Project Description: The genomic basis for dysregulation of protein abundance in maize

Research and Training Plan

Understanding, predicting, and modifying phenotypes are among the fundamental aims of biological science. An organism's proteomic state is directly responsible for phenotype, and can be considered a phenotype itself. The central dogma of molecular biology, which characterizes the flow of information from DNA, to RNA, to protein, implies that protein abundance can be inferred from transcript abundance[1]. Though a useful model, the elegance of the central dogma belies the complexity of protein synthesis and degradation. The correlation between transcript abundance and protein abundance can be modest and variable[2–7], indicating that protein abundance is likely dependent on other genomic, cellular, temporal, or environmental influences. Genomic influences include DNA variants associated with differential protein abundance between individuals (pQTL)[5,8–11]. There is some evidence that protein abundance is buffered against changes in transcript abundance[9], and the apparent relationship between transcript and protein abundance is likely also complicated by differential degradation rates, differential transport between tissues, and temporal cycling[2]. Evidence of pQTL that control protein abundance independently of transcript abundance[5], combined with a well established pattern between rare variants and expression dysregulation[12,13], lead us to hypothesize that **rare or deleterious variants are predictive of protein abundance dysregulation**. We propose to model genome-wide patterns of how rare or deleterious variants cause changes in protein structure and function that disrupt protein translation, accumulation, and degradation.

Experiments studying the proteomes of genetically diverse individuals have been conducted in several species[4,5,8,9,11], but no such study has yet been conducted in an agronomically important species and the single study in a plant (Arabidopsis)[9] utilized closely related individuals from a biparental cross. With the exception of one[5], all the above studies isolated proteins and mRNA from different samples, which is likely to have introduced latent error in comparisons between transcriptome and proteome. None of the above studies were designed to account for temporally delayed relationships between transcript and protein abundance. In this project description, we propose to utilize the first vertical integration of the central dogma - genomic, transcriptomic, and proteomic data - to develop a model characterizing genomic influence on protein accumulation in maize (*Zea mays*). Maize is well suited for this study - it is the most-produced crop in the world[14] and an essential source of calories for humans and livestock. As such, research findings in maize can have immediate and far-reaching impact on global food production. Ideal for genetic studies, it has well-characterized inbred populations, exhibits enormous polymorphism and diversity[15], has rapid linkage disequilibrium decay, is replicable due to inbreeding tolerance, and has a well-established research community that has developed an abundance of genomic resources. In this study, we will characterize protein and transcript abundance in 27 diverse maize inbreds and investigate genome-wide dysregulation of post-transcriptional protein abundance.

Generating transcriptomic and proteomic data sets

Germplasm utilized in this study will be 27 maize inbred lines that are the parents of the maize nested association mapping (NAM) population[16]. These lines were chosen to encompass the genetic diversity of maize, representing temperate, tropical, flint, and popcorn materials, as well as historically important lines[16]. They have been genotyped at more than 80 million single nucleotide polymorphism (SNP) and insertion/deletion (indel) variants[17], and extensive phenotypic data are available for both the parents and

5,000 of their inbred progeny[18–24]. We will grow randomized replicates of each line in growth chambers at Cornell University and collect seedling vegetative tissue at five time points during the three-leaf stage. We will extract RNA and protein from the latest-sampled tissue, and RNA alone from tissues sampled 6, 12, 18, and 24 hours earlier. We chose to sample vegetative tissue from three-leaf stage seedlings to ensure uniform developmental age and because data from previous proteomic[3] and transcriptomic[25,26] studies in maize identified a large number of unique transcripts and proteins in leaf and seedling tissue relative to other tissues. By extracting proteins and RNA from the same final tissue samples, we will reduce latent variability that complicates comparisons of transcript and protein generated by different experiments, as has been the case in previous studies[3,9,11]. 3' RNA sequencing will be performed by the Cornell Genomic Diversity Facility, using the Lexogen Quantseq FWD kit. Protein quantification will be performed using a high-throughput tandem mass spectroscopy protocol developed by the Vierstra Lab at Washington University[27] on their in-house Thermo-Fisher Q-Exactive Plus mass spectrometer connected online to a Dionex Ultimate 3000 HPLC equipped with an Acclaim PepMap RSLC C18 reverse-phase column. High quality *de novo* genome assemblies of the NAM parents will be publicly available soon (R. Kelly Dawe and M. Hufford, personal communication), which will allow inbred-specific alignment of RNAseq reads and prediction of peptides. Variability in presence and location of genes between the 27 genome assemblies will complicate comparisons between these inbred lines. To alleviate this issue for downstream analyses we will use the Practical Haplotype Graph (PHG), developed by our group and approaching public release. The PHG is constructed from nodes that represent haplotypes, and the genic space of an individual is specified as a path through the PHG. This facilitates comparison of genes across the 27 reference genomes.

Initial assessments of proteomic diversity in the assayed lines will be conducted visually by principal component analysis (PCA) or multidimensional scaling (MDS). PCA performed using genotypic and transcriptomic data separately show that the NAM parents are diverse by both measures, and that population structure is globally similar but locally variable depending on the type of data used (**Figure 1**). We expect that the PCA of the NAM parents' proteomes will preserve global patterns of similarity that reflect underlying population structure, but will be characterized by local variability. Due to buffering between genotypic variants and ultimate phenotype[9,28], we expect that there will be decreasing variability among the NAM parents in the genotypic, transcriptomic, and proteomic PCAs.

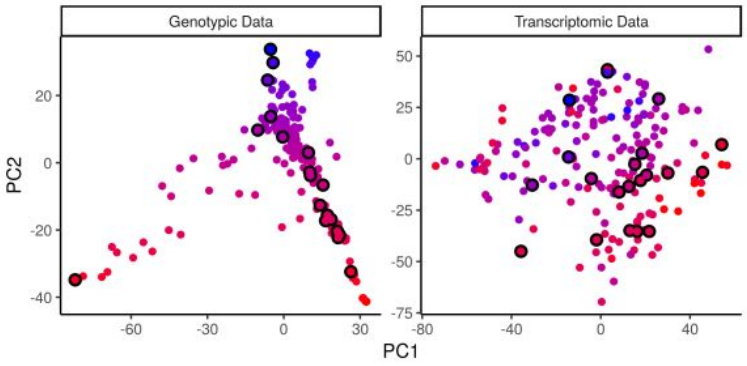

Figure 1: Population structure of 282 maize inbred lines, visualized by PCA on genotypic data (left) or expression data from mature leaves (right). NAM parents, indicated by black outlines, are representative of the present diversity regardless of data type. Coloring in both plots is according to the second genotypic PC, demonstrating consistent global population structure regardless of data type.

Characterizing the impact of deleterious mutations on protein translation and abundance

According to the central dogma, in the absence of genetic, cellular, or environmental factors transcript and protein abundances should be highly correlated. However, estimates of the correlation between transcript abundance and protein abundance are 0.4-0.5 in maize[2]. Evidence of pQTL in other species[4,9,11]

includes many that appear to act on protein abundance independently of transcript abundance[5]; that is, many pQTL effects take place after transcription. This indicates that the variable relationship between transcript and protein abundances is partially due to genetic effects. Both translation and accumulation can be impacted by protein structure, and the effects of mutations in the protein coding sequence are often additive such that individuals with a greater number of disruptive mutations will exhibit greater changes in protein structure or function[29]. Most newly occurring mutations are harmful, or deleterious. Deleterious mutations are abundant and ubiquitous[30,31], yet persist in populations at low frequencies[32]. These rare variants are quite predictive of transcript abundance dysregulation[12,13] and have large effects on phenotype relative to common variants[33,34]. A large portion of phenotypic variability attributable to rare variants comes from those in the coding regions and untranslated regions (UTRs) of genes[35]. Based on these patterns, **we hypothesize that rare or deleterious variants are predictive of protein abundance dysregulation**. We propose a model in which rare or deleterious alleles in the 5' and 3' UTRs, as well as the coding regions, of genes cause dysregulation of protein abundance by affecting translation, transport, and degradation via changes to protein structure and function. We will investigate rare and deleterious alleles for their impacts on post-transcriptional protein accumulation by comparing protein abundance to transcript abundance at five different time points. We will subsequently characterize the rare variants that cause protein abundance dysregulation for their effects on protein structure and function. To study the contribution of rare variants to post-transcriptional protein dysregulation, we will first quantify deviations in protein abundance from what would be expected given transcript abundance alone. We can then test for a relationship between these deviations and the number of nearby rare or deleterious variants.

To adequately assess post-transcriptional effects on protein abundance, it will be necessary to first appropriately control for transcript abundance. This study will improve on previous experiments by 1) extracting mRNA and protein from individuals grown as part of the same experiment and 2) by sampling mRNA at multiple time points leading up to protein sampling. A previous study compared the maize transcriptome and proteome during development, revealing a modest correlation of 0.4-0.5 between between transcript and protein abundance[2]. In addition, many genes exhibited high protein abundance but little to no detectable mRNA[2]. The authors hypothesized, and found supporting evidence, that this could be due to 1) relatively consistent protein levels while mRNA cycles diurnally, 2) mRNA being degraded faster than protein, or 3) protein being produced elsewhere and subsequently transported into the sampled tissue[2]. Diurnal cycling of gene expression in plants is well documented, and studies in maize[36] and Arabidopsis[37] both estimate that more than 1 in 10 genes cycle diurnally. Because such a large proportion of genes are diurnally expressed, sampling RNA at multiple time points over 24 hours is likely to improve correlations between transcript and protein abundance for a large number of genes. By sampling transcript abundance across time, as well as obtaining RNA and protein from the same experiment, we anticipate a higher global correlation between transcript and protein abundance than previously reported in maize. Once the relationship between transcript and protein abundance is established for each gene, we can proceed to developing the model describing genomic impacts on post-transcriptional protein abundance.

For a given gene, we will have protein and transcript abundance measurements for each of the 27 unique NAM parent lines at one time point (T0), and transcript abundance for each NAM parent at four previous time points (6, 12, 18, and 24 hours before T0). To estimate each parent's deviation from expected protein levels, we will compute externally studentized residuals from the regression of protein abundance against transcript abundance at all five time points. Conceptually, externally studentized

residuals are obtained by leaving out one parent and fitting a line between transcript and protein abundance in the remaining 26 parents (**Figure 2**). The difference between the held-aside parent's true protein level and its protein level predicted from the fit line is divided by the standard error of the residuals. This is done for all 27 lines independently and repeated across all genes. The resulting externally studentized residuals are expected to follow a T distribution, simplifying parametric analysis of the deviations; we will also consider nonparametric, rank-based methods for comparing protein abundance deviations to the presence of rare or deleterious variants.

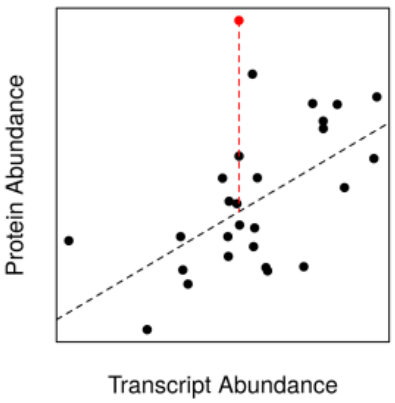

Figure 2: Demonstration of studentized residual calculation. The black regression line was fit to 26 (black) points, and the residual calculated for the 27th (red) point is represented by the dashed red line. Data were simulated such that the correlation of all points is 0.5, to reflect expectation based on previous studies.

SNPs and indels will be classified independently as rare and/or deleterious, resulting in two sets of variants for analysis. Both SNPs and indels will be considered when determining rare variants, which will be based on their prevalence within a panel of 1218 inbred maize lines that have been genotyped at over 80 million sites[17]. In addition to identifying rare variants we will also classify SNPs as deleterious or not using genomic evolutionary rate profiling (GERP) scores[38]. GERP identifies constrained regions where sequence is conserved across related species, reflecting purifying selection across evolutionary time. Variants within constrained regions are putatively deleterious. Using regression or non-parametric approaches, the number of rare/deleterious variants in each individual, at each gene, will be compared to the externally studentized residuals for protein abundance for the same individuals and genes.

Whether or not we observe global trends for the relationship between rare/deleterious variants and protein abundance deviation, we will further investigate genes that have strong local signal. Genetic variants can alter protein abundance after transcription either by regulating translation or by post-translation effects on, for example, protein stability, degradation, or transport. Translation of mRNA to protein can be regulated by both the 5' the 3' UTRs[39,40]. Variants in the coding sequence can affect translation[40] as well as protein secondary and tertiary structure[41]. Whereas transcription can be dysregulated by variants in regulatory regions upstream of genes[12], we expect that variants with post-transcriptional effects on protein abundance will be enriched in UTRs and exons of the affected genes.

In addition to studying the intragenic position of variants that affect post-transcriptional protein abundance, novel methods for predicting changes in protein structure will be leveraged to study the effects of influential variants on protein structure and behavior. Additive effects of multiple UTR or coding-sequence variants may have increasingly severe consequences[29] that manifest as changes in structure or folding and subsequently affect translation[42], transport[43], or degradation[44]. We will use machine learning methods such as rawMSA[45] and VSL2B[46], which can accurately predict protein secondary structure, solvent accessibility, contact between residues, and intrinsic disorder, to identify the effects of influential variants on protein structural characteristics.

The primary objective of this study is to develop models describing the biological mechanisms of protein abundance dysregulation. A logical next step is to use those biological models to enable improvements in plant breeding. Proteomic profiling is currently not scalable to the level of breeding programs, but genotyping is. The biological models developed here can subsequently be used to inform

machine learning methods for predicting protein abundance between diverse inbred lines from genotypic data, particularly genetic variants with known effects on protein structure and function. If successful, such models could make proteomics scalable to the level needed for application in plant breeding programs.

Broader Impacts

By studying the model of protein abundance dysregulation described here, we will expand the current domain of knowledge regarding how protein accumulation is regulated on a genomewide scale. We expect that many of the findings of this study, though performed in maize, will be translate to other organisms. Data generated by this study will be a valuable resource for the maize genetics community. For example, parts of our proposed analysis are dependent on high-quality genome annotation of the NAM parents, and our proteomic data will be useful for refining the results of initial genome annotations.

The broader impacts of this study will not be limited to distribution of data and findings; I also intend to train and mentor graduate and undergraduate students during my tenure as an NSF-PRFB fellow. I became a scientist by a circuitous route, which instilled in me an appreciation for patient and open-minded mentors. I hope to provide similar mentorship to students in the coming years and throughout my career. During this fellowship, I will begin a regular tradition of participating in Skype a Scientist (www.skypeascientist.com), which pairs scientists with classrooms of children for question and answer discourse. This program makes conversation with a scientist possible for all classrooms, increasing accessibility for children who may not otherwise be able to visit museums or meet scientists. Both mentorship and Skype a Scientist help make science friendly and accessible, which is essential to enabling involvement in STEM for all who are interested.

Training Objectives and Career Development

During my M.S. in Biometry and Ph.D. in Plant Breeding and Plant Genetics at the University of Wisconsin - Madison, my research resulted in publications on the topics of phenotyping, phenotypic stability, selection, and genetic mapping. My background in statistics and quantitative analysis will be a strong foundation for success in the project described here, which is orthogonal to my previous research with respect to studying proteomics and translational regulation. This project will supply me with experience in bioinformatics, molecular genetics, proteomics, transcriptomics, application of machine learning, and functional genomics. Visits to Washington University and mentorship from Dr. Richard Vierstra will allow me to learn new molecular biology and proteomic techniques, and expand my professional network. Engaging in these new fields will provide me with the opportunity to expand my professional network, while continuing to strengthen relationships within the quantitative genetics community. Being selected as an NSF-PRFB and participating in the program's annual meetings will provide exposure to current research and facilitate networking with the next generation of biological researchers.

Results from these studies will be disseminated in the form of posters and oral presentations at meetings with broad attendance such as the Maize Genetics Conference, the Plant and Animal Genome Conference, and the Gordon Research Conference on Proteins. The ultimate results from these studies will be published in peer-reviewed journals. Presenting and publishing my findings will provide me with valuable experience communicating results to a broad audience.

Though I am keeping an open mind to the possibility of pursuing many different career paths, I have wanted to become a tenure-track professor at an R1 or R2 university since I was young. This NSF-PRFB

will give me experience with managing a budget and directing my own project, while expanding my knowledge and skill set into new disciplines. While at Cornell, I will have the opportunity to improve my leadership through participation in the Cornell Postdoctoral Leadership Program, as well as engage in numerous other professional development and networking programs offered by the Cornell Office of Postdoctoral Studies and the School of Integrative Plant Science Postdoctoral Association. These challenges and opportunities will strengthen my ability to lead and operate a research group in the future.

Justification of Sponsoring Scientists and Host Institution

The Buckler lab at Cornell University has a long standing record of high quality quantitative genetics research in maize, and will provide a peerless environment in which to study proteomic diversity, receive mentorship, and develop as a scientist. Dr. Buckler was the principal investigator responsible for developing choosing NAM parents and developing the NAM[16]. Since then the Buckler lab has continued to be an innovative source of cutting-edge research in quantitative genetics, making contributions in mixed models[47,48], genetic architecture of complex traits[24], high throughput genotyping[49], computational software[50], and functional genomics[12,51,52]. Dr. Buckler has a long history of securing funding from a diverse array of sources and has extensive mentoring experience, having advised 16 graduate students and 30 postdocs, many of whom have gone on to lead their own research groups.

The Vierstra lab at Washington University has extensive experience with quantitative proteomics in Arabidopsis[53,54] and maize[27]. Dr. Vierstra has performed groundbreaking work studying protein recycling via autophagy[55] and ubiquitination[56], as well as light signalling[57,58]. Recent development of a high throughput, 2-hour proteome method[27] makes the Vierstra lab uniquely capable of proteome quantification on a large number of maize samples. Dr. Vierstra's expertise in molecular biology compliments my background in quantitative genetics and will add new dimension to my postdoctoral training.

Cornell University and Washington University both offer state-of-the-art core facilities, and the Vierstra lab is equipped with all in-house instrumentation necessary for protein quantification. Cornell, the primary host institution, provides excellent leadership and professional development opportunities through the Office of Postdoctoral Studies. The Buckler lab is associated with both the USDA-ARS, which has a branch on campus, and the School of Integrative Plant Science, which consists of more than 80 faculty members and has its own Postdoctoral Association. Being a part of such a large and diverse community of plant scientists affords extensive opportunities for intellectual interactions and collaboration.

Timetable

	Year 1	Year 2	Year 3
Plant growth; protein and mRNA extraction and quantification			
Characterize the impact of rare/deleterious variants on translation			
Prepare manuscript(s)			
Attend conferences and present results			