UNIVERSITY OF CALIFORNIA SAN DIEGO

Subcellular Localization and Prediction of Qualitative Expression of the Proteome of Sorghum and Maize

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Biology

by

Laura de Boer

Committee in charge:

Professor Steven P. Briggs, Chair
Professor Vineet Bafna
Professor Eric Bennett
Professor Alisa Huffaker
Professor Scott Rifkin
Professor Yunde Zhao

2019

The Dissertation of Laura de Boer is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

I would like to dedicate this dissertation to my husband, Patric de Boer, for his immutable love and support. He embodies the natural curiosity, optimism, and kindness all scientists hope to achieve.

I would also like to dedicate this dissertation to the undergraduate students who not only assisted in this work, but also provided inspiration every day. I am honored and humbled by the time, energy, and trust you have given me. Thank you for giving me the chance to relive the experience of learning research through you.

EPIGRAPH

We are in the right spot, somehow, like a breath

Entering a singer’s chest, that shapes itself

For the song that is to follow.

Alicia Ostriker, Excerpt from *Move*

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ATP: Adenosine Triphosphate

AUC: Area Under the Curve

Bd: Brachypodium distachyon

BP: Biological Process

CC: Cellular Compartment

Chloro: Chloroplast

DTT: Dithiothreitol

DNA: Deoxyribonucleic acid

EPC: Express-ability Protein Classifier

ER: Endoplasmic Reticulum

ERC: Express-ability RNA Classifier

FDR: False Discovery Rate

FGS: filtered gene set

Glyoxy: Glyoxysome

GO: Gene Ontology

HDAC: Histone de-acetylase

HPLC: high-performance liquid chromatography

IMP: Inosine monophosphate

Mito: Mitochondria

MS: Mass spectrometry or mass spectrometer

PEG: Polyethylene glycol

PM: Plasma Membrane

PMSF: Phenylmethylsulfonyl fluoride

PR: Precision/Recall

PTS1: Peroxisomal Targeting Signal1

PTS2: Peroxisomal Targeting Signal2

RNA: Ribonucleic acid

RPKM: Reads per kilobase measured

ROC: Receiver Operating Characteristic

Sb: Sorghum bicolor

Sv: Setaria viridis

TMT: Tandem Mass Tags

WGS: Working Gene Set

Zm: Zea mays

LIST OF SUPPLEMENTAL FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor Dr. Steven Briggs, for his support and guidance. I additionally thank all Briggs lab members with whom I have worked over these last six years, especially Dr. Zhouxin Shen, Ann Tong, and Dr. Ryan Sartor. I would also like to thank Briggs lab undergraduate volunteers Bonny Pham, Kevin Celniker, Riran Kawakami, Serena Zhang, and Martin Galvan for their assistance. Lastly, I would like to thank the remainder of my thesis committee: Dr. Yunde Zhao, Dr. Eric Bennett, Dr. Alisa Huffaker, Dr. Scott Rifkin, and Dr. Vineet Bafna, for providing excellent leadership and scientific advice which has shaped the following work.

I would like to acknowledge the Cellular and Molecular Genetics Training Grant, and Chancellor’s Research Excellence Scholarship for funding and support.

Chapter 1, in full, is currently being prepared for submission for publication of the material. de Boer, Laura; Shen, Zhouxin; Putarjunan, Aarthi; Paschold, Anja; Dorchak, Alexandria; Helzer, Kyle T.; Sartor, Ryan C.; Facette, Michelle; Hochholdinger, Frank; Olsen, Laura J.; Rodermel, Steven; Smith, Laurie G.; Briggs, Steven P. “Direct Evidence for the Sub-cellular Localization of Proteins Encoded by 3,378 Genes of Maize.” *In preparation.* The dissertation author was the primary investigator and author of this material.

Chapter 2, in full, is currently being prepared for submission for publication of the material. de Boer, Laura; Sartor, Ryan C.; Shen, Zhouxin; Noshay, Jaclyn; Springer, Nathan M.; Schmelz, Eric A.; Huffaker, Alisa; Schnable, James; Briggs, Steven P. “Discovery of species-specific expressible genes via machine learning with omic data.”

In preparation. The dissertation author was the primary investigator and first author of this material.

Chapter 3, in part, is currently being prepared for submission for publication of the material. de Boer, Laura; Sartor, Ryan C.; Shen, Zhouxin; Noshay, Jaclyn; Springer, Nathan M.; Schmelz, Eric A.; Huffaker, Alisa; Schnable, James; Briggs, Steven P. “Discovery of species-specific expressible genes via machine learning with omic data.” *In preparation.* The dissertation author was the primary investigator and first author of this material.

.

VITA

2012	Bachelor of Science, University of Massachusetts, Amherst
2019	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Duan, Q., Kita, D., Johnson, E.A., Aggarwal, M., Gates, L., Wu, H.-M., and Cheung, A.Y. (2014). Reactive oxygen species mediate pollen tube rupture to release sperm for fertilization in Arabidopsis. Nature Communications 5:3129

Li, C., Yeh, F.-L., Cheung, A.Y., Duan, Q., Kita, D., Liu, M.-C., Maman, J., Luu, E.J., Wu, B.W., Gates, L., Jalal, M., Kwong, A., Carpenter, H., Wu, H.M. (2015). Glycosylphosphatidylinositol-anchored proteins as chaperones and co-receptors for FERONIA receptor kinase signaling in Arabidopsis. ELife 4:e06587

ABSTRACT OF THE DISSERTATION

Subcellular Localization and Prediction of Qualitative Expression of the Proteome of Sorghum and Maize

by

Laura de Boer

Doctor of Philosophy in Biology

University of California San Diego, 2019

Professor Steven P. Briggs, Chair

Study of the regulation and location of the protein products of genes is essential for understanding the phenotype of the organism. The quantitative and qualitative control of protein production, as well as post-translational modification and subcellular localization of the protein, in part determine the effect of a gene on a biological system.

As it becomes more apparent that complex traits are controlled by effects from many loci, it has become more imperative that we seek a proteome-wide understanding of protein regulation and localization. The proteomes of four maize subcellular organelles were characterized by comparison to their source tissues, defining both organelle-enriched and depleted protein sets. High confidence localizations to organelle were obtained for plasma membranes (2154), mitochondria (1079), glyoxysomes (461), and plastids (539). Many cases of localization were novel or revised existing annotations, including that of nearly 40% of localized maize classical genes. Many proteins localized to multiple compartments, including a large overlap between mitochondrial and chloroplast proteins, whereas few proteins were shared between the chloroplast and non-mitochondrial organelles. A machine learning approach was used to identify the expressible gene sets of sorghum and gene set annotations from versions 2 and 4 of the maize genome. These gene sets were identified by a classifier trained using only DNA methylation data as model features. Synteny was leveraged to identify species-specific expressible genes, revealing enrichment of biotic and abiotic stress-associated genes in the species-specific pools. Expressible gene sets also provide evidence for express-ability of gene models absent from the maize version 2 annotated filtered gene set or the maize version 4 annotated gene set.

CHAPTER 1

Direct evidence for the sub-cellular locations of proteins encoded by 3,378 genes of maize

1.1 Introduction

Proteins are localized to subcellular compartments that enable them to carry out their cellular functions. A full understanding of protein functions and biochemical pathways depends on knowledge of their spatial distributions within the cell. While the subcellular locations of individual proteins have been determined empirically, high throughput methods are needed to assign protein subcellular locations on a proteome-wide scale. Protein subcellular locations can be predicted by homology-based methods or via prediction of transit peptides, but these methods have limitations. While they can predict true organelle proteins, some experimentally validated organellar proteins lack canonical transport peptides, and membrane proteins often use an alternate import mechanism into organelles (Mayerhofer, 2016; Reumann et al., 2009). It can be difficult to infer localization based on sequence homology to proteins of distant species, and inference can only be as complete as the annotation of the comparator species. Both predictive approaches are biased against discovery of novel or species-specific organellar proteins. Additionally, computational prediction of all putative organellar proteins does not provide information regarding changes in organelle contents in specific tissues or stages of development. Information about protein post-translational modifications that may affect protein activity, such as phosphorylation, also cannot currently be determined via computational prediction. Thus, computational prediction of the subcellular localization of plant proteins cannot fully replace experimental proteomics, especially in non-*Arabidopsis* species or for understudied organelles.

However, proteomic methods face many challenges, including difficulty purifying intact organelles, co-purification of organelles, challenges with identification of membrane proteins particularly via gel-based methods, and the presence of highly expressed enzymes that obscure detection of lower abundance proteins.

Experimental proteomics studies have identified proteins from the major organelles of *Arabidopsis thaliana* (see Hooper et al., 2017 for collective resource). Fewer studies have been performed in the food crop model organism maize. These include investigations of the mitochondria (Dahal et al., 2012; Hochholdinger et al., 2004; Wang et al., 2018), plasma membrane (Hopff et al., 2013; Voothuluru et al., 2016a; Zhang et al., 2013), and the chloroplast at various stages of development and from different cell types (Friso et al., 2010; Huang et al., 2013; Majeran et al., 2005, 2008). No proteomics studies have been published on the peroxisome from monocots or on any C4-photosynthetic plants. Membrane proteins are underrepresented in proteomic studies, particularly when proteins are separated by 2D-gel electrophoresis (Tan et al., 2008). Characterization of plant organelles can be significantly improved by more comprehensive identification of their membrane protein contents. Recent advances in mass spectrometry and protein sample preparation have enabled greater depth in proteomic profiling and have enhanced detection of low-solubility membrane proteins (Walley et al., 2016).

Given the essential role of both the peroxisome and mitochondria in photorespiration, it is necessary to characterize differences in the peroxisome and mitochondrial proteome between C3 and C4 plants to fully understand the C4 photosynthetic pathway. Glyoxysomes are specialized peroxisomes involved in lipid beta-oxidation; they are found in the cotyledons of germinating seeds. Glyoxysomes contain enzymes of the glyoxylate cycle, which bypass decarboxylation steps of the citric acid cycle to produce carbohydrates from mobilized lipids. No studies of this sub-functionalized peroxisome have previously been published from maize. The maize plasma membrane is under-characterized despite its key roles in transport and response to the environment. Etioplasts are non-pigmented plastids from dark grown tissues, which rapidly green in response to light. The maize leaf displays a gradient from photosynthetically immature at the leaf base to mature, C4 photosynthetic tissue at the leaf tip. The transition from photosynthetically inactive plastids to C4, photosynthetically mature chloroplasts is associated with corresponding changes in the chloroplast proteome. Further study of these maize organelles will provide a valuable resource for plant researchers.

We present here the proteome of four major maize subcellular organelles: plasma membrane, mitochondria, and chloroplast; as well as the only peroxisome proteome from a C4 species. Etioplasts from five stages of greening were profiled. Intact organelles were isolated from their source tissues by density-based centrifugation and total protein was extracted from both the purified organelles and their corresponding source tissues. Additionally, phosphopeptides were enriched and identified from the plasma membrane protein samples. Extracted peptides were subjected to tandem mass spectrometry with label-free quantification. Protein abundance was compared between source tissue and purified organelle samples to identify proteins enriched in organellar fractions. These include known organelle markers as well as novel proteins.

1.2 Results

1.2.1 Isolation and proteomics of organellar proteins

All organelle and source tissue samples were obtained from *Zea mays* inbred line B73. The maize plasma membrane-enriched fraction was obtained from a five-day old whole seedling extract via a dextran-PEG two-phase partitioning system (SI Fig. 1-6). Maize mitochondria were isolated from non-pollinated ears. Glyoxysomes were isolated from the scutella of seedlings three days after germination. Mitochondria and glyoxysomes were purified using density-based centrifugation through a percoll gradient. Chloroplasts from one month old eight-leaf stage plants were isolated using density-based centrifugation from the bottom, middle, and top of the eighth leaf (SI Fig. 1-7A-B). Greening etioplasts were isolated from dark grown 14-day old seedlings at 0, 2, 4, 12, and 24 hours of light exposure (SI Fig. 1-7C). Trypsin-digested protein extracts from purified organelles and source tissues were analyzed independently by mass spectrometry. Phosphopeptides were enriched from a subset of the plasma membrane peptide samples prior to analysis. Mass spectra were searched against the B73 Ref.Gen. V4 working gene set (Jiao et al., 2017), identifying 15582 unique proteins from 15033 gene models (Table S1). The proteomes of intact leaf and ear used as source tissues for the chloroplast and mitochondria samples were previously published with mass spectra searched against B73 Ref.Gen.V2 working gene set (Walley et al., 2016). Mass spectra from these ear and leaf samples, were re-searched against Ref.Gen.V4 along with the organelle, scutella, and germinating seedling samples. We identified 4095 phosphopeptides from 4035 gene models in the phosphopeptide-enriched plasma membrane samples (Table S2).

1.2.2 Identification of organelle-enriched and -depleted proteins

To distinguish organelle-localized proteins from proteins of contaminating organelles, we compared the spectral counts observed for a given protein in the organelle-enriched sample versus the intact source tissue using the single-tailed Fisher's exact test. The resultant p-values were adjusted using the Benjamini-Hochberg correction for multiple comparisons. A p-value cutoff of 0.05 was used to identify proteins for which significantly more spectral counts were observed in the organelle samples versus the source tissue from which organelles were derived. The converse process was repeated to identify proteins depleted from the organelle fraction. Additionally, the fold-enrichment of spectral counts observed from the organelle versus source tissue was also calculated. A fold-enrichment or fold-depletion threshold of two was chosen to further constrain the set of organelle enriched or depleted proteins. The plasma membrane was compared to germinating kernel; the glyoxysome was compared to the scutellum; and the mitochondria were compared to the immature ear. Chloroplast samples derived from the tip, middle, and bottom of the leaf were each compared separately to intact mature leaf (SI Fig. 1-7A). The organelle protein enrichment and depletion scores are available in Table S3. Volcano plots of enrichment scores showed a substantial portion of proteins detected in the organelle samples have significantly fewer spectral counts observed in the organellar fraction versus the source tissue (Fig. 1-1A-B; SI Fig. 1-8A-L).

1.2.3 Comparison of organellar proteins to gold standard proteins

One challenge of organelle isolation is the co-enrichment of contaminating organelles. Mitochondria, chloroplasts, and peroxisomes are frequently co-isolated. To assess the level of purity of the organelle-enriched protein samples, organellar proteins were compared to a manually curated list of maize proteins with previously published localizations based on direct observation (Table S4). While many false positive gold standards were detected in the organelle fractions, setting thresholds for enrichment decreased the false positive rate by many fold (Fig 1-1C). Volcano plots of enrichment and depletion p-value versus fold enrichment for all detected proteins across organelles show discrimination between false and true positives as fold enrichment increases and p-values decrease (Fig. 1-1A-B). This trend is consistent across all organelles (SI Fig. 1-8A-L). Increasing the fold-change threshold for significance beyond two reduced the true positive rate for the chloroplast protein sets, whereas decreasing increases the false positive rate of the glyoxysome proteins (SI Fig. 1-9A). The plasma membrane and chloroplast true positive rates were sensitive to decrease in the threshold for p-value from 0.05 to 0.01, whereas the false positive rates remained similar at both thresholds (SI Fig. 1-9B). Given the decrease in organelle protein set size at more stringent p-value and fold change thresholds, the threshold for significance of p-value less than 0.05 and greater than two-fold enrichment or depletion is optimal (SI Fig. 1-9C-D). We located 2154, 1079, and 461 proteins to the plasma membrane, mitochondria, and glyoxysome, respectively (Table 1-1). Across all chloroplast over-enriched sets, there were 539 unique proteins. Out of all three chloroplast types, chloroplasts from the leaf bottom had to most over-enriched proteins, 381, but had the fewest depleted proteins, 569. The chloroplasts from the leaf top had 301 enriched proteins and the most depleted proteins. The chloroplasts from the leaf middle had the fewest enriched proteins. From plasma membrane phosphopeptide enriched samples, 4095 unique phosphopeptides were identified, including 524 phosphopeptides from genes encoding plasma membrane localized proteins (Tables S2-S3). Among greening etioplast sample proteins, 539 were from genes encoding chloroplast localized proteins (Table S1 and Table S3).

Table 1-1: Organelle over-enriched and depleted proteins: Organelle over-enriched and depleted proteins. Number of genes for each observed protein from organelle-enriched samples compared to their background tissues, p-value < 0.05; greater than 2-fold enrichment. Chloroplast (All) represents genes from all unique proteins significantly enriched or depleted in any of the three chloroplast samples. We detected 524 phosphopeptides from genes which produced plasma membrane proteins.

Organelle	Enriched Proteins	Depleted Proteins
Plasma Membrane (phospho)	2154 (524)	1224
Glyoxysome	461	964
Mitochondria	1079	2219
Chloroplast (Bottom)	381	569
Chloroplast (Mid)	265	992
Chloroplast (Top)	301	1146
Chloroplast (All)	539	1206

Many known organelle “gold standard” proteins were enriched in the corresponding target organelle, including 90% of glyoxysome, 54% of mitochondria, and 93% of plasma membrane gold standards detected in the organelle or source tissue (Fig. 1-1D). Only one gold standard protein from the mitochondria and none from any other contaminating organelles were enriched in the chloroplast, whereas 42% of gold standard chloroplast proteins were enriched indicating that the chloroplast preparations were pure. Less than 9% of detected chloroplast gold standard proteins and only a single peroxisome gold standard was enriched in the mitochondria, indicating that these organelles were not co-enriched during mitochondrial isolation. Only a single gold standard from other possible contaminating organelles was enriched in the mitochondrial fraction. Plasma membrane-enriched proteins included low percentages of detected gold standard proteins from contaminating organelles, apart from

Figure 1-1: Comparison of Organelle Enriched Proteins to Manually Curated Gold Standard Proteins.

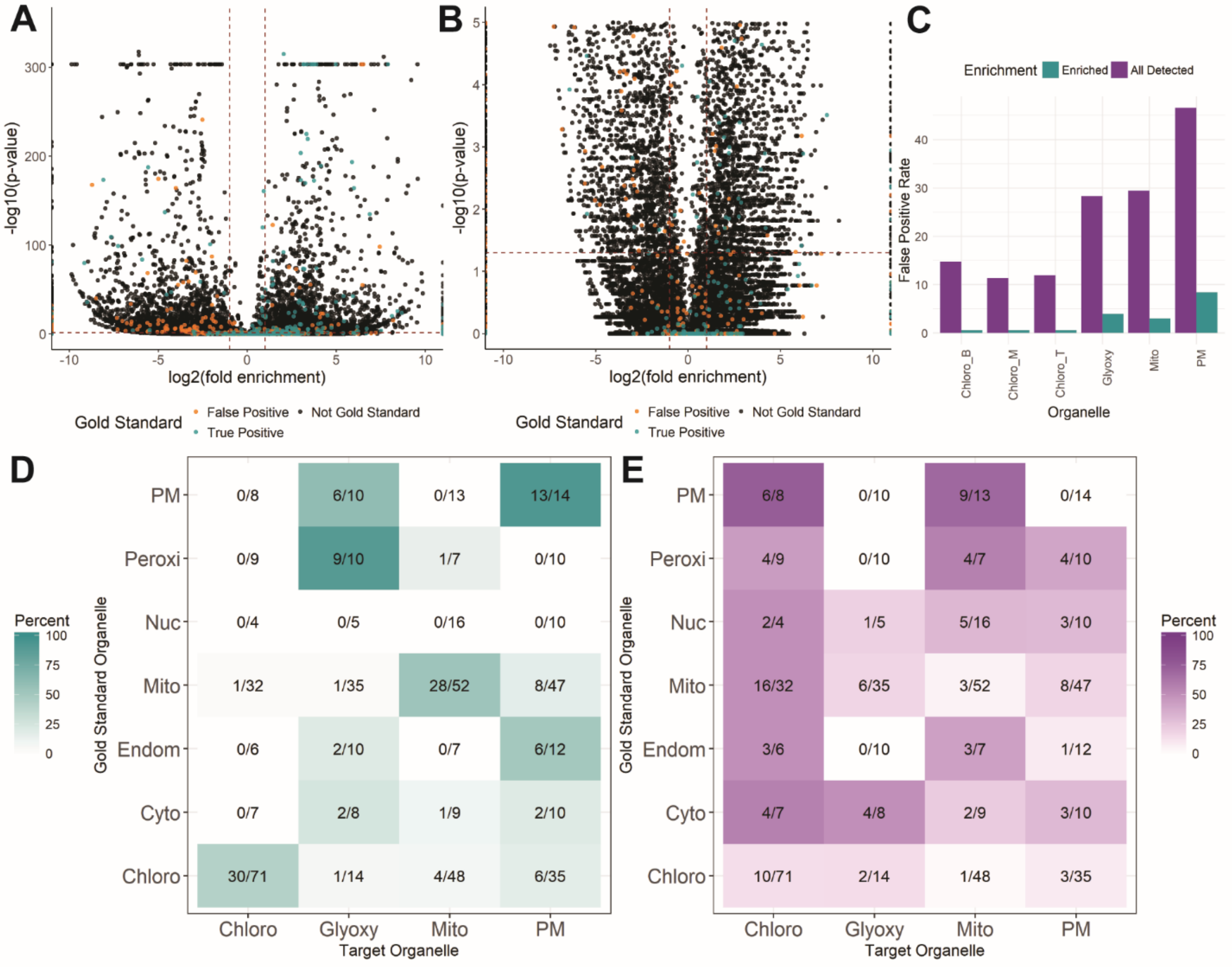


Figure 1-1: Comparison of Organelle Enriched Proteins to Manually Curated Gold Standard Proteins. A-B: volcano plots of proteins detected in combined organelle dataset. Fold enrichment represents proportion of normalized spectral counts for each protein observed in the organelle sample versus the source tissue. P-values derived from one-tailed fishers exact test comparing spectral counts observed in organelle versus source tissue for individual protein compare to spectral counts observed for all proteins. For proteins with negative fold-enrichment, depletion p-values were used; for proteins with positive fold-enrichment, enrichment p-values were used. Points colored based on gold standard true or false positives for target organelle. **A:** All detected proteins. **B:** subset of detected proteins with -log10(p-value) between 0 and 5. **C:** False positive rate of all organelle detected proteins versus all significantly enriched proteins for each organelle. False positive rate determined based on comparison to all detected (in organelle or source tissue) gold standards. **D-E:** Heatmap of percent of manually curated gold standard marker proteins localized to various organelles enriched (**D**) or depleted (**E**) in each target organelle fraction. Rows represent organelle fraction, with columns representing maize organelle gold standard proteins. Cell color fill represents percentage of detected gold standards over-enriched in each organelle. Cell label fractions represent number of enriched proteins over number of detected proteins.

endomembrane gold standards, for which 50% of detected proteins were enriched. The glyoxysome-enriched proteins included 60% of plasma membrane detected gold standard proteins. While the percentages of detected gold standard proteins from the endomembrane system enriched in the plasma membrane, and from the plasma membrane enriched in the glyoxysome, are high, proportionately few gold standards from other contaminating organelles were enriched in the target organelles. Few gold standard true positives were depleted from the target organelles (Fig. 1-1E).

1.2.4 Overlap of enriched and depleted proteins across target organelles

Six proteins were enriched in all four organelles (Fig. 1-2A). Comparatively few proteins enriched in the plasma membrane were enriched in the other organelles, with 82% uniquely enriched in the plasma membrane. The highest number of enriched proteins shared between the plasma membrane and another organelle was shared with the glyoxysome, representing 11.4% of plasma membrane enriched proteins and 53% of glyoxysome enriched proteins. A minority of proteins enriched in the glyoxysome were also enriched in the chloroplast or mitochondria. There was a large overlap in mitochondria and chloroplast enriched proteins, representing 30% of chloroplast proteins and 15% of mitochondrial proteins. Few proteins were shared between the chloroplast and other, non-mitochondria organelles. The plasma membrane, glyoxysome, mitochondria and chloroplast depleted proteins were 30%, 51%, 9% and 24% unique, respectively (Fig. 1-2B). Between leaf bottom, middle, and top-derived chloroplast samples, 27% of proteins were over-enriched in chloroplasts from all three tissues (Fig. 1-2C). The chloroplasts from the middle leaf shared more enriched proteins with the bottom than with the top derived chloroplasts. Few chloroplast depleted

Figure 1-2: Maize proteins with localization in multiple organelles.

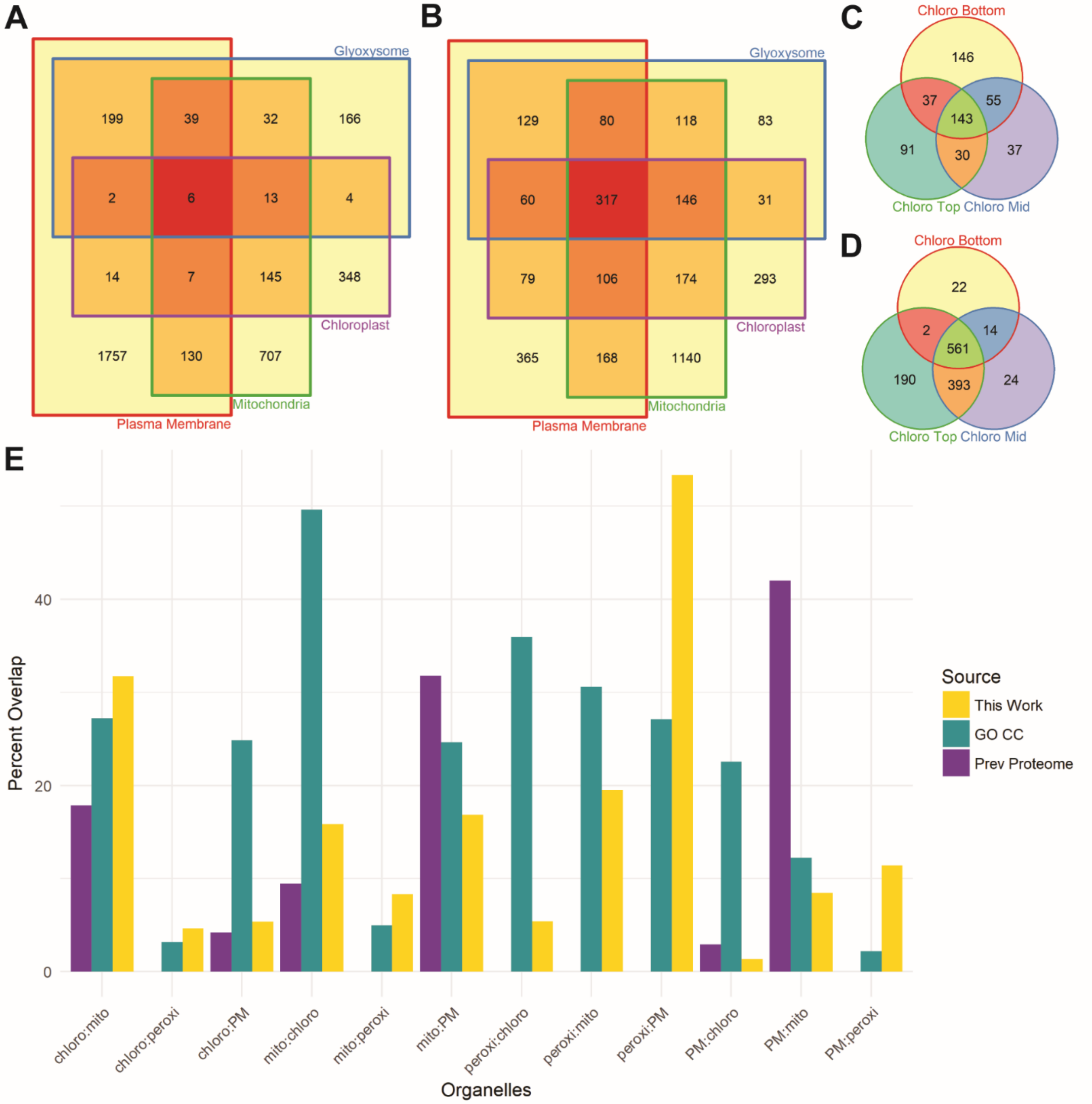


Figure 1-2: Maize proteins with localization in multiple organelles. A. Venn diagram showing overlap of significantly over-enriched proteins between organelles. **B.** Venn diagram showing overlap of significantly depleted proteins between organelles. **C:** Venn diagram showing overlap of significantly over-enriched proteins between leaf bottom-, middle-, and top-derived chloroplasts. **D:** Venn diagram showing overlap of significantly depleted proteins between leaf bottom-, middle-, and top-derived chloroplasts. **E:** Percent overlap of each organelle combination in this work (yellow), maize GAMER GO CC annotations (teal), and previously published organelle proteomes (purple). Percent calculated comparing overlap set size to total organelle localized protein set size of the first organelle. As there are no previously published proteomes from the peroxisome, bars are absent for overlaps containing peroxisome.

proteins were unique to the bottom and middle leaf derived chloroplasts, representing less than 5% of depleted proteins from each, while over 16% of the leaf top chloroplast depleted proteins were unique (Fig. 1-2D).

1.2.5 Observed overlap between organelle protein sets differs from overlap of gene ontology location annotations and previously published organellar proteins

We compared our protein localization data to published annotations from the maize GAMER Gene Ontology (GO) cellular compartment (CC) database (Wimalanathan et al., 2018) and to previously published proteomes of the plasma membrane (Voothuluru et al., 2016b; Zhang et al., 2013), mitochondria (Dahal et al., 2012, 2016; Hochholdinger et al., 2004; Wang et al., 2018), and chloroplast (Friso et al., 2010; Fristedt et al., 2012; Huang et al., 2013; Majeran et al., 2012; Zörb et al., 2009). As no maize peroxisome proteomes have been published, the glyoxysome proteins were compared only to GO CC annotations. The overlap between mitochondrial and chloroplast proteins in this work exceeded the overlap of previously published mitochondrial and chloroplast proteins (Fig. 1-2E; SI Fig. 1-10A-B). The observed overlap was similar or smaller than was annotated in the GO chloroplast and mitochondrial CC (Fig. 1-2E; SI Fig. 1-10C-D). Of the observed overlap between mitochondria and chloroplast proteins, 37% were annotated as localized in both target organelles (SI Fig. 1-10D). In contrast, the overlap between the glyoxysome and plasma membrane proteins observed was larger than the corresponding overlap in GO CC annotations (Fig. 1-2E; SI Fig 1-10C). The overlap between genes annotated in GO CC was larger than that observed in this work or previous proteomes for the chloroplast and plasma membrane, the peroxisome and chloroplast, and the peroxisome and mitochondria (Fig. 1-2E; SI Fig 1-10C). Only the overlap between the mitochondria and plasma membrane was higher between previously published proteome than between GO CC or this work (Fig. 1-2E; SI Fig. 1-10A).

1.2.6 Distribution of enrichment scores of GO CC annotated and previously published organelle proteins

Given the significant difference of organelle shared protein sets between this work, previously published proteomes, and GO CC annotations, we then looked at the distribution of enrichment scores for proteins with previous localization information. In the volcano plots, previously published plasma membrane, mitochondria, and chloroplast proteins detected in this work are biased towards enrichment versus depletion (Fig. 1-3A-C; SI Fig. 1-11A-G). However, many previously published proteins are significantly depleted in this work. Compared to the organellar proteins from this work, which were defined in part by a fold-enrichment threshold of two, the distribution of previously published organellar proteins was biased towards lower fold enrichment (Fig. 1-3D). The density curve of previously published mitochondrial proteins showed two peaks, one with positive fold enriched and the other with negative fold enrichment. We then compared the enrichment scores of proteins annotated in GO CC target organelles and contaminating organelles of interest, as determined by proportion of false positive gold standards enriched (Fig. 1-3E-H, SI Fig. 1-12). The mitochondria and chloroplast, which had few enriched false positive gold standards, but for which many proteins are shared, were compared to each other (Fig. 1-3G-H; SI Fig. 1-12). For all four organelles of interest, we see a bias towards enrichment for proteins with GO CC annotation in the target organelle. For the plasma membrane and mitochondria, proteins

Figure 1-3: Distribution of enrichment scores of previously published and GO CC annotated organellar proteins

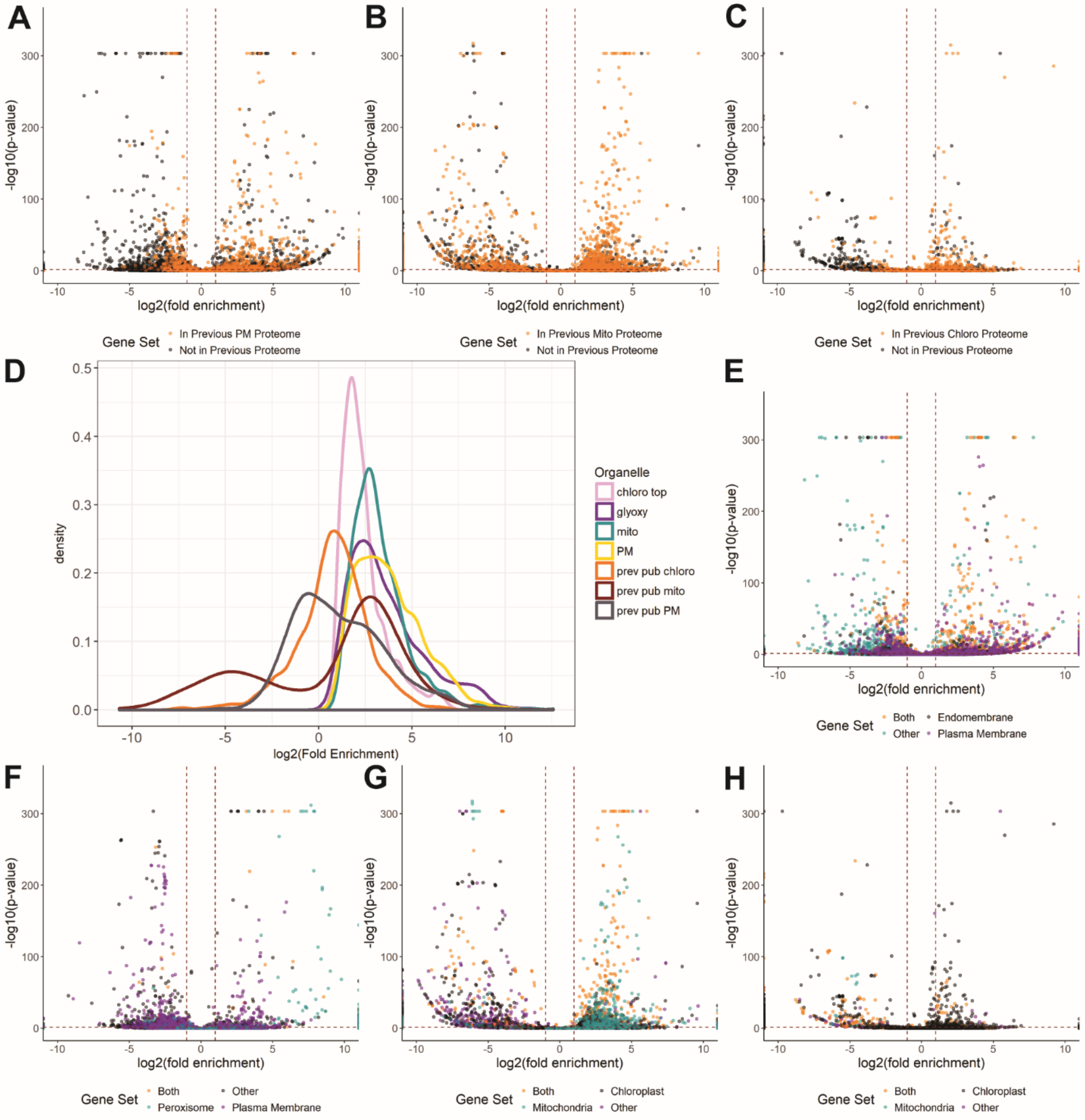


Figure 1-3: Distribution of enrichment scores of previously published and GO CC annotated organellar proteins. A-C: Volcano plots of enrichment scores of detected proteins from organelle datasets. Color based on localization in previously published proteome of plasma membrane (**A**) mitochondria (**B**) and chloroplast, compared to leaf top (**C**). **D:** Density plot showing fold enrichment of normalized spectral counts observed in target organelle versus source tissue for organelle enriched proteins and previously published organellar proteins. **E-H:** Volcano plots of enrichment scores of detected proteins from target organelle datasets. Color based on GO CC annotation in target organelle or contaminating organelle of interest for plasma membrane (**E**) glyoxysome (**F**) mitochondria (**G**) and chloroplast derived from the leaf top (**H**).

with GO CC annotation in the contaminating organelle of interest (endomembrane for plasma membrane, chloroplast for mitochondria) also show a bias towards enrichment. In contrast, plasma membrane GO CC annotated proteins detected in the glyoxysome were both enriched and depleted. Mitochondrial GO CC annotated proteins detected in the chloroplast were biased towards depletion.

1.2.7 Comparison of organelle enriched and depleted proteins to previously reported localization

More than 80% of chloroplast localized proteins were annotated as localized in the target GO CC (Fig. 1-4A). More than half of plasma membrane and mitochondria proteins and just over a quarter of glyoxysome proteins were also annotated as localized to the target CC. More proteins were annotated as localized to the plasma membrane, mitochondria, and chloroplast than were experimentally observed (Fig. SI 1-10C). Interestingly, fewer proteins were predicted to be localized to the peroxisome than were experimentally enriched in the glyoxysome (317 versus 461). Just over 30% of chloroplast and mitochondrial proteins and over 70% of plasma membrane proteins were novel compared to previously published subcellular proteomics data (Fig. 1-4B).

We then compared the GO CC annotation to our experimental observation for all genes with detected proteins in the work, as well as for the subset of detected genes predicted to produce membrane proteins from the ARAMEMNON database (Schwacke et al., 2003). For all detected proteins, 36% had no localization evidence, neither experimental nor in GO CC annotations. Among detected proteins with localization evidence, over 35% showed revised localization in this work (Fig. 1-4C). A smaller percentage of predicted membrane proteins than all detected proteins had novel

Figure 1-4: Comparison of localization of enriched or depleted proteins with GO CC localization annotation or localization in previous organellar proteomes.

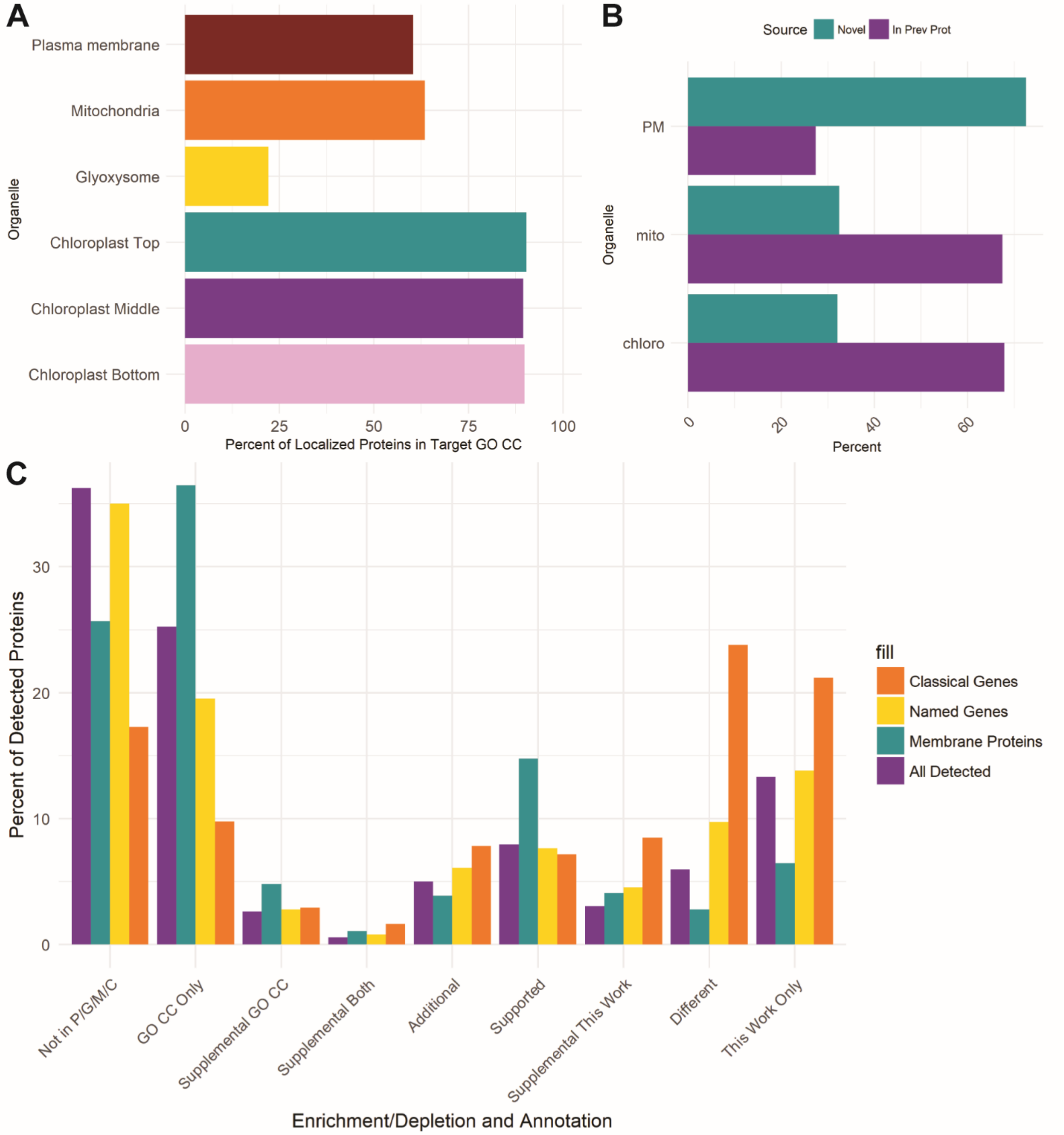


Figure 1-4: Comparison of localization of enriched or depleted proteins with GO CC localization annotation or localization in previous organellar proteomes. A: Percent of enriched proteins annotated as localized to the target GO CC. **B:** Percent of enriched proteins which are novel or supported by localization in previously published organelle proteome. **C:** Percent of proteins localized in this work or in GO CC annotation out of all detected proteins (purple), ARAMEMNON predicted membrane proteins (teal), maizeGDB “named” genes (yellow), and maize classical genes (orange).

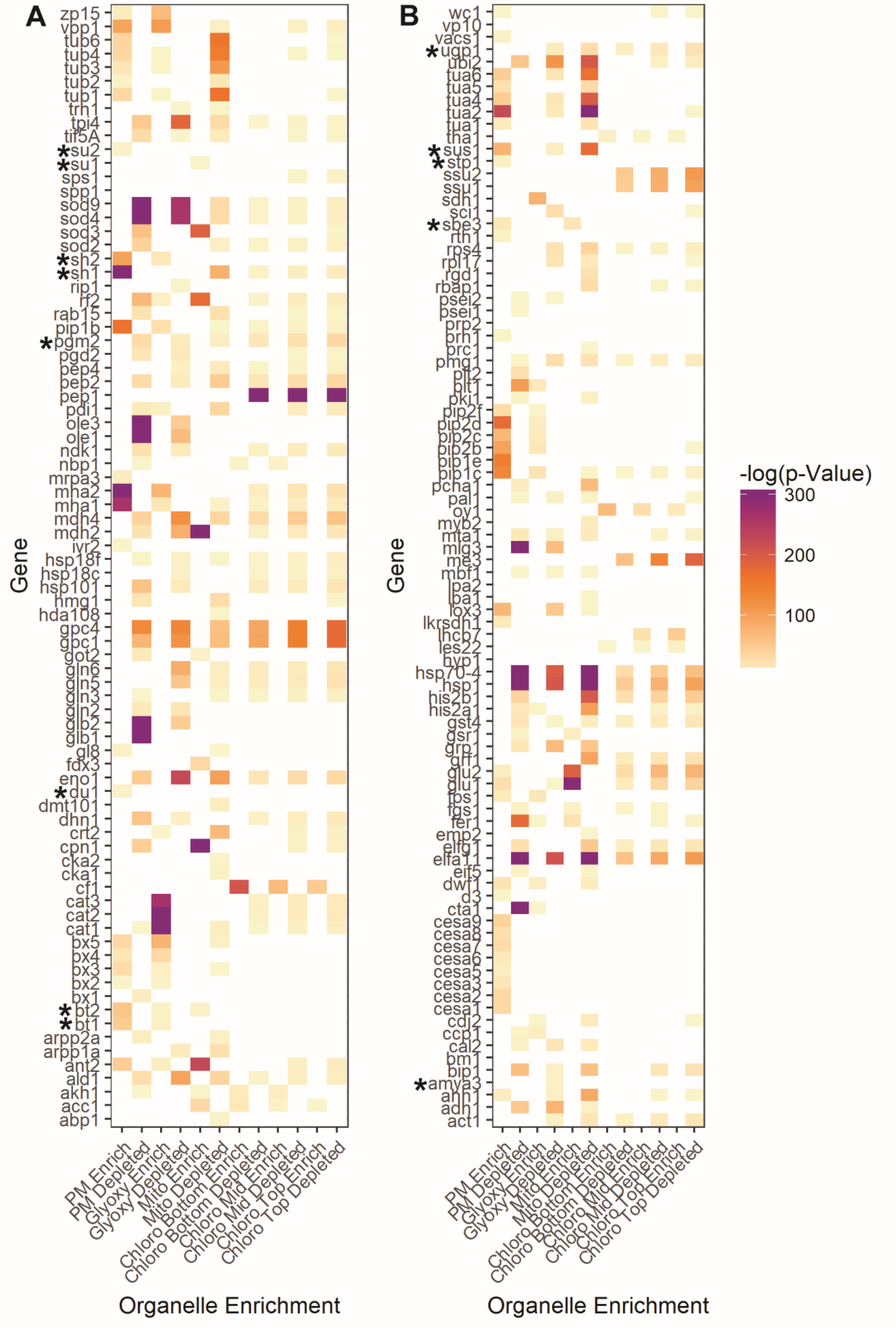
localization in this work. Maize “named” and maize classical genes are genes which are overrepresented in the literature (Schnable and Freeling, 2011). A lower percentage of these best studied subsets of all detected genes have no localization information in this work or GO CC annotations. Surprisingly, experimental results presented here offer evidence of new and revised localization for 42% of maize “named” genes and nearly 70% of proteins encoded by maize classical genes (Fig. 1-4C).

1.2.8 Subcellular localization of maize classical genes

For a more detailed comparison of our subcellular localization of maize classical genes to published results, we hand-annotated the subcellular localization of all detected maize classical genes, using maizeGDB annotations and experimental evidence from the literature (Table S5). Of the 268 maize classical gene-encoded proteins identified in our data, 51 had no annotated localization and were not significantly over-enriched or depleted in any organelle in this work. 225 were significantly over-enriched or depleted from profiled organelles, 120 had localization annotation in maizeGDB, and 100 had experimental evidence for localization in the literature (SI Fig. 1-13). All three sources provided information about the localization of products of 50 classical genes, while 99 classical gene-encoded proteins were localized only in this work (Fig. SI Fig. 1-13). Of the detected classical genes with localization both in this work and in either maizeGDB or the literature, the annotated localization of 15 was supported by this work (Fig. 1-5A). For nine proteins, our data indicates significant depletion from the annotated organelle. Additionally, 102 proteins show over-enrichment or depletion in at least one additional organelle beyond the annotation, for a total of 111 classical gene proteins with revised localization (Fig. 1-5B).

Figure 1-5: Subcellular localization of maize classical gene-encoded proteins. A-B: Heatmap of enrichment p-values across subcellular organelles for maize classical genes. Fill denotes negative log transformed p-value, with outliers fixed at 300. Asterisk denotes starch biosynthesis and sugar transport associated genes. **A**. Maize classical genes for which this work provides additional organelle localization annotation beyond maizeGDB or literature. **B.** Subset of maize classical genes without localization annotation in the literature or maizeGDB, but with significant enrichment in this work and subset of maize classical genes for which localization annotation in maizeGDB or literature differs from localization in this work.

Figure 1-5: Subcellular localization of maize classical gene-encoded proteins:



Classical genes localized in this work included many involved in starch biosynthesis, degradation, or sugar transport (Fig 1-5A-B; asterisks)(Walley et al., 2013; MaizeCyc). The detected starch biosynthetic enzymes were not enriched in the chloroplast fraction, with UGP1, SH1, and PGM2, significantly depleted from the chloroplast. PGM2 is additionally significantly depleted from the plasma membrane, mitochondria, and glyoxysome. Surprisingly, proteins from several starch biosynthetic enzymes were enriched in the plasma membrane fraction: BT1, BT2, DU1, SH1, SH2, SU2, SUS1, STP1, and SBE3. SH2, BT1, and BT2 are also enriched in the glyoxysome. BT2, SU1, and SBE3 are enriched in the mitochondria.

1.2.9 Identification of gene-splicing isoform-specific subcellular localization

Examples of gene isoform-specific localization were identified by comparing the enrichment of isoforms across subcellular samples. Gene models from which one isoform was significantly over-enriched, while another was significantly under-enriched, in an individual organelle (p-value<0.01), were distinguished as genes with evidence of isoform specific localization (Table S6). There were five examples of isoform specific localization, including protein products from genes on the maize classical or maizeGDB curated gene lists, AHH1 (ADENOSYL HOMOCYSTEINE HYDROLASE1), PDK1 (PYRUVATE ORTHOPHOSPHATE DIKINASE1), and ELFa2 (ELONGATION FACTOR ALPHA2). PDK1 had nine detected isoforms, all of which were significantly enriched in the plasma membrane and chloroplasts from the leaf middle and top, while depleted from the glyoxysome (Table S3). Interestingly, six of the PDK1 isoforms were significantly depleted while two isoforms were significantly enriched from the bottom leaf derived chloroplasts. The two detected isoforms of AHH1 were both significantly depleted from the glyoxysome, mitochondria, and chloroplast. One isoform of AHH1 was significantly enriched, while the other was significantly depleted from the plasma membrane. Likewise, the two detected isoforms of ELFa2 showed the same localization pattern. The other two examples showed evidence of mitochondrial and plastid isoform-specific localization respectively.

1.2.10 Functional enrichment analysis of organelle proteins

To determine statistical significance for organelle proteins annotated in GO CCs, gene ontology enrichment analysis was performed on organelle enriched or depleted proteins compared to all detected proteins. For all organelles, the over-enriched protein set had highest cellular compartment GO enrichment for the target organelle (Table 1-2; Tables S7-S12). All three chloroplast protein sets were under-enriched for proteins from contaminating organelles such as the nucleus, cytosol, plasma membrane and endomembrane system, among others (Tables S10-S12). The chloroplast bottom protein set was also slightly enriched for mitochondrial proteins, but under enriched for mitochondrial membrane and peroxisome proteins. There was no significant enrichment of mitochondrial and peroxisome proteins from the chloroplast middle nor chloroplast top. Interestingly, the chloroplast depleted proteins were over-enriched for mitochondrial, chloroplast, peroxisome, plasma membrane, and endomembrane system proteins, but under enriched for chloroplast membrane proteins (Tables S18-S20). The mitochondria protein set was significantly under-enriched for plasma membrane, nuclear, and endomembrane system organelle proteins, but highly enriched for plastid proteins (Table S9). While the mitochondrial depleted proteins were over-enriched for nuclear, cytosolic, plasma membrane and cell wall proteins, they were also under-enriched for plastid parts, plastid membrane, and mitochondrial proteins (Table S17). The glyoxysome localized protein set was over-enriched for proteins annotated in membranes of multiple organelles, including the vacuole and plasma membrane (Table S8). The glyoxysome localized protein set was also over-enriched for chloroplast and mitochondrial proteins. The glyoxysome protein set was under-enriched for nuclear or cytosolic proteins (Table S8). Glyoxysome depleted proteins were over-enriched for plastid, vacuole, nuclear, cytosol and plasma membrane proteins, while being under-enriched for endosome, endoplasmic reticulum, and plastid and mitochondrial membrane proteins (Table S16). The plasma membrane protein set was over-enriched for endomembrane system proteins, including the golgi, ER, and vacuole, as well as membrane proteins from the mitochondria and chloroplast (Table S7). The plasma membrane protein set was under-enriched for nuclear and cytosolic proteins (Table S7). Plasma membrane depleted proteins were over-enriched for the other target organelle and vacuole proteins, but under-enriched for endomembrane system proteins (Table S15). Given the significant overlap between both mitochondrial and chloroplast enriched proteins and plasma membrane and glyoxysome enriched proteins, GO cellular compartment enrichment analysis was performed on the shared enriched proteins compared to the set union of detected proteins in both organelle datasets. The chloroplast and mitochondria shared enriched proteins were over-enriched for both chloroplast and mitochondrial annotated proteins (Table S13). Additionally, the chloroplast and mitochondrial shared proteins were under-enriched for nuclear and endomembrane system proteins. The peroxisome and plasma membrane shared proteins were over-enriched for vacuole, plasma membrane, plastid, mitochondrial and peroxisome proteins, and under-enriched for cytosolic and nuclear proteins (Table S14).

Table 1-2: Top Enriched GO Categories of Organelle Proteins: GO cellular compartment and biological process categories most highly enriched in organelle protein sets.

Organelle	Top GO CC	Top GO BP
Plasma Membrane	plasma membrane	Transport, signaling
Glyoxysome	Peroxisome	transmembrane transport, proton transport
Mitochondria	Mitochondrion	small molecule metabolic process, generation of precursor metabolites and energy
Chloro (Bottom)	plastid part	plastid organization, organic acid metabolic process
Chloro (Mid)	plastid part	photosynthesis, plastid organization
Chloro (Top)	plastid part	photosynthesis, plastid organization

For GO biological process annotations (Tables S21-S28, “BP”), glyoxysome localized proteins were enriched for known glyoxysomal processes, such as the glyoxylate cycle, reactive oxygen species metabolism, photorespiration, and lipid oxidation (Table S22). Glyoxysomal proteins were also enriched for biotic and abiotic stress related processes, including salt, osmotic and water-deprivation stress. Interestingly, glyoxysomal proteins were highly enriched for multiple types of membrane transport (Table 1-2; Table S22). Mitochondria proteins were significantly over-enriched for expected metabolic processes and photorespiration (Table S23). Mitochondrial proteins were also over-enriched for inosine monophosphate (IMP) metabolism, and biosynthesis of poly-unsaturated fatty acids; both metabolic pathways with ambiguous or unknown subcellular localization. Enriched biological processes among plasma membrane proteins include expected categories such as response to stimulus, signaling, and transport (Table S21). Additional enriched plasma membrane categories, however, included regulation of cell size and cell developmental growth. Plasma membrane enriched development-associated GO subcategories included root morphogenesis, root epidermal cell differentiation and trichoblast differentiation (GO:0010015, GO:0010053, GO:0010054, respectively).

The most enriched GO BP categories were shared across all three chloroplast samples: photosynthesis, plastid organization, glyceraldehyde-3-phosphate metabolic process, and isoprenoid metabolism, in addition to many others (Tables S24-S26). To investigate the differences between photosynthetically mature, leaf tip-derived chloroplasts and photosynthetically immature leaf-base chloroplasts, functional enrichment analysis was also performed on chloroplast enriched proteins which are unique to either chloroplast sample (Tables S27-S28). Few of the most enriched GO BP categories were shared between groups of unique proteins, indicating that not only the identities, but also the biological roles of these proteins are unique.

Enrichment of GO biological process categories was performed on protein sets localized in two organelles (Fig 2A; Tables S29-S30). Proteins enriched in both the plasma membrane and glyoxysome were highly enriched for multiple transport-associated GO categories, as well as nucleoside phosphate metabolism, ATP synthesis, and response to multiple stresses (Table S30). More than 58% of plasma membrane and glyoxysome dual localized proteins were annotated as being involved in response to stress, with over-enrichment for response to salt and osmotic stress (Table S30). Chloroplast and mitochondria dual localized proteins were over-enriched for a wide variety of small molecule metabolic processes and chloroplast protein import and membrane transport associated GO categories (Table S29). Interestingly, plant ovule development was enriched among chloroplast and mitochondrial dual localized proteins, along with response to salt, osmotic, and oxidative stresses.

1.3 Discussion

The total number of enriched proteins for the plasma membrane, glyoxysome, and mitochondria was consistent with predicted organelle proteome sizes, as the peroxisome proteome is predicted to contain several hundred proteins, whereas the plasma membrane, mitochondria, and chloroplast are each predicted to have thousands of proteins (Emanuelsson et al., 2000). Fewer chloroplast proteins were localized than were predicted, annotated in GO CC, or detected in previous proteomes (Friso et al., 2010; Fristedt et al., 2012; Huang et al., 2013; Majeran et al., 2012; Zörb et al., 2009). Additionally, the percentage of gold standard true positives enriched in the chloroplast was lower than other target organelles. However, the chloroplast enriched proteins showed few enriched gold standard false positives, indicating the that chloroplast samples were relatively free of contaminating organelles. Thus the chloroplast protein set shows high specificity but at the cost of lower sensitivity. Relaxing the fold enrichment or p-value threshold for chloroplast localization did not substantially increase the true positive rate for the Comparison to gold standard false positives indicated low cross-contamination between the mitochondria, chloroplast, and glyoxysome samples, which are commonly co-purified. Individual gold standard proteins detected in subcellular fractions other than the target organelle may represent single protein contaminants, or bona fide dual-localized proteins.

In this work, the organelle proteins sets were defined with the threshold for enrichment of two-fold. In contrast, detected proteins from previously published mitochondria, chloroplast, and plasma membrane proteomes had overall lower fold enrichment, including proteins with fewer spectral counts observed in the organelle than the source tissue (Fig. 1-3A-D). The distribution of proteins in the volcano plot, where many proteins detected in the organelle have negative fold enrichment, indicates that detection of a protein within an organelle enriched subcellular fraction is not adequate to discriminate organelle proteins from contaminants. This was true even for the purest organelle based on comparison to gold standards, the chloroplast. As the sensitivity of proteomics methods increases, the detection of low abundance proteins in individual samples will also increase. To distinguish between low abundance contaminants and low abundance true organelle proteins requires a statistical measure of enrichment. False positive gold standards were detected in the organelle fraction of each organelle (Fig. 1-1A-B; SI Fig 1-8). Many of these false positives are excluded by comparison to the source tissue. Fold-enrichment alone does not reflect reproducibility nor variability, while proteins with low fold-enrichment may still receive a significant p-value. Filtering the organelle protein sets based on both p-value and fold enrichment gave the lowest false positive rate. Only one of the previously published maize subcellular proteomes compared in this work used comparison to a source tissue to filter the organelle protein set, using a fold-enrichment threshold only (Huang et al., 2013).

Although we cannot exclude the possibility of organelle cross-contamination, given the low percentage of contaminating gold standard proteins enriched in chloroplast and mitochondrial organelle protein sets, proteins present in both organelles may represent dual-localized proteins. Functional enrichment analysis among chloroplast and mitochondrial dual localized proteins did not show a strong bias towards single organelle specific processes. The prevalence of salt, osmotic, and oxidative stress associated genes among dual localized mitochondrial and chloroplast proteins suggests complex multi-organelle involvement in these pathways. The observed overlap between chloroplast and mitochondrial proteins was larger than the overlap of previously published chloroplast and mitochondrial proteins. Similarly, the overlap between plasma membrane and glyoxysome enriched proteins was much larger than the overlap between GO CC annotated plasma membrane and peroxisome proteins (Fig. 1-2C). However the observed overlap between the plasma membrane and chloroplast proteins, and between peroxisome and chloroplast proteins was smaller than in the GO CC annotation (Fig. 1-2C). As each organelle was isolated from a different source tissue, some proteins enriched in more than one organelle may have tissue-specific or condition-specific subcellular localization, and thus may not exist in both organelles within the same cell.

Comparison to a source tissue also allows for definition of an organelle depleted set. Few chloroplast depleted proteins were unique from the bottom and middle leaf, while a much higher percentage were unique from the leaf top derived chloroplasts. Additionally, more leaf middle depleted proteins were shared between the leaf top than with the leaf bottom. Organelle-specific variation in number of unique depleted proteins could be related to variation in the number or complexity of contaminating organelles. Scores for both enrichment and depletion from a given organelle enabled identification of five examples of protein isoform specific subcellular localization. Isoforms of PDK1 are have differential splicing and transcript initiation sites, resulting in both cytosolic and plastid localized proteins (Sheen, 1991). Nine protein isoforms of PDK1 were detected, with isoform specific enrichment and depletion from leaf bottom-derived chloroplasts. The subcellular localization of the other isoform-specific localized proteins was previously unknown.

Comparison to previously published subcellular proteomes and to GO CC annotations showed both supported and novel subcellular localization of proteins detected in this work (Fig. 1-4). Interestingly, the proportion of proteins with novel localization compared to GO CC annotations increased when looking at the best studied subset of maize genes: the classical genes (Fig. 1-4C). Comparison to manually annotated subcellular localization of maize classical genes identified many classical genes with additional or revised localization in this work. While starch biosynthesis from sucrose is thought to occur in both the cytosol and plastid (Walley et al., 2013), several starch biosynthetic enzymes were enriched in multiple organelles. The role of the peroxisome in starch biosynthesis is not fully understood, but several starch biosynthetic enzymes were over-enriched in the glyoxysome. These results suggest broad participation of the profiled subcellular organelles in starch biosynthesis. PGM2 is significantly depleted from all target organelles, consistent with the annotated cytosolic localization.

Due to difficulties with purification and detection of insoluble proteins, organelle membrane proteins are less studied than their soluble counterparts. Computational prediction of organelle membrane proteins is also challenging, for example while transport of many proteins to the peroxisome matrix involves recognition of PTS1 or

PTS2-type peroxisomal transit peptides, peroxisomal membrane proteins require an alternative import mechanism (Mayerhofer, 2016). Transport associated GO BP categories were enriched in all organelle protein sets, suggesting success of enrichment for and detection of organelle membrane proteins. Greater than 10% of each organelle protein set was in the ARAMEMNON database of predicted maize membrane proteins (SI Fig. 1-14). A similar percentage of glyoxysomal proteins were predicted to be membrane proteins in the ARAMEMNON database, compared to plasma membrane proteins (SI Fig. 1-14). However, high overlap between the plasma membrane and glyoxysome proteins, as well as the presence of plasma membrane gold standard proteins in the glyoxysome enriched protein set, revealed possible contamination of the glyoxysome with plasma membrane proteins. The plasma membrane and glyoxysome were both isolated from seedling tissues at similar stages of development. These similar source tissues could have similar potential contaminants during the glyoxysome and plasma membrane isolation. It is unknown what percentage of shared glyoxysome and plasma membrane proteins represent true dual-localized proteins, contamination between the two organelles, or shared contaminants from another organelle.

We present here evidence for the subcellular localization of proteins from 3,378 maize genes, including proteins localized to the plasma membrane, glyoxysome, mitochondria, and chloroplast. Organelle localized proteins include both novel and supported localization, in comparison to both gene ontology and previously published work. Revised and additional localization information is provided for over a hundred maize classical genes. Enrichment and depletion scores for all detected proteins from the newest reference genome of maize across each organelle are provided as a unique resource for the maize community.

1.4 Materials and Methods

1.4.1 Plasma Membrane Isolation

For plasma membrane isolation, B73 seeds were sterilized in 100% bleach and 0.01% Triton X 100. Seeds were vacuum infiltrated for 5 minutes, incubated for an additional 30 minutes with shaking, then washed 5X with sterile water. The seeds were placed in pyrex dishes lined with a layer of moistened Gel Blot Paper (GB003,Whatman) and a layer of wet germination paper (Anchor Paper, GP1015), and covered with another layer of wet germination paper. The dishes were wrapped in aluminum foil and placed in the dark chamber at 25 °C for 5 days. At harvest, seedlings had coleoptiles 1-3 cm long, primary roots of 3-8 cm, and most seeds had crown roots 1-3 cm long.

Two-phase partitioning was adapted from Marmagne et al., 2006 (SI Fig. 1-6). Two-phase partitioning systems were prepared, mixed well, and allowed to equilibrate overnight. Systems contained 6.3% dextran, 6.3% PEG 3350, 300 mM Sucrose, 5 mM Potassium Phosphate buffer pH 7.8, 4 mM KCl, 1 mM DTT and 50 µM EDTA. 140 g of germinated seedlings (including seeds) and 200 ml of ice cold buffer H (100mM Hepes-KOH pH 7.5, 0.2% N-Z amine B, 300mM sucrose, 10% (w/v) glycerol, 5mM EDTA free acid, 15mM EGTA, 0.6% PVP K-25, 5 mM Ascorbic Acid [Fisher Scientific]) supplemented with freshly added 1mM DTT, a protease inhibitor cocktail (20mM NaF

(Fisher), 0.5% Protease Inhibitor Cocktail (Sigma), 1 mM PMSF(EMD), 10 µM leupeptin (EMD)) and a phosphatase/HDAC inhibitor cocktail (5 mM β-glycerophosphate (CalBiochem), 1 nM calyculin A (Cell Signaling Technology), 1 µM Trichostatin A (Sigma) and 10 mM Nicotinomide (Sigma)) were homogenized in a Blender (Waring) in two pulses of 5 sec and 5 pulses of 10 sec. The homogenate was filtered through Miracloth (Millipore) and then centrifuged at 10 000xg for 10 minutes. The pellet was discarded. The resulting supernatant is the whole seedling extract. The remaining supernatant was filtered through a 250 micron polypropylene mesh (Small Parts, Inc.) and centrifuged at 100 000xg in a Ti60 rotor (Beckman). The supernatant was kept as the soluble fraction. The pellet (microsomal fraction) was thoroughly resuspended in buffer R (5 mM Potassium Phosphate buffer, pH 7.8, 300 mM sucrose, 0.1 mM EDTA, 4 mM KCl, 5 mM Ascorbic Acid – all chemicals from Fisher) with DTT, protease and inhibitor cocktails as above using a homogenizer (Omni TH). Two-phase portioning was performed using the two-phase partitioning systems as outlined in Figure S1. After phase separation, the upper phase (PM) was diluted ~4X and the lower phase (endomembranes) was diluted ~6X in buffer R (containing DTT and phosphatase/HDAC inhibitors but not protease inhibitors) and pelleted in a Ti60 rotor at 200 000xg for 60 min. The pellets were snap frozen in liquid nitrogen and stored at -80°C. Within one week, the pellets were thawed on ice, resuspended in ~2 ml (plasma membranes) or ~10 ml (endomembranes) ice cold buffer R with phosphatase inhibitors, and pelleted at 100 000xg in a SW50.1 rotor (Beckman) for 45 minutes. The resulting pellets were snap frozen in liquid nitrogen and stored at -80 °C.

1.4.2 Intact Glyoxysome Isolation

Maize (Zea mays, B73 hybrid line) kernels were sown onto moist vermiculite and incubated at 25°C for 3 days in the dark. The seedlings were rinsed in water and the scutellum tissue was excised using a razor blade. The tissue was then finely minced and ground with an ice-cold mortar and pestle in 3 mL/g tissue grinding buffer (400mM sucrose; 170mM tricine; 2mM EDTA; 0.1mM BSA; 10mM KCl; 1mM $MgCl_2$; 5mM DTT; 100µM PMSF; 31 µg/mL benzamidine; 26 µg/mL ε-aminocaproic acid; pH 7.5). The homogenate was filtered through miracloth and the filtrate was collected in an ice-cold tube. The crude extract was centrifuged at 480xg (2000 rpm, SS34 rotor) for 10 min at 4°C and the supernatant was transferred to a fresh tube and then centrifuged at 10800xg (9500 rpm, SS34 rotor) for 10 min at 4 °C. The supernatant was decanted and the pellet was resuspended in 2 mL 36% sucrose. The resuspension was then loaded onto a sucrose gradient (1.0 mL 85% sucrose; 1.0 mL 60% sucrose; 0.5 mL 55.2% sucrose; 0.5 mL 50.5% sucrose; 2.0 mL 48.5% sucrose; 1.0 mL 46.0% sucrose; 1.0 mL 43.7% sucrose; 1.0 mL 41.2% sucrose) and centrifuged for 40 min at 25000 rpm (WX80 ultracentrifuge). Eleven fractions from the top (F1) to the bottom (F11) of the sucrose gradient were collected. Protease inhibitors were added to each fraction (1mM PMSF; 1 µg/ml pepstatin A; 310 µg/ml benzamidine; 260 µg/ml ε-animocaproic acid; 1 µg/ml aprotinin; 1 µg/ml leupeptin). The samples were subjected to SDS-PAGE, electroblotted onto PVDF membranes and probed with separate antibodies raised against catalase and porin (SI Fig. 1-15). For maximum purity and yield, fraction 8 (F8) was routinely harvested as the glyoxysome fraction to be used for proteomics analysis.

A total of six glyoxysome isolations were performed and five samples were sent for further proteomic analysis.

To determine the relative purity of the six glyoxysome preparations, each sample was first assayed for protein concentration using the BCA™ Protein Assay Kit (Thermo Scientific). Then an equal amount (60 µg) of glyoxysomal proteins was concentrated using methanol/chloroform precipitation, resuspended in SDS-PAGE sample buffer, separated by SDS-PAGE, and transferred onto PVDF membranes. The blots were probed with antibodies raised against catalase (for presence of glyoxysomal proteins) and porin (for presence of mitochondrial proteins). Exposure times of 1 and 30 minutes were used to allow comparisons between the relative abundance of catalase and porin. The preparations were ranked from 1 to 5 with 1 being the most pure and 5 being the least (SI Fig. 1-16). The samples ranked 1-3 and 5 were sent for further proteomic analysis.

1.4.3 Intact Mitochondria Isolation

Mitochondria were isolated as described previously (Hochholdinger et al., 2004). In brief, for mitochondrial isolation, 100 g unpollinated maize B73 ears were homogenized in buffer A (30 mM MOPS, 1 mM EDTA, 400 mM sucrose, 4 mM cysteine; pH 7.5). Vlieseline filtered extract was spun at 3500 rpm for 10 min, the supernatant was then spun again at 8500 rpm for 20 min. The subsequent pellet was resuspended in 10 ml buffer B (400 mM sucrose, 30 mM MOPS pH 7.5) and spun again for 15 min at 9000 rpm. The crude mitochondria pellet was resuspended in 2 ml of buffer C (300 mM mannitol, 30 mM MOPS pH 7.5) and layered on top of a 13.5%:21%:45% percoll gradient in 1 mM sucrose and 10 mM MOPS pH 7.5. Mitochondria were spun through the percoll gradient at 7200 rpm for 30 min and the mitochondria were collected from the 21%:45% percoll interphase. Mitochondria were washed with buffer (300 mM mannitol, 1 mM EDTA, 30 mM MOPS pH 7.2) and collected by spinning at full speed for 15 min.

1.4.4 Intact Chloroplast Isolation

For maize chloroplast isolation, B73 seedlings were grown in soil in a greenhouse (with an average daytime temperature of 28 °C and an average nighttime temperature of 16 °C) with a 16-hr-light (supplemented to 500 µmol m-2 sec-1) and 8-hr-dark cycle. Etiolated seedlings were grown in the dark for 14 days at 28 °C. After removal from the dark, the etiolated seedlings were placed in the green house under continuous light for 2, 4, 12 and 24 hours before isolation of etioplast samples at various time points.

For chloroplast isolation from 1-month-old plants, plants were grown in a greenhouse to the 8-leaf stage (1 month). The partially emerged eighth leaf was dissected from the plant into top, middle and bottom sections (SI Fig. 1-7A)(Cahoon et al., 2008). The leaf tissue (~25g) from the 3 sections was homogenized in 1x Grinding Buffer (0.33 M sorbitol, 50 mm HEPES (pH 8.0), 2mM EDTA, 1mM $MgCl_2$, 1mM $MnCl_2$, 5mM sodium ascorbate and 1% BSA). The homogenate was filtered through two layers of Miracloth (Calbiochem, La Jolla, CA), and the filtrate was centrifuged at 4000 rpm for 8 min. The pellet was resuspended in 5 ml of 1x Grinding buffer and was loaded onto a continuous percoll gradient. The gradient was centrifuged at 6000 rpm for 15 min using a swinging bucket rotor. Intact chloroplasts sink to the bottom of the gradient and the broken cell debris can be found on top. The bottom band containing intact chloroplasts was washed twice in 15 ml of re-suspension buffer (0.33 M sorbitol, 50 mM HEPES (pH 8.0)), and finally resuspended in 1 ml of re-suspension buffer, following which the samples were flash frozen in liquid nitrogen for further analysis. For etioplast samples, 14-day-old etiolated maize seedlings were harvested at 0, 2, 4, 12, and 24 hours after exposure to light. The etioplasts were isolated using the same procedure used for intact chloroplast isolation.

1.4.5 Proteomics of subcellular organelle enriched fractions

Proteins were extracted from source tissue and glyoxysome, mitochondria, and chloroplast fractions, trypsin digested, and prepared for mass spectrometry as described previously (Walley et al., 2013). Plasma membrane mass spectrometry samples underwent alternative sample preparation (Supplemental materials and methods). In brief, plasma membrane proteins were extracted in 8M urea/tris buffer and trypsin digested. Plasma membrane phosphopeptides were enriched as described previously (Walley et al., 2013). Plasma membrane mass spectra were acquired using Q-exactive HF mass spectrometer. All mass spectra were searched against B73 RefGen_v4 Working Gene Set. Global peptide false discovery rate compared to forward:reverse decoy database was constrained at 0.1% and the protein FDR constrained at 1%. Proteomes of mature leaf and ear used as source tissue for chloroplast and mitochondria samples are previously published, with mass spectra searched against working gene set from B73 RefGen_V2 (Walley et al., 2016)

1.4.6 Identification of subcellular localized proteins

The proteins identified in each organelle were analyzed separately against the proteins from the source tissue of the organelle prep. The proportion of raw spectral counts from each unique protein group (Walley et al., 2013) was compared between source and organelle samples using one-tailed Fishers Exact Test to identify over-enriched proteins. The converse process was then applied, using the opposite tail, to identify under-enriched proteins. The resultant p-values were adjusted for multiple comparisons using Benjamini-Hochberg p-value adjustment using the base R functions fisher.test() and p.adjust(). The fold difference of protein normalized spectral counts of the organelle enriched fraction versus source tissue was also calculated. The thresholds for organelle enrichment and depletion were >2 fold difference in spectral counts between organelle and source tissue with fishers exact test p-value of less than 0.05.

1.4.7 Gene ontology categorical enrichment analysis of organelle over-enriched proteins

GO enrichment analysis of organelle over-enriched (p-value<0.05) proteins was executed using previously published annotation of GO “cellular compartment”, “biological process”, and “molecular function” categories for maize genes from the maize GAMER database (Wimalanathan et al., 2018). Terms for the GO annotations were accessed using the GO.db R package (Carlson, 2017). Enrichment analysis was performed using a custom R Script, as described previously, with basic modifications to utilize GO annotations instead of MapMan annotations (Walley et al., 2016). In brief, each list of organelle-enriched proteins was compared to each GO category present in the dataset. A hypergeometric test using phyper() function from the stats package was implemented, comparing number of organelle localized proteins annotated in the GO category to all detected proteins annotated in the GO category. The resultant p-values for each category were adjusted using the Benjamini & Hochberg correction.

1.4.8 Additional data analysis

Venn diagrams were created using Vennerable R Package (Swinton, 2009). Other plots were created using R Packages reshape2 and ggplot2 unless otherwise mentioned (Wickham, 2007, 2009). For comparison to previously published proteomes which used NCBI genbank “gi” accessions, organelle protein amino acid sequences were retrieved from the NCBI website and used as queries for a local blast search against maize Ref.Gen.v4 accessions. Heatmap of subset of previously published maize classical genes (Schnable and Freeling, 2011) was created using R package ggplot2. Enrichment p-values were negative log transformed. Outlier log-transformed p-values of greater than 300 were fixed at a value of 300.

1.5 Acknowledgements

The authors would like to acknowledge Joshua Osborn for assistance with sample processing and mass spectrometry analysis.

Chapter 1, in full, is currently being prepared for submission for publication of the material. de Boer, Laura; Shen, Zhouxin; Putarjunan, Aarthi; Paschold, Anja; Dorchak, Alexandria; Helzer, Kyle T.; Sartor, Ryan C.; Facette, Michelle; Hochholdinger, Frank; Olsen, Laura J.; Rodermel, Steven; Smith, Laurie G.; Briggs, Steven P. “Direct Evidence for the Sub-cellular Localization of Proteins Encoded by 3,378 Genes of

Maize.” *In preparation.* The dissertation author was the primary investigator and author of this material.

1.6 References

Cahoon, A.B., Takacs, E.M., Sharpe, R.M., and Stern, D.B. (2008). Nuclear, chloroplast, and mitochondrial transcript abundance along a maize leaf developmental gradient. Plant Mol. Biol. *66*, 33–46.

Carlson, M. (2017). GO.db: A set of annotation maps describing the entire Gene Ontology.

Dahal, D., Mooney, B.P., and Newton, K.J. (2012). Specific changes in total and mitochondrial proteomes are associated with higher levels of heterosis in maize hybrids: *2D DIGE of heterosis in maize*. Plant J. *72*, 70–83.

Dahal, D., Newton, K.J., and Mooney, B.P. (2016). Quantitative Proteomics of *Zea mays* Hybrids Exhibiting Different Levels of Heterosis. J. Proteome Res. *15*, 2445–2454.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J. Mol. Biol. *300*, 1005–1016.

Friso, G., Majeran, W., Huang, M., Sun, Q., and van Wijk, K.J. (2010). Reconstruction of Metabolic Pathways, Protein Expression, and Homeostasis Machineries across Maize Bundle Sheath and Mesophyll Chloroplasts: Large-Scale Quantitative Proteomics Using the First Maize Genome Assembly. Plant Physiol. *152*, 1219–1250.

Fristedt, R., Wasilewska, W., Romanowska, E., and Vener, A.V. (2012). Differential phosphorylation of thylakoid proteins in mesophyll and bundle sheath chloroplasts from maize plants grown under low or high light. Proteomics *12*, 2852–2861.

Hochholdinger, F., Guo, L., and Schnable, P.S. (2004). Cytoplasmic regulation of the accumulation of nuclear-encoded proteins in the mitochondrial proteome of maize: Proteomic analysis of NA- and T-cytoplasm maize. Plant J. *37*, 199–208.

Hooper, C.M., Castleden, I.R., Tanz, S.K., Aryamanesh, N., and Millar, A.H. (2017). SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations. Nucleic Acids Res. *45*, D1064–D1074.

Hopff, D., Wienkoop, S., and Lüthje, S. (2013). The plasma membrane proteome of maize roots grown under low and high iron conditions. J. Proteomics *91*, 605–618.

Huang, M., Friso, G., Nishimura, K., Qu, X., Olinares, P.D.B., Majeran, W., Sun, Q., and van Wijk, K.J. (2013). Construction of Plastid Reference Proteomes for Maize and *Arabidopsis* and Evaluation of Their Orthologous Relationships; The Concept of Orthoproteomics. J. Proteome Res. *12*, 491–504.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature *546*.

Majeran, W., Cai, Y., Sun, Q., and van Wijk, K.J. (2005). Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. Plant Cell *17*, 3111–3140.

Majeran, W., Zybailov, B., Ytterberg, A.J., Dunsmore, J., Sun, Q., and van Wijk, K.J. (2008). Consequences of C4 differentiation for chloroplast membrane proteomes in maize mesophyll and bundle sheath cells. Mol. Cell. Proteomics *7*, 1609–1638.

Majeran, W., Friso, G., Asakura, Y., Qu, X., Huang, M., Ponnala, L., Watkins, K.P., Barkan, A., and van Wijk, K.J. (2012). Nucleoid-Enriched Proteomes in Developing Plastids and Chloroplasts from Maize Leaves: A New Conceptual Framework for Nucleoid Functions. Plant Physiol. *158*, 156–189.

Marmagne, A., Salvi, D., Rolland, N., Ephritikhine, G., Joyard, J., and Barbier-Brygoo, H. (2006). Purification and Fractionation of Membranes for Proteomic Analysis. In Arabidopsis Protocols, J. Salina, and Sanchez-Serrano, eds. (Humana Press), pp. 403–420.

Mayerhofer, P.U. (2016). Targeting and insertion of peroxisomal membrane proteins: ER trafficking versus direct delivery to peroxisomes. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1863*, 870–880.

Reumann, S., Quan, S., Aung, K., Yang, P., Manandhar-Shrestha, K., Holbrook, D., Linka, N., Switzenberg, R., Wilkerson, C.G., Weber, A.P.M., et al. (2009). In-Depth Proteome Analysis of Arabidopsis Leaf Peroxisomes Combined with in Vivo Subcellular Targeting Verification Indicates Novel Metabolic and Regulatory Functions of Peroxisomes. Plant Physiol. *150*, 125–143.

Schnable, J.C., and Freeling, M. (2011). Genes Identified by Visible Mutant Phenotypes Show Increased Bias toward One of Two Subgenomes of Maize. PLoS ONE *6*, e17855.

Schwacke, R., Schneider, A., van der Graaff, E., Fischer, K., Catoni, E., Dsimone, M., Frommer, W.B., Flugge, U.-I., and Kunze, R. (2003). ARAMEMNON, a Novel Database for Arabidopsis Integral Membrane Proteins. Plant Physiol. *131*, 16–26.

Swinton, J. (2009). Venn Diagrams in R with the Vennerable Package.

Tan, S., Tan, H.T., and Chung, M.C.M. (2008). Membrane proteins and membrane proteomics. Proteomics *8*, 3924–3932.

Voothuluru, P., Anderson, J.C., Sharp, R.E., and Peck, S.C. (2016a). Plasma membrane proteomics in the maize primary root growth zone: novel insights into root growth adaptation to water stress: Plasma membrane proteomics in water-stressed roots. Plant Cell Environ. *39*, 2043–2054.

Voothuluru, P., Anderson, J.C., Sharp, R.E., and Peck, S.C. (2016b). Plasma membrane proteomics in the maize primary root growth zone: novel insights into root growth adaptation to water stress: Plasma membrane proteomics in water-stressed roots. Plant Cell Environ. *39*, 2043–2054.

Walley, J.W., Shen, Z., Sartor, R., Wu, K.J., Osborn, J., Smith, L.G., and Briggs, S.P. (2013). Reconstruction of protein networks from an atlas of maize seed proteotypes. Proc. Natl. Acad. Sci. *110*, E4808–E4817.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., et al. (2016). Integration of omic networks in a developmental atlas of maize. Science *353*, 814.

Wang, W.-Q., Wang, Y., Zhang, Q., Møller, I.M., and Song, S.-Q. (2018). Changes in the mitochondrial proteome of developing maize seed embryos. Physiol. Plant.

Wickham, H. (2007). Reshaping Data with the reshape Package. J. Stat. Softw. *21*.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wimalanathan, K., Friedberg, I., Andorf, C.M., and Lawrence-Dill, C.J. (2018). Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER). Plant Direct *2*, e00052.

Zhang, Z., Voothuluru, P., Yamaguchi, M., Sharp, R.E., and Peck, S.C. (2013). Developmental Distribution of the Plasma Membrane-Enriched Proteome in the Maize Primary Root Growth Zone. Front. Plant Sci. *4*.

Zörb, C., Herbst, R., Forreiter, C., and Schubert, S. (2009). Short-term effects of salt exposure on the maize chloroplast protein pattern. PROTEOMICS *9*, 4209–4220.

1.7 Supplemental Information

Figure 1-6:

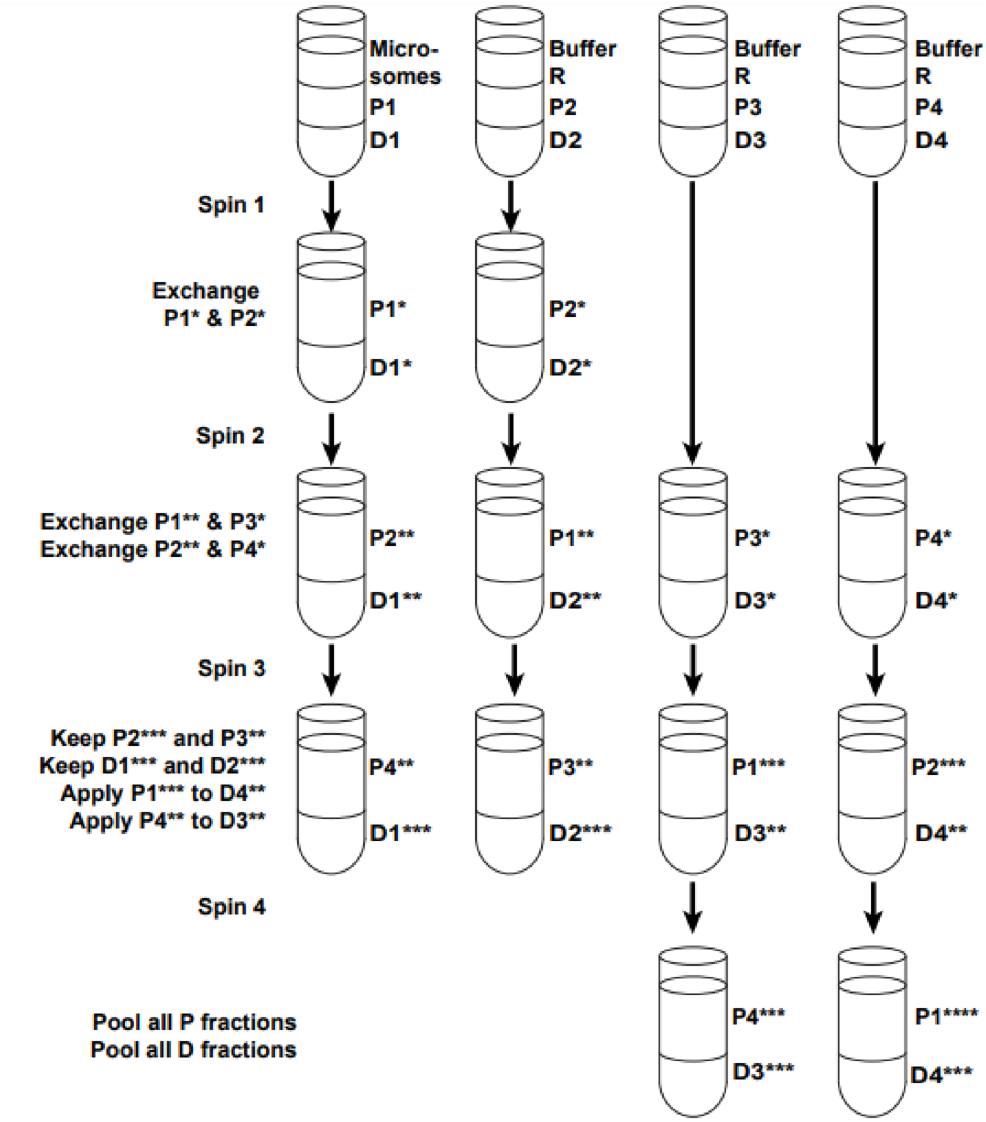


Figure 1-6: Purification schema for two phase partitioning of plasma membranes and endomembranes. Four systems were initially prepared. Prior to the first spin, the dextran layer (D) is on the bottom, with PEG layered (P) on top. The microsomal extract is applied to system 1 and Buffer R is applied to systems 2, 3, and 4. The systems are mixed thoroughly by shaking and spun at 2500 RPM in a JS4.3 rotor with no brake at 4 degrees celcius for 5 minutes. This results in a top PEG phase containing enriched plasma membranes and a bottom dextran layer containing enriched endomembranes. The top and bottom layers are exchanged as outlined, and then finally all peg layers and dextran layers are pooled for subsequent dilution and ultracentrifugation as outlined in the materials and methods. See materials and methods for buffer composition.

Figure 1-7: Chloroplast and Etioplast isolation

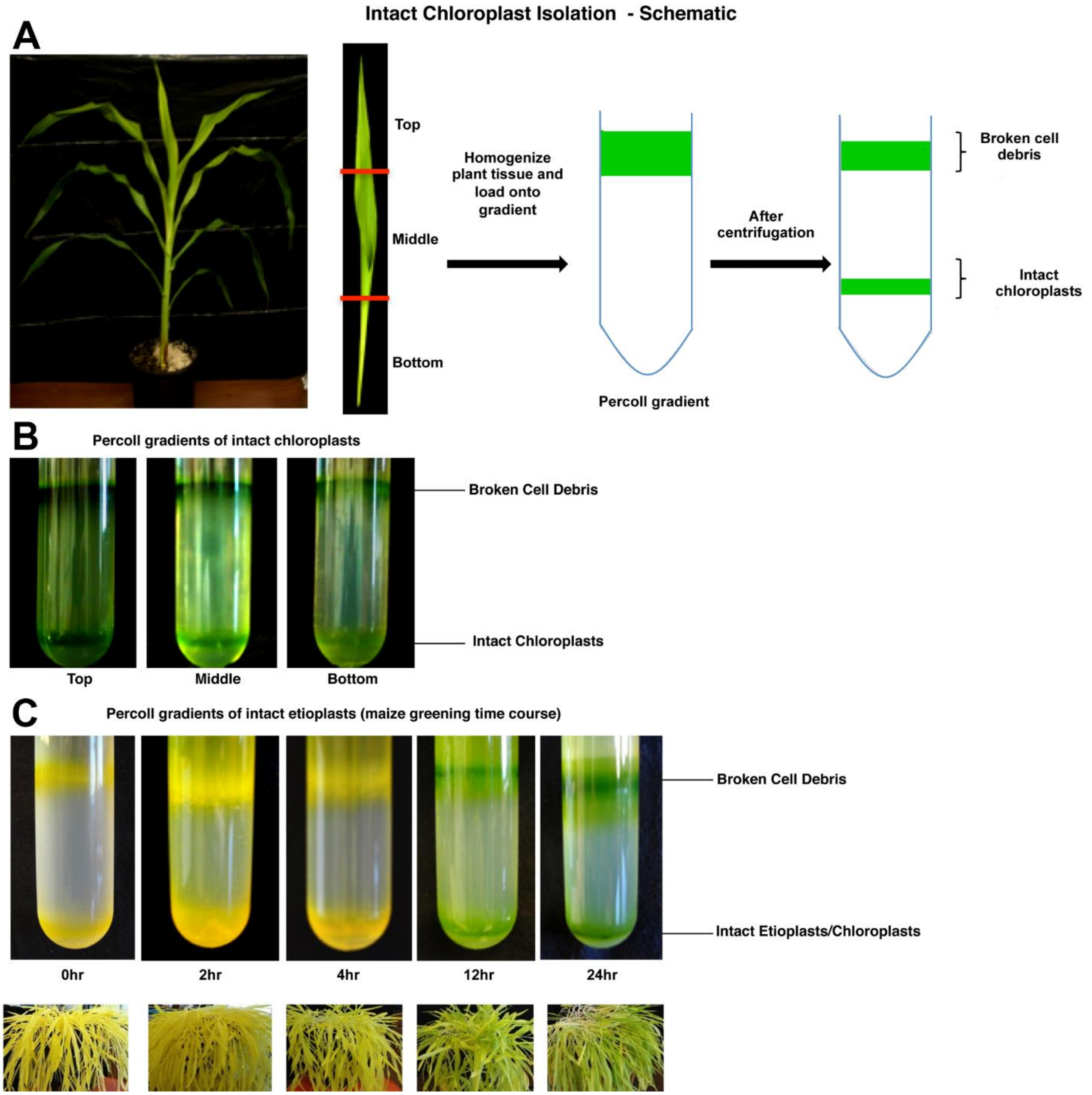

A
Intact Chloroplast Isolation - Schematic
Top
Middle
Bottom
Homogenize plant tissue and load onto gradient
Percoll gradient
After centrifugation
Broken cell debris
Intact chloroplasts
B
Percoll gradients of intact chloroplasts
Broken Cell Debris
Intact Chloroplasts
Top
Middle
Bottom
C
Percoll gradients of intact etioplasts (maize greening time course)
Broken Cell Debris
Intact Etioplasts/Chloroplasts
0hr
2hr
4hr
12hr
24hr

Figure 1-7: Chloroplast and Etioplast isolation. A: Chloroplast isolation via density-based centrifugation through percoll gradient. Chloroplasts were isolated from leaf eight of eight-leaf-stage plants. Leaf eight was divided into three zones as shown (top, middle, and bottom) and chloroplasts were isolated from each zone. **B:** Chloroplast isolation percoll gradients from leaf zones depicted in S2A showing layer of intact chloroplasts. **C:** Etioplast isolation percoll gradients from etioplasts at five stages of greening, from five durations of light exposure: 0 hr, 2 hr, 4 hr, 12 hr, and 24 hr. Bottom panel shows etiolated seedlings after each stage of light exposure.

Figure 1-8: Volcano plots of enrichment scores of proteins detected in organelle datasets: Volcano plots of proteins detected in combined organelle dataset. Fold enrichment represents proportion of normalized spectral counts for each protein observed in the organelle sample versus the source tissue. P-values derived from one-tailed fishers exact test comparing spectral counts observed in organelle versus source tissue for individual protein compare to spectral counts observed for all proteins. For proteins with negative fold-enrichment, depletion p-values were used; for proteins with positive fold-enrichment, enrichment p-values were used. Points colored based on gold standard true or false positives for target organelle. A-F: Volcano plots of all detected proteins in the organelle datasets from the plasma membrane (A), glyoxysome (B), mitochondria (C), and chloroplasts from leaf bottom (D), leaf middle (E), and leaf top (F). G-L: Volcano plots of detected proteins from organelle datasets with log-transformed p-values between 0 and 5 from the plasma membrane (G), glyoxysome (H), mitochondria (I), and chloroplasts from leaf bottom (J), leaf middle (K), and leaf top (L).

Figure 1-8: Volcano plots of enrichment scores of proteins detected in organelle datasets

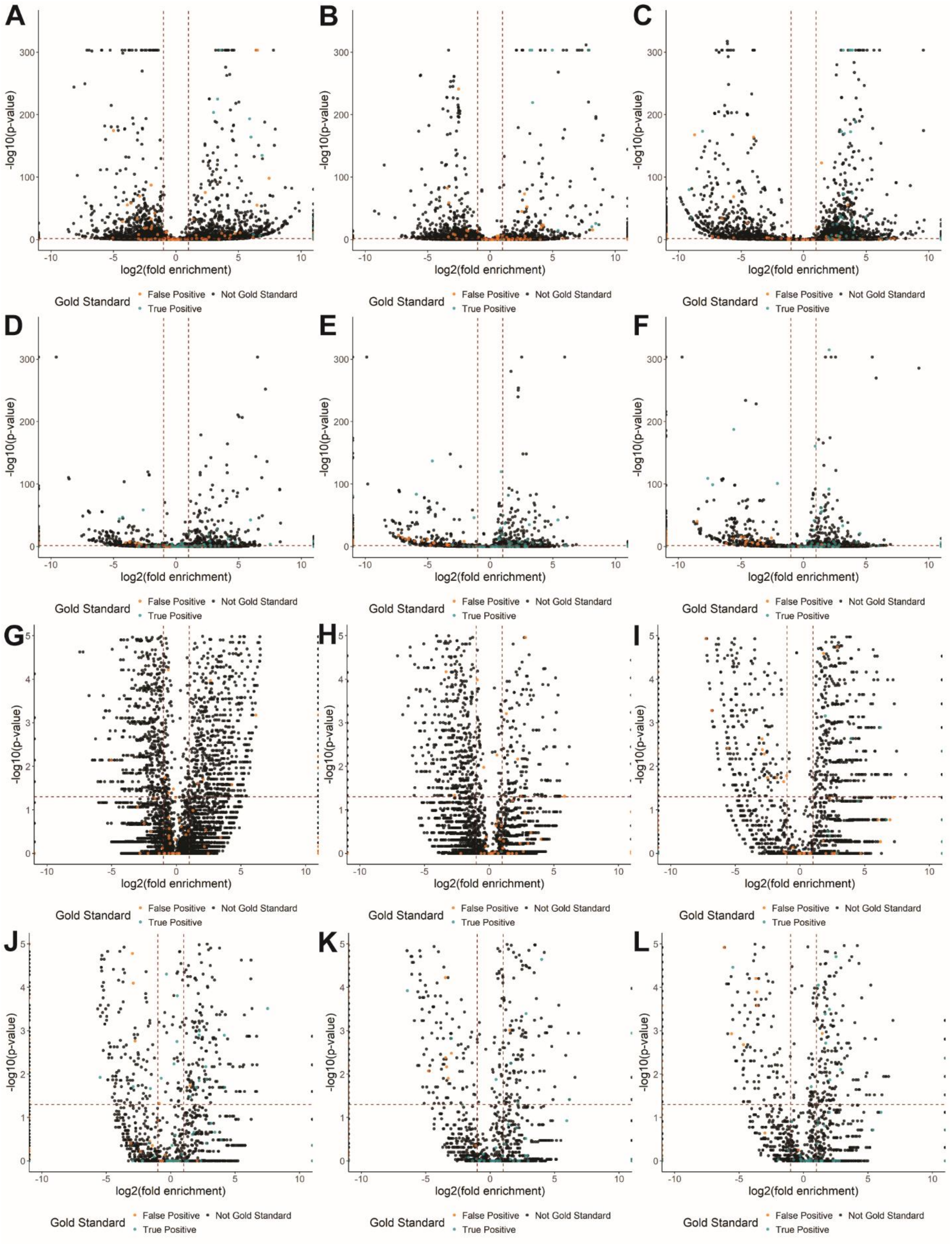


Figure 1-9: Organelle set size and true and false positive rates across fold-enrichment and p-value thresholds

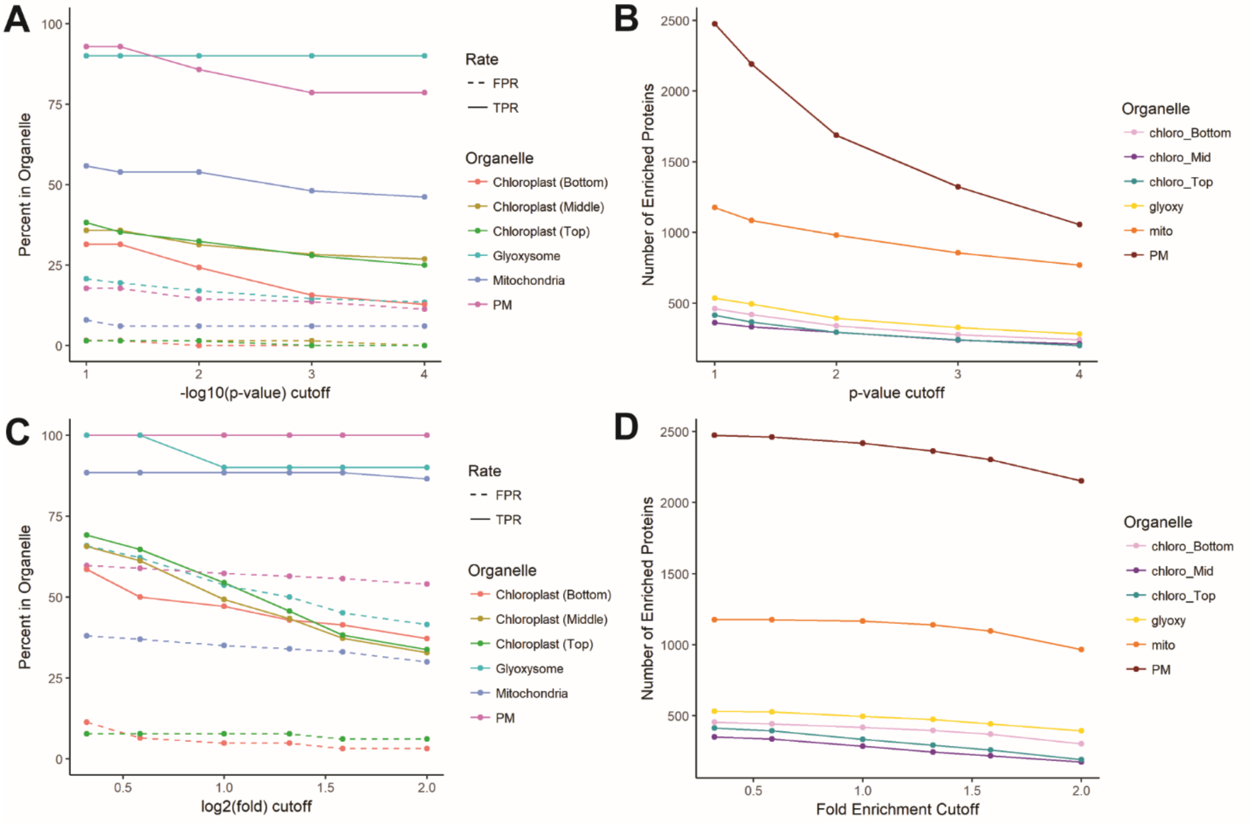


Figure 1-9: Organelle set size and true and false positive rates across fold-enrichment and p-value thresholds. A: True and false positive rates of organelle protein sets in comparison to gold standards across p-value thresholds. **B:** Organelle protein set size across p-value thresholds. **C:** True and false positive rates of organelle protein sets in comparison to gold standards across fold enrichment thresholds. **D:** Organelle protein set size across fold enrichment thresholds.

Figure 1-10: Overlap of GO CC annotated and previous organelle proteome protein sets

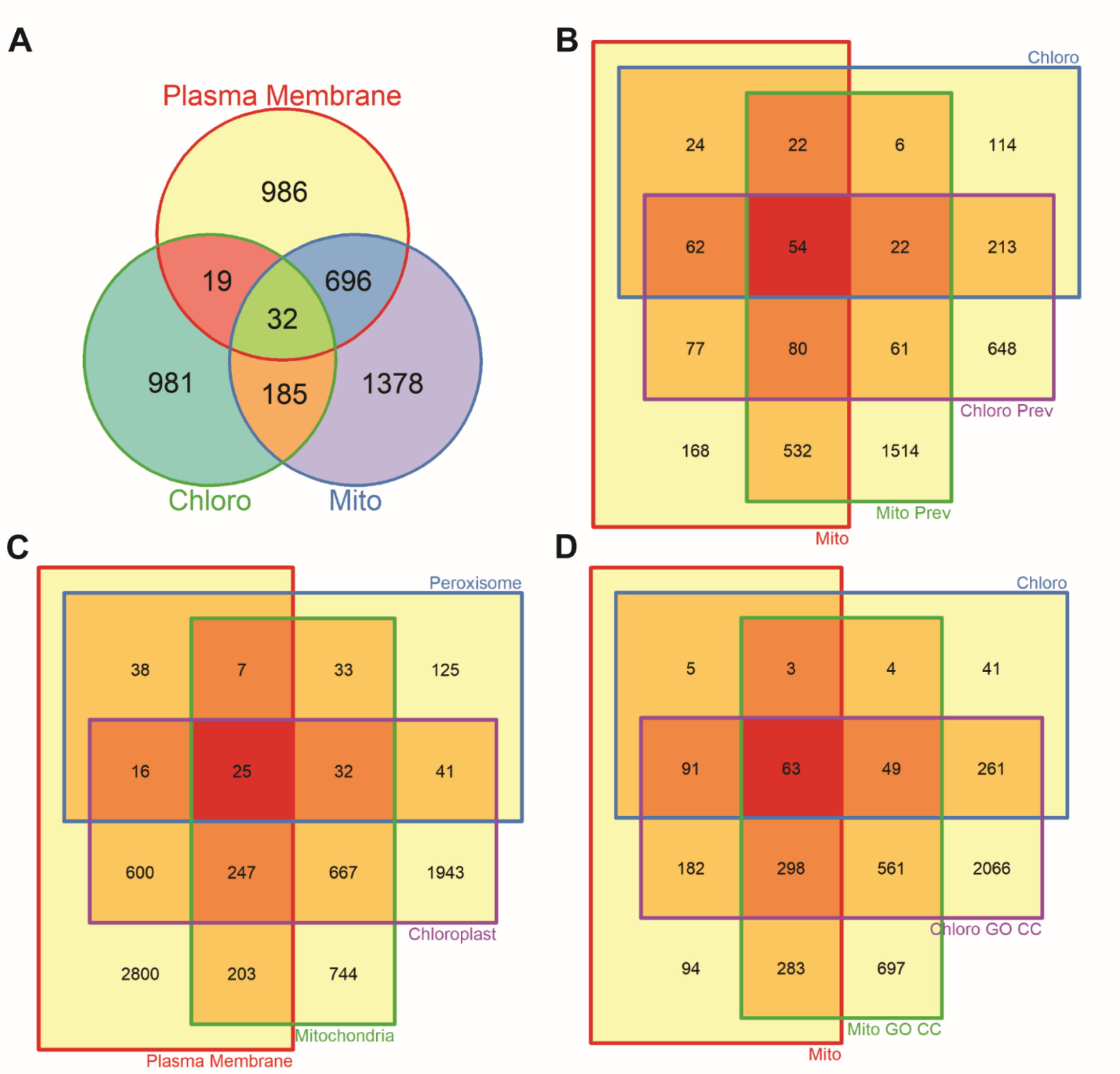


Figure 1-10: Overlap of GO CC annotated and previous organelle proteome protein sets. A: Venn diagram of overlap between previously published organelle protein sets. **B:** Venn diagram overlap between previously published mitochondrial and chloroplast protein sets and chloroplast and mitochondria enriched protein sets from this work. **C:** Venn diagram of overlap of GO CC annotated protein sets. **D:** Venn diagram of overlap of GO CC annotated mitochondria and chloroplast proteins and chloroplast and mitochondria enriched protein sets from this work.

Figure 1-11: Volcano plots of enrichment scores of previously published organelle proteome proteins. Volcano plots of proteins detected in combined organelle dataset. Fold enrichment represents proportion of normalized spectral counts for each protein observed in the organelle sample versus the source tissue. P-values derived from one-tailed fishers exact test comparing spectral counts observed in organelle versus source tissue for individual protein compare to spectral counts observed for all proteins. For proteins with negative fold-enrichment, depletion p-values were used; for proteins with positive fold-enrichment, enrichment p-values were used. Points colored based on presence or absence from previously published proteome. **A-B:** Volcano plots of all detected proteins in the organelle datasets from chloroplasts from the leaf bottom (**A**), and leaf top (**B**). **C-G:** Volcano plots of detected proteins with log-transformed p-values between 0 and 5 from the plasma membrane (**C**), mitochondria (**D**), and chloroplasts derived from the leaf bottom (**E**), leaf middle (**F**), and leaf top (**G**).

Figure 1-11: Volcano plots of enrichment scores of previously published organelle proteome proteins

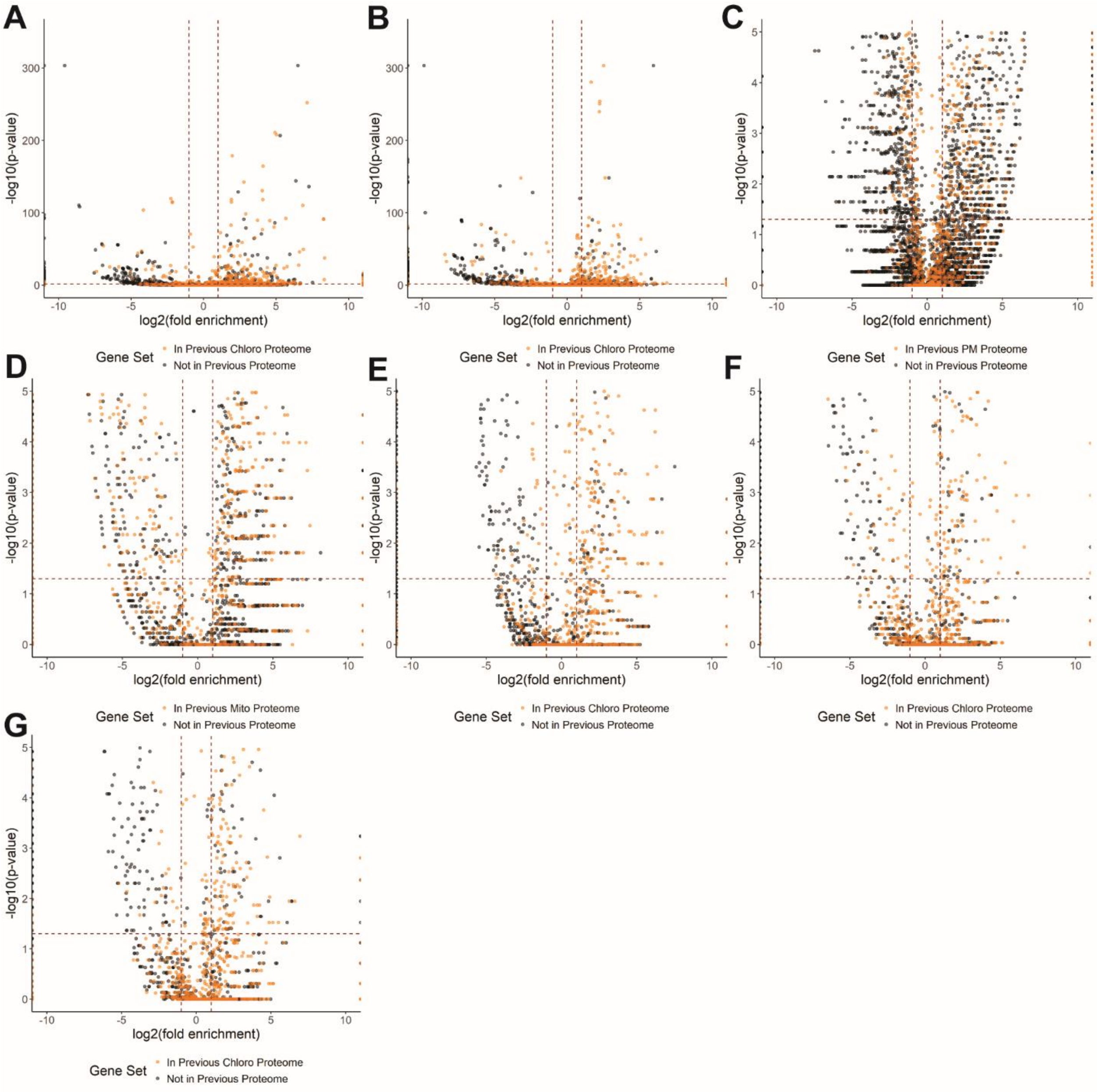


Figure 1-12: Volcano plots of enrichment scores of proteins annotated in GO CC target and contaminating organelle. Volcano plots of proteins detected in combined organelle dataset. Fold enrichment represents proportion of normalized spectral counts for each protein observed in the organelle sample versus the source tissue. P-values derived from one-tailed fishers exact test comparing spectral counts observed in organelle versus source tissue for individual protein compare to spectral counts observed for all proteins. For proteins with negative fold-enrichment, depletion p-values were used; for proteins with positive fold-enrichment, enrichment p-values were used. Points colored based annotation in GO CC compartments. **A-B:** Volcano plots of all detected proteins in the organelle datasets from chloroplasts from the leaf bottom (**A**), and leaf top (**B**). **C-H:** Volcano plots of detected proteins with log-transformed p-values between 0 and 5 from the plasma membrane (**C**), glyoxysome (**D**), mitochondria (**E**), and chloroplasts derived from the leaf bottom (**F**), leaf middle (**G**), and leaf top (**E**).

Figure 1-12: Volcano plots of enrichment scores of proteins annotated in GO CC target and contaminating organelle

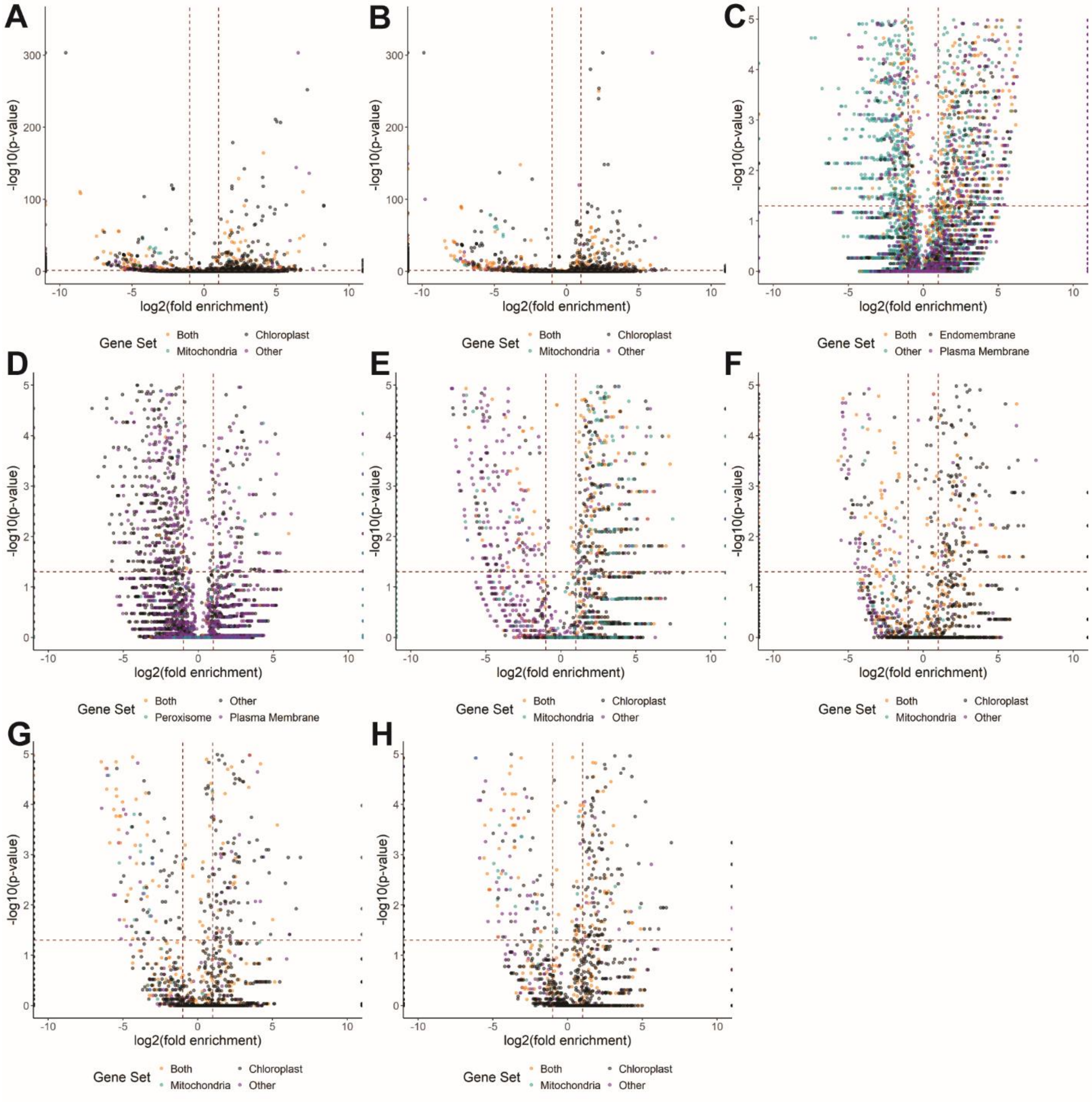


Figure 1-13: Localization of classical genes in this work, maizeGDB annotations, or literature

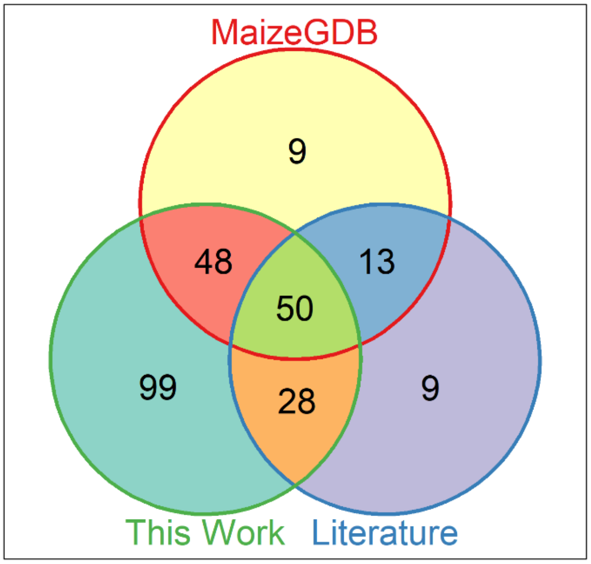


Figure 1-13: Localization of classical genes in this work, maizeGDB annotations, or literature. Venn diagram showing overlap of detected classical genes with localization information from this work, maizeGDB annotations, or the literature.

Figure 1-14: Percent of organelle enriched proteins in ARAMEMNON database of predicted membrane proteins

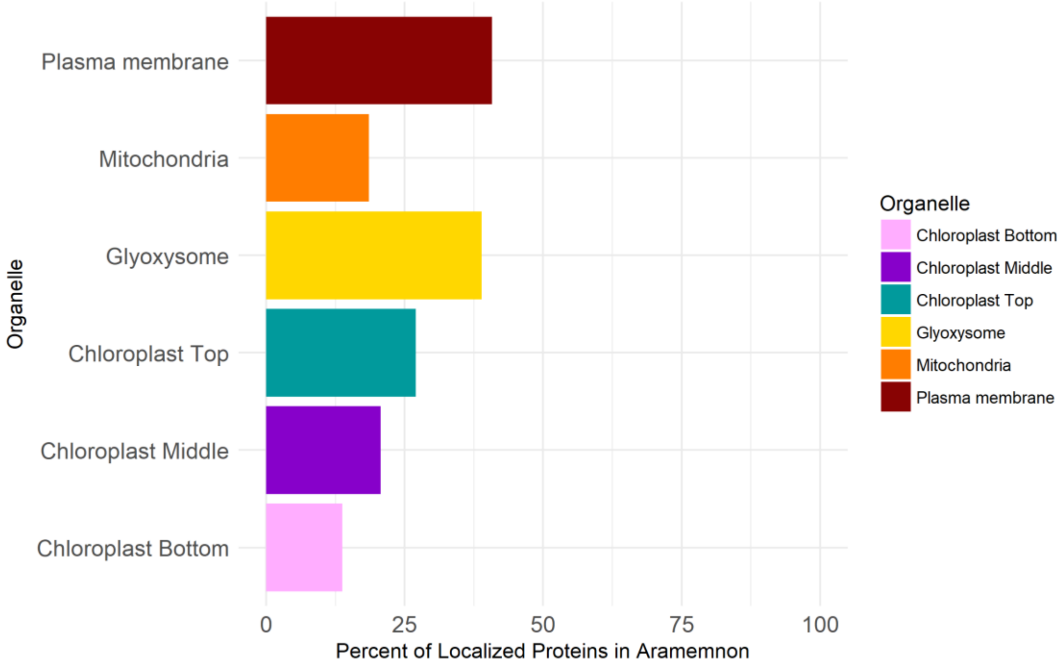


Figure 1-14: Percent of organelle enriched proteins in ARAMEMNON database of predicted membrane proteins.

Figure 1-15: Immunoblot detection of Catalase and Porin in glyoxysome isolation step samples

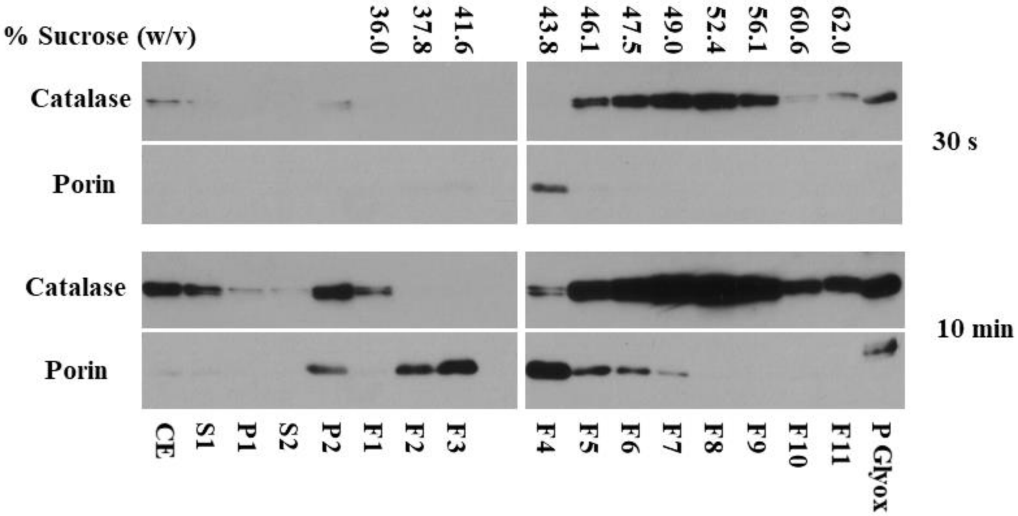


Figure 1-15: Immunoblot detection of Catalase and Porin in glyoxysome isolation step samples. Immunoblot detection of Catalase and Porin distribution in isolation step samples and samples from fractions of the sucrose gradient collected during a representative B73 maize glyoxysome preparation. BCA protein determinations using Bovine Serum Albumin as a standard were performed to facilitate methanol/chloroform precipitations of equal amounts of total protein from samples of crude extract (CE), supernatants 1 and 2 (S1 and S2, respectively) and pellets 1 and 2 (P1 and P2, respectively), taken during the glyoxysome isolation protocol, as well as fractions 1-11 (F1-F11) collected from the top (F1) to the bottom (F11) of the sucrose gradient. Precipitated proteins (25 µg per lane) were resuspended in sample buffer, separated on acrylamide gels,and electroblotted onto PVDF. Anti-Catalase immunoblotting confirmed the presence of glyoxysomes in the second pellet (P2) and between fractions 5-9 of the sucrose gradient. The sucrose concentration of equivalent fractions collected from a blank, balance gradient were determined by refractometer and are noted above. Anti-Porin immunoblotting verified that mitochondria co-purify with glyoxysomes in the second pellet and were distributed among fractions 1-7 of the sucrose gradient. Pumpkin glyoxysomes (P Glyox) served as a positive control for both immunoblots. Exposure times of 30s and 10 min were used to demonstrate the relative abundance of Catalase as compared to Porin in the glyoxysome-enriched fractions. To maximize purity and yield, fraction 8 was routinely harvested as the glyoxysome fraction.

Figure 1-16: Purity analysis of B73 maize glyoxysome preparations

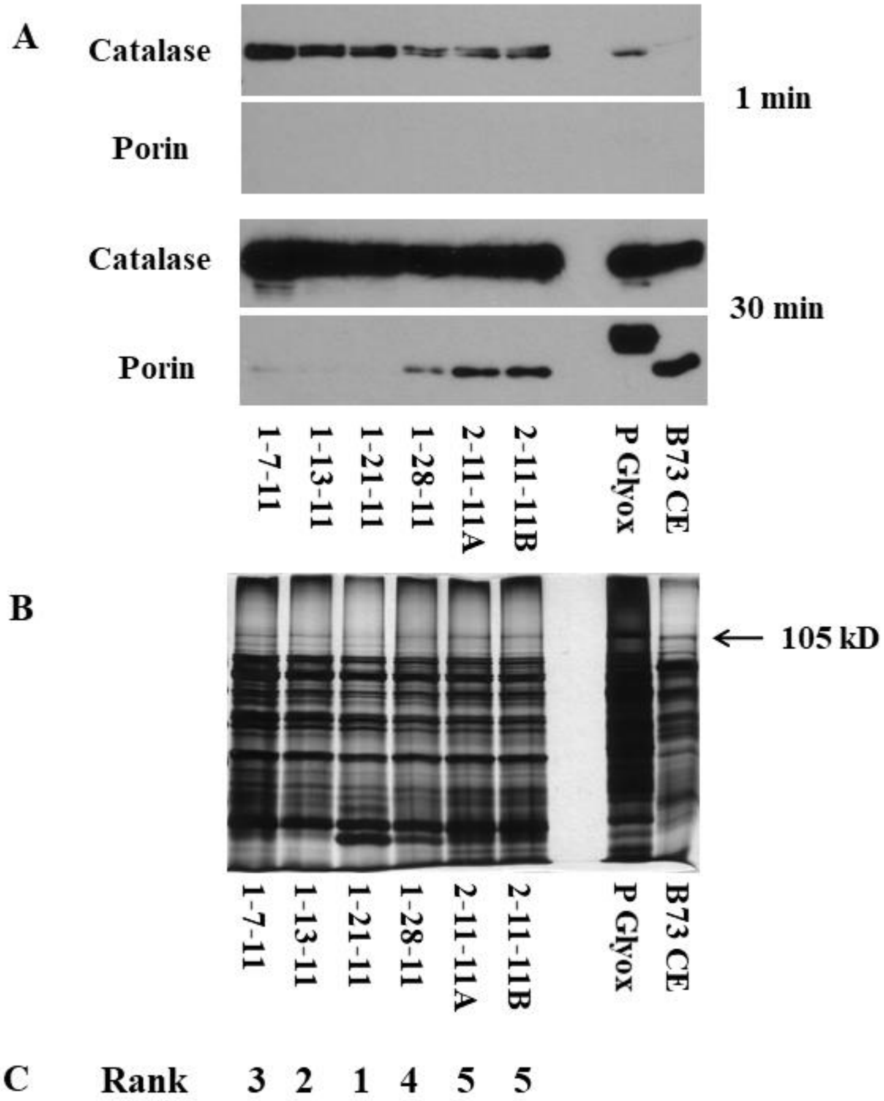


Figure 1-16: Purity analysis of B73 maize glyoxysome preparations. Glyoxysome proteins (24 µg per lane for immunoblots, 5 µg per lane for silver staining) were concentrated by methanol/chloroform precipitation, resuspended in sample buffer, and separated on acrylamide gels. Anti-Catalase immunoblotting (A) and silver staining verified the total amount of loaded proteins and confirmed the presence of glyoxysomes. Anti-Porin immunoblotting (A) and silver staining (B) were used to determine the relative amounts of mitochondrial Porin and P-protein, respectively. Exposure times of 1 and 30 minutes were used to demonstrate the relative abundance of Catalase as compared to Porin. Using the relative amounts of Porin, the preparations were ranked from 1-5 (C), with 1 being the most pure and 5 the least. Pumpkin glyoxysomes (P Glyox) and crude extract of B73 maize (B73 CE) served as positive controls for Catalase and Porin immunoblots and the presence of the 105 kD mitochondrial P-protein.

1.7.1 Supplemental Figure Legends

Table S1: Nonmodified proteomes of organelles and source tissues. Abundance values are spectral counts.

Table S2: Phosphopeptides from plasma membrane enriched samples. Abundance values are spectral counts

Table S3: Over-enrichment and depletion scores for detected proteins across target organelles. P-values and fold enrichment for protein over-enrichment and depletion in organelle fraction. P-values calculated using one-tailed fisher's exact test comparing protein spectral counts in organelle fraction versus intact organelle source tissue. Fold enrichment calculated as proportion of normalized spectral counts from organelle to source tissue.

Table S4: Gold standard organelle proteins with experimental evidence of localization in the literature. List of "gold standard" proteins with high confidence localization annotation in various organelles.

Table S5: Localization of detected maize classical genes. Subset of maize classical genes detected in this work with localization annotation from the literature and/or maizeGDB, if applicable.

Table S6: Protein isoform specific localization. Subset of Table S3 showing examples of isoform specific localization.

Table S7-S14: Gene ontology Cellular Compartment (GO CC) enrichment analysis for organelle over-enriched proteins

Table S7: GO CC Plasma membrane enriched proteins

Table S8: GO CC Glyoxysome enriched proteins

Table S9: GO CC Mitochondria enriched proteins

Table S10: GO CC Chloroplast (leaf bottom) enriched proteins

Table S11: GO CC Chloroplast (leaf middle) enriched proteins

Table S12: GO CC Chloroplast (leaf top) enriched proteins

Table S13: GO CC Chloroplast and mitochondria shared enriched proteins

Table S14: GO CC Plasma membrane and glyoxysome shared enriched proteins

Table S15:S20: Gene ontology cellular compartment (GO CC) enrichment analysis for organelle depleted proteins

Table S15: GO CC Plasma membrane depleted proteins

Table S16: GO CC Glyoxysome depleted proteins

Table S17: GO CC Mitochondria depleted proteins

Table S18: GO CC Chloroplast (leaf bottom) depleted proteins

Table S19: GO CC Chloroplast (leaf middle) depleted proteins

Table S20: GO CC Chloroplast (leaf top) depleted proteins

Table S21-S30: Gene ontology biological process (GO BP) enrichment analysis of organelle enriched proteins

Table S21: GO BP Plasma membrane enriched proteins

Table S22: GO BP Glyoxysome enriched proteins

Table S23: GO BP Mitochondria enriched proteins

Table S24: GO BP Chloroplast (leaf bottom) enriched proteins

Table S25: GO BP Chloroplast (leaf middle) enriched proteins

Table S26: GO BP Chloroplast (leaf top) enriched proteins

Table S27: GO BP Chloroplast (leaf bottom unique) enriched proteins

Table S28: GO BP Chloroplast (leaf top unique) enriched proteins

Table S29: GO BP Chloroplast and mitochondria shared proteins

Table S30: GO BP Plasma membrane and glyoxysome shared proteins

1.7.2 Supplemental Materials and Methods

Plasma Membrane Protein purification and separation: Plasma protein pellets are suspended in extraction buffer (8M Urea/Tris/phosphatase inhibitors, pH 7). After cysteine reduction and alkylation with Tris(2- carboxyethyl)phosphine (TCEP) and iodoacetamide, proteins are quantified using a Bradford assay. Protein solution is diluted 8 times to 1M urea and twice digested with Lys-C and trypsin. Digested peptides are purified on a Waters Sep-Pak C18 cartridges, eluted with 60% acetonitrile. An Agilent 1100 HPLC system is used to deliver a flow rate of 600 nL min-1 to a custom 3-phase capillary chromatography column through a splitter. Column phases are a 30 cm long reverse phase (RP1, 5 µm Zorbax SB-C18, Agilent), 8 cm long strong cation exchange (SCX, 3 µm PolySulfoethyl, PolyLC), and 40 cm long reverse phase 2 (RP2, 3.5 µm BEH C18, Waters), with the electrospray tip of the fused silica tubing pulled to a sharp tip (inner diameter <1 µm). Peptides are loaded onto RP1, and the 3 sections are joined and mounted on a custom electrospray adapter for on-line nested elutions, with a new set of columns is used for each LC-MS/MS analysis. Peptides are eluted from RP1 section to SCX section using a 0 to 80% acetonitrile gradient for 120 min, and then are fractionated by the SCX column section using a series of 19 step salt gradients of ammonium acetate over 20 min, followed by high-resolution reverse phase separation on the RP2 section of the column using an acetonitrile gradient of 0 to 80% for 210 min.

Data acquisition: Spectra are acquired on a Q-exactive-HF mass spectrometer (Thermo Electron Corporation, San Jose, CA) operated in positive ion mode with a source temperature of 275 °C and spray voltage of 3kV. Automated data-dependent acquisition was employed of the top 20 ions. The mass resolution is set at 60,000 for MS and 30,000 for MS/MS scans, respectively. Dynamic exclusion is used to improve the duty cycle.

Data analysis: The raw data are extracted and searched using Spectrum Mill vB.06 (Agilent Technologies). MS/MS spectra with a sequence tag length of 1 or less are considered to be poor spectra and were discarded. The remaining MS/MS spectra are searched against maize B73 v4 gene set. Search parameters are set to Spectrum Mill's default settings with the enzyme parameter limited to full tryptic peptides with a maximum mis-cleavage of 1. A 1:1 concatenated forward-reverse database will be constructed to calculate the false discovery rate (FDR). Cutoff scores are dynamically assigned to each dataset to obtain the false discovery rates (FDR) of 0.1% for peptides, and 1% for proteins. Proteins that share common peptides are grouped using principles of parsimony to address protein database redundancy.

1.7.3 Supplemental References

Ach, R.A., Durfee, T., Miller, A.B., Taranto, P., Hanley-Bowdoin, L., Zambryski, P.C., and Gruissem, W. (1997). RRB1 and RRB2 encode maize retinoblastoma-related proteins that interact with a plant D-type cyclin and geminivirus replication protein. Molecular and Cellular Biology *17*, 5077–5086.

Acosta, I.F., Laparra, H., Romero, S.P., Schmelz, E., Hamberg, M., Mottinger, J.P., Moreno, M.A., and Dellaporta, S.L. (2009). tasselseed1 is a lipoxygenase affecting jasmonic acid signaling in sex determination of maize. Science *323*, 262–265.

Alba, M.M., Culianez-Macia, F.A., Goday, A., Freire, M.A., Nadal, B., and Pages, M. (1994). The maize RNA-binding protein, MA16, is a nucleolar protein located in the dense fibrillar component. The Plant Journal *6*, 825–834.

Arcalis, E., Stadlmann, J., Marcel, S., Drakakaki, G., Winter, V., Rodriguez, J., Fischer, R., Altmann, F., and Stoger, E. (2010). The Changing Fate of a Secretory Glycoprotein in Developing Maize Endosperm. PLANT PHYSIOLOGY *153*, 693–702.

Asghar, R., Fenton, R.D., DeMason, D.A., and Close, T.J. (1994). Nuclear and cytoplasmic localization of maize embryo and aleurone dehydrin. Protoplasma *177*, 87–94.

Bailey, B.A., and Larson, R.L. (1991). Maize Microsomal Benzoxazinone N-Monooxygenase. PLANT PHYSIOLOGY *95*, 792–796.

Bailey-Serres, J., Tom, J., and Freeling, M. (1992). Expression and distribution of cytosolic 6-phosphogluconate dehydrogenase isozymes in maize. Biochemical Genetics *30*, 233–246.

Bass, H.W., Webster, Cecelia, OBrian, Gregory R., Roberts, Justin K.M., and Boston, Rebecca S. (1992). A Maize Ribosome-lnactivating Protein 1s Controlled by the Transcriptional Activator Opaque-2. The Plant Cell *4*, 225–234.

Baum, J.A., and Scandalios, J.G. (1979). Developmental Expression and Intracellular Localization of Superoxide Dismutases in Maize. Differentiation *13*, 133–140.

Beick, S., Schmitz-Linneweber, C., Williams-Carrier, R., Jensen, B., and Barkan, A. (2008). The Pentatricopeptide Repeat Protein PPR5 Stabilizes a Specific tRNA Precursor in Maize Chloroplasts. Molecular and Cellular Biology *28*, 5337–5347.

Bellmann, R., and Werr, W. (1992). Zmhox1a, the product of a novel maize homeobox gene, interacts with the Shrunken 26 bp feedback control element. The EMBO Journal *11*, 3367–3374.

Campos, Narciso, Schell, Jeff, and Palme, Klaus (1994). In Vitro Uptake and Processing of Maize Auxin-Binding Proteins by ER-Derived Microsomes. Plant and Cell Physiology *35*, 153–161.

Chang, C.-C., Sheen, J., Bligny, M., Niwa, Y., Lerbs-Mache, S., and Stern, D.B. (1999). Functional Analysis of Two Maize cDNAs Encoding T7-like RNA Polymerases. The Plant Cell *11*, 911–926.

Cook, W.B., and Walker, J.C. (1992). Identificiation of a maize nucleic acid-binding protein (NBP) belonging to a family of nuclear-encoded chloroplast proteins. Nucleic Acids Research *20*, 359–364.

da Costa e Silva, O., Lorbiecke, R., Garg, P., Müller, L., Waßmann, M., Lauert, P., Scanlon, M., Hsia, A.-P., Schnable, P.S., Krupinska, K., et al. (2004). The *Etched1* gene of *Zea mays* (L.) encodes a zinc ribbon protein that belongs to the transcriptionally active chromosome (TAC) of plastids and is similar to the transcription factor TFIIS. The Plant Journal *38*, 923–939.

Fisk, D.G., Walker, M.B., and Barkan, A. (1999). Molecular cloning of the maize gene crp1 reveals similarity between regulators of mitochondrial and chloroplast gene expression. The EMBO Journal *18*, 2621–2630.

Forestan, C., Farinati, S., Rouster, J., Lassagne, H., Lauria, M., Dal Ferro, N., and Varotto, S. (2018). Control of Maize Vegetative and Reproductive Development, Fertility, and rRNAs Silencing by *HISTONE DEACETYLASE 108*. Genetics *208*, 1443–1466.

Goodman, C.D., Casati, P., and Walbot, V. (2004). A Multidrug Resistance-Associated Protein Involved in Anthocyanin Transport in Zea mays. THE PLANT CELL ONLINE *16*, 1812–1826.

Gowri, G., and Campbell, W.H. (1989). cDNA Clones for Corn Leaf NADH:Nitrate Reductase and Chloroplast NAD(P)+:Glyceraldehyde-3-Phosphate Dehydrogenase’. Plant Physiology *90*, 792–798.

Grasser, K.D., Maier, U.G., Haass, M.M., and Feix, G. (1990). Maize high mobility group proteins bind to CCAAT and TATA boxes of a zein gene promoter. Journal of Biological Chemistry *265*, 4185–4188.

Gray, J., Janick-Buckner, D., Buckner, B., Close, P.S., and Johal, G.S. (2002). Light-Dependent Death of Maize lls1 Cells Is Mediated by Mature Chloroplasts. PLANT PHYSIOLOGY *130*, 1894–1907.

Han, C., Coe, E.H., and Martienssen, R.A. (1992). Molecular cloning and characterization of iojap (ij) a pattern striping gene of maize. The EMBO Journal *11*, 4037–4046.

Hinchliffe, D.J., and Kemp, J.D. (2002). ß-Zein protein bodies sequester and protect the 18 kDA δ-zein protein from degradation. Plant Science *163*, 741–752.

Huang, B., Hennen-Bierwagen, T.A., and Myers, A.M. (2014). Functions of Multiple Genes Encoding ADP-Glucose Pyrophosphorylase Subunits in Maize Endosperm, Embryo, and Leaf. PLANT PHYSIOLOGY *164*, 596–611.

Huang, M., Slewinski, T.L., Baker, R.F., Janick-Buckner, D., Buckner, B., Johal, G.S., and Braun, D.M. (2009). Camouflage Patterning in Maize Leaves Results from a Defect in Porphobilinogen Deaminase. Molecular Plant *2*, 773–789.

Jahrmann, T., Bastida, M., Pineda, M., Gasol, E., Ludevid, M.D., Palacín, M., and Puigdomènech, P. (2005). Studies on the function of TM20, a transmembrane protein present in cereal embryos. Planta *222*, 80–90.

Je, B.I., Xu, F., Wu, Q., Liu, L., Meeley, R., Gallagher, J.P., Corcilius, L., Payne, R.J., Bartlett, M.E., and Jackson, D. (2018). The CLAVATA receptor FASCIATED EAR2 responds to distinct CLE peptides by signaling through two downstream effectors. Elife *7*, e35673.

Jones, A.M., and Herman, E.M. (1993). KDEL-containing auxin-binding protein is secreted to the plasma membrane and cell wall. Plant Physiology *101*, 595–606.

Kelley, P.M., and Freeling, M. (1984). Anaerobic Expression of Maize Fructose-1,6-diphosphate aldolase. Journal of Biological Chemistry *259*, 14180–14183.

Kimata, Y., and Hase, T. (1989). Localization of Ferredoxin Isoproteins in Mesophyll and Bundle Sheath Cells in Maize Leaf. Plant Physiology *89*, 1193–1197.

Konishi, T., Shinohara, K., Yamada, K., and Sasaki, Y. (1996). Acetyl-CoA carboxylase in higher plants: most plants other than gramineae have both the prokaryotic and the eukaryotic forms of this enzyme. Plant and Cell Physiology *37*, 117–122.

Kopylov, M., Bass, H.W., and Stroupe, M.E. (2015). The Maize (*Zea mays* L.) *Nucleoside Diphosphate Kinase1 (ZmNDPK1)* Gene Encodes a Human NM23-H2 Homologue That Binds and Stabilizes G-Quadruplex DNA. Biochemistry *54*, 1743–1757.

Kroeger, T.S., Watkins, K.P., Friso, G., van Wijk, K.J., and Barkan, A. (2009). A plant-specific RNA-binding domain revealed through analysis of chloroplast group II intron splicing. Proceedings of the National Academy of Sciences *106*, 4537–4542.

Lal, S.K., Kelley, P.M., and Elthon, T.F. (1994). Purification and differential expression of enolase from maize. Physiologia Plantarum *91*, 587–592.

Łebska, M., Ciesielski, A., Szymona, L., Godecka, L., Lewandowska-Gnatowska, E., Szczegielniak, J., and Muszyńska, G. (2010). Phosphorylation of Maize Eukaryotic Translation Initiation Factor 5A (eIF5A) by Casein Kinase 2: IDENTIFICATION OF PHOSPHORYLATED RESIDUE AND INFLUENCE ON INTRACELLULAR LOCALIZATION OF eIF5A. Journal of Biological Chemistry *285*, 6217–6226.

Li, Q., Eichten, S.R., Hermanson, P.J., Zaunbrecher, V.M., Song, J., Wendt, J., Rosenbaum, H., Madzima, T.F., Sloan, A.E., Huang, J., et al. (2014). Genetic Perturbation of the Maize Methylome. The Plant Cell Online *26*, 4602–4616.

Liu, F., Cui, X., Horner, H.T., Weiner, H., and Schnable, P.S. (2001). Mitochondrial aldehyde dehydrogenase activity is required for male fertility in maize. The Plant Cell *13*, 1063–1078.

Liu, F., Romanova, N., Lee, E.A., Ahmed, R., Evans, M., Gilbert, E.P., Morell, M.K., Emes, M.J., and Tetlow, I.J. (2012). Glucan affinity of starch synthase IIa determines binding of starch synthase I and starch-branching enzyme IIb to starch granules. Biochemical Journal *448*, 373–387.

Ma, Z., and Dooner, H.K. (2004). A mutation in the nuclear-encoded plastid ribosomal protein S9 leads to early embryo lethality in maize. The Plant Journal *37*, 92–103.

McMillin, D.E., Roupakias, D.G., and Scandalios, J.G. (1979). Chromosomal location of two mitochondrial malate dehydrogenase structural genes in Zea mays using trisomics and BA translocations. Genetics *92*, 1241–1250.

Mohanty, A., Luo, A., DeBlasio, S., Ling, X., Yang, Y., Tuthill, D.E., Williams, K.E., Hill, D., Zadrozny, T., Chan, A., et al. (2009). Advancing Cell Biology and Functional Genomics in Maize Using Fluorescent Protein-Tagge Lines. Plant Physiology *149*, 601–605.

Napier, R.M., Trueman, S., Henderson, J., Boyce, J.M., Hawes, C., Fricker, M.D., and Venis, M.A. (1995). Purification, sequencing and functions of calreticulin from maize. Journal of Experimental Botany *46*, 1603–1613.

Nieto-Sotelo, J., Martinez, L.M., Ponce, G., Cassab, G.I., Alagon, A., Meeley, R.B., Ribaut, J.-M., and Yang, R. (2002). Maize HSP101 Plays Important Roles in Both Induced and Basal Thermotolerance and Primary Root Growth. The Plant Cell *14*, 1621–1633.

Nikus, J., Daniel, G., and Jonsson, L.M. (2001). Subcellular localization of β-glucosidase in rye, maize and wheat seedlings. Physiologia Plantarum *111*, 466–472.

Ostheimer, G.J., Rojas, M., Hadjivassiliou, H., and Barkan, A. (2006). Formation of the CRS2-CAF2 Group II Intron Splicing Complex Is Mediated by a 22-Amino Acid Motif in the COOH-terminal Region of CAF2. Journal of Biological Chemistry *281*, 4732–4738.

Peracchia, G., Jensen, A.B., Culiáñez-Macià, F.A., Grosset, J., Goday, A., Issinger, O.-G., and Pagès, M. (1999). Characterization, subcellular localization and nuclear targeting of casein kinase 2 from Zea mays. Plant Molecular Biology *40*, 199–211.

Perrot-Rechenmann, C., Vidal, J., Brulfert, J., Burlet, A., and Gadal, P. (1982). A comparative immunocytochemical localization study of phosphoenolpyruvate carboxylase in leaves of higher plants. Planta *155*, 24–30.

Prasad, Tottempudi K., and Steward, Cecil R. (1992). cDNA clones encoding Arabidopsis thaliana and Zea mays mitochondrial chaperonin HSP60 and gene expression during seed germination and heat shock. Plant Molecular Biology *18*, 873–885.

Pysh, L.D., Aukerman, M.J., and Schmidt, R.J. (1993). OHP1: A maize basic domain/leucine zipper protein that interacts with opaque2. The Plant Cell *5*, 227–236.

Riera, M., Irar, S., Vélez-Bermúdez, I.C., Carretero-Paulet, L., Lumbreras, V., and Pagès, M. (2011). Role of Plant-Specific N-Terminal Domain of Maize CK2β1 Subunit in CK2β Functions and Holoenzyme Regulation. PLoS ONE *6*, e21909.

Roth, R., Lisa, H., Brutnell, T.P., and Langdale, J.A. (1996). bundle sheath defective2, a Mutation that Disrupts the Coordinated Development of Bundle Sheath and Mesophyll Cells in the Maize Leaf. The Plant Cell *8*, 915–927.

Roy, L.M., and Barkan, A. (1998). A SecY Homologue Is Required for the Elaboration of the Chloroplast Thylakoid Membrane and for Normal Chloroplast Gene Expression. The Journal of Cell Biology *141*, 385–395.

Sakakibara, H., Fujii, K., and Sugiyama, T. (1995). Isolation and Characterization of a cDNA That Encodes Maize Glutamate Dehydrogenase. Plant and Cell Physiology *36*, 789–797.

Santi, S., Locci, G., Monte, R., Pinton, R., and Varanini, Z. (2003). Induction of nitrate uptake in maize roots: expression of a putative high-affinity nitrate transporter and plasma membrane H+-ATPase isoforms. Journal of Experimental Botany *54*, 1851–1864.

Scandalios, J.G. (1974). Subcellular localization of catalase variants coded by two genetic loci during maize development. Journal of Heredity *65*, 28–32.

Scandalios, J.G., Tong, W.-F., and Roupakias, D.G. (1980). Cat3, a third gene locus coding for a tissue-specific catalase in maize: genetics, intracellular location, and some biochemical properties. Molecular and General Genetics MGG *179*, 33–41.

Sheen, J.-Y., and Bogorad, L. (1986). Differential expression of six light-harvesting chlorophyll a/b binding protein genes in maize leaf cell types. Proceedings of the National Academy of Sciences *83*, 7811–7815.

Skadsen, R.W., and Scandalios, J.G. (1987). Translational control of photo-induced expression of the Cat2 catalase gene during leaf development in maize. Proceedings of the National Academy of Sciences *84*, 2785–2789.

Subbaiah, C.C., and Sachs, M.M. (2000). Maize cap1 encodes a novel SERCA-type Calcium ATPase with a calmodulin-Binding Domain. Journal of Biological Chemistry *275*, 21678–21687.

Subbaiah, C.C., Palaniappan, A., Duncan, K., Rhoads, D.M., Huber, S.C., and Sachs, M.M. (2006). Mitochondrial Localization and Putative Signaling Function of Sucrose Synthase in Maize. Journal of Biological Chemistry *281*, 15625–15635.

Sullivan, T.D., and Kaneko, Y. (1995). The maize brittle1 gene encodes amyloplast membrane polypeptides. Planta *196*, 477–484.

Theodoris, G., Inada, N., and Freeling, M. (2003). Conservation and molecular dissection of ROUGH SHEATH2 and ASYMMETRIC LEAVES1 function in leaf development. Proceedings of the National Academy of Sciences *100*, 6837–6842.

Tian, Q., Olsen, L., Sun, B., Lid, S.E., Brown, R.C., Lemmon, B.E., Fosnes, K., Gruis, D. (Fred), Opsahl-Sorteberg, H.-G., Otegui, M.S., et al. (2007). Subcellular Localization and Functional Domain Studies of DEFECTIVE KERNEL1 in Maize and *Arabidopsis* Suggest a Model for Aleurone Cell Fate Specification Involving CRINKLY4 and SUPERNUMERARY ALEURONE LAYER1. The Plant Cell *19*, 3127–3145.

Ting, J.T.L., Lee, keunmyoung, Ratnayake, C., Platt, K.A., Balsamo, R.A., and Huang, A.H.C. (1996). Oleosin genes in maize kernels having diverse oil contents are constitutively expressed independent of oil contents. Planta *199*, 158–165.

de Vetten, N.C., Lu, G., and Feri, R.J. (1992). A maize protein associated with the G-box binding complex has homology to brain regulatory proteins. The Plant Cell *4*, 1295–1307.

Vladimirou, E., Li, M., Aldridge, C.P., Frigerio, L., Kirkilionis, M., and Robinson, C. (2009). Diffusion of a membrane protein, Tat subunit Hcf106, is highly restricted within the chloroplast thylakoid network. FEBS Letters *583*, 3690–3696.

Wendel, J.F., Stuber, C.W., Goodman, M.M., and Beckett, J.B. (1989). Duplicated plastid and triplicated cytosolic isozymes of triosephosphate isomerase in maize (Zea mays L.). Journal of Heredity *80*, 218–228.

Wright, A.J., Gallagher, K., and Smith, L.G. (2009). discordia1 and alternative discordia1 Function Redundantly at the Cortical Division Site to Promote Preprophase Band Formation and Orient Division Planes in Maize. The Plant Cell *21*, 234–247.

Xu, X., Dietrich, C.R., Lessire, R., Nikolau, B.J., and Schnable, P.S. (2002). The Endoplasmic Reticulum-Associated Maize GL8 Protein Is a Component of the Acyl-Coenzyme A Elongase Involved in the Production of Cuticular Waxes. PLANT PHYSIOLOGY *128*, 924–934.

Yamagata, T., Kato, H., Kuroda, S., Abe, S., and Davies, E. (2003). Uncleaved legumin in developing maize endosperm: identification, accumulation and putative subcellular localization. Journal of Experimental Botany *54*, 913–922.

Yang, N.-S., and Scandalios, J.G. (1974). Purification and biochemical properties of genetically defined malate dehydrogenase in maize. Archives of Biochemistry and Biophysics *161*, 335–353.

Yu, Y., Mu, H.H., Mu-Forster, C., and Wasserman, B.P. (1998). Polypeptides of the maize amyloplast stroma: stromal localization of starch-biosynthetic enzymes and identification of an 81-kilodalton amyloplast stromal heat-shock cognate. Plant Physiology *116*, 1451–1460.

Zelazny, E., Borst, J.W., Muylaert, M., Batoko, H., Hemminga, M.A., and Chaumont, F. (2007). FRET imaging in living maize cells reveals that plasma membrane aquaporins interact to regulate their subcellular localization. Proceedings of the National Academy of Sciences *104*, 12359–12364.

Zhang, Z., Yang, J., and Wu, Y. (2015). Transcriptional Regulation of Zein Gene Expression in Maize through the Additive and Synergistic Action of opaque2, Prolamine-Box Binding Factor, and O2 Heterodimerizing Proteins. The Plant Cell *27*, 1162–1172.

Zhu, D., and Scandalios, J.G. (1995). The maize mitohondrial MnSODS encoded by multiple genes are localized in the mitochondrial matrix of transformed yeast cells. Free Radical Biology & Medicine *18*, 179–183.

CHAPTER 2

Discovery of species-specific expressible genes via machine learning with omics data

2.1 Introduction

Determining which predicted gene models produce functional products is an exciting challenge in genome-wide biology. While the number of predicted gene models from plant genomes can exceed 60,000, the plant research community has detected transcript products from only a subset of these genes, and an even smaller subset have detected protein products. Gathering direct experimental evidence of proteins from all expressible genes is challenging, due to technical difficulties with detection of low abundance or low solubility proteins and the infeasibility of profiling all cell types under all environmental conditions or combination of conditions. There is a crucial need to annotate which portions of plant genomes are likely to produce functional products and what genomic features determine the ability of a gene to be expressed at the transcript or protein level. Gene homology and expression evidence is often used to curate a high confidence group of genes from full predicted gene model sets, e.g. the filtered gene set of maize, or the high confidence gene set of sorghum. An open question is whether genes outside of these curated high confidence sets are expressible at the protein level. It is also unknown what portion of the high confidence gene models without previously detected protein products are expressible. The gene model set of a species serves as the search space for genome-wide technologies such as proteomics. Inclusion of extraneous gene models increases the false discovery rate, while exclusion of true gene models prevents discovery of products of those genes.

The lineages that gave rise to maize (*Zea mays)* and sorghum (*Sorghum bicolor*) diverged approximately twelve million years ago, after which the maize lineage underwent an additional whole genome duplication event (Swigoňová et al., 2004). This whole genome duplication event resulted in two subgenomes of maize, which can be defined in relation to sorghum. One of these two subgenomes is more highly expressed at the RNA level and has retained more genes (Schnable et al., 2011). It is unknown whether expression at the protein level is also biased towards one of the two subgenomes. Genes which have both sequence and colinear context conservation between two species are known as syntenic genes. Nonsyntenic genes emerged after the split between sorghum and maize lineages and are likely not represented in the ancestral genome. While the most of maize “working gene set” genes are nonsyntenic, maize genes which were identified by mutant phenotype and genes which are overrepresented in the maize literature (“Classical” genes) are enriched for syntenic genes (Schnable and Freeling, 2011). Previous work provided potential molecular explanations for this phenomenon: the majority of maize genes with detected protein products are syntenic, and the gene bodies of maize syntenic genes are hypomethylated, compared to nonsyntenic genes (Eichten et al., 2011; Walley et al., 2016). The relationship between gene methylation and synteny, and between synteny and express-ability, suggests a possible relationship between methylation and express-ability. Due to the close evolutionary relationship between the sorghum and maize lineages, the gene contents between these two species are similar, with most sorghum genes homologous to two or more genes of maize. Defining differences in the expressible sets may define the underlying proteomic differences leading to species specific traits.

Prediction of high confidence gene models has been previously performed in sorghum using a combination of transcript expression data, synteny information, and DNA methylation data to train a J48 decision tree classifier (Olson et al., 2014). This classifier required all three data types for accurate gene classification. Previously, it has been shown that accurate classification of maize gene express-ability is possible using only DNA methylation data (Sartor et al., 2018 [under review]). 41,056 and 32,979 maize genes are predicted to be expressible at the RNA and the protein level, respectively. As the total number genes and the proportion of syntenic genes differs significantly between sorghum and maize, it is unknown if the proportion or regulation of express-ability differs for this closely related grass species. We report here accurate gene express-ability classification using only DNA methylation data for the sorghum genome. Leveraging both the previously defined expressible gene sets of maize and the syntenic orthologs between grass species allowed for comparison not just of the gene contents of sorghum and maize, but of the protein expressible subsets of both genomes.

2.2 Results

2.2.1 Identification of protein expressible genes of sorghum stems under fungal elicitation

Protein was extracted from slit sorghum stems treated with heat-killed *Fusarium venenatum* or water-soaked cloth. Protein was extracted from flash frozen stem samples after 44 hours of treatment. Tryptic digested peptides were labeled with Tandem Mass Tags (TMT) isobaric tags for relative quantitation (Thompson et al., 2003). Labeled peptides were separated and analyzed by HPLC MS/MS as described in the supplemental methods. Mass spectra were searched against a forward:reverse decoy database and the false discovery rate was constrained at 0.1% and 1% at the peptide and protein levels, respectively (Table S31). Positively identified proteins served as the positive class for protein express-ability classifier training.

2.2.2 Creation of sorghum gene express-ability classifiers

Sorghum gene models from version 1.4 of the sorghum genome (Paterson et al., 2009) were annotated into groups based on expression at both transcript and protein levels. For the first classifier, which sought to classify genes as express-able or silent at the RNA level (ERC), genes were classified as highly expressed (average RPKM>=1; positive case), or non-detected (no reported RNA; negative case) using previously published RNAseq data (Dugas et al., 2011). For the second classifier, for classification of genes as expressible or silent at the protein level (EPC), genes were grouped into the non-expressed (negative case) category if they had no detectable protein or RNA, and the expressed (positive case) category if they had both detected protein and high RNA expression. Thus we had distinct gene groupings for two models, each involving different positive (expressible) and negative (non-expressible) cases.

Each annotated sorghum gene model was divided into five equally sized “bins”. Previously published bisulfite sequencing data from both sorghum shoots and roots were mapped to these bins (Turco et al., 2017). Each sequence context, CG, CHG, or CHH, and bisulfite sequencing sample tissue were mapped separately to each gene

Figure 2-1: Creation of gene express-ability classifier

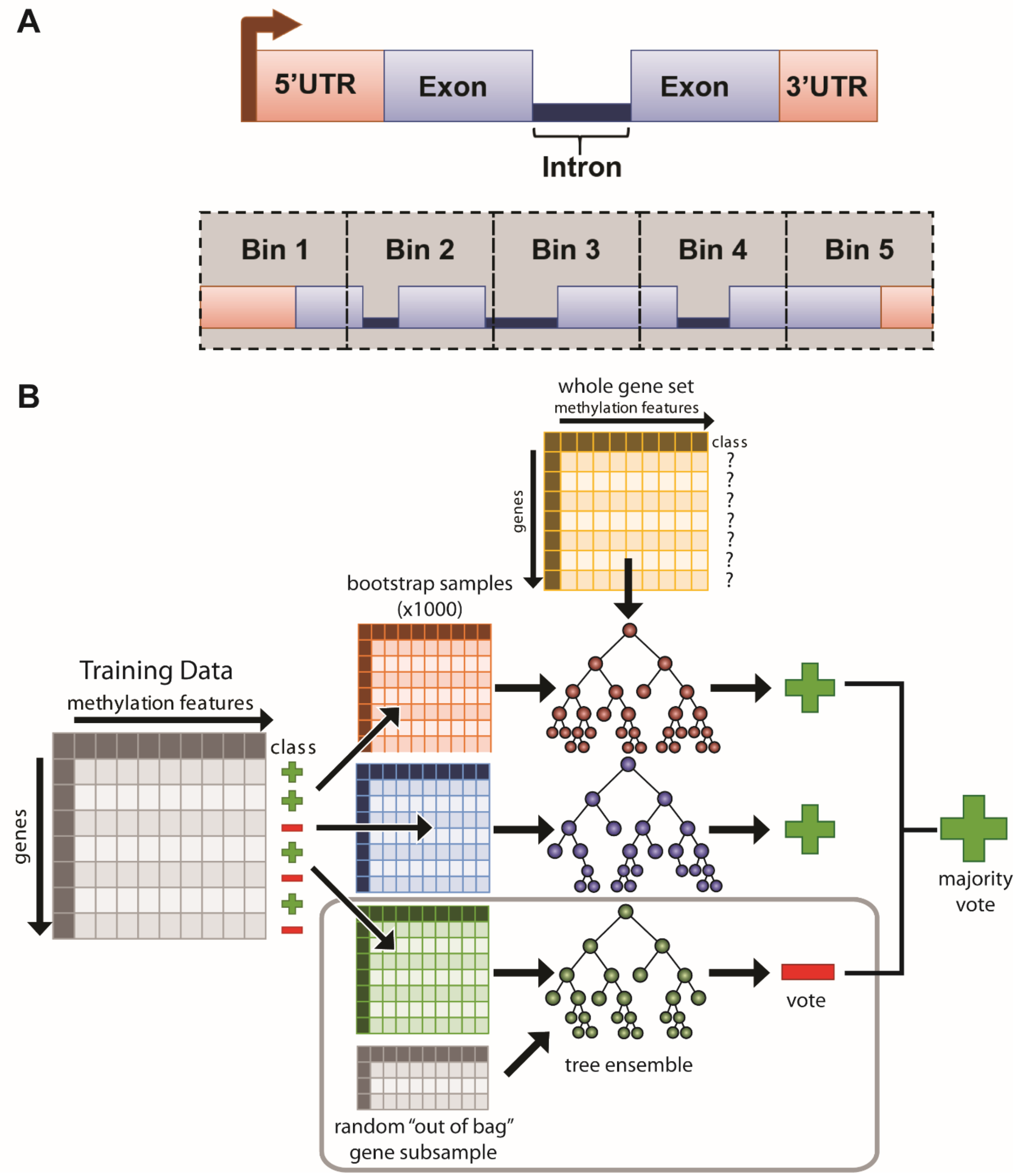


Figure 2-1: Creation of gene express-ability classifier. A: Definition of genomic regions used for calculation of methylation features for gene express-ability classifier. Full gene length was divided into five equal bins. Methylation levels were calculated for entire bins and for each genomic feature within each bin. **B:** Workflow for building, testing, and applying express-ability classifiers.

bin. The methylation level, as the proportion of methylated cytosines to unmethylated cytosines, was calculated for each feature. The average methylation level was also calculated for introns and exons within these bins (Fig. 2-1). This resulted in ninety DNA methylation features for each gene model (Table S32). For each classifier, these DNA methylation features, along with the positive and negative expression classifications described above, were used to train a random forest classification algorithm (Fig. 2-1) (Breiman, 2001).

Model performance was measured using the random out-of-bag cross-validation prediction of express-ability, compared to genes with known class labels. The proportion of positive votes to negative votes was compared to the known class labels to calculate the number of false positive and false negatives for our models. These results were plotted in receiver operating characteristic (“ROC”) and precision-recall (“PR”) curves (Fig. 2-2A-B). For models with perfect performance, we expect an area under the curve of one for both the ROC and PR curves, whereas a model which performs no better than random will have a ROC area under the curve (AUC) of 0.5. The AUC for both the ROC and PR curves for the ERC and EPC exceeded 0.95.

2.2.3 Model feature importance

ERC and EPC model feature importance was calculated from the average decrease in accuracy upon random permutation of each individual variable as implemented in the random forest R package (Breiman, 2001). The normalized sum of this value for a given set of methylation features was used to calculate the unsigned feature importance of each feature type used for model building (Fig. 2-3A). Across both models, features derived from methylation data from the root and shoot had similar

Figure 2-2: Classifier testing: Receiver Operating Characteristic (ROC) and precision recall (PR) curves for classifier models.

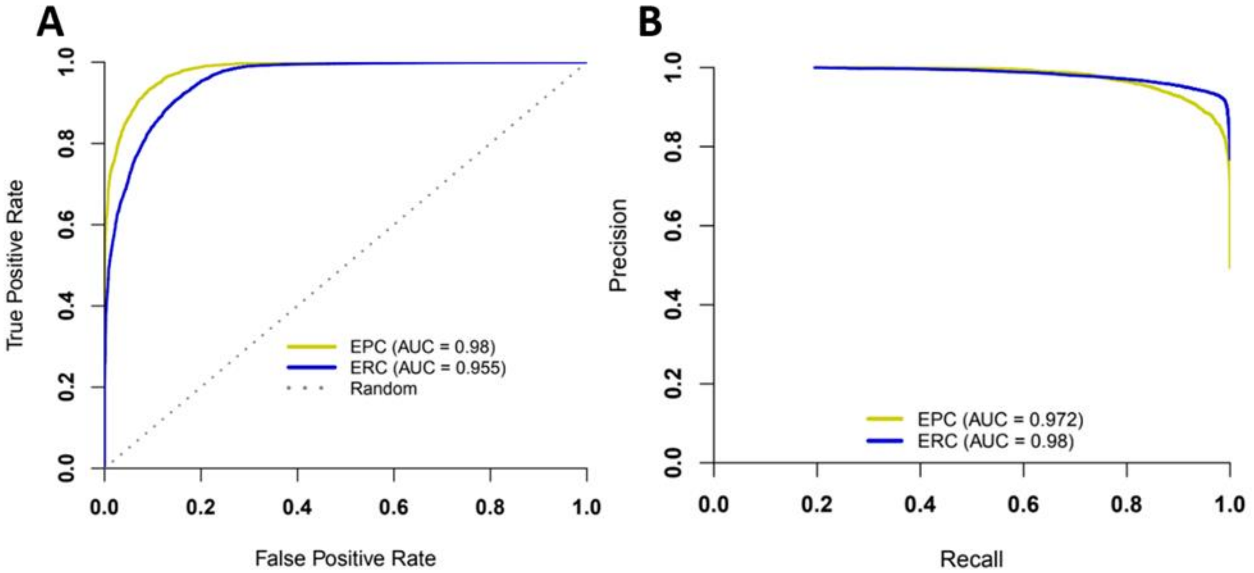


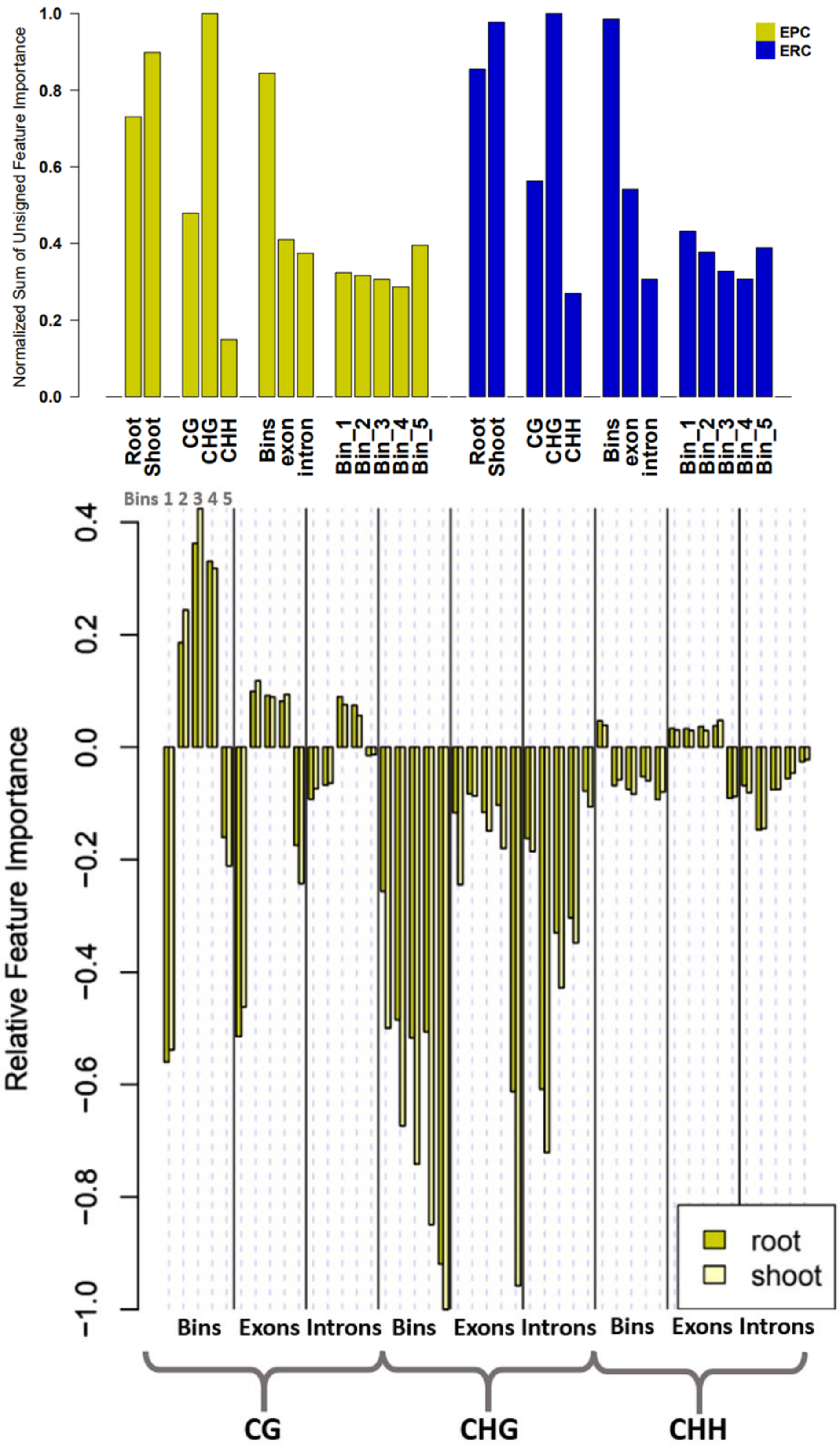
Figure 2-2: Classifier testing: Receiver Operating Characteristic (ROC) and precision recall (PR) curves for classifier models. ROC (**A**) and PR (**B**) curves were plotted using successful and unsuccessful classifications from random forest model out-of-bag cross-validation votes compared to known classifications. EPC curves shown in gold, ERC curves shown in blue.

importance, with the shoot-derived methylation data slightly outperforming the root. EPC and ERC had extremely similar importance across features, with CG and CHG methylation having higher feature importance than CHH methylation, and the 5' and 3' gene ends having higher feature importance than the middle of the gene. For the ERC, methylation of the 5' end of the gene was more important to model performance than methylation of the 3' end, the converse was true for the EPC. Overall, methylation level of the equal sized bins was more important than methylation of either the exon or intron features.

Since the model feature importance is by nature unsigned, to discover whether high or low methylation of a given feature is associated with expression, the sign of the feature importance was assigned by calculating a t-statistic between the positive and

Figure 2-3: Classifier feature importance. A: Sum of the unsigned feature importance of EPC and ERC models. Feature importance was taken from the mean decrease in model accuracy upon random permutation of each feature variable, summed across each feature type. EPC feature importance shown in gold, ERC in blue. **B:** Relative signed feature importance of across EPC features. The sign of the feature importance was assigned by taking the sign of the t-statistic between each feature and the positive and negative expression classes.

Figure 2-3: Classifier feature importance:



negative class and each feature. The sign of the t-statistic was used to assign a direction to each feature importance value. Again, the feature importance between EPC (Fig. 2-2B) and ERC (SI Fig. 2-7) was similar. For both models, CG methylation at the first and last bin was negatively associated with expression, whereas methylation in the middle of the gene was positively associated with expression. For CHG methylation, methylation of all bins was negatively associated with expression. CHH methylation had minimal feature importance and therefore is minimally associated with expression. While overall, intron methylation features were less important to model accuracy than exon or bin features, for CHG methylation, methylation of gene interior introns had higher feature importance than interior exons.

2.2.4 Prediction of sorghum expressible gene sets

Our trained models were then used to categorize all sorghum annotated genes with methylation data, 34,496 genes total, as expressible or non-expressible at the protein and transcript levels. This resulted in a set of 21,132 genes predicted to be expressed at the protein level, with an additional 6929 predicted to be expressed at the transcript level. 6,435 genes were predicted to be silent. All genes predicted to be expressed at the protein level were also predicted to be expressed at the transcript level (Fig 2-4). The EPC expressible gene set was significantly different from both the previously predicted high confidence gene set (Olson et al., 2014) and the expert curated high confidence gene set of sorghum (Fig. 2-4).

Figure 2-4: Comparison of predicted expressed gene sets to previously annotated high confidence gene sets.

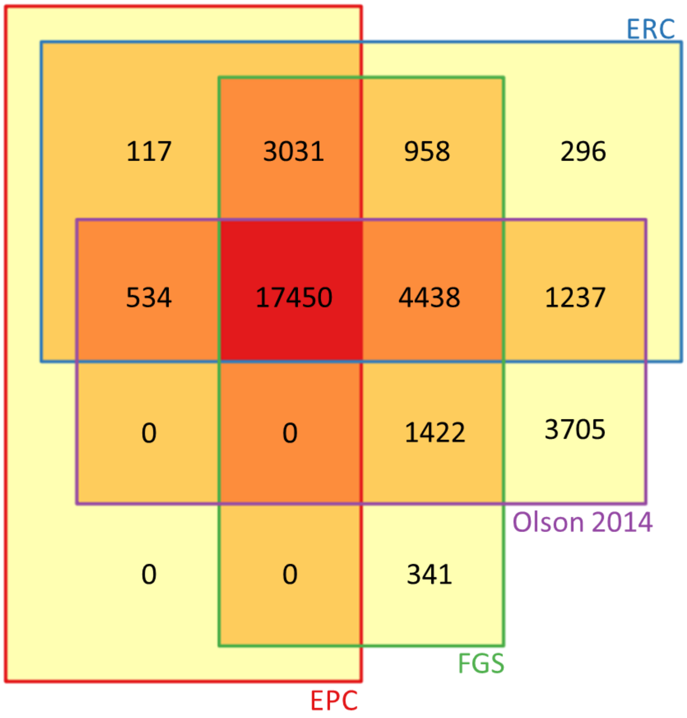


Figure 2-4: Comparison of predicted expressed gene sets to previously annotated high confidence gene sets. Venn diagram of overlap between sorghum genes predicted as expressible by the EPC and ERC and previously annotated high confidence gene sets. "Olson 2014" from high confidence set from Olson et al., 2014.

2.2.5 Comparison of sorghum expressible genes to syntenic gene sets identifies uniquely expressible genes of sorghum and maize

Given the enrichment of syntenic genes within protein detectable and classical gene sets (Schnable and Freeling, 2011; Walley et al., 2016), we then compared the sorghum expressible gene sets, previously published maize expressible gene sets (Sartor et al., 2018 [Under Review]) and the syntenic gene sets between sorghum, maize, rice, setaria, and brachypodium (Table S34-35; Fig. 2-5). Across all genes, a higher percentage of sorghum genes were syntenic than maize. Correspondingly, at all levels of express-ability, including silent genes, sorghum shows a higher percentage of syntenic genes than maize (Fig. 2-5A). Overall, a higher percent of sorghum genes are expressible than maize genes. However, a lower percentage of sorghum syntenic genes are expressible than maize syntenic genes, and more sorghum nonsyntenic genes are expressible (Fig. 2-5B). The two maize subgenomes show nearly identical express-ability (fig. 2-5B). To identify potential genes responsible for species-specific traits, the unique expressible genes between sorghum and maize were identified, using previously published predictions of maize protein expressible genes (Sartor et al., 2018 [Under Review]). These unique genes included EPC or ERC expressible syntenic orthologs which were expressible in only one of the two species (Fig. 2-5C-D). The majority of syntenic expressible genes were shared between sorghum and maize. RNA expressible gene sets showed fewer species-specific expressible genes, compared to protein expressible gene sets (Fig. 2-5C-D). Functional enrichment analysis based on MapMan gene ontology annotations was performed on the unique protein expressible genes plus protein expressible nonsyntenic genes (Thimm et al., 2004). For sorghum, there were 4,512 protein expressible genes nonsyntenic to maize, and for maize there were 11,353 protein expressible genes nonsyntenic to sorghum. The most over-enriched category for both sorghum and maize species-specific expressible gene sets was “unknown.unassigned” (Table S36-S37). Both sorghum and maize uniquely expressible genes were under-enriched for MapMan categories of RNA, regulation of transcription, redox, homeobox transcription factors and protein post-translational modification (Table 2-1; Table S36-37). Over-enriched categories from both species included terpenoid secondary metabolism and jacalin myrosinases. Interestingly, while sorghum unique protein expressible genes are highly over-enriched from NAC-family transcription factors, maize unique expressible genes are significantly under-enriched for the same category (Table 2-1). Maize unique expressible genes were also under-enriched for several other transcription factor families (Table S37). Maize unique expressible genes were additionally over-enriched for biotic stress, receptor-like kinases, AP2/EREB transcription factors, and phenylpropanoid and flavonoid secondary metabolism (Table S37).

Figure 2-5: Comparison of maize and sorghum expressible gene sets

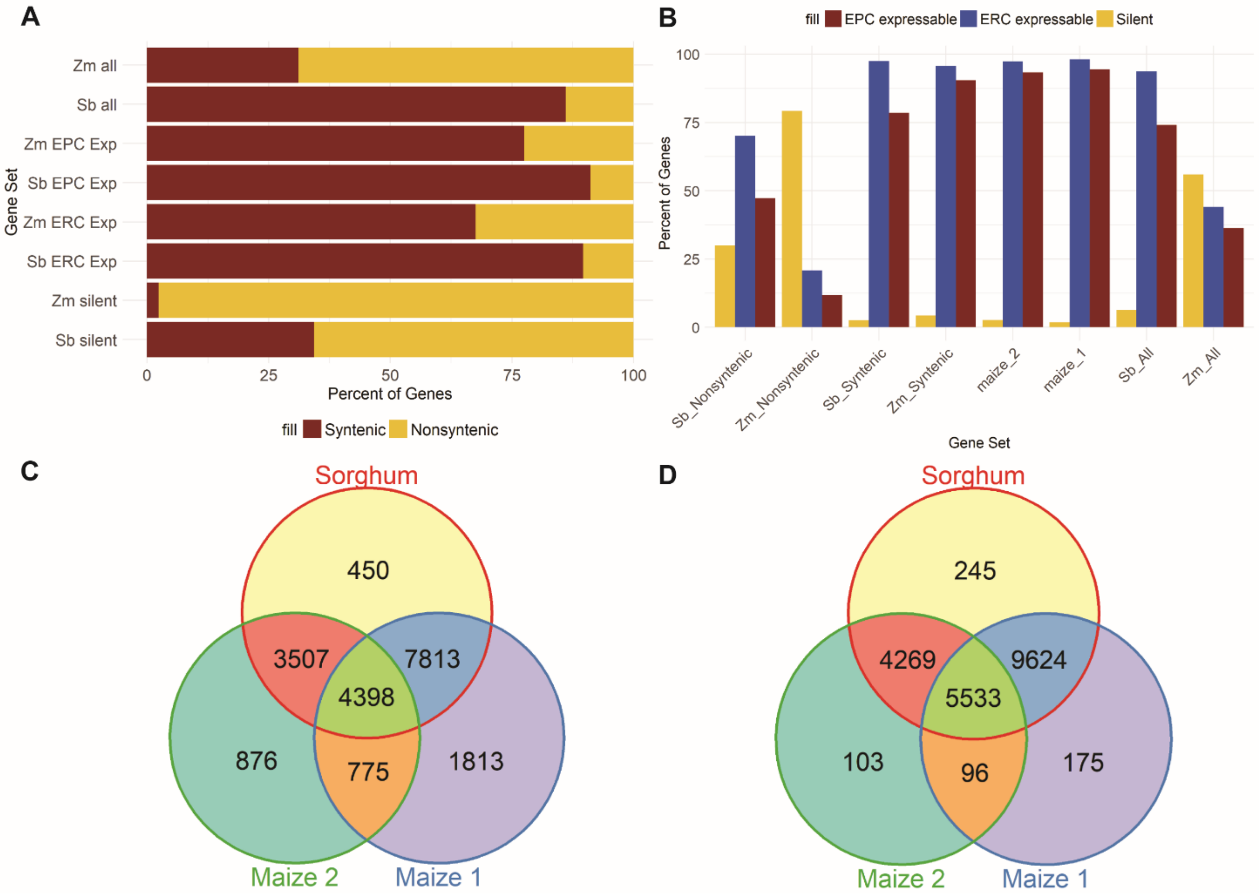


Figure 2-5: Comparison of maize and sorghum expressible gene sets. A: Percent of sorghum and maize expressible genes which are syntenic to grass species rice, brachypodium, setaria, maize, and sorghum. **B:** Percent of syntenic, nonsyntenic, and maize subgenome 1 and 2 genes which are expressible at the protein or RNA level. Silent genes are genes which are not predicted to be expressible at the RNA or protein level. **C:** Venn diagram of overlap of maize and sorghum syntenic EPC expressible genes. **D:** Venn diagram of overlap of maize and sorghum syntenic ERC expressible genes.

Table 2-1: Species-specific over- or under-enriched mapman categories: Mapman annotated transcription factor families over or under enriched in sorghum or maize species-specific protein expressible gene sets. Categories are in vertical order of statistical significance. Species-specific expressible genes include expressible nonsyntenic genes and expressible syntenic genes for which the corresponding syntelog is silent.

Sorghum Over	Sorghum Under	Maize Over	Maize Under
NAC domain B3	Homeobox	AP2/EREB	bZIP ARF NAC domain Squamosa WRKY Homeobox

2.2.6 Express-ability of sorghum syntenic genes across grass species

Given that the majority of protein and RNA sorghum expressible genes are syntenic, we then separately compared the express-ability of sorghum genes which are syntenic to maize, rice, Brachypodium, or Setaria. When looking at all syntenic gene pairs, species of comparison creates no sizable effect in the percentage of sorghum syntenic genes which are expressible (fig. 2-6A). However, when looking at the species-unique syntenic pairs, e.g. pairs which are unique between sorghum and Brachypodium only, sorghum genes which are uniquely syntenic to rice or Setaria are less expressible than all sorghum syntenic genes or sorghum genes uniquely syntenic to maize. Sorghum genes uniquely syntenic to Brachypodium had intermediate express-ability (fig. 2-6B).

Figure 2-6: Effect of species of comparison on express-ability of syntenic genes

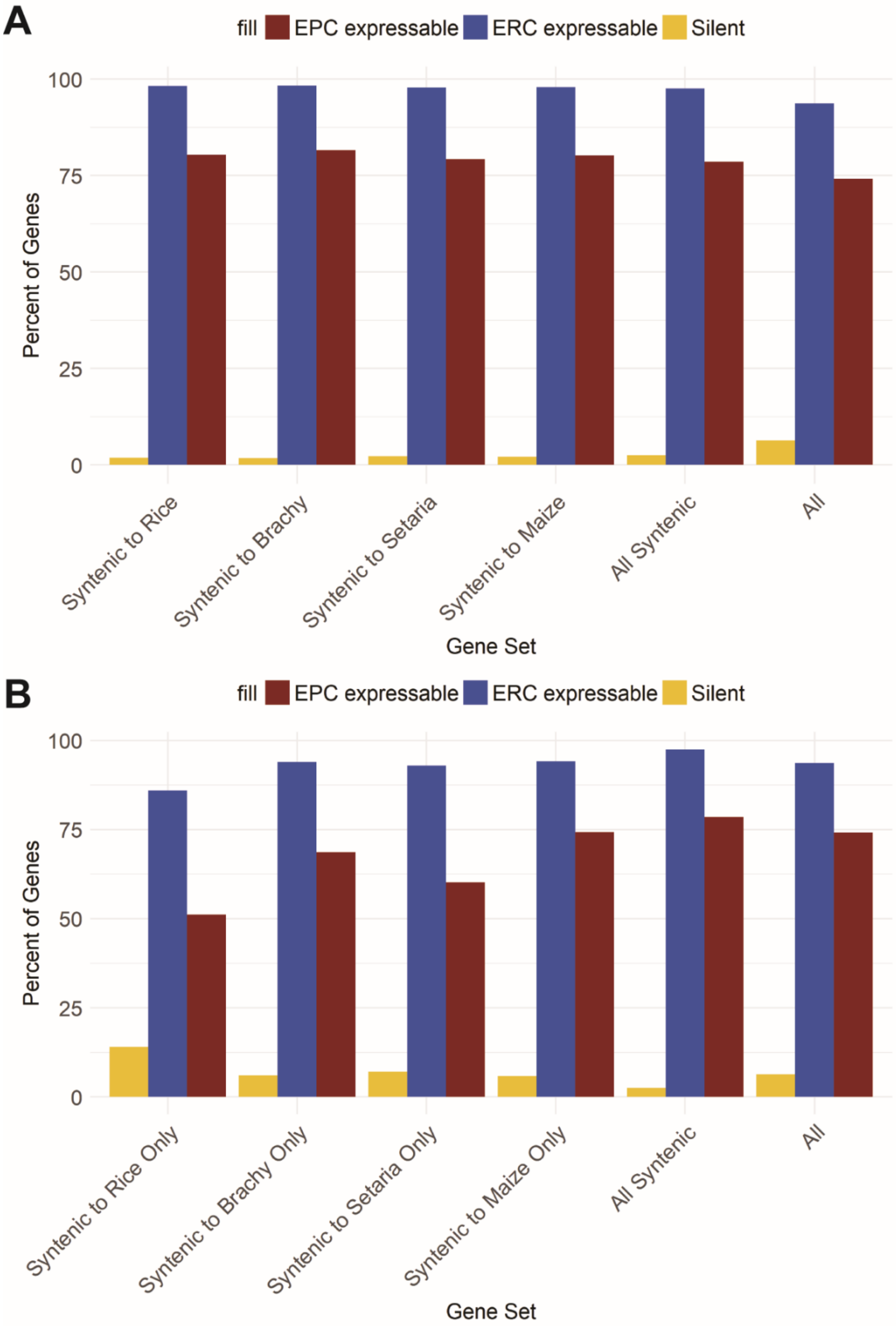


Figure 2-6: Effect of species of comparison on express-ability of syntenic genes. A: Percent express-ability of sorghum genes syntenic to various grass species. All genes syntenic to a given species are used. **B:** Percent express-ability of sorghum genes uniquely syntenic to specific grass species.

2.3 Discussion

As their lineages diverged only recently, maize and sorghum plants show very similar morphology and development prior to the reproductive stages. Similar phenotypes suggest substantial similarities between the proteomes of these two species. Approximately two thirds of sorghum genes are predicted to be expressible at the protein level, resulting in just over 21,000 protein coding genes. Despite having nearly three times as many annotated genes in the working gene set, only 32,979 maize genes are predicted to be protein coding. This suggests that while the total predicted gene contents of the maize genome dwarf that of sorghum, the expressible gene sets are of comparable size. Additionally, a large proportion of maize and sorghum syntenic genes are shared in their express-ability. This is especially pronounced at the RNA level, where there are very few species-specific syntenic expressible genes (Fig. 2-5D). Many of these RNA expressible genes are not predicted to be expressible at the protein level, but the total number of species-specific protein expressible syntenic genes is larger at the protein level than at the RNA level.

While the two species are similar in appearance, the origin of their domestication, with sorghum from North Africa and maize from Mesoamerica, indicates exposure to different stresses during evolution and domestication. Maize and sorghum show differential tolerance to abiotic and biotic stresses. The over-enrichment of multiple stress associated gene ontology categories in the species-specific expressible genes indicates that some of this differential stress tolerance may be controlled at the epigenetic level. The NAC family of transcription factors is a large plant transcription factor family with diverse roles in development and mediation of biotic and abiotic stress response (Olsen et al., 2005). Genes annotated in this family of transcription factors was over-enriched in the sorghum unique expressible gene set and under-enriched in the maize unique expressible gene set. For both species, the most over-enriched category among species-specific expressible genes was “unknown.unassigned”, representing nearly half of the unique genes in both sorghum and maize. As species-specific genes are less likely to have functional annotation, further characterization of these genes is necessary for a holistic understanding of phenotypic differences between sorghum and maize. In contrast, maize genes without a syntenic ortholog in sorghum were not over-enriched for stress associated categories (Table S38). Thus, looking at the gene presence/absence variants and gene express-ability variants of sorghum and maize provides different information about potential species-specific phenotypes.

While the proportion of maize syntenic genes which are EPC expressible exceeds 90%, the proportion of sorghum syntenic EPC expressible genes is substantially smaller. Conversely, while few maize nonsyntenic genes are EPC expressible, a much larger percentage of sorghum nonsyntenic genes, nearly 30%, are EPC expressible. This indicates a less strong association between synteny and expressibility in sorghum, compared to maize. This difference may result from variation in gene expression regulation, which may have arisen as a response to the relatively large number of nonsyntenic genes and transposable elements which have proliferated in the maize lineage. The maize lineage specific autopolyploidy event resulted in two subgenomes of maize, of which one has undergone nearly two times as much gene loss as the other (Schnable et al., 2012). Despite the more extensive gene loss of maize subgenome 2, the two maize subgenomes show nearly identical express-ability. Similar proportions of genes from each maize subgenome were detectable at the protein level in maize roots, consistent with classifier predictions (SI Fig. 2-8).

Comparisons between sorghum syntenic genes syntelogous to single grass species showed lower express-ability of sorghum genes uniquely syntenic to rice or Setaria, versus those uniquely syntenic to maize. The lineage that gave rise to rice and Brachypodium diverged from the lineage that gave rise to sorghum approximately 70 million years ago, more than 50 million years prior to the divergence of maize and sorghum lineages (Swigoňová et al., 2004; Wang et al., 2015). The Setaria lineage diverged from the sorghum and maize lineage fifty million years ago. Further comparison to maize expressible gene sets could illuminate whether loss of express-ability of sorghum genes uniquely syntenic to rice occurred prior to divergence of the sorghum and maize lineages. Sorghum genes uniquely syntenic to Brachypodium had intermediate express-ability. However, the number of sorghum genes uniquely syntenic to Brachypodium (83) or to rice (135) was low compared to Setaria (904) or maize (1431). This may be too few unique pairs to compare proportional expressibility.

2.4 Materials and Methods

2.4.1 Data sources

Gene models were taken from version 1.4 of the Sorghum bicolor reference genome (Paterson et al., 2009), accessed from the Joint Genome Institute on 04/04/2018. Total cytosines within different genomic contexts (CG, CHG, and CHH) were counted within feature bins from Sorghum bicolor reference genome version 1.4 accessed from phytozome on 04/15/2018.

2.4.2 Sorghum slit-stem fungal elicitation assay

Sorghum plants were grown in a greenhouse at the University of California, San Diego, under 12 hour light:dark cycles with minimum of 300 μmol m^{-2} s^{-1} of photosynthetically active radiation supplied by supplemental lighting, 70% relative humidity, and temperature cycle of 24 °C at night and 28 °C during the day (Schmelz et al., 2009). Plants were treated at the 8-9 leaf stage. For fungus elicited plants, commercially available heat-killed *Fusarium venenatum* (strain PTA-2684, Monde Nissin Corporation Co.) mixed in water. Fusarium-soaked, or water-only cloths were inserted into the stem by a three-inch incision at the first internode. Stems were wrapped in tape to prevent drying. Samples were harvested at 40 hours post treatment and fungus adjacent tissue was removed with a clean razor blade before freezing in liquid nitrogen for storage at -80°C.

2.4.3 Proteomics of sorghum fungal and non-fungal treated tissue

Proteins were extracted from Sorghum stems and peptides were prepared for mass spectrometry as described previously (Walley et al., 2016). Mass spectra were searched against the protein database from Sorghum bicolor V3.1.1 (McCormick et al., 2018). False discovery rate was constrained as 1% at the protein level, as described previously (Walley et al., 2016). The proteomics results are in Table S31. The detected V3 gene models with V1 gene model equivalents were used as the positive case for protein express-ability classifier, resulting in identification of 6478 V1 proteins with binned gene models.

2.4.4 Methylation Feature Construction

For current model construction, gene models were binned into five equal sized fragments for bins 1 through 5. The regions of exons and introns contained within each whole gene length bin were used. Due to low numbers of gene models with annotated untranslated regions, gaps between the annotated exons were used for intron features. This produced a total of 15 genomic intervals for each gene model.

Methylated cytosines within each sequence context were separately mapped to the intervals described above. Published methylation data from both tissues (root and shoot) were handled separately (Turco et al., 2017). All cytosines with each sequence context were counted within the same intervals. The proportion of methylated to unmethylated cytosines was calculated for each context, each bin, and each bisulfite sequencing sample to produce 90 distinct features for each gene. For any features without annotation within the gene model (E.g. gene models without exons in an individual bin, or gene models without introns) a methylation level of 0.5 was substituted, representing neither hypo- nor hyper-methylation. Methylation features are in Table S32.

2.4.5 Classification of training data

Published RNA-seq data was taken from Dugas et al., 2011. Transcript abundance in RPKMs for each gene was averaged across all 24 samples. Gene models from which no RPKMs are reported were defined as “not detected”. Gene models with average RPKM of greater than or equal to one were defined as high abundance, whereas gene models with reported average RPKMs of less than one were defined as low abundance. Proteins with positive identification from the sorghum slit-stem assay

2.4.6 Construction of Classification Models

Classification models were built as described previously (Sartor et al., 2018 [under revision]) using random forest machine learning algorithm (Breiman, 2001). In brief, random forest models were built using methylation matrices described above as the training data features and gene expression classification at the RNA level (ERC) or protein level (EPC) as the classification factors. For the ERC, the expressed gene set consisted of gene models with high gene expression (RPKM=>1) and unexpressed gene set consisted of gene models with undetected RNA. For the EPC, the expressed gene set consisted of gene models with high RNA as well as detected protein, and the unexpressed gene set consisted of gene models with high RNA without detected protein.

2.4.7 Discovery of expressible gene sets

Genes were defined as expressible or silent based on the proportion of votes of each classifer (Table S33). Genes with a proportion of votes >0.5 were defined as expressible. Genes expressible at the RNA and protein level were separately defined using the proportion of votes from the ERC and EPC classifiers, respectively.

2.4.8 Data analysis

Syntenic gene sets were created as described previously (Schnable and Freeling, 2011).Venn diagrams were created using the R packages “VennDiagram” and “Vennerable” (Chen, 2018; Swinton, 2009). Data tables were read and processed using R packages “Readr” and “reshape2” (Wickham, 2007; Wickham et al., 2017). Barplots for figures 5-6 were created using R package “ggplot2” (Wickham, 2009). ROC and PR curves were plotted and AUC were calculated using R packages “ROCR” and “stringr” (Sing et al., 2005; Wickham, 2018).

2.5 Acknowledgements

Chapter 2, in full, is currently being prepared for submission for publication of the material. de Boer, Laura; Sartor, Ryan C.; Shen, Zhouxin; Noshay, Jaclyn; Springer, Nathan M.; Schmelz, Eric A.; Huffaker, Alisa; Schnable, James; Briggs, Steven P. “Discovery of species-specific expressible genes via machine learning with omic data.” *In preparation.* The dissertation author was the primary investigator and first author of this material.

2.6 References

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Chen, H. (2018). VennDiagram: Generate High-Resolution Venn and Euler Plots.

Dugas, D.V., Monaco, M.K., Olson, A., Klein, R.R., Kumari, S., Ware, D., and Klein, P.E. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. BMC Genomics *12*, 514.

Eichten, S.R., Swanson-Wagner, R.A., Schnable, J.C., Waters, A.J., Hermanson, P.J., Liu, S., Yeh, C.-T., Jia, Y., Gendler, K., Freeling, M., et al. (2011). Heritable Epigenetic Variation among Maize Inbreds. PLoS Genet. *7*, e1002372.

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B.D., McKinley, B., et al. (2018). The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. Plant J. *93*, 338–354.

Olsen, A.N., Ernst, H.A., Leggio, L.L., and Skriver, K. (2005). NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. *10*, 79–87.

Olson, A., Klein, R.R., Dugas, D.V., Lu, Z., Regulski, M., Klein, P.E., and Ware, D. (2014). Expanding and Vetting Gene Annotations through Transcriptome and Methylome Sequencing. Plant Genome *7*, 0.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. Nature *457*, 551–556.

Schmelz, E.A., Engelberth, J., Alborn, H.T., Tumlinson, J.H., and Teal, P.E.A. (2009). Phytohormone-based activity mapping of insect herbivore-produced elicitors. Proc. Natl. Acad. Sci. *106*, 653–657.

Schnable, J.C., and Freeling, M. (2011). Genes Identified by Visible Mutant Phenotypes Show Increased Bias toward One of Two Subgenomes of Maize. PLoS ONE *6*, e17855.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. *108*, 4069–4074.

Schnable, J.C., Freeling, M., and Lyons, E. (2012). Genome-Wide Analysis of Syntenic Gene Deletion in the Grasses. Genome Biol. Evol. *4*, 265–277.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 7881.

Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. Genome Res. *14*, 1916–1923.

Swinton, J. (2009). Venn Diagrams in R with the Vennerable Package.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. *37*, 914–939.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., and Hamon, C. (2003). Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. Anal. Chem. *75*, 1895–1904.

Turco, G.M., Kajala, K., Kunde-Ramamoorthy, G., Ngan, C.-Y., Olson, A., Deshphande, S., Tolkunov, D., Waring, B., Stelpflug, S., Klein, P., et al. (2017). DNA methylation and gene expression regulation associated with vascularization in *Sorghum bicolor*. New Phytol. *214*, 1213–1229.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., et al. (2016). Integration of omic networks in a developmental atlas of maize. Science *353*, 814–818.

Wang, X., Wang, J., Jin, D., Guo, H., Lee, T.-H., Liu, T., and Paterson, A.H. (2015). Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events. Mol. Plant *8*, 885–898.

Wickham, H. (2007). Reshaping Data with the reshape Package. J. Stat. Softw. *12*, 1–20.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wickham, H. (2018). stringr: simple, consistent wrappers for common string operations.

Wickham, H., Hester, J., and Francois, R. (2017). readr: Read Rectangular Text Data.

Wickham, H. (2018). stringr: simple, consistent wrappers for common string operations.

2.7 Supplemental Material

Figure 2-7: Relative signed feature importance of ERC classifier

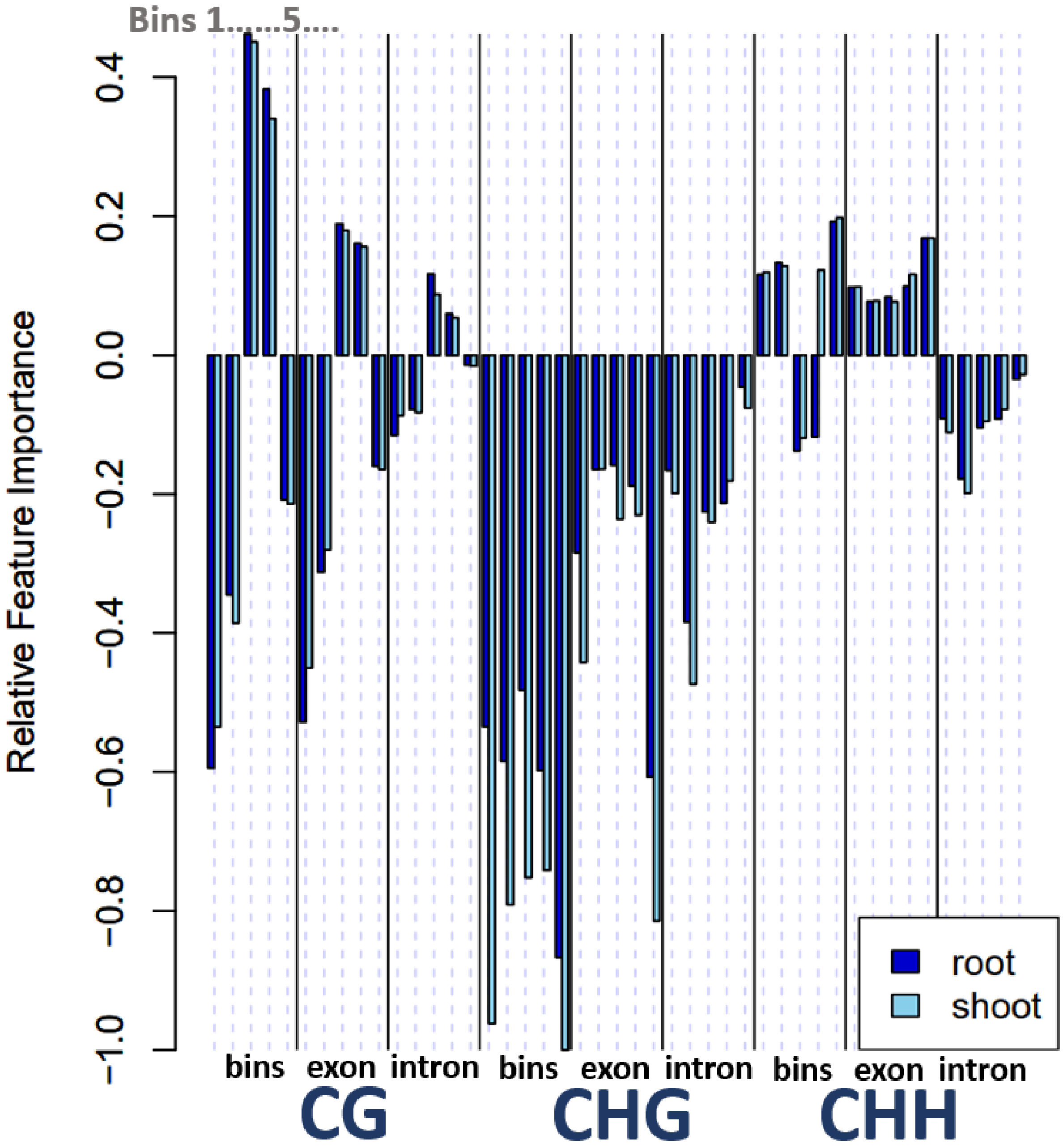


Figure 2-7: Relative signed feature importance of ERC classifier. Relative signed feature importance of across EPC features. The sign of the feature importance was assigned by taking the sign of the t-statistic between each feature and the positive and negative expression classes.

Figure 2-8: Percent of maize subgenome 1 and 2 genes detectable as protein

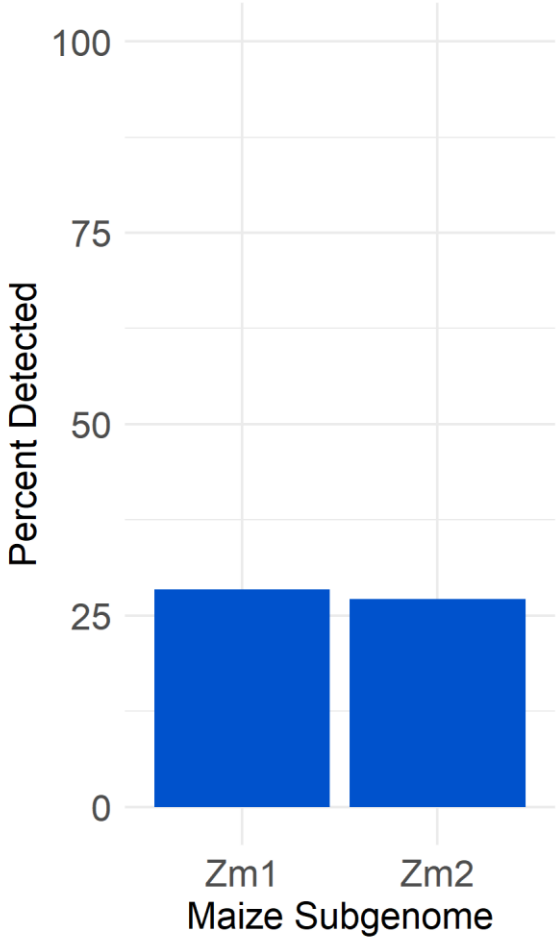


Figure 2-8: Percent of maize subgenome 1 and 2 genes detectable as protein. Percent of maize subgenome 1 and subgenome 2 genes which were detected at protein level in maize root from Walley et al., 2016

CHAPTER 3

Comparison of expressible gene sets across maize genome annotation versions

3.1 Introduction

Accurate annotation of the true set of protein coding genes is crucial for drawing accurate biological conclusions from analyses. The annotated set of protein coding genes serves as the search space for proteomics, therefore any bias that exists in the annotation can be propagated to bias in the identification of peptides via database search of mass spectra. The total set of protein coding genes also often serves as the “background” for gene ontological functional enrichment or similar analyses. Successful genome-wide association and quantitative trait locus (QTL) studies lead to the identification of a genetic interval associated with a particular phenotype or phenotypes. Further identification of the true casual gene or genes requires accurate annotation of the gene contents of that region. Both inclusion of extraneous genes and exclusion of true protein coding genes can lead to false conclusions from these studies.

The predicted number of protein coding genes in genomes varies significantly based on method of annotation and the amount of evidence supporting individual gene models. As more evidence is gathered, the gene model set is refined. For example, prior to sequencing of the human genome, some estimates of the number of protein coding genes exceeded 100,000. Following the release of the first draft of the human genome, the estimates dropped significantly to 35,000. The set of human protein coding genes continues to be refined as more evidence is gathered, supporting a subset of these genes. The number of predicted protein coding genes can also increase with additional evidence, as recently occurred for the human genome, with the predicted number of genes increasing from the lower estimate of 19,000 to 20,000. The continual change in the number of genes in, arguably, one of the best studied higher organisms highlights the challenge of accurate gene annotation. While evidence of transcript or protein products has been identified for a subset of genes from model organisms, it remains difficult to rule out protein coding genes based on lack of expression evidence. This is due in part to inducible expression of a subset of genes, which may only be expressed in a small number of cell types or under a narrow range of conditions.

In the maize reference inbred B73, the current number of predicted protein coding gene models is 39,324. These gene models were derived from the most recent release of the maize reference genome, taking advantage of long-read sequencing (Jiao et al., 2017). Prior to this recent revision, the maize community maintained two sets of gene models, utilizing sanger sequencing data from the first version of the maize reference genome (Schnable et al., 2009). The largest of the two gene sets from genome annotation version 2 (v2) is a comprehensive collection of more than 110,000 possible protein coding gene models known as the “working gene set” (WGS). An unbiased proteogenomics approach, using a splice graph and six-frame translation database of maize, detected few protein products from genes outside the WGS (Castellana et al., 2014). A higher confidence subset of the WGS, known as the “filtered gene set” (FGS), contains 39,695 gene models. Gene model annotation reference version 3 represents an incremental improvement on the v2 gene model set, without the addition of new genome sequencing data. There is no large WGS equivalent for v4.

Many v2 WGS annotated genes have no annotated v4 equivalent. It is unknown what portion of the v4 annotated gene set is expressible.

Recently, the expressible gene sets of maize were accurately predicted using only gene body methylation levels (Sartor et al., 2019 [in review]). The express-ability of all maize v2 WGS genes with methylation data was predicted at both the transcript and protein level using random forests machine learning algorithm (Breiman, 2001). This resulted in the identification of 32,979 and 41,056 protein and RNA expressible genes, respectively. These expressible gene sets differed from the FGS, with classification of some FGS genes as silent and some genes absent from the FGS as expressible. The expressible gene set for the most recent annotation of the maize genome, v4, has not been defined. However, comparison of the v4 filtered gene set genes with v2 equivalents to the predicted expressible sets of v2 WGS revealed more than 7,900 genes predicted to be protein expressible in v2 but without an annotated v4 equivalent.

We present here accurate classification of genes from a collective database of B73 v2 and v4 annotated gene models as expressible at the protein and RNA level. These classifications were created via a machine learning approach using only DNA methylation data. Previous successful classification of maize express-ability used 23 tissues of maize proteomics data and single nucleotide resolution whole genome bisulfite sequencing (WGBS) data for classifier training (Sartor et al., 2019 [in review]). The classifiers reported here show comparable performance using only a single tissue of proteomics data and 100 base-pair (BP) tiling resolution for WGBS data.

3.2 Results

3.2.1 Detectable protein products observed from v2 gene models without annotated v4 equivalents

To discover if any maize B73 v2 gene models absent from the new v4 annotation had detectible protein products, we created a collective database of all v2 WGS and all v4 annotated gene model predicted peptides. This served as the database for peptide identification of mass spectra via database search, using peptides extracted from B73 roots. This resulted in the identification of 11,158 proteins (Table S39). For predicted peptides which are shared between v2 and v4, identified peptides from both v2 and v4 are in the same protein group. We defined the v2 protein detected genes as any genes for which only v2-unique peptides were discovered, or for which both v2 and v4 peptides are present and the v2 database hit peptide identification score is at least 90% of the v4 hit. The same process was applied to identify the v4 protein detected genes. This resulted in the identification of 10,668 protein detected from v2 genes and 10,991 protein detected from v4 genes. This included detection of 866 proteins for which a v2 gene model had a better peptide match score than a v4 gene model, and 1,481 v2 gene models for which there is no annotated v4 equivalent.

3.2.2 Discovery of expressible gene sets for v2 and v4

Given the detection of proteins from v2 gene models without annotated v4 equivalent, and due to previously discovered express-ability of v2 gene models without v4 equivalents (Sartor et al., 2019 [In Review]), we then created a protein express-ability classifier for both v2 WGS and v4 annotated gene models collectively. Previously published 100 bp-tiling WGBS data from maize reference inbred B73 leaf was used to create methylation features for training and application of v2/v4 express-ability classifiers (Li et al., 2015). All v2 WGS and all v4 gene models were divided into five

Figure 3-1: Creation of v2/v4 express-ability classifier

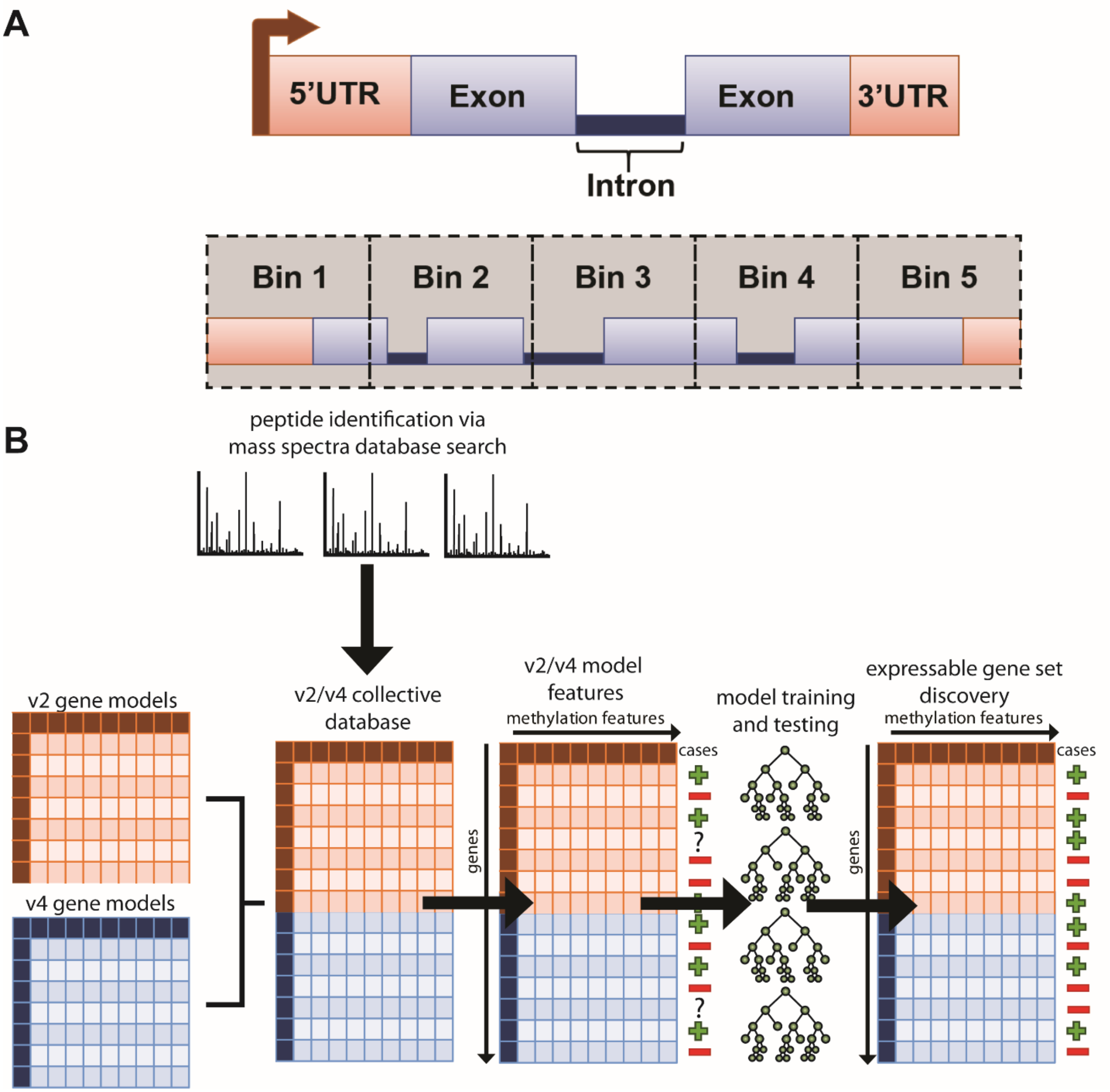


Figure 3-1: Creation of a v2/v4 express-ability classifier. A: Definition of gene regions used for quantification of methylation level for classifier features. **B:** Workflow used for creation of v2/v4 express-ability classifier. The set union of all v2 WGS and v4 gene models served as a database for peptide identification of mass spectra. Methylation features were created for all v2 and v4 gene models with WGBS data. These methylation features were used, in combination with observation of gene products at protein and RNA level, to train random forest machine learning classifier.

equally sized bins. DNA methylation levels of the individual bins and of gene features within each bin were separately quantified (Fig. 3-1A) (Table S40). Two express-ability classifiers were separately created, to classify express-ability of maize v2/v4 genes at the protein (EPC) and RNA (ERC) level Previously published RNA-seq data for v2 and v4 were used to define the negative case (average RPKM = 0) and positive case (average RPKM > 1) for the EPC and ERC. For the EPC, the v2 and v4 protein detectable genes described above were additionally used to define the positive case. The positive and negative cases as defined above, as well as the collective DNA methylation features of both v2 and v4 gene models were used to train two random forest machine learning classifiers (Breiman, 2001) (Fig. 3-1B).

Classifier performance was tested using random out-of-bag cross-validation. The random out-of-bag cross-validated results were compared to the true observed positive and negative cases to identify true and false classifications. These true and false classifications were used for creation of Receiver Operating Characteristic (ROC) and Precision-recall curves (PR) (Fig. 3-2A-B). For perfect classification, we anticipate an area under the curve of one, for both the ROC and PR curves. For random classification, we anticipate an area under the ROC curve of 0.5. The area under the ROC curves for the v2/v4 ERC and EPC was 0.951 and 0.994 respectively (Fig. 3-2A). The area under the PR curves for the v2/v4 ERC and EPC was 0.965 and 0.987 respectively (Fig. 3-2B). Taken together, the high value of the area under the curves indicates highly accurate v2/v4 express-ability classifier performance.

Figure 3-2: v2/v4 classifier testing: Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves for classifier models

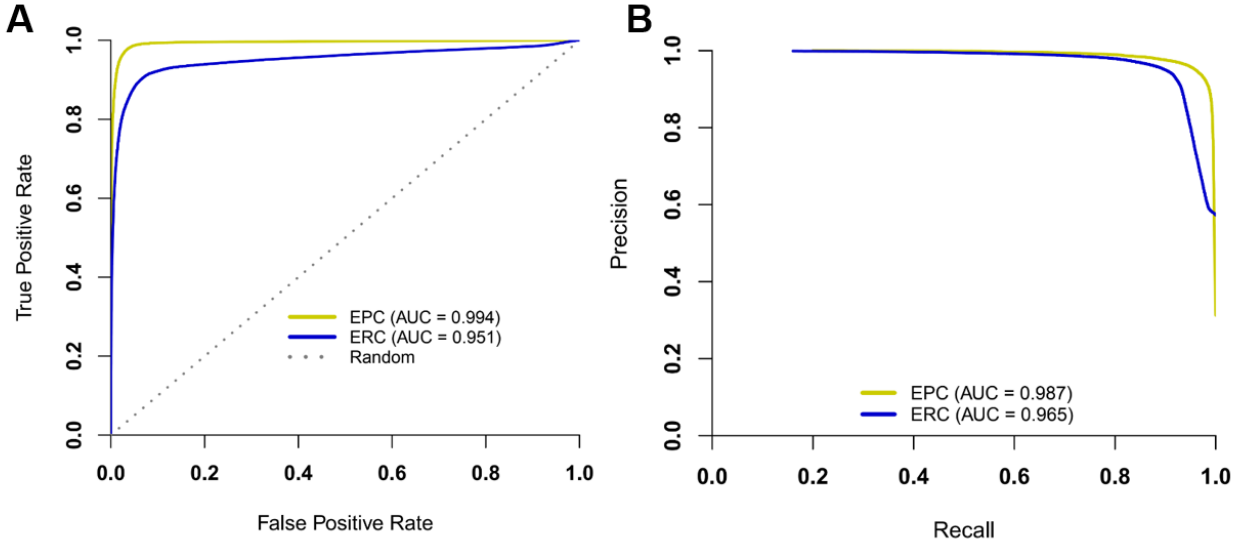


Figure 3-2: v2/v4 classifier testing: Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves for classifier models. ROC (**A**) and PR (**B**) curves were plotted using rates of successful and unsuccessful classifications from random forest model out-of-bag cross-validation votes compared to known classifications. EPC curves shown in gold, ERC curves shown in blue.

3.2.3 Comparison of protein express-ability of maize annotated gene sets from v2 and v4

The classifiers created as described above were then used to expand express-ability classification to the genes beyond the training and testing set. This resulted in the identification of 27,302 protein expressible and 30,346 RNA expressible genes from v4, and 30,779 protein expressible and 40,993 RNA expressible genes from v2 WGS. We then compared the protein expressible gene set of v2/v4 to the v4 gene set and v2 FGS (Fig 3-3). Interestingly, we discovered 3,223 protein expressible genes which are absent from either v4 or v2 FGS. We also identified 14,059 genes which are in the v2 FGS which are not protein expressible, and 12,057 genes present in v4 which are not protein expressible. The majority of protein expressible genes are shared between all gene sets (Fig. 3-3).

Figure 3-3: Comparison of v2/v4 expressible gene set to previously annotated high confidence sets of maize

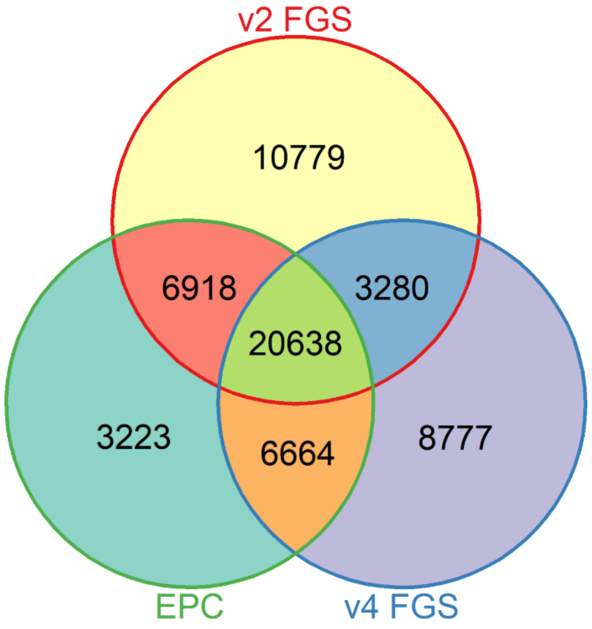


Figure 3-3: Comparison of v2/v4 expressible gene set to previously annotated high confidence sets of maize. Venn diagram showing overlap of v2/v4 protein expressible genes (EPC) and the expert curated v4 and v2 FGS. Equivalent genes determined based on maizeGDB annotated equivalents.

3.2.4 Filtering by protein express-ability reduces number of candidate genes in QTL studies

Given the significant differences between the v2/v4 protein expressible gene set and the v4 and v2 FGS, we then wanted to investigate if our novel annotation of the protein coding gene set changes the biological conclusions drawn from experimentation. Previously published high throughput phenotyping of 106 traits in a maize recombinant inbred population identifies two QTL hotspots significantly associated with maize growth and biomass associated traits (Zhang et al., 2017). There are 28 and 53 genes within the QTL hotspots located on maize chromosomes 7 and 10. However, comparison of the QTL hotspot genes with the protein expressible gene set further narrowed down the number of candidate genes in these regions to 22 and 31 genes respectively. This is a significant reduction, 21% and 41%, in the number of candidate genes.

3.3 Discussion

For many v2 WGS gene models, there is no annotated v4 equivalent gene. To investigate if any potential protein coding gene models were excluded from v4, we used an unbiased proteomics approach using a collective database of v2 and v4 gene models as the database for peptide identification. This led to the detection of thousands of proteins for which the v2 gene model peptide had a better match to the spectra observed or for which there is no equivalent v4 gene model.

To further identify potential differences between the true protein coding genes of maize and both v2 and v4 annotations, we then used the detected proteins to form the positive case to train a maize protein express-ability classifier. This classifier, and its RNA equivalent, was able identify protein and RNA expressible genes with high accuracy using only gene body methylation levels as model features (Fig. 3-2). Previous maize express-ability classifiers used single nucleotide resolution WGBS data and 23 tissues of proteomics and RNA-seq data spanning development for model training. The classifiers reported here used only a single tissue of proteomics data and 100-bp tiling quantitation of WGBS, while the same number RNA-seq tissues was used. This suggests that only a single proteomics experiment may be sufficient for creation of the positive case for EPC classifier training, drastically reducing the cost and effort associated with performing classification with other species or inbreds.

The predicted protein expressible gene set differs from the v4 and v2 filtered gene sets (Fig 3-3). Both v2 and v4 gene sets include gene models which are not predicted to be protein expressible, and lack 3,223 gene models which are predicted to be expressed. The presence of expressible genes outside of the expert curated sets suggests that some genes which have the potential to affect traits are absent from the newest annotation of the maize genome. These 3,223 models include v2 WGS genes for which there is no annotated v4 equivalent. We propose addition of these models to the current filtered gene set of maize, or, where appropriate, annotation of the missing v4 equivalent. Thousands of genes present in the filtered gene sets were not predicted to be expressible at the protein level. Accurate annotation of the protein coding gene set is critical for defining the search space for omics methods and population-based studies. Peptides from missing gene models will be unidentified in proteomics experiments, and peptides which are shared between missing and present gene models may be misidentified as unique. Our EPC can also be used to reduce the number of candidate genes in GWAS or QTL studies. Genes which are not protein expressible are less likely to affect phenotype. Assessment of the potential contribution towards a trait of each gene candidate within a genetic interval is costly. We were able to reduce the number of gene candidates of a published QTL study by 21% and 41%. This included gene candidates from a QTL hotspot which was also identified as a QTL controlling maize primary metabolism (Wen et al., 2015; Zhang et al., 2017). Accurate definition of the gene pool used as the search space can reduce both type I and type II errors in experimentation.

3.4 Materials and Methods

3.4.1 Plant materials

B73 plants were grown and roots were sampled from field grown plants, 0 days after pollination, as described previously (Zhou et al., 2018). Protein extraction was performed on the same root tissue as used for previously published RNA-seq analysis (Zhou et al., 2018).

3.4.2 Proteomics

Tissue powders were suspended in extraction buffer (8M Urea/100mM Tris/5mM Tris(2- carboxyethyl)phosphine (TCEP)/phosphatase inhibitors, pH 7). Proteins were precipitated by adding 4 volumes of cold acetone and incubated at 4o C for 2 hours. Samples were centrifuged at 4,000xg, 4o C for 5 minutes. Supernatant was removed and discarded. Proteins were re-suspended in urea extraction buffer and precipitated by cold acetone one more time. Protein pellets were washed by cold methanol with 0.2mM Na3VO4 to further remove non-protein contaminants and re-suspended in the original extraction buffer. Proteins were then digested with Lys-C (Wako Chemicals, 125-05061) at 37o C for 15 minutes then diluted 8-fold with 1M urea containing 100mM Tris and secondarily digested with trypsin (Roche, 03 708 969 001) for 4 hours. Digested peptides were purified on Waters Sep-Pak C18 cartridges and eluted with 60% acetonitrile. TMT-10 labelling was performed in 50% acetonitrile/150mM Tris, pH 7 and checked by LC-MS/MS to confirm > 99% efficiency. Labelled peptides from each time point sample were pooled together for 2D-nanoLC-MS/MS analysis. An Agilent 1100 HPLC system was used to deliver a flow rate of 600 nL min-1 to a custom 3-phase capillary chromatography column through a splitter. Column phases employed consisted of a 30 cm long reverse phase (RP1: 5 µm Zorbax SB-C18, Agilent), 8 cm long strong cation exchange (SCX: 3 µm PolySulfoethyl, PolyLC), and 40 cm long reverse phase 2 (RP2: 3.5 µm BEH C18, Waters) coupled with an electrospray tip of fused silica tubing pulled to a sharp point (inner diameter <1 um). Peptide mixtures were loaded onto RP1, the 3 column sections were joined and mounted on a custom electrospray adapter for on-line nested elution. Peptides were eluted from the RP1 section to SCX section using a 0 to 80% acetonitrile gradient for 60 minutes, and then are fractionated by the SCX column section using a series of 20 step salt gradients of ammonium acetate over 20 min, followed by high-resolution reverse phase separation on the RP2 section of the column using an acetonitrile gradient of 0 to 80% for 150 minutes. Spectra were acquired on a Q-exactive-HF mass spectrometer (Thermo Electron Corporation, San Jose, CA) operated in positive ion mode with a source temperature of 275 °C and spray voltage of 3kV. Automated data-dependent acquisition was employed of the top 20 ions with an isolation window of 1.0 Da and collision energy of 30. The mass resolution was set at 60,000 for MS and 30,000 for MS/MS scans, respectively. Dynamic exclusion was used to improve the duty cycle. The raw data was extracted and searched using Spectrum Mill vB.06 (Agilent Technologies). MS/MS spectra with a sequence tag length of 1 or less were considered to be poor spectra and were discarded. The remaining MS/MS spectra were searched against maize a collective database of all B73 v2 WGS genes and all v4 genes. Search parameters were set to Spectrum Mill's default settings with the enzyme parameter limited to full tryptic peptides with a maximum miscleavage of 1. A 1:1 concatenated forward-reverse database was constructed to calculate the false discovery rate (FDR). Proteins that share common peptides were grouped using principles of parsimony to address protein database redundancy. Proteins observed are in Table S39.

3.4.3 Quantitation of gene body methylation level and creation of model methylation features

Whole genome bisulfite sequencing data generated from maize inbred B73 leaf was previously published (Li et al., 2015). B73 leaf-derived WGBS sequencing data previously published mapped to maize reference genome v2, was then remapped to maize reference genome version v4. Data mapped to different genome versions was handled separately. Gene body methylation level was calculated as described previously with slight modifications (Sartor et al., 2019 [In Review]). In brief, the proportion of methylated cytosines to unmethylated cytosines was calculated in 100 bp non-overlapping tiles across the maize B73 and Mo17 genomes. Genes with methylation coverage across less than 60% of the gene body length were discarded. Gene models were taken from maize B73 5a WGS (RefGen_v2) and B73 AGPv4.36 (RefGen_v4; accessed ensemblPlants 10/4/18) annotations. For B73 v4, both annotated genes and long noncoding RNAs were used as gene models. Gene models were binned into non-overlapping fifths (“bins”). Methylation level of 100 bp tiles overlapping with gene model bins was averaged to calculate methylation level of individual bins. Methylation levels of annotated exons and introns within each bin were separately calculated by the same process. Gene models with missing genomic features (e.g. no introns in a given bin) or no methylation data within single features were assigned a methylation level of 0.5 for the missing features, representing neither hypo- nor hyper-methylation. The methylation level across gene model genomic features served as the observed features for random forest model creation. Model features for B73 v2 and v4 were collectively joined to create a feature set encompassing both v2 and v4 gene models. The quantified methylation features for v2/v4 express-ability classifiers are found in Table S40.

3.4.4 Identification of class variables for model training and testing

For each inbred, two classifiers were created, the expressible protein classifier (“EPC”) and the expressible RNA classifier (“ERC”). For the B73 EPC, the positive class consisted of genes with proteins identified from proteomics data from the primary root, searched against a collective database containing both B73 v2 and v4 gene models, and for which mRNA abundance was high (RPKM>=1) (Figure 1, Table S39). The negative case of the v2/v4 B73 EPC consisted of genes without detectable protein nor detectable mRNA (v2 RNA: Walley et al., 2016; V4 RNA: Zhou et al., 2018).

3.4.5 Creation of classifiers

Classification models were built as described previously (Sartor et al., 2019 [under revision]) using random forest machine learning algorithm (Breiman, 2001). In brief, random forest models were built using methylation matrices described above as the training data features and gene expression classification at the RNA level (ERC) or protein level (EPC) as the classification factors. For the ERC, the expressed gene set consisted of gene models with high gene expression (RPKM=>1) and unexpressed gene set consisted of gene models with undetected RNA. For the EPC, the expressed gene set consisted of gene models with high RNA as well as detected protein, and the unexpressed gene set consisted of gene models with high RNA without detected protein. Classifier performance was determined by using random out-of-bag cross-validation, to determine true and false classifications, as implemented in the random forest R package (Breiman, 2001).

3.4.6 Classification of expressibility of maize v2/v4 genes

Genes were defined as expressible or silent based on the proportion of votes of each classifier (Table S41). Genes with a proportion of votes >0.5 were defined as expressible. Genes expressible at the RNA and protein level were separately defined using the proportion of votes from the ERC and EPC classifiers, respectively.

3.4.7 Additional data analysis

Maize v2/v4 “equivalent” genes were determined based on maizeGDB.com conversion table between for v2 accessions (accessed 12/14/18). Venn diagrams were created using the “Vennerable” R Package (Swinton, 2009). ROC and PR curves were plotted and the areas under the curve were calculated using R packages “Stringr” and “ROCR” (Sing et al., 2005; Wickham, 2018)

3.5 Acknowledgements

Chapter 3, in part, is currently being prepared for submission for publication of the material. de Boer, Laura; Sartor, Ryan C.; Shen, Zhouxin; Noshay, Jaclyn; Springer, Nathan M.; Schmelz, Eric A.; Huffaker, Alisa; Schnable, James; Briggs, Steven P. “Discovery of species-specific expressible genes via machine learning with omic data.” *In preparation.* The dissertation author was the primary investigator and first author of this material.

3.6 References

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Castellana, N.E., Shen, Z., He, Y., Walley, J.W., and Cassidy, C.J. (2014). An Automated Proteogenomic Method Uses Mass Spectrometry to Reveal Novel Genes in. 11.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M.C., Wang, B., Campbell, M.S., Stein, J.C., Wei, X., Chin, C.-S., et al. (2017). Improved maize reference genome with single-molecule technologies. Nature *546*.

Li, Q., Song, J., West, P.T., Zynda, G., Eichten, S.R., Vaughn, M.W., and Springer, N.M. (2015). Examining the Causes and Consequences of Context-Specific Differential DNA Methylation in Maize. Plant Physiol. *168*, 1262–1274.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science *326*, 1112–1115.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 7881.

Swinton, J. (2009). Venn Diagrams in R with the Vennerable Package.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., et al. (2016). Integration of omic networks in a developmental atlas of maize. Science *353*, 814–818.

Wen, W., Li, K., Alseekh, S., Omranian, N., Zhao, L., Zhou, Y., Xiao, Y., Jin, M., Yang, N., Liu, H., et al. (2015). Genetic Determinants of the Network of Primary Metabolism and Their Relationships to Plant Performance in a Maize Recombinant Inbred Line Population. Plant Cell *27*, 1839–1856.

Wickham, H. (2018). stringr: simple, consistent wrappers for common string operations.

Zhang, X., Huang, C., Wu, D., Qiao, F., Li, W., Duan, L., Wang, K., Xiao, Y., Chen, G., Liu, Q., et al. (2017). High-Throughput Phenotyping and QTL Mapping Reveals the Genetic Architecture of Maize Plant Growth. Plant Physiol. *173*, 1554–1564.

Zhou, P., Hirsch, C.N., Briggs, S.P., and Springer, N.M. (2018). Dynamic Patterns of Gene Expression Additivity and Regulatory Variation throughout Maize Development. Mol. Plant.