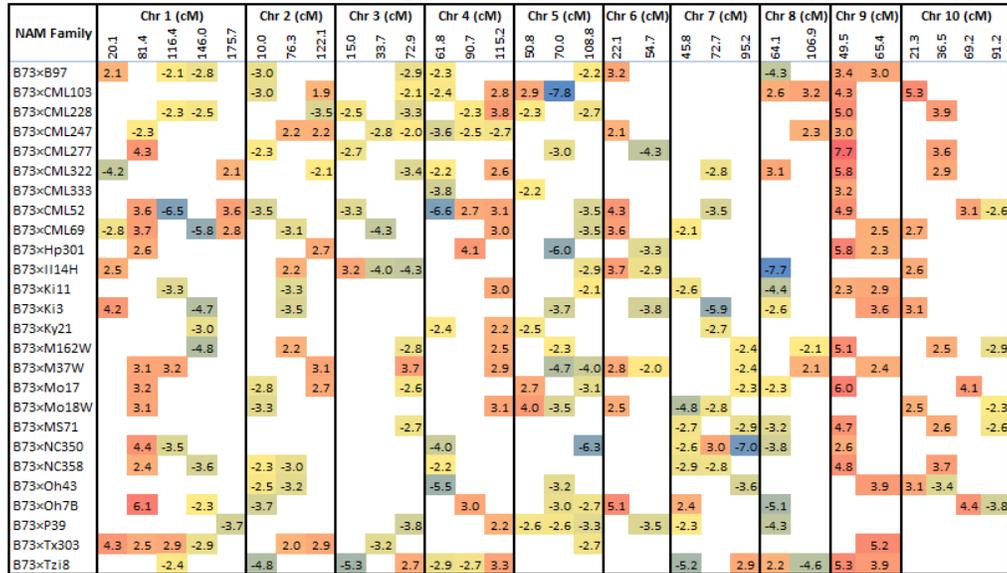
Joint-linkage mapping of total height and related traits was performed by a stepwise QTL model selection approach from a set of 1,106 markers nested within each of the 25 NAM families as previously described (Buckler et al., 2009). As a measure of QTL robustness, family-stratified parametric bootstrapping was employed to attain estimates of the resample model inclusion probability (RMIP) (Valdar et al., 2009) for putative QTL. Given the complexity of height and correlated phenotypes, the NAM panel's effective population size, genetic map size, marker density, and the QTL selection mapping method, over 39 robust QTL with a RMIP >0.05 were identified for every surveyed phenotype.

These QTL models captured the major proportion of heritable phenotypic variation for most of the traits (median height 77% captured – median days to anthesis 88% captured). However, the identified total plant height QTL were all of small effect with the largest explaining only approximately 2.1% of the total heritable phenotypic variance for height (Figure

4.4). Differences were noted in the magnitude and directionality of QTL captured for the phenotypes within each of the 25 NAM families; however, over 90% of the mapped QTL for each phenotype were identified as significant and shared across at least three NAM families (Figure 4.5). Over 70% of these shared QTL contained allele series possessing both positive and negative effects relative to the common reference parent, B73. Positive and negative effects for each trait explained comparable phenotypic variance. Their distributions were symmetric and of nearly equal variance. Given the polygenicity of plant height, flowering time, and node counts, all parental genomes possessed repulsion phase QTL for every trait at the mapping resolution noted within each NAM family. Similarly, all NAM families exhibited transgressive segregation for the complex traits surveyed.



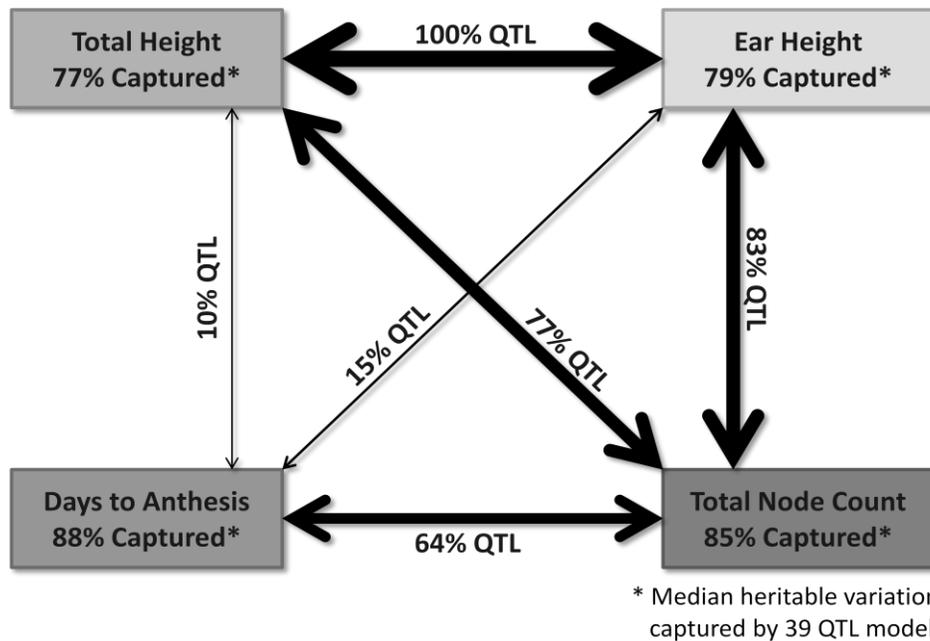
**Figure 4.4 Distribution of heritability height variation in NAM by QTL model selection**  
The distribution of height variation across the maize genome revealed both the polygenicity and complexity of height. The locus controlling the largest proportion of variation was identified on chromosome 9L, yet only captured approximately 2.1% of the variation.



**Figure 4.5 Effect sizes of most significant height QTL by NAM family**  
 Estimate of allelic effects across the maize genome revealed numerous allele series on all chromosomes. Most effects were less than five centimeters in size.

*Pleiotropy Among Phenotypes in NAM Panel*

Total plant height and measures of flowering time have long been considered tightly regulated phenomena. Analysis of strong genetic correlation (days to anthesis  $r = 0.62$ , days to silk  $r = .57$ ) between these traits across the RILs in the NAM panel initially supported this supposition. Nonetheless, after dissecting the QTL underlying these traits across all 25 NAM families, only four QTL present within a 39 QTL model were identified possessing significantly correlated ( $r > 0.4$ ) allelic effect estimates with days to anthesis. The allelic effects of these same four QTL were also significantly correlated with days to silk measurements. In contrast, all the allelic effects of QTL mapped for total plant height were found to strongly correlate with those identified when the model was fitted to the heritable variation for primary ear height. Even the allelic effects of thirty of the thirty nine total height QTL were significantly correlated with total node counts (Figure 4.6).



**Figure 4.6 Pleiotropy of QTL capturing height variation in NAM**

All the allelic effects of those QTL mapped for height were highly significantly correlated with their effect on primary ear height when the same QTL model was fitted to both traits. In contrast, the allelic effects of only four of the thirty nine height QTL were significantly correlated with days to anthesis. Comparable results were identified for days to silk.

*Joint-Linkage-Assisted GWAS: Partitioning heritable variation across NAM families*

Joint linkage mapping provides a powerful yet low resolution view of the genetic architecture underpinning plant height and correlated traits. Genotype-to-phenotype associations identified in linkage analysis remain limited to recently recombined linkage blocks segregating within each of the NAM families. This precludes our ability to specify the precise location of a genetic effect with much certainty beyond an interval of approximately 2-3 centimorgans in length.

To capitalize on ancestral recombination and further fractionate these QTL, background QTL on all but a single linkage group were fit in a NAM family nest QTL regression model and residual heritable height variation attributable to the absent linkage group was mapped to the polymorphisms by resampled forward regression GWAS. This joint-linkage assisted GWAS

method was conducted across the NAM families using HapMap snps and cnvs discovered in the NAM parents and imputed onto their RIL progeny as previously described. From the 26 million tested snps and cnvs, hundreds of significant associations (RMIP >0.05) were identified for plant height and correlated traits. Many of these associations were found to co-segregate across NAM families with the allelic effects of their nearest QTL for each trait; however, for several of these associations the directionality opposed the QTL's main allelic effects. Most pleiotropic QTL possessed significant GWAS associations for their underlying traits within a two centimorgan interval. For plant height and many of the correlated phenotypes, the distribution of significant associations across NAM families revealed approximately symmetric distributions with equal densities of positive and negative effects relative to a common parent; however, effect sizes were notably smaller than those observed during joint linkage mapping of QTL for all traits.

#### *Co-localization of natural genetic diversity and cloned height loci*

We possess substantial understanding of the molecular dynamics underpinning several biochemical pathways governing plant height such as those responsible for regulation and biosynthesis of gibberellins, brassinosteroids, and auxin hormones. However, the basis for most of our knowledge of these pathways does not stem from studies of naturally segregating genetic diversity. Surprisingly, few significant robust associations were found in linkage disequilibrium with established genes in these canonical pathways or in most cases within a 250,000 base pair window surrounding them. Marker density within these regions was not significantly diminished compared to the genome-wide distribution or those regions surrounding significant associations. Further analysis of over thirty five previously cloned plant height loci, similarly found little co-localization with those significant joint-linkage assisted GWAS hits identified in the NAM panel (Table 4.3).

**Table 4.3 Co-localization of candidate height genes and joint-linkage GWAS**

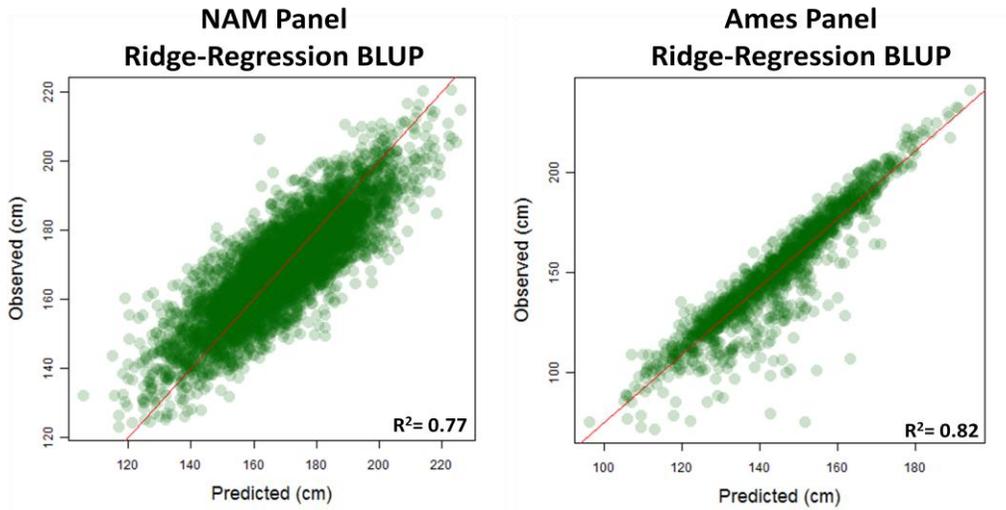
All the allelic effects of those QTL mapped for height were highly significantly correlated with their effect on primary ear height when the same QTL model was fitted to both traits. In contrast, the allelic effects of only four of the thirty nine height QTL were significantly correlated with days to anthesis. Comparable results were identified for days to silk.

Candidate	Distance	Median Effect (cm)	Significance (RMIP)
<b>Gibberellin-Regulated Protein 2</b>	47Kb upstream	-1.4	50
<b>Gibberellin-Receptor-Like Protein</b>	93Kb downstream	-1.2	24
<b>Gibberellin-Responsive-Like Protein</b>	78Kb upstream	-0.9	17
<b>Phytosulfokine Receptor Protein</b>	Intronic	1.2	80
<b>Brassinosteroid Synthesis Protein</b>	78Kb downstream	-1.3	17
<b>Brassinosteroid LRR Receptor Kinase</b>	0.413Kb upstream	1.8	8

#### *Genomic prediction by ridge regression BLUP*

Recently, a substantial proportion of the heritable variation in human height (45%), unidentified by previous single marker GWAS, was reportedly captured in a mixed linear model framework utilizing all marker data to define genetic relatedness among individuals (Yang et al., 2010) instead of defining significant polymorphisms to construct a multiple regression which was found to capture less than 5% of the heritable phenotypic variance. Employing the same ridge regression modeling framework, we captured approximately 77% of the heritable variation in plant height noted within the NAM panel from polymorphisms scored in the first generation maize HapMap (Gore et al., 2009) (Figure 3.7). This was comparable to the proportion of heritable height variation captured in a QTL model selection framework. Despite a similar capacity to capture heritable height variation through both QTL model selection and regression methods, the manner in which variation was partitioned across the genome differed substantially

between the approaches. While the AMES panel captures more genetic diversity at lower minor allele frequencies than the NAM panel, ridge regression in the AMES panel was found to capture a comparable 82% of heritable height variation from the polymorphisms in both panels.



**Figure 3.7 Genomic prediction of height in NAM and AMES panels**  
Comparable prediction accuracy was attained from both QTL model selection approaches and ridge regression models. Both NAM and AMES populations possessed similar prediction accuracies based on HapMap 1 polymorphisms.

## DISCUSSION

Height is both one of the most heritable and most complex of all maize traits. In spite of this complexity, we sought to dissect the natural phenotypic variation in maize height and genetically correlated traits and partition it into components of genetic, environmental, and environmentally conditional genetic variance. Following the characterization of heritable height variation, this genetic variation was further partitioned across the maize genome and the polygenicity as well as the pleiotropy of QTL mapped for height and related traits including node counts and flowering time were assessed to discern their independence at the genomic level. Cloned height loci and candidate genes already implicated in well-established height related hormonal pathways such as the auxin, brassinosteroid, and gibberellin pathway were compared to the joint-linkage mapped QTL as well as those significant polymorphisms identified in joint-linkage assisted GWAS to discern if previously identified loci adequately capture the natural heritable height variation existing in the NAM panel. Two unique regression approaches were

used to differentially model the distribution and size of allele effects across the maize genome: a model selection approach seeking to capture the most heritable variation and attribute it to the fewest QTL, and a ridge regression model wherein all polymorphisms were assumed to possess a marginal influence on maize height variation.

#### *Proportion of heritable height variation*

As in node count and flowering time measurements, most height variation in both the NAM and AMES panels was explainable. Heritable diversity in all surveyed traits captured the most substantial proportion of phenotypic variation. To overcome the confounding effects of photoperiod (Coles et al., 2010) in determining both plant height and flowering time, only temperate field environments were included in this study. Given the termination of apical growth upon flowering, inclusion of both tropical and temperate environments may have greatly increased the proportion of height and flowering time variation attributable to environmental effects and affected the capture of environmentally conditional genetic variation. Conversion of flowering time to growing degree days and thus controlling for temperature differences among fields did not appear to influence proportions of variation attributable to environment or the relationships between flowering time and plant height. However, similar environmental variables likely influence estimates of heritability and the proportion of variance reported for each trait. In the instance of plant height measurements this may be augmented as a substantially larger proportion of variation was attributable to environmentally conditional genetic variation than that observed in other traits, most notable both node count measures.

In addition to environments of study, the heritability of each trait is strongly influenced by the allelic composition of the population in which it is phenotyped. Substantially variance in estimates of heritability was observed between both NAM families and the AMES inbred panel for height and related traits. Interestingly, correlations between the heritable variance of the traits across the NAM families were not well paralleled by the covariance of the traits across all the NAM RILs. Given the significant proportion of heritable variation in these traits captured by

differences between NAM families, this was unexpected. A shared reduction in genetic variance of two traits such as total height and flowering time across the 25 NAM families was not found to proportionally reduce their correlation across all the NAM RILs. While correlations of heritable variation between total height and node counts across NAM families were increased compared to flowering time, they were reduced across all the NAM RILs. Similarly, correlations between traits within each of the NAM families were found to significantly differ. This was further evidenced by variation in the NAM family's predicted multi-trait responses to selection. Further analyses of the variation in these responses could not be explained by estimates of kinship among the lines indicating the total genetic relatedness was not a powerful indicator of multi-trait response to selection.

#### *Distribution of heritable height variation across the maize genome*

Partitioning the heritable height variation across the maize genome revealed the substantial effect of modeling method in attributing variation to polymorphisms. To characterize the genetic architecture of maize height the employed stepwise NAM family nested QTL selection approach estimated a minimal number of QTL which could capture the most variation in height. This approach revealed substantially less variation in total height and primary ear height existed between NAM families than that observed for both flowering time and node count measures. Moreover, the proportion of heritable variation captured by the largest effect QTL while small in all the complex traits analyzed was smallest for plant height at only approximately 2.1% of the heritable variation. Moreover, the distribution of variation per QTL was not substantially more uniform for height than that observed for the other surveyed traits. These factors led the proportion of total heritable height variation captured by these models to be reduced relative to the other complex traits and suggested an even more polygenic pattern of inheritance.

Given the polygenicity of height, ridge regression was employed to assess the ability of all polymorphisms genotyped across both NAM families and the AMES inbred panel to capture

height variation. In contrast to model selection approaches, polymorphisms were not nested within each of the NAM families, and no term was fitted for the proportion of variation captured between families. This ridge regression approach sought to capture more variation in height and reduced the probable overestimation of allelic effects or Beavis effect (Beavis, 1994). While allele effect estimates were substantially smaller than that observed by model selection approaches, the total proportion of heritable variation captured by both methods was comparable and approximated 77% of the total heritable height variation. Although no multiple regression was performed in the AMES panel, ridge regression was employed and captured a comparable 82% of the heritable phenotypic variation.

Differences in genetic architecture estimates from both methods were substantial; however, given the number of polymorphisms scored relative to the number of genotypes on which we possess phenotypic data, we are left with an ill-posed problem or a lack of degrees of freedom to accurately estimate QTL effects. With insufficient degrees of freedom, we have an unconstrained solution space with an infinite number of potential QTL allele effect estimates that are equally valid from a numerical but perhaps not biological perspective. In order to discern the most biologically appropriate model, additional information or assumptions beyond phenotypes and genotypes is needed to constrain the solution space of possible effect estimates. The QTL model selection approach sought to do so by invoking Occam's razor and assuming the minimal number of QTL capturing the most height variation was the most accurate model of genetic architecture. Ridge regression sought to constrain the solution space by limiting maximum effect sizes and assuming all QTL effects must be shrunken equally to 0. While numerous other methods have been successfully applied to genomic prediction (Jannink et al., 2010), all seek to either limit the number of effects or shrink their squared or absolute effect size either equally across all predicted QTL or differentially as in several hierarchical Bayesian approaches based on repeated sampling of the probability distributions. Although many methods may predict phenotypes and capture heritable variation, the most biologically accurate model of genetic architecture often remains to be determined.

### *Pleiotropy with genetically correlated traits*

Pleiotropy remains an aspect of genetic architecture and is critical to predicting the independence of evolvability among traits in a selection regime. The design of the NAM panel provides a unique opportunity to characterize the pleiotropy of QTL through correlation of the allelic effects of a locus across the 25 NAM families. Using this approach we identified substantially reduced pleiotropy between both measures of flowering time and measures of total height and ear height than expected by comparison to their genetic correlations of ( $r = 0.58 - 0.62$ ). Upon further review we found a substantial correlation ( $r = 0.51$ ) between NAM families for days to anthesis and total height indicating that while the mapped QTL variation within each NAM family were not shared between these traits, heritable height variation captured between NAM families was significantly pleiotropic with flowering time measurements. This suggests larger effect loci may be independently evolvable for measures of flowering time and plant height; however, numerous small effect loci may ensure these traits continue to coevolve.

### *GWAS and co-localization of candidate loci*

Joint-linkage assisted GWAS of total height and ear height across the NAM families revealed a substantial number (345 and 351 respectively) of significantly associated ( $RMIP > 5$ ) polymorphisms across the entire maize genome. Many of these were identified as co-localizing with those QTL mapped during joint-linkage analysis; however, no associations were identified near previously cloned height loci such as *Anther ear*, *Brachytic 1, 2, 3*, *Brevis plant 1, 2*, *Clumped Tassel 1, 2*, *Crinkly Leaves*, *Dwarf 1, 3, 8, 9, 10, 12*, *Etched N617*, *Lilliputian*, *Nana Plant 1, 2*, *Pygmy*, *Terminal ear*, or *Yellow dwarf 1, 2*. Similarly, no genes centrally involved in the auxin, gibberellin, and brassinosteroid pathways were in linkage disequilibrium with the GWAS associations. These results indicate the previously identified heritable height variation does not well explain natural heritable variation in height. Given most previously identified height variation resulted in severe stunting of plants, it is possible these large effect loci are primarily conserved and have already reached fixation in natural populations.

According to the Fisher's geometric model (Fisher, 1930) large effect loci are only beneficial if a population is far from its fitness maximum. The closer a population approaches its fitness maximum the smaller effects must be to become adaptive. Height, unlike many kernel traits (Brown et al., 2011), has not been under recent direct selection in most maize populations. As such, few large effect loci likely remain segregating within these populations and instead have been purged or have reached fixation. The remaining small effect loci influencing height variation likely exist in proximity to genes less central to those hormonal and biochemical pathways regulating height, or persist in transcriptional regulation sites which only weakly regulate these central genes. In agreement with this supposition, the only significant associations identified were located in linkage disequilibrium or close proximity to candidate genes only tangentially regulating or regulated by centrally established hormonal networks that have been previously related to maize height variation through extensive molecular work (Table 3.3).

## REFERENCES

- Atwell S., Huang Y.S., Vilhjalmsdottir B.J., Willems G., Horton M., Li Y., Meng D., Platt A., Tarone A.M., Hu T.T., Jiang R., Mulyati N.W., Zhang X., Amer M.A., Baxter I., Brachi B., Chory J., Dean C., Debieu M., de Meaux J., Ecker J.R., Faure N., Kniskern J.M., Jones J.D.G., Michael T., Nemri A., Roux F., Salt D.E., Tang C., Todesco M., Traw M.B., Weigel D., Marjoram P., Borevitz J.O., Bergelson J., Nordborg M. (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627-631.
- Benjamini Y., and Hochberg, Y. . (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57:289-300.
- Brachi B., Faure N., Horton M., Flahauw E., Vazquez A., Nordborg M., Bergelson J., Cuguen J., Roux F. Linkage and Association Mapping of *Arabidopsis thaliana* Flowering Time in Nature. *PLoS Genet* 6:e1000940.
- Bradbury P.J., Zhang Z., Kroon D.E., Casstevens T.M., Ramdoss Y., Buckler E.S. (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633-2635.
- Buckler E.S., Holland J.B., Bradbury P.J., Acharya C.B., Brown P.J., Browne C., Ersoz E., Flint-Garcia S., Garcia A., Glaubitz J.C., Goodman M.M., Harjes C., Guill K., Kroon D.E., Larsson S., Lepak N.K., Li H., Mitchell S.E., Pressoir G., Peiffer J.A., Rosas M.O., Rocheford T.R., Romay M.C., Romero S., Salvo S., Villeda H.S., Sofia da Silva H., Sun Q., Tian F., Upadyayula N., Ware D., Yates H., Yu J., Zhang Z., Kresovich S., McMullen M.D. (2009) The Genetic Architecture of Maize Flowering Time. *Science* 325:714-718.
- Cheverud J.M., Marroig G. (2007) Research Article Comparing covariance matrices: random skewers method compared to the common principal components model. *Genetics and Molecular Biology* 30:461-469.
- Chia J.-M. (2012) A Second Generation Maize HapMap. *Nature Genetics*.
- Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6:e19379.
- Endelman J.B. (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Gen.* 4:250-255.
- Gilmour A.R., Thompson R., Cullis B.R. (1995) Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics* 51:1440-1450.
- Gore M.A., Chia J.-M., Elshire R.J., Sun Q., Ersoz E.S., Hurwitz B.L., Peiffer J.A., McMullen M.D., Grills G.S., Ross-Ibarra J., Ware D.H., Buckler E.S. (2009) A First-Generation Haplotype Map of Maize. *Science* 326:1115-1117.
- Hamblin M.T., Warburton M.L., Buckler E.S. (2007) Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. *PLoS ONE* 2:e1367.
- Huang X., Wei X., Sang T., Zhao Q., Feng Q., Zhao Y., Li C., Zhu C., Lu T., Zhang Z., Li M., Fan D., Guo Y., Wang A., Wang L., Deng L., Li W., Lu Y., Weng Q., Liu K., Huang T., Zhou T., Jing Y., Li W., Lin Z., Buckler E.S., Qian Q., Zhang Q.-F., Li J., Han B. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961-967.
- Khush G.S. (2001) Green revolution: the way forward. *Nat Rev Genet* 2:815.
- Lander E.S. (1996) The New Genomics: Global Views of Biology. *Science* 274:536-539.
- Lee M., Sharopova N., Beavis W.D., Grant D., Katt M., Blair D., Hallauer A. (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. *Plant Molecular Biology* 48:453-461.

- Lin Y.R., Schertz K.F., Paterson A.H. (1995) Comparative Analysis of QTLs Affecting Plant Height and Maturity Across the Poaceae, in Reference to an Interspecific Sorghum Population. *Genetics* 141:391-411.
- Logsdon B., Hoffman G., Mezey J. (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* 11:58.
- McMullen M.D., Kresovich S., Villeda H.S., Bradbury P., Li H., Sun Q., Flint-Garcia S., Thornsberry J., Acharya C., Bottoms C., Brown P., Browne C., Eller M., Guill K., Harjes C., Kroon D., Lepak N., Mitchell S.E., Peterson B., Pressoir G., Romero S., Rosas M.O., Salvo S., Yates H., Hanson M., Jones E., Smith S., Glaubitz J.C., Goodman M., Ware D., Holland J.B., Buckler E.S. (2009) Genetic Properties of the Maize Nested Association Mapping Population. *Science* 325:737-740.
- Messina C.D., Podlich D., Dong Z., Samples M., Cooper M. (2011) Yield–trait performance landscapes: from theory to application in breeding maize for drought tolerance. *Journal of Experimental Botany* 62:855-868.
- Peng J., Richards D.E., Hartley N.M., Murphy G.P., Devos K.M., Flintham J.E., Beales J., Fish L.J., Worland A.J., Pelica F., Sudhakar D., Christou P., Snape J.W., Gale M.D., Harberd N.P. (1999) 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* 400:256-261.
- Piperno D.R., Ranere A.J., Holst I., Iriarte J., Dickau R. (2009) Starch grain and phytolith evidence for early ninth millennium B.P. maize from the Central Balsas River Valley, Mexico. *Proceedings of the National Academy of Sciences* 106:5019-5024.
- R D.C.T. (2011) R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.
- Remington D.L., Thornsberry J.M., Matsuoka Y., Wilson L.M., Whitt S.R., Doebley J., Kresovich S., Goodman M., Iv E.S.B. (2001) Structure of Linkage Disequilibrium and Phenotypic Associations in the Maize Genome. *Proceedings of the National Academy of Sciences of the United States of America* 98:11479-11484.
- Risch N., Merikangas K. (1996) The Future of Genetic Studies of Complex Human Diseases. *Science* 273:1516-1517.
- Romay M.C. (2012) Ames Inbred Diversity Resource.
- Salas Fernandez M.G., Becraft P.W., Yin Y., Lübberstedt T. (2009) From dwarves to giants? Plant height manipulation for biomass yield. *Trends in plant science* 14:454-461.
- SAS. (2002-2004) SAS 9.1.3, Cary, NC: SAS Institute Inc.
- Sasaki A., Ashikari M., Ueguchi-Tanaka M., Itoh H., Nishimura A., Swapan D., Ishiyama K., Saito T., Kobayashi M., Khush G.S., Kitano H., Matsuoka M. (2002) Green revolution: A mutant gibberellin-synthesis gene in rice. *Nature* 416:701-702.
- Stone E.A., Ayroles J.F. (2009) Modulated Modularity Clustering as an Exploratory Tool for Functional Genomic Inference. *PLoS Genet* 5:e1000479.
- Tian F., Bradbury P.J., Brown P.J., Hung H., Sun Q., Flint-Garcia S., Rocheford T.R., McMullen M.D., Holland J.B., Buckler E.S. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* advance online publication.
- Tian F., Bradbury P.J., Brown P.J., Hung H., Sun Q., Flint-Garcia S., Rocheford T.R., McMullen M.D., Holland J.B., Buckler E.S. (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159-162.
- Valdar W., Holmes C.C., Mott R., Flint J. (2009) Mapping in Structured Populations by Resample Model Averaging. *Genetics* 182:1263-1277.
- Wagner G.P., Zhang J. (2011) The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat Rev Genet* 12:204-213.
- Winkler R.G., Freeling M. (1994) Physiological genetics of the dominant gibberellin-nonresponsive maize dwarfs. *Planta* 193:341-348.

- Wisser R.J., Kolkman J.M., Patzoldt M.E., Holland J.B., Yu J., Krakowsky M., Nelson R.J., Balint-Kurti P.J. (2011) Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a GST gene. *Proceedings of the National Academy of Sciences* 108:7339-7344.
- Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E., Visscher P.M. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42:565-569.
- Yu J., Pressoir G., Briggs W.H., Vroh Bi I., Yamasaki M., Doebley J.F., McMullen M.D., Gaut B.S., Nielsen D.M., Holland J.B., Kresovich S., Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203-208.